An Experimental Study of Prisoners' Dilemma and Stag Hunt Games Played by Teams of Players *

Jeongbin Kim† Thomas R. Palfrey‡

December 7, 2023

Abstract

We compare behavior in six variations of prisoners' dilemma and stag hunt games played by 5-member teams to behavior in the same games played by individuals. The experimental design is informed by a theory of team equilibrium that generates hypotheses about systematic differences between teams and individuals in these games. We also compare three different team collective choice procedures: majority rule; majority rule preceded by a poll; and majority rule preceded by chat. In all six games, we observe significant bandwagon effects that drive consensus in the poll and chat treatments, but voting procedures have no significant effects on team decision frequencies except for the poll treatment of the stag hunt games. Surprisingly, teams do not always defect more than individuals. In prisoners' dilemma games with higher incentives to cooperate, teams cooperate more than individuals. A one-parameter logit specification of the team equilibrium model provides a good fit to the prisoners' dilemma data across all treatments. Teams coordinate more successfully than individuals in the stag hunt games. Additional data are gathered using Amazon Mechanical Turk, including data for much larger (25-member) teams. The observed MTurk behavior differs quantitatively from laboratory behavior, but with similar qualitative comparative statics.

JEL Classification Numbers: C72

Keywords: Team Games, Collective Choice, Laboratory Experiments, MTurk

*We acknowledge the financial support of the National Science Foundation (grant #SES-1426560), the Social Science Experimental Laboratory (SSEL) at Caltech, and the The Ronald and Maxine Linde Institute of Economic and Management Sciences at Caltech. We are grateful to John Duffy, Michael McBride, and the staff of the Experimental Social Science Laboratory (ESSL) at UC Irvine for their support and for granting access to the laboratory and subject pool. The paper has also benefited from suggestions by referees and comments by audiences at KAIST, University of Southern California, University of Utah, the 2018 Meeting of the Economic Science Association in Berlin, Germany and the 2018 Southwest Experimental and Behavioral Economics workshop at Caltech.

†Department of Economics, Florida State University. jkim33@fsu.edu.

‡Division of the Humanities and Social Sciences, California Institute of Technology. trp@hss.caltech.edu.

1 Introduction

In many, if not most, strategic situations that are studied in economics, the interacting parties, "players" in game-theoretic language, are actually groups of individuals - what we refer to here as teams. Prominent examples include models of international conflict where the players are nation-states, models of oligopoly and imperfect competition where the players are firms, and bargaining models where the objective is to understand strategic interactions between unions and management or political parties. In all of these applications, the choice of strategy for each player in the game is itself the product of a collective choice procedure to reach a team decision. Experimental economists and social psychologists have recognized this fact and have conducted experiments in which games are played between teams of individuals with identical payoffs, and each team uses a collective decision-making procedure to decide on their strategy in the game. These experiments have been, for the most part, purely exploratory and have sought to determine whether there are systematic differences in observed behavior depending on whether the strategy choices are made by teams or by individuals, and if so, what the nature of these differences is.

The typical finding in these experiments is that teams act more rationally than individuals (and this is also true in non-strategic choice tasks), and outcomes tend to be closer to Nash equilibrium or best replies. This finding is not universal, but has been found in the overwhelming majority of such experiments. We discuss some of this existing literature in more detail in the next section.

There have been a number of qualitative explanations based on general psychological principles, but there has been a dearth of formal models with which to study team games, as implemented in these experiments, in rigorous analytical frameworks. As a first step toward filling this gap, Kim et al. (2021) propose a framework, *team equilibrium*, that combines the formal elements of noncooperative game theory to capture the strategic interaction *between teams* with a general social choice theoretic framework to model the team collective choice procedure *within teams* to decide on a strategy. The framework is developed in sufficient generality such that it can be applied to any finite game in strategic or extensive form and can incorporate a broad range of collective choice procedures.

A key element of the team equilibrium framework in Kim et al. (2021) is that members of the same team are heterogeneous in their expectations about the expected payoff of each available strategy, but these expectations are, on average, equal to equilibrium expected

payoffs of the underlying game. Each team then uses a collective choice rule to aggregate the members' diverse expectations and implement a strategy. Disagreements between team members about which strategy is better can be due to heterogeneity of beliefs about the expected payoffs of the possible strategies or random payoff disturbances as in quantal response equilibrium (McKelvey and Palfrey 1995, 1998) and in the same spirit as Harsanyi (1973).[1] Because of this heterogeneity, team choices are not deterministic. In a strategic environment, this leads to equilibrium effects, in which the distribution of strategy choices that emerge from one team's choice procedure will affect other teams' choices. A team equilibrium is then a mutually consistent distribution of strategy choices for each team, taking into account the collective choice procedures of each team.

Two results that emerge from the framework are as follows: (1) for a wide class of collective choice procedures, team equilibrium will closely approximate Nash equilibrium if teams are sufficiently large; and (2) for many games, small teams will be *further from Nash equilibrium* than individuals (i.e., single-member teams), as predicted by team equilibrium.

These two results have different interpretations in terms of the existing findings from laboratory experiments. On the one hand, the first result (Nash convergence) seems to match up nicely with the typical experimental finding that games played by teams produce outcomes that are closer to Nash equilibrium. On the other hand, Nash convergence is a limiting theoretical result, while most experimental findings have been based on small teams of two or three members. The second theoretical result suggests that it may have only been by coincidence that the games played by teams in the laboratory happen to also be games in which the behavior of even very small teams will theoretically be closer to Nash equilibrium than with games played by individuals. The theory allows one to identify specific games in which the nonmonotonicity results (2) are likely to be observed, making it possible to design an experiment in which the "surprising" finding that teams are further from Nash equilibrium is, at the same time, the "expected" finding in the team equilibrium theoretical framework.

This paper compares choice behavior in six different 2×2 games using majority rule

[1] Kim et al. (2021) refer to these disturbances as estimation errors, but note that equivalently, one can assume that players have correct beliefs about the other team's mixed strategy, but each member of the team has a random idiosyncratic additive private payoff term associated with each of his or her team's strategies, and these idiosyncratic payoffs are distributed identically and independently. The theoretical properties of the two alternative specifications of team equilibrium are identical. In this paper we adopt the payoff disturbance interpretation.

with 1-member, 5-member, and 25-member teams, including four variations of the prisoners' dilemma game and two variations of the stag hunt coordination game. Some of the games are designed so that team equilibrium behavior with 5-member teams is further from the Nash equilibrium than with 1-member teams, and behavior with 25-member teams is even further from the Nash equilibrium.[2]

In addition to comparing team and individual behaviors in these six games, we vary the collective choice procedure used within each team to reach a decision about which strategy to choose. The first procedure is simple majority rule (*Majority*). Each member of the team simultaneously and independently casts a vote for one of the strategies and the strategy with the majority of votes is chosen. The second procedure is similar to the first, except that it is preceded by a straw vote (*Poll*). The result of the straw vote is reported to all members, who then cast a final binding vote. The third procedure is similar to the first, except that the vote is preceded by a discussion stage during which the members can communicate with each other (*Chat*). In principle, the use of different pre-play communication protocols can lead to significantly different behavior in collective decision making, as has been observed in many past experiments with voting rules in strategic environments.[3]

As the third element of our experimental design, we conduct an Amazon Mechanical Turk (MTurk) experiment with much larger (25-member) teams, which was not feasible in the laboratory setting. We also include both the individual and 5-member team treatments with MTurk to obtain a baseline comparison with the laboratory data. One challenge with MTurk is to have multiple interactions among subjects and to give them feedback about outcomes between interactions. To render the experimental design as comparable as possible to the experimental design in the lab, we perform a longitudinal experiment over five consecutive days.

We find systematic differences in behavior between the 1×1 games, the 5×5 games, and the 25×25 games we study. For prisoners' dilemma games, the incentives for defection determine the direction of differences. In games with high incentives for defection, teams defect more than individuals, but the opposite holds true in games with low incentives for defection. This result is in sharp contrast to the results of previous experiments that found that teams always defect more than individuals. In coordination games, teams

[2]In our team game equilibrium framework, the equilibrium with 1-member teams coincides with a quantal response equilibrium of the game.

[3]See, for example, Guarnaschelli et al. (2000), Agranov and Tergiman (2014, 2019), Martinelli and Palfrey (2020), Palfrey and Pogorelskiiy (2019), and Palfrey et al. (2017).

successfully coordinate more frequently than individuals, and conditional on coordination, teams coordinate more frequently on the payoff-dominant outcome than individuals.

Second, we find that the different collective choice procedures affect the voting decisions of individual team members. Specifically, the presence of a straw poll or pre-vote discussion leads to more one-sided votes, indicating the information aggregation and consensus-formation roles of communication in the group. While these significantly affect the distribution of vote margins, pre-play communication has little impact on the distribution of team action choices, except in coordination games. We also observe a bandwagon effect in the poll treatment, according to which team members voting in the minority in the straw poll are more likely to switch their votes than members voting in the majority in the straw poll.

Third, we fit the data from the PD and WPD games to a logit specification of the team equilibrium model by estimating a single logit precision parameter for all games and team treatments.[4] We find that the one-parameter logit specification of the team equilibrium model fits the data well and matches the qualitative comparative statics for the row and column players' defection frequencies. An OLS regression of the fitted defect frequencies on the observed defect frequencies produces an $R^2 = .69$. In contrast the standard Nash equilibrium model predicts no treatment effects, with 100% defection in all treatments of all PD and WPD games for both the row and column players/teams.

Fourth, we find some difference in behavior between subjects in the MTurk experiment and subjects in the laboratory sessions. While most of the qualitative comparisons across games and across different team sizes are similar in the two environments, there are some notable differences. For instance, in prisoners' dilemma games, we observe more cooperation in the MTurk experiment than in the lab. The effect is large in magnitude and statistically significant and applies to both the 1-member and 5-member teams. A second difference is that in coordination games, MTurk teams are more successful than teams in the laboratory. Mturk teams are not only less likely to miscoordinate, but, in addition, conditional on coordination they are more likely to coordinate on the payoff-dominant outcome.

Fifth, we examine the comparative static predictions regarding the team size variable using the MTurk data, which allows a comparison of 5-member and 25-member teams.

[4]The PD and WPD games have a unique team equilibrium for all values of the logit parameter. We do not estimate the team equilibrium model for the SH games because there are multiple team equilibria in both SH games.

For prisoners' dilemma games, as subjects gain experience, teams in the 25×25 games cooperate significantly more than teams in the 5×5 games. In coordination games, 25-member teams coordinate more successfully than 5-member teams. In fact, in one of the two coordination games, the observed coordination rate is 100% for 25×25 games - and always at the payoff-dominant equilibrium.

Section 2 provides an overview of previous experimental results concerning observed behavioral differences between games played by teams of players and games played by individuals. Section 3 presents the model of team games developed in Kim et al. (2021), specialized to the case of two-by-two games. Section 4 presents the six games used in the experiment and draws hypotheses based on the team equilibrium model. Section 5 describes the procedures in the laboratory experiment, with the results reported in Section 6. Section 7 describes the procedures and results of the MTurk experiment, and compares those results with the laboratory experiment.

2 Related literature

The bulk of the evidence on differences between teams and individuals in PD games comes from the social psychology literature, referred to as the "discontinuity" effect, according to which teams defect more than individuals.[5]. Most of these experiments use communication and consensus (a form of unanimity rule) to decide on a team decision. For instance, Insko et al. (1988) and Wildschut et al. (2001) allowed three members in a team three minutes of face-to-face communication to reach a final consensus to either defect or cooperate. More recently, there were two economics experiments that studied team behavior in a one-shot PD, but the questions addressed by and motivations for these studies are different from ours. Subjects in Cason et al. (2019) played a single PD game without repetition, with pairs of 3-member teams matched against each other, where each team decided to cooperate or defect. Team decisions were made in three stages: first, a poll; second a chat stage; and third, a binding majority vote. The main focus was to investigate whether induced group identity would affect behavior in these games. They showed successful coordination in a six-person coordination game played by the two three-member teams can increase individuals' concerns for the welfare of their out-group and increase cooperation in the subsequent one-shot prisoner's dilemma played

[5]This literature fairly extensive. See Kugler et al. (2012) and the references therein

by the two three-member teams. Bauer et al. (2018) investigated PD game behavior using a non-standard pool of subjects: adolescents aged 12-18 years old in Slovakia and Uganda. Subjects were randomly assigned a team of three members and communicated to reach a team consensus decision. Their findings were similar to those of the psychology experiments.

There have been two studies of coordination games with teams. Behavior in team coordination games was studied in Feri et al. (2010). Subjects were randomly assigned a three-member team and five teams engaged in several different coordination games, including weakest-link and average opinion games with 7 or 14 numerical action choices using a two-round chat/unanimity voting rule. In each round, team members first chatted and then each team member voted for a number. If all of the entered numbers by team members were identical, this number became the team choice. If such unanimity failed in both rounds, all of the team members earned zero payoff, and that team's decision was excluded from the 5-team coordination game. Each coordination game was repeated 20 times. The main finding was that teams were more likely to choose higher numbers (which leads to more efficient outcomes) than individuals. They also had a treatment using the median voting rule instead of unanimity, and they obtained similar results.

Charness and Jackson (2007) studied the effects of voting rules on 2-member team behavior in a 2-player network formation game, with the same incentive structure as a stag hunt game. Each player (team) could pay a cost to attempt to form a link with the other player. A link was formed, yielding a bonus for both sides, if and only if both players (teams) attempted to form a link. Thus, attempting to form a link is similar to choosing stag, and not trying to form a link is similar to choosing hare. There was no communication, and they compared two different unanimity rules within each team. In the first, a team tried to form a link only if both members voted to try to form the link. In the second, a team tried to form a link if at least one member voted in favor. The authors found that individual subjects were more likely to vote in favor under the second voting rule. Feri et al. (2010) also conducted a stag hunt game framed as a 2-number minimum game but did not find significant differences between 3-person teams and individuals.

While most previous studies of team play in strategic games used a consensus rule through communication for collective decision making, Gillet et al. (2009) is a notable exception in which majority and unanimity voting rules were compared in a common pool resource dilemma. They found that 3-member teams with majority rule behave more competitively than individuals, while 3-member teams with unanimity do so only

with repetition.[6] There have been many other studies of team play in games using games other than PD and coordination games.[7] See Kim et al. (2021) and the survey paper by Charness and Sutter (2012) for more details.

3 Team Equilibrium in 2×2 Games

The experimental design, hypotheses, and analysis of results are closely linked to the theory of team equilibrium in games (Kim et al. 2021). Therefore, we first describe the theory of team equilibrium in this section, specialized to the case of 2×2 games with equal-sized odd-numbered teams, where each team independently chooses a strategy by majority rule voting.

3.1 Team equilibrium with majority rule

Let $\Gamma = [A, , u]$ be a 2×2 game, where $A = A^1 \times A^2 = \{a_1^1, a_2^1\} \times \{a_1^2, a_2^2\}$ is the set of strategy profiles and $u = (u^1, u^2)$ are the payoff functions with $u^t : A^t \to \Re$, $t = 1, 2$. Each player in the *team game* associated with $\Gamma = [A, , u]$ consists of a team, t, of an odd number, n, of individual members.[8] Each team $t = 1, 2$ chooses one of the two available actions in A^t. Given a team action profile $a = (a^1, a^2)$, all members of the same team, t, receive the same payoff $u^t(a)$. Denote by σ^1 the probability that team 1 chooses action a_1^1, and let σ^2 the probability that team 2 chooses action a_1^2 and denote the expected payoff to team t from choosing action a_k^t by $U_k^t(\sigma^{-t}) = \sigma^{-t} u^t(a_k^t, a_1^{-t}) + (1 - \sigma^{-t}) u^t(a_k^t, a_2^{-t})$. Using an approach similar to quantal response equilibrium, each member i of team t has an additive random payoff disturbance. We denote the disturbed expected payoff to member i to action a_k^t if team $-t$ uses mixed strategy σ^{-t} by $\widetilde{U}_{ik}^t(^{-t}) = U_k^t(^{-t}) + \varepsilon_k^t$ where the ε_k^t's are iid draws from a distribution with full support on the real line, and a continuous, strictly increasing cdf. Because there are only two actions for each team, we simplify the notation by letting $\varepsilon_i^t = \varepsilon_{i1}^t - \varepsilon_{i2}^t$ equal the difference in disturbance terms for individual

[6] Blinder and Morgan (2005) found no difference between majority and unanimity rules in a 5-member team monetary policy experiment without strategic interaction.

[7] For instance, Cason and Mui (1997) and Luhan et al. (2009) for dictator games, Sutter (2005), Kocher and Sutter (2005), and Kocher et al. (2006) for beauty contest games, Ambrus et al. (2015) for gift-giving games, Bornstein and Yaniv (1998) and Elbittar et al. (2011) for ultimatum games, Cox (2002) and Kugler et al. (2007) for trust games, and Cooper and Kagel (2005) for limit-pricing games.

[8] We assume the two teams have equal size, since this is the case in our experiment. See Kim et al. (2021) for the general model with different team sizes.

i on team t and denote by H^t the distribution of this difference, ε_i^t. Hence H^t is strictly positive and continuously differentiable on $(-\infty, \infty)$ and symmetric around 0 ; i.e., for all $x \in \Re$, $H^t(x) = 1 - H^t(-x)$ and $F^t(0) = \frac{1}{2}$. Thus, given any mixed strategy σ, and any team member i in team t, the difference in expected utility to member j of team t is:

$$\Delta\widetilde{U}_i^t(\sigma) \equiv \widetilde{U}_{i1}^t(\sigma) - \widetilde{U}_{i2}^t(\sigma) = U_1^t(\sigma) - U_2^t(\sigma) + \varepsilon_i^t.$$

The probability that a member of team t prefers action a_1^t to a_2^t is

$$\begin{aligned} p_n^t(\sigma_n^{-t}) &= \Pr\{\Delta\widetilde{U}_i^t(\sigma_n^{-t}) > 0\} \\ &= H(U_1^t(\sigma_n^{-t}) - U_2^t(\sigma_n^{-t})) \end{aligned}$$

where we call p_n^t the a_1^t *vote probability* for team t since a member of team t votes for a_1^t if $\Delta\widetilde{U}_i^t(\sigma_n^{-t}) > 0$. Since the team collective choice rule is *majority rule,* the probability team t chooses action a_1^t in response to σ^{-t}, $P_n^t(\sigma_n^{-t})$, is equal to the probability that $|\{i \in t | \Delta\widetilde{U}_i^t(\sigma_n^{-t}) > 0\}| > \frac{n-1}{2}$, which is given by the cumulative binomial distribution:

$$P_n^t(\sigma_n^{-t}) = \sum_{j=\frac{n+1}{2}}^{n} \binom{n}{j} p_n^t(\sigma_n^{-t})^j (1 - p_n^t(\sigma_n^{-t}))^{n-j}.$$

These team response functions define a mapping from the set of team mixed strategy profiles into itself, and any fixed point of this mapping is an equilibrium of the team game. Formally:

Definition 1. *Given a 2×2 team game under majority rule, $\Gamma = [A, u, n, H]$ a Team Equilibrium of Γ is a mixed strategy profile $\sigma_n^* = (\sigma_n^{*1}, \sigma_n^{*2})$ such that $\sigma_n^{*t} = P_n^t(\sigma_n^{*-t})$ for $t = 1, 2$, with corresponding vote probabilities equal to $p_n^* = (p_n^{*1}, p_n^{*2})$ such that $p_n^{*t}(\sigma_n^{*-t}) = \Pr\{\Delta\widetilde{U}_i^t(\sigma_n^{*-t}) > 0\}$ for $t = 1, 2$*

It is straightforward to see that this definition reduces to regular quantal response equilibrium of the underlying game if $n = 1$.[9] The following two results about the properties of team equilibrium were proved in Kim et al. (2021).[10]

[9] See Goeree et al. (2016, p. 19-20).

[10] In addition, Kim et al. (2021) proved that team response functions in 2×2 games under majority rule satisfy the same properties of stochastic rationality as regular quantal response functions (payoff responsiveness and rank dependence).

Theorem 1. *For any 2×2 team game under majority rule, $\Gamma = [A, u, n, H]$, a* ***Team Equilibrium*** *exists.*

Theorem 2. *Consider an infinite sequence of team games under majority rule, $\{\Gamma_m\}_{m=1}^{\infty}$ such that (1) $A_m^t = A^t = \{a_1^t, a_2^t\}$ for all m; (2) $u_m^t = u^t$ for all m; (3) n_m is odd for all m; (4) $n_{m+1} > n_m$ for all m; and $H_m = H$ for all m. Let $\{\sigma_m^*\}_{m=1}^{\infty}$ be a sequence of team equilibria where $\lim_{m\to\infty} \sigma_m^* = \sigma^*$. Then σ^* is a Nash equilibrium of the underlying strategic form game $[A, u]$.*

The second of these results only considers the limit when team sizes become arbitrarily large.[11] There is no general monotonicity property of convergence to Nash equilibrium. Thus, for example, it is possible for small n that the team equilibrium of a game drifts further away from Nash equilibrium as n increases, even though for large n all team equilibria are approximate Nash equilibria. We provide some examples of this below, which suggest that the typical findings in past experiments where games are played by teams may be rather special and dependent on the specific choice of game parameters in those experiments.

4 Experimental Design and Hypotheses

4.1 Games

The experiment was designed to compare behavior in several variations on prisoners' dilemma and stag hunt games, played between individual players (1-player teams), to behavior in the same games played between 5-player teams and 25-player teams. In the canonical prisoners' dilemma (PD) game, there is a unique strictly dominant strategy equilibrium that yields an inefficient outcome, (Defect, Defect). We also study a modified version of the prisoners' dilemma game that is strictly dominance solvable in two steps. One player has a strictly dominant strategy to defect, while the other player's best reply to defection is to defect and best reply to cooperation is to cooperate. Thus, these modified prisoners' dilemma games also have (Defect, Defect) as the unique rationalizable strategy

[11] For the case of two teams and two actions for each team, this resembles a form of the Condorcet jury theorem, except there is strategic interaction across the two teams. Because heterogeneity within a team is due to payoff disturbances, sincere voting for each team member is a weakly dominant strategy - in contrast to the usual approach with heterogeneous information, where strategic voting considerations can have an effect (Austen-Smith and Banks 1996).

profile and strict pure strategy Nash equilibrium. We call these *weak prisoners' dilemma (WPD)* games.

Stag hunt (SH) games have two strict equal-payoff Pareto-ranked Nash equilibria, with one relatively safe strategy with low coordination payoffs and one riskier strategy with high coordination payoffs. Thus, all of the games we investigate here share a similar properties in that there is an unambiguous, efficient and fair outcome, but they differ in that players face the problem of cooperation in PD and WPD games and the problem of coordination in SH games.

To check for robustness of our results and to test the comparative static predictions of the theory, we include two payoff variations of each of the three classes of games. The two PD games that we study differ in the gain from defection. There is a larger gain from defection in PD1 than in PD2. The two WPD games have a similar difference. In WPD1, the player with the dominant strategy to defect gains twice as much from defection than in WPD2, and the second player (without a dominant strategy) also gains twice as much in WPD1 by defecting in response to defection compared to WPD2. The two SH games differ only in the payoffs from the safe strategy. In SH2 the safe strategy is both riskier and has lower payoffs than in SH1. In both SH1 and SH2, playing safe is the risk dominant equilibrium, while both playing the risky strategy is the payoff-dominant equilibrium. The efficiency gain from playing the payoff-dominant equilibrium in SH2 is more than three times greater than in SH1. The payoff matrices are displayed in Figure 1.

PD1	D	C
D	35,35	98,11
C	11,98	60,60

PD2	D	C
D	28,28	89,11
C	11,89	78,78

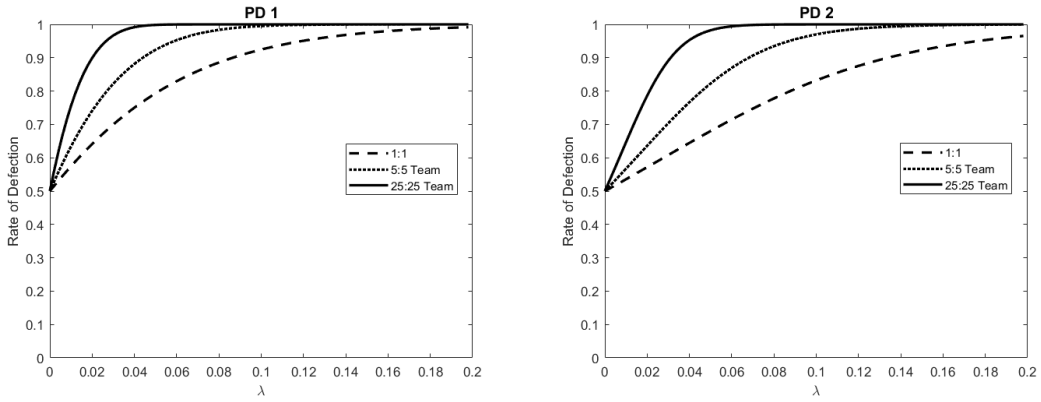
SH1	S	H
S	77,77	13,75
H	75,13	65,65

SH2	S	H
S	77,77	13,66
H	66,13	38,38

WPD1	D	C
D	35,35	77,16
C	16,49	58,58

WPD2	D	C
D	19,19	77,10
C	10,49	68,68

Figure 1: Payoff matrices for the games used in the experiment



(a) PD1 (b) PD2

Figure 2: Team Equilibrium in PD games

4.2 Hypotheses

In this section, we propose hypotheses about behavior in the six games and how it depends on the team sizes, 1:1, 5:5 and 25:25.[12]

4.2.1 Specific Hypotheses for the Prisoners' Dilemma Games

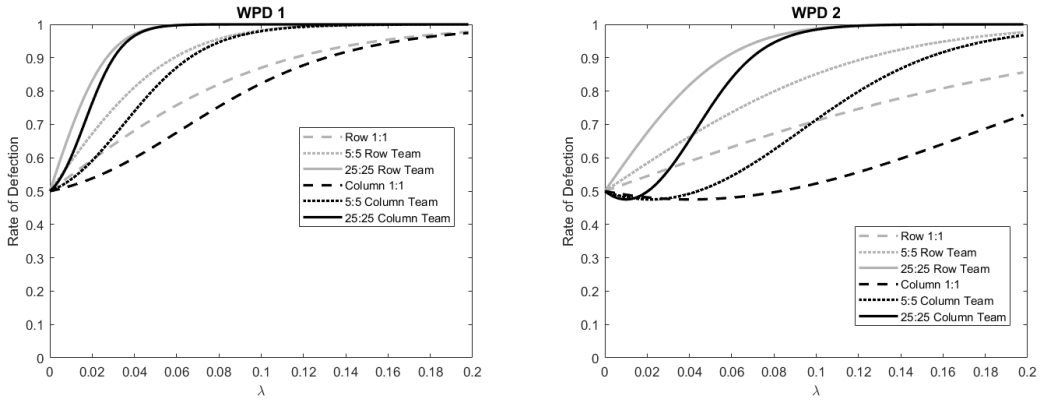
Figure 2 compares the equilibria in the 1:1, 5:5, and 25:25 member team games for the two PD games as a function of the precision of the distribution of payoff disturbances, λ.[13] The precision is on the horizontal axis, and the team equilibrium probability of defection is on the vertical axis.

In both games, the 5:5 team equilibrium defection rates are higher than the defection rate in the 1:1 treatment played by individual players, and the 25:25 team equilibrium defection rates are even higher. Second, for all values of $\lambda > 0$ the equilibrium defection rate is higher in PD1 than in PD2 for both the 5:5 team games and the 1:1 individual games, due to the stronger incentives to defect in PD1. These observations lead to the following hypotheses for the two PD games.

Hypothesis 1. *(a) For all team treatments and the 1:1 treatment, the rate of defection will be higher in PD1 than in PD2. (b) For both PD games, teams will defect more*

[12]For clarity we use the following terminology. "Team" refers to a team in the team treatments. For individual voting behavior in the team treatments, we use the term "team member". Subjects in the 1:1 treatment are called "individuals in the 1:1 treatment".

[13]All figures in this section are created using a logistic distribution of the payof disturbances, $H(x) = \frac{1}{1+e^{-\cdot\lambda x}}$, and graphs the team equilibrium for precisions in the range of $\lambda \in [0, .2]$.



(a) WPD1 (b) WPD2

Figure 3: Team Equilibrium in WPD games

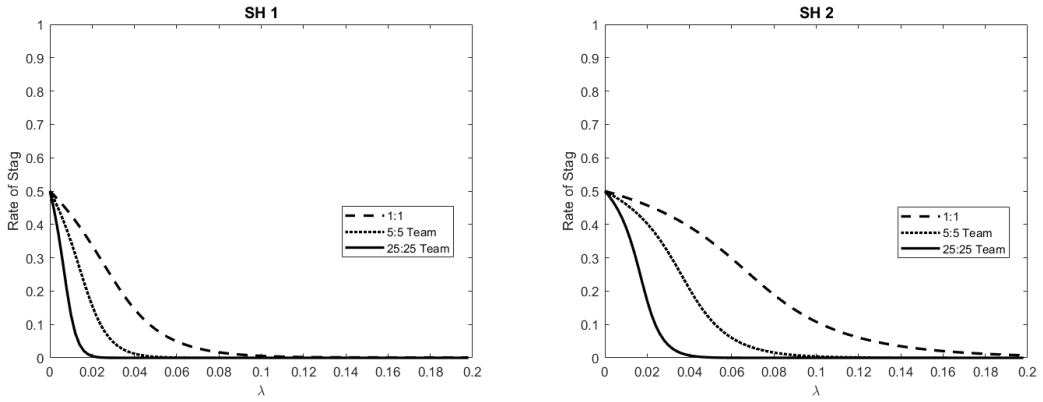
frequently than individuals in the 1:1 treatment, with 25-member teams defecting more than 5-member teams.

4.2.2 Specific Hypotheses for the Weak Prisoners' Dilemma Games

Figure 3 displays the equilibrium in the 1:1 games and 5:5 team games for the two WPD games as a function of the precision of the distribution of payoff disturbances. The equilibrium choice frequencies for row are represented by the lighter curves, with the 25:25 team equilibrium shown as the solid curve, the 5:5 team equilibrium shown as the dotted curve and the 1:1 equilibrium shown as the dashed curve. The corresponding darker curves display the column player's team equilibrium choice frequencies.

The theory generates a number of comparative static predictions. As in the PD games, the incentive to defect is stronger in WPD1 than in WPD2; hence, the equilibrium defection rate is higher in WPD1 than in WPD2 for all team sizes, for both the row and column player, and for all $\lambda > 0$.

This difference between WPD1 and WPD2 is particularly striking for the column player, whose best reply depends on the probability of row defection. The equilibrium defect probability for column is always greater than column's cooperate probability in WPD1, which is not the case in WPD2. For relatively low values of λ, the equilibrium row defection probability is sufficiently close to $\frac{1}{2}$ so that *column's best reply is to cooperate in WPD2*. Hence, for all three team treatments (25:25, 5:5, and 1:1), the model predicts column's equilibrium defect probability to be less than $\frac{1}{2}$. Hence, for these relatively low values of λ, the theory predicts that *equilibrium team behavior is further from Nash*



(a) SH1 (b) SH2

Figure 4: Team Equilibrium converging to risk-dominant Nash equilibrium in the SH games

equilibrium than individuals. In fact, 25:25 teams will defect even less than the smaller 5:5 teams.[14] This is a reversal of the predictions for WPD1, PD1, and PD2. The qualitative effect of team size on row defection frequencies goes in the same direction as in WPD1, PD1, and PD2, with larger teams predicted to defect more than smaller teams.

For both games, there is also a comparison between equilibrium row and column defection rates. For all team sizes and for both WPD1 and WPD2, row players are predicted to defect more than column players, and this holds for all $\lambda > 0$.

Hypothesis 2. *(a) For all team treatments and the 1:1 treatment and for both the row and column players, the rate of defection will be higher in WPD1 than in WPD2. (b) In WPD1, row and column defection rates are increasing with team size. (c) In WPD2, row defection rates are increasing with team size, but column defection rates are decreasing with team size for relatively low values of λ. (d) In both treatments (1:1 and 5:5) of both games, row players/teams will defect more than column players/teams.*

4.2.3 Specific Hypotheses for the Stag Hunt Games

Figure 4 displays the team equilibrium Stag frequencies that converge toward the risk-dominant Nash equilibrium (H,H) for the 1:1 games (dashed curve), the 5:5 games (dotted curve), and the 25:25 games (solid curve) for the two SH games as a function of the error precision. As is clear from the figure, the probability of coordination increases with team

[14]These properties of team equilibrium for low values of λ are not special to the logit distribution of payoff disturbances, and hold for all H functions.

size along these curves for both games and for all $\lambda > 0$. There is also a second component of the team equilibrium graph (not shown) corresponding to the payoff-dominant (S,S) equilibrium that exists for sufficiently large values of λ. This second component converges to the (S,S) equilibrium as λ goes to infinity, and the probability of coordinating on (S,S) in the two SH games increases with team size on this component.

The intuition for the observation that larger teams are more likely to coordinate than smaller teams can be understood in terms of stability of the two pure strategy equilibrium. Consider a stag hunt game in which the individual member vote frequency for Stag is p_n^t in a team t with the team size n. For a large n, an interesting observation is that p_n^t must converge to 0.5. Suppose to the contrary $p_n^t \neq 0.5$ in the limit. Then, it is easy to see that the team choice probability will converge to either 0 or 1 by the consensus effect. That is, a small change in the individual member vote frequency will push the team strategy toward a pure strategy equilibrium. Since the boundary limits of this dynamic are indeed stable Nash equilibria, this leads to the hypothesis that teams will coordinate more successfully than individuals.[15]

Hypothesis 3. *In both SH games, the rate of coordination will be higher in the 5:5 team treatment than in the 1:1 treatment, and higher yet in the 25:25 team treatment.*

5 Laboratory Experiment: Procedures

The experimental sessions were conducted at the Experimental Social Science Laboratory (ESSL) located on the campus of the University of California, Irvine. Subjects were recruited from the general undergraduate population, from all majors. Experiments were conducted using ZTree software (Fischbacher, 2007). We conducted 22 sessions, using a total of 420 subjects. No subject participated in more than one session. The experiments lasted, on average, one and a half hours, and subjects' average earnings were $28.15, including the $8 show-up fee (max. $38 and min. $18).

Each session was divided into two phases. In Phase 1 of each session, each subject made a single individual decision for each of the six games over a sequence of six rounds. For each game, the subject was randomly matched with another subject to determine

[15] This stability argument does not address which of the two pure strategy equilibria is more likely to be played by teams. Intuition suggests that it would depend on the relative coordination payoffs for S and H and the risk associated with S. For our two games, this dependency would suggest that coordination in the payoff-dominant equilibrium is more likely in SH2 than SH1, although we do not state this as a hypothesis.

the payoff, and subjects were informed of this fact. That is, Phase 1 consisted of the six one-shot games played once each in sequence by 1-person teams. Subjects received no feedback about the play of their randomly selected opponent until the very end of the session. One round in Phase 1 was randomly selected at the end for subject payment. We also used two different sequences of presenting the games, with half of the sessions using one sequence and half of them a different sequence.[16]

Phase 2 of each session was run using one of four different treatment conditions. In each of these treatment conditions, subjects played each of the games for 5 rounds with anonymous random rematching and feedback between rounds, for a total of 30 rounds. That is, each subject would play five rounds of game X, followed by 5 rounds of game Y, and so on. In four sessions, called the *1:1* treatment, the games were played exactly as in Phase 1, except that there was feedback after every round. That is, the games were played by 1-person teams that were randomly rematched after each round. Phase 2 of the three "team" treatments, with 6 sessions of 20 subjects for each treatment, which were played by 5-person teams.[17] Team membership was always reassigned at the beginning of each round by randomly dividing the 20 subjects into 4 different teams, and these four teams for the round were then randomly matched into pairs of teams, with team playing one of the roles in the game.[18] Subjects were fully informed of this matching protocol. The three team treatments differed in the collective choice procedure used by teams to reach a decision on an action choice. In the *Majority* condition, each subject voted for one of the two actions, and for each team, the action that received a majority of votes by members of the team was implemented as the team's action. In the *Poll* condition, the majority vote was preceded by a straw vote. [19] In the *Chat* condition, the majority vote was preceded by computerized chat among the members of each group.[20]. For all of the team treatments, the vote totals for both teams were included in the feedback after each round. For payment in Phase 2, exactly one round of each of the six games played in part

[16]We used the following two sequences: PD2-SH1-WPD1-PD1-SH2-WPD2 and WPD2-SH2-PD1-WPD1-SH1-PD2. Each session also included additional rounds for strictly competitive games with a unique mixed strategy equilibrium.

[17]The 25:25 treatment was not feasible to run in the laboratory, but was run in the MTurk environment. See Section 7.

[18]In the asymmetric WPD games, each subject's role was fixed for all five rounds of the game. Hence, rematching for the 5-person team sessions was conducted with two populations of 10 subjects in each role.

[19]Before voting for an action, the subjects received feedback about their team's straw vote totals.

[20]The length of time allowed for chat was 80 seconds for the first round and 40 seconds for rounds 2-5 of each game.

two was randomly selected for payment.[21]

The reason for employing two phases was to allow for a direct comparison within the team treatments between naive decisions in the 1-person team games (Phase 1) and naive decisions in the 5-person team games (Round 1 of each game in Phase 2). For this reason there was no feedback in Phase 1. This comparison was not relevant for the 1:1 treatment. The reason for having 5 plays of each game in Phase 2 is that our benchmark theory is an equilibrium theory, so we expect learning to occur.

The games were presented to subjects in matrix form, exactly as displayed in Figure 1 with three differences. In all of the games, the top action for rows was always labeled action A, and the bottom action was always labeled action B. Similarly, the left action for the column player was always labeled action A and the right action was always labeled action B. Second, the matrices were configured so that all subjects made decisions as row players, so in the asymmetric games, the payoff matrix for column players was displayed as the transpose of the matrix shown in Figure 1. Third, each subject saw a matrix that randomly flipped the payoffs entries for actions A and B and reversed the left and right actions to control for possible framing or labeling effects (such as a bias to choose "A" or top). For the 1:1 treatment, four sessions were conducted with a total of 60 subjects. For each of the three team treatments, six sessions were run with 20 subjects in each session.

6 Laboratory Experiment: Results

The strategic considerations in SH games are quite different from those in PD and WPD games, especially because communication within a team could have different effects in the case of multiple equilibria. Therefore, we analyze the results separately, beginning with the four prisoners' dilemma variants, followed by an analysis of the results from coordination games. In each subsection, our primary focus is given to the difference between the action choices in the 1:1 games and the team choice frequencies in the 5:5 games of Phase 2.

[21] A sample copy of the instructions for the *Poll* treatment appears in the Appendix. Instructions for the other three treatments are similar.

6.1 PD and WPD Games

6.1.1 Effects of the collective choice procedure

We begin the data analysis by reporting the effects of the different voting procedures on behavior. The purpose of examining the effect of voting procedures is beyond testing the robustness of majority voting in team games. Given the fact that most of previous experiments use unanimity voting through face-to-face or chat communication, it is important to know whether any difference between the result of our experiment and that of the previous experiments is due to the different voting rules (majority vs. unanimity) or to the existence of communication opportunities in the collective decision making procedure. Since the poll and chat treatments differ from the majority treatment only to the extent that team members are able to communicate with each other before making a final voting decision, comparing voting procedures will allow us to examine the effect of communication on team choice frequencies in a minimal manner.

Table 1: Comparison of team defection frequencies in PD and WPD games under different Voting rules (Phase 2)

Game	Majority	Poll	Chat
PD 1	0.90	0.93	0.90
PD 2	0.66	0.74	0.58*[1]
WPD 1 Row	0.85	0.90	0.88
WPD 1 Column	0.83	0.85	0.85
WPD 2 Row	0.62	0.48	0.42*[2]
WPD 2 Column	0.43	0.28	0.35

* Significant at the 5% level.
[1] Poll vs. Chat.
[2] Majority vs. Chat.

The effect of the collective choice procedure on team decision frequencies Table 1 displays the final team defection frequencies across the three voting procedures in different PD games for all rounds of Phase 2. The table shows that communication has little, if any, effect on the team defection frequencies. We conducted three pairwise comparisons of the frequencies (Majority vs. Poll, Majority vs. Chat, and Poll vs. Chat) for each row. There are only two statistically significant differences[22] out of a total of

[22]Unless otherwise noted, all significant differences reported in this paper are at the 5% level, using Fisher's exact test.

18 comparisons. However, there is a high likelihood of false positives since 18 separate tests were simultaneously conducted. There are several ways to correct for this. Applying either Bonferroni's or the Dunn-Sidak correction for multiple comparisons, we cannot reject the null hypothesis that none of the comparisons are significant. Alternatively, we fail to reject the null using the BH step-up procedure to control for false discoveries (Benjamini and Hochberg, 1995).

This leads to the first result: there is little, if any, effect of the voting procedure on the final team decisions. Hence, in the remainder of the analysis of the PD and WPD games, we pool the team decision frequencies across the three treatments.

Result 1. *There is no significant difference in team defection frequencies across the three collective choice treatments in the PD and WPD games.*

The effect of communication on consensus formation In addition to considering the effect of collective choice procedures on the team decision frequencies, there is a separate (possibly related) issue regarding the effects of collective choice procedures on voting and consensus formation. The relevant procedural variable to address this question is the deliberation process leading to the team decision choice. Our experimental design, with three different deliberation processes - no deliberation, straw votes, and pre-vote discussion - allows us to provide some insight into this issue. For this purpose we consider two direct measures of consensus formation: the distribution of margins of victory in the elections in both the poll treatment and the chat treatment, compared with the no-deliberation treatment; and vote-switching between the straw vote and the final vote in the poll treatment.

One-sided votes In the team treatments, we have three collective decision-making rules to determine team behavior—majority voting, majority voting after poll, and majority voting after chat. The observation from the previous section is that the decisions of teams do not significantly differ across the collective choice treatments. A possible explanation is that, even though the team decision frequencies are the same, the vote results are more one-sided (closer to unanimous) in the Poll and the Chat treatments than in the Majority treatment. This is a subtle form of an equilibrium effect, which can be measured in terms of individual vote decisions, even if team decision probabilities are not significantly affected. Thus, we first examine the distribution of the one-sided votes.

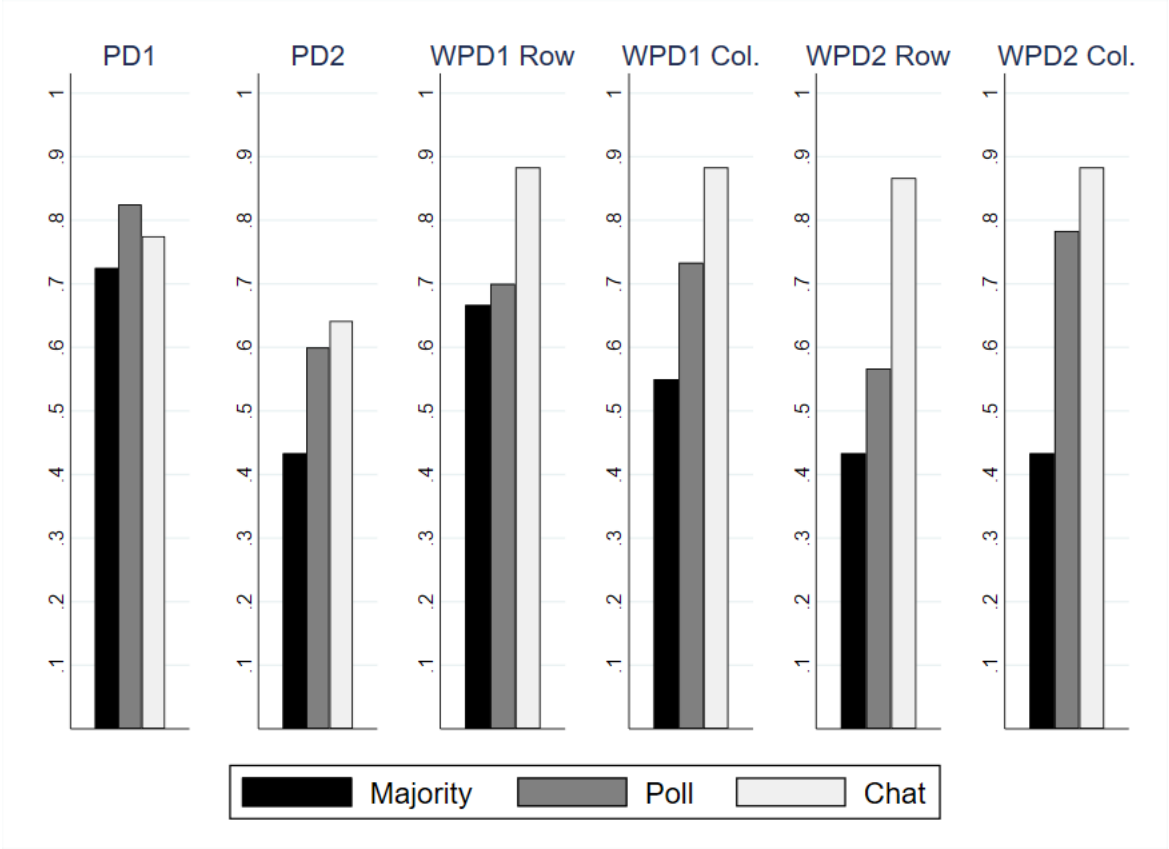


Figure 5: Frequency of one-sided votes

We define a *one-sided vote* as occurring if the vote on the team decision is decided by supermajority, i.e., either a 4-1 margin or a 5-0 margin. Measuring the frequency of one-sided votes provides a simple way to check whether the three different collective decision-making procedures produce different levels of consensus in teams. Figure 5 shows the distribution of one-sided votes across games and team treatments. In all cases, one-sided votes are less frequent in the Majority treatment without communication than either of the treatments with communication. Furthermore, in most comparisons the effect of communication on consensus is statistically significant.[23] Finally, we note that in three out of four games one-sided votes occurred more frequently in the Chat treatment than

[23]The three exceptions where the effect was not significant were all in the prisoners' dilemma games with strong incentives to defect (PD1 and WPD1). Recall that in these games the vast majority of subjects voted to defect, so there were many one-sided votes in all team treatments. Specifically in PD1, the comparisons between majority and chat and majority and poll showed no significant difference in one-sided votes ($p = 0.46$ and $p = 0.09$, respectively). In WPD1, the comparison between majority and chat was significant ($p < 0.01$) but the comparison between majority and Poll was not significant ($p = 0.10$).

Table 2: The rate of switching behavior in the poll treatment

Game	Majority	Minority
PD1	0.05	0.36
PD2	0.08	0.30
WPD1	0.04	0.34
WPD2	0.06	0.26

in the Poll treatment, and this difference is significant in both WPD games.

Result 2. *The effects of straw polls and pre-vote chat communication lead to more group consensus than in the Majority treatment. Open chat communication leads to greater consensus than straw polls.*

This result indicates that individual voting behavior in teams is affected by the collective decision making rule, and in particular, communication leads to more agreement or consensus in teams, even if the effect of the pre-vote communication on the final decision by the teams is weak (Result 1). Thus we do not find that procedures that lead to greater consensus ultimately lead to different outcomes, at least for these simple games. Such a conclusion is not supported by our data, which reveal no significant effects of the collective choice procedure on strategy choice frequencies decided by teams.

Bandwagon effects The Poll treatment allows for a simple quantitative test of the *process* by which consensus is reached in a team. This provides another piece of evidence and additional insight into consensus building in teams. Given that we have data for both the straw vote and the final vote of each team member, this sequence allows us to measure the extent to which an individual team member's final vote choice is affected by the feedback from outcomes of the straw vote. We say that a *bandwagon effect* occurs if team members who cast a minority vote in the straw poll switch to vote with the majority in the final vote.[24] To control for possible random effects of individual vote switches, we compare this to switches from the majority to the minority. We classify team members as *majority* and *minority* poll voters if their choice in the straw vote agrees or disagrees, respectively, with the opinions of majority straw vote outcome in their group.

Table 2 shows the rate of switching behavior in the poll treatment depending on whether a voter was in the majority or the minority in the straw poll. In every game,

[24]There are several possible explanations for why a voter might switch between the poll and the final vote. For example, Callander (2007) shows that bandwagon effects can arise if voter have a direct preference for conformity, in the sense of preferring to vote with the winning side.

the differences are systematic, highly significant ($p < 0.01$), and large in magnitude. The probability that a minority poll voter switches to the majority side across the four games ranges from four to eight times greater than the probability that a majority voter switches to the minority side. Between 26 and 36 percent of the minority voters switched their straw vote to the majority side in the final vote, while only 4 to 8 percent of majority voters switched to the minority. This result demonstrates a clear and strong bandwagon effect indicating a convergence toward consensus via straw polling. The bandwagon effect also explains why we observe more one-sided votes in the poll treatment than in the majority treatment.

Result 3. *There is a significant bandwagon effect in the PD and WPD games.*

Taken together, we conclude that the effect of the voting procedure is prominent on individual voting behavior in teams, but the final team decisions are robust to the voting procedure. Hence, in the remainder of the analysis of the PD and WPD games, we pool the team decision frequencies across the three treatments.

6.1.2 Testing Hypotheses 1 and 2 about predicted behavior in the PD and WPD games

Figure 6 displays the frequency of 5:5 team defection rates and individual 1:1 defection rates in the PD and WPD games. For the WPD games, which are asymmetric, the defection rates are reported separately for row and column players.

We first compare defection rates between PD1 and PD2 and between WPD1 and WPD2. Team equilibrium predicts higher defection rates in PD1 than in PD2 for both the 1:1 treatment and the 5:5 treatment (Hypothesis 1a), and this is supported in the data, where these differences have the correct sign and are statistically significant in all cases ($p < 0.01$). Team equilibrium also predicts higher defection rates in WPD1 than in WPD2 for both the 1:1 treatment and the 5:5 treatment *and* for both row and column players (Hypothesis 2a), and this is supported in the data, where the differences have the correct sign in all cases. These differences are statistically significant ($p < 0.01$) except for the comparison of row players in the 1:1 treatment ($p = 0.059$).

We next compare defection rates between the 1:1 treatments and 5:5 treatments. Surprisingly, we find that teams in the 5:5 treatment do not always defect more than individuals in the 1:1 treatment. Rather, the effect of team size depends on the details of the payoff matrix, and in particular on how strong are the incentives to defect. For

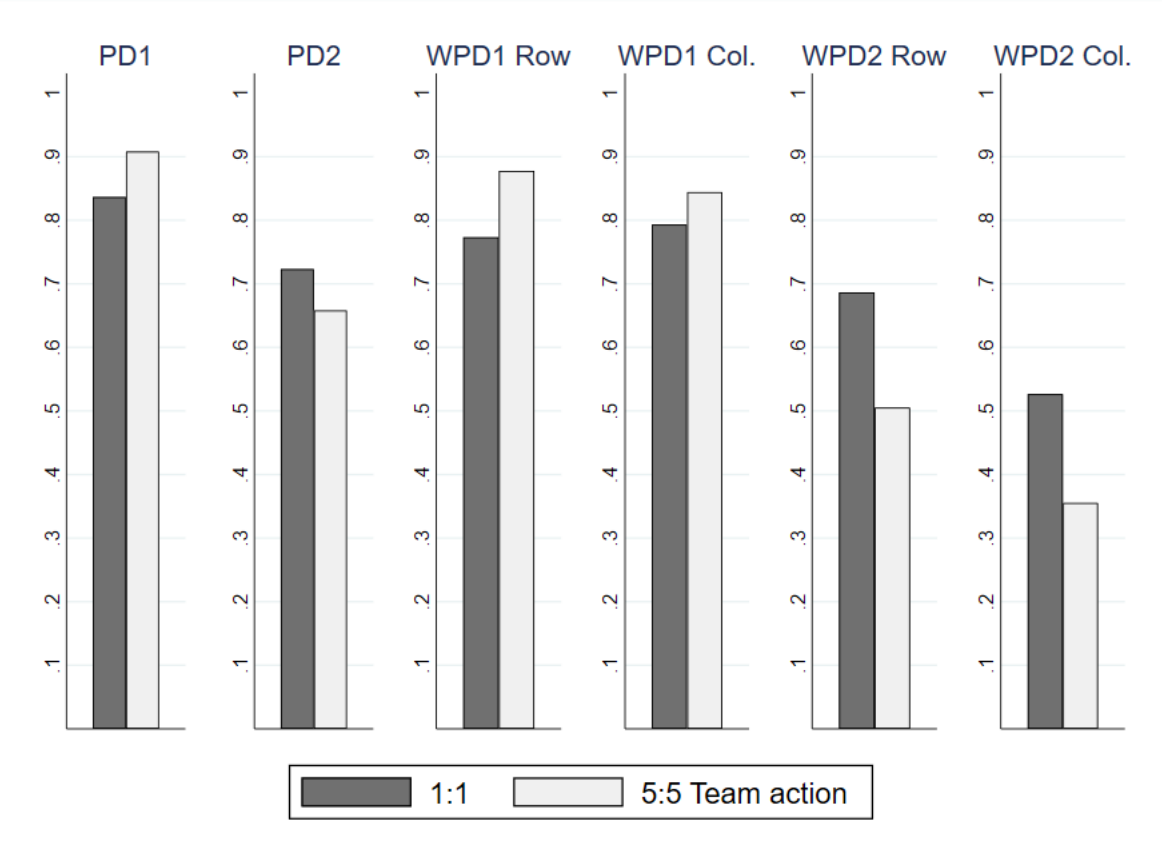


Figure 6: Comparison of 1:1 and 5:5 defection frequencies.

PD1 and WPD1, the Nash equilibrium strategy of defect is chosen more often by 5-member teams than in the 1:1 treatment, and these differences are statistically significant except for the WPD1 column players (p-values: PD1$<$ 0.01, WPD1 Row$=$ 0.013, WPD1 Column$=$ 0.25). This finding is consistent with findings from previous one-shot PD experiments that generally found that teams were closer to Nash equilibrium, and provides support for Hyotheses 1b and 2b.

However, in the PD2 and WPD2 games, where the payoffs provide weaker incentives to defect, we find the *opposite*. For those games, we find *less* defection in the 5:5 treatment than in the 1:1 treatment. This finding for PD2 and for row players in WPD2 contradict Hypothesis 1b, although the difference is not statistically significant at the 5% level ($p = 0.077$). The effects in WPD2 are statistically significant (p-values: Row$<$ 0.01, and Column$<$ 0.01) but this does not contradict Hypothesis 2c, where this "reversal" is predicted if λ is not too high. In fact, in WPD2, column players have a best response to cooperate if they expect row players to cooperate at least 32% of the time, which is

consistent with the data, where row teams cooperate 49% of the time. So, given the behavior of row players in the experiment, cooperation was in fact the best reply for column players, which can explain why column teams in WPD2 cooperate nearly twice as often as they defect. In contrast, column players in the 1:1 treatment cooperate only half the time.

The data also provides partial support Hypothesis 2d, that row teams defect more than column teams in the WPD games. This comparison significantly holds for WPD2 for both the 1:1 treatment and the 5:5 treatment ($p < 0.01$). However, in WPD 1, row and column player behaviors for both the 1:1 treatment and the 5:5 treatment are essentially the same and the differences are not significant ($p = 0.39$ and $p = 0.22$, respectively).

Recall from the discussion of past experiments that, in nearly all past experiments that have found differences between team and individual play in games, team play was closer to Nash equilibrium than individual play. In fact, this observation was the original motivation for our theoretical paper. However, a result of that theoretical paper is that, while team play is predicted to be closer to Nash equilibrium in games with very large teams, it can go either way for small teams, depending on the exact parameters of the game and the error rates. That is what we find in our games.

Result 4. *In PD and WPD games, there are significant differences between team decisions in the 5:5 treatments and individual decisions in the 1:1 treatment. Teams are closer to Nash equilibrium in games with strong incentives to defect but not in games with weaker incentives to defect.*

6.2 Estimating the logit team equilibrium model from the PD and WPD data

The analysis of the data from PD and WPD games above examined team equilibrium predictions and comparative statics about the *qualitative* effect of payoffs, team size, and collective choice rules. In fact, the logit equilibrium specification also makes precise quantitative predictions about behavior, as a function of the logit error parameter, λ. We know, for example, that for very low values of λ in the neighborhood of 0, there are essentially no effects of team size, payoffs, and the collective choice rule, since all choice probabilities (vote probabilities in the team treatments) are approximately 0.5. At the other extreme, when λ is very large, the predictions also coincide for all payoff and team treatments, i.e., defect choice probabilities by both row and column players

should equal 1 in all treatments, since that is the Nash equilibrium in all cases. However, for intermediate values of λ, as typically found in experimental data with logit QRE estimation, the equilibrium defect choice probabilities differ greatly across the treatments, and are predicted to differ between row and column players in the WPD games. Thus, it is natural to estimate the error parameter for the logit team equilibrium model to see if the predicted defect probabilities at that estimated value match the patterns found in the data across payoffs, player roles (in the WPD games) and team sizes.

To avoid overfitting, we estimate *a single* logit error parameter using the row and column data from all payoff and team size treatments, in the following way. For each non-negative value of $\lambda \in [0, \infty)$, there is a unique logit equilibrium probability of defection for each payoff and team treatment.[25] For each λ and for each payoff matrix g ($PD1, PD2, WPD1, WPD2$), team size treatment s (1 or 5) and team t (row or col), denote the equilibrium choice (or vote) defection probability by $p^*_{gst}(\lambda)$. Given the observed defection frequencies in our data, f_{gst} and the number of observations of each gst, n_{gst}, we can write the log likelihood function as follows:

$$L(f|\lambda) = \sum_{g,s,t} \left[f_{gst} \ln p^*_{gst}(\lambda) + (n_{gst} - f_{gst}) \ln(1 - p^*_{gst}(\lambda)) \right].$$

The value of λ that maximizes $L(f|\lambda)$ is solved for computationally, which involves an inner loop for each value of lambda to compute the unique team equilibrium, $p^*_{gst}(\lambda)$ for each gst and for each value of λ.[26]

The estimated value of λ is 0.057. The results are presented in Table 3, with the treatment/role in the first column, the observed relative frequency of defection in the second column, and the model estimated equilibrium defection rates in the third column.

Figure 7 displays a scatter plot of the observed relative frequencies of defection for each gst. The 5x5 (1x1) data and estimates are shown as triangles (squares). As a contrast, the figure also displays (as circles) the scatter plot of observed vs. Nash defection probabilities, where all the Nash defection probabilities equal 1. The team equilibrium model, with only one free parameter (λ) provides a good fit to the data across all these different treatments.

[25] For the 1:1 games, the unique team equilibrium coincides with the unique logit QRE.

[26] The 5x5 data on individual vote choice has six times as many observations as the 1x1 data on individual vote decisions. In order to avoid inflating the significance and to avoid overweighting the 5x5 vote data, the number of 5x5 observations and defect frequencies is deflated by a factor of six in the estimation, which equals the number of observations in the 1x1 treatments.

Table 3: Team Equilibrium Estimation Results

Game and Role	Team Size	Observed	Estimated
PD1	1×1	0.84	0.82
	5×5	0.81	0.80
PD2	1×1	0.72	0.70
	5×5	0.61	0.71
WPD1Row	1×1	0.77	0.75
	5×5	0.79	0.75
WPD1Col	1×1	0.79	0.66
	5×5	0.75	0.71
WPD2Row	1×1	0.69	0.63
	5×5	0.49	0.63
WPD2Col	1×1	0.53	0.48
	5×5	0.38	0.52

A linear regression of all the points in the figure (dashed line) produces an $R^2 = 0.69$.[27]

6.3 Coordination games

6.3.1 Effects of the collective choice procedure in coordination games

As with the previous analysis of the PD and WPD games, we first examine the effects of the collective choice procedure on team behavior in coordination games. Table 4 represents different aspects of team decision frequencies across the three voting procedures in the coordination games.

In contrast to the PD and WPD games, there are significant effects of communication on team behavior in coordination games. The first two rows show that within-team communication via a poll leads to higher frequencies of Stag in team decisions compared to simple majority rule without pre-play communications, while chat does not have significant effects. However, the rate of one-sided votes is the highest in the chat treatment, implying that greater consensus has resulted through chat. While the rates of coordination in the poll and the chat treatments are not significantly different from that in the Majority treatment in SH 1, pre-play communication through either via a poll or chat

[27]Social preferences could also account for some of the observed cooperation, so we also estimated a two-parameter model, with an additional *altruism* parameter, A, in addition to the error parameter, λ. The altruism parameter is statistically significant at the 1% level, but the fitted estimates of the defection probabilities change only slightly. The altruism model and the resulting scatter plot figure (similar to Figure 7) is in the appendix.

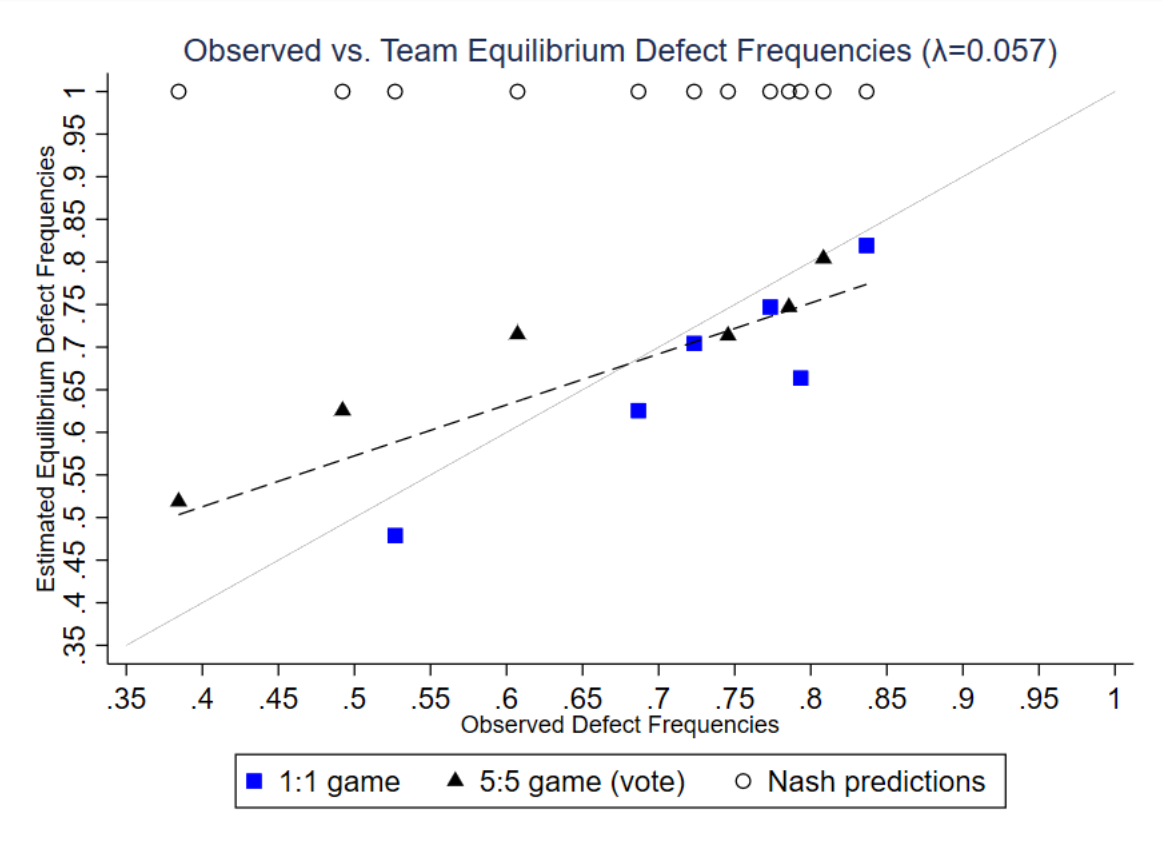


Figure 7: Fitted equilibrium defect frequencies vs. observed defect frequencies.

significantly increases coordination rates in SH 2 where Stag is relatively less risky.[28] As shown in the last two rows, the Poll treatment is significantly more successful in achieving the payoff-dominant (S,S) outcome conditional on coordination than the Majority treatment without communication, but the Chat treatment has no significant effect on equilibrium selection.

In terms of the effect of polls on consensus building within teams, the findings are similar to the PD and WPD games. There was a clear bandwagon effect in the poll treatment for both coordination games. In SH1, 18% of the team members who voted in the minority in the poll switched to the majority in the final binding vote, but only 3% of the team members who sided with the majority in the poll switched their vote, a difference that is significant at $p < 0.01$. In SH2, the respective switching frequencies are

[28]It might be tempting to dismiss this finding as trivial, since pre-play communication in coordination games *between teams* can obviously improve coordination by allowing correlated strategies. However, communication in our experiments is limited to *within-team* communication, either by a poll or by chat. There is never communication allowed *between* teams in any of our 5:5 treatments.

Table 4: Comparison of Team decisions in Stag Hunt Games (Phase 2)

	Game	Majority	Poll	Chat
Stag	SH 1	0.36	0.51*[1]	0.36
	SH 2	0.63	0.81*[1]	0.68
One-sided votes	SH 1	0.59	0.57	0.69
	SH 2	0.60*[3]	0.81	0.87
Coordination	SH 1	0.65	0.55	0.72
	SH 2	0.53*[3]	0.82	0.88
Stag-Stag (S,S)	SH 1	0.28	0.52	0.30
	SH 2	0.75	0.88*[2]	0.70

*Significant at the 1% or 5% level.
[1] Poll vs. Majority/Chat.
[2] Poll vs. Chat.
[3] Majority vs. Poll/Chat.

29% and 4%, also significant at $p < 0.01$.

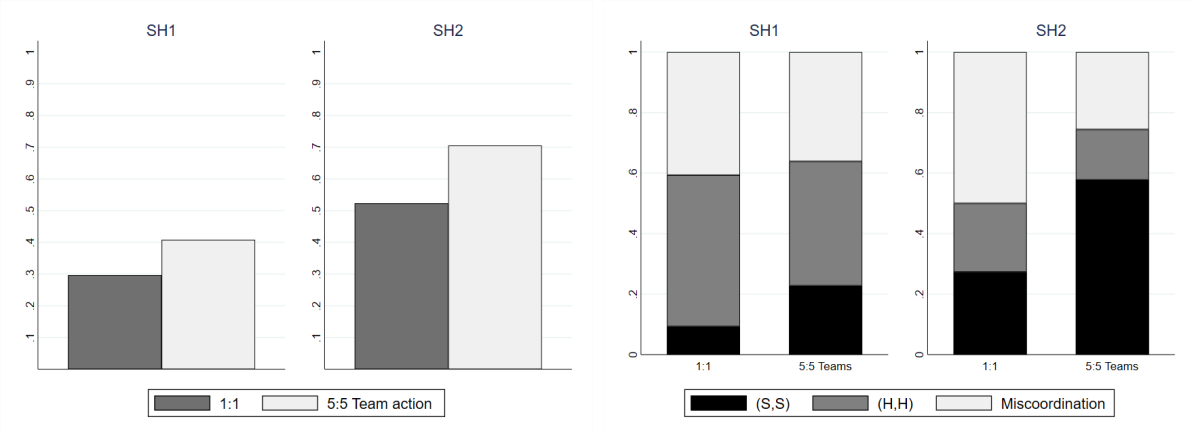
6.3.2 Comparison between 1:1 behavior and 5:5 team behavior in coordination games

For the two stag hunt games in our laboratory study, we compare 1:1 individual choices and 5:5 team decisions in three different ways: (1) Are there differences in the frequency of choosing the riskier strategy (stag) associated with the payoff-dominant Nash equilibrium? (2) Are there differences in the frequency of successful coordination on either the risky or the safe Nash equilibrium? (3) Conditional on coordination, is there a difference in the frequency of the payoff-dominant Nash equilibrium?

The left panel of Figure 8 shows the action choice frequencies in the 1:1 and team treatments. In both games, teams are significantly more likely to choose the riskier strategy than individuals in the 1:1 treatment ($p < 0.01$).[29]

We next turn to the questions concerning *coordination* and *equilibrium selection*, questions (2) and (3). The right panel of Figure 8 compares the coordination rates between the 1:1 and 5:5 treatments. In both coordination games, teams are more successful in coordinating on a symmetric equilibrium, either (H,H) or (S,S). The difference is highly significant ($p < 0.01$) in SH2, while the difference is not significant in SH1.

[29]In SH1, if we exclude the data from the Poll treatments, the qualitatively similar results still hold, but the difference between the choice frequencies in the 1:1 treatments and the two team treatments (Majority and Chat) becomes insignificant ($p = 0.139$).



(a) Rates of Stag in SH (b) Rates of Coordination in SH

Figure 8: Team behavior and coordination in the SH games

Result 5. Coordination: *Teams coordinate more frequently than individuals.*

The right panel of Figure 8 also addresses the equilibrium selection question. Teams are more likely to coordinate on (S,S), conditional on coordination. The differences are significant and large in magnitude for both games. In SH1 the (S,S) coordination rate is 36% in the 5:5 treatments and 20% in the 1:1 treatment ($p < 0.01$), and in SH2 the (S,S) coordination rate increases from 57% to 78% ($p < 0.01$).[30]

A Stag decision in SH2 is less risky and therefore relatively more attractive than in SH1. Interestingly, given that the frequency that teams choose stag is much higher in SH2 than in SH1 (0.71 vs. 0.41, $p < 0.01$)[31], the rates of coordination are quite similar between the two SH games. Specifically, for both individuals in the 1:1 treatment and teams, coordination rates in SH1 compared to SH2 are not significantly different ($p = 0.66$ and $= 0.13$, respectively). However, the payoff-dominant coordination outcome is far more frequent in SH2 than in SH1, conditional on coordination, and such differences are significant for both individuals in the 1:1 treatment and teams ($p < 0.01$). That is, in our experiment, the changes in the riskiness of efficient coordination affect the content of coordination (intensive margin), but not the overall rate of coordination (extensive margin), and this principle applies to both individuals and teams.

Result 6. Equilibrium selection: *Conditional on successful coordination, teams are more likely than individuals to coordinate on the payoff-dominant equilibrium, (S,S), than (H,H),*

[30]The statistical significance does not depend on the inclusion of the data from the Poll treatments.

[31]Team member vote frequencies in Phase 2 of SH1 and SH2 are 0.42 and 0.70, respectively. This difference is also significant ($p < 0.01$).

compared with individual play in both stag hunt games.

7 The MTurk Experiment

In a laboratory experiment, because of the limited subject capacity of laboratory facilities, it was not feasible to implement the 25:25 treatment.[32] To overcome these difficulties, we conducted sessions through Amazon Mechanical Turk (MTurk) with many subjects and team sizes as large as 25 members per team.[33]

While the MTurk platform allows us to run the same 2×2 games with much larger teams at an affordable cost, it also presents a number of logistical challenges. The basic design of our MTurk experiment is as comparable as possible to the design of Phase 2 of the 1:1 and the Majority treatments in the lab experiment: there are 5 rounds, and in each round a subject makes one decision for each game, with feedback about the outcome of a previous round.[34] There were a number of hurdles to overcome. The major hurdle, which required a novel implementation of MTurk games, was having multiple rounds with random matching and feedback between the rounds. The games in the laboratory involve simultaneous interactions among subjects with immediate feedback about outcomes, which is difficult in the MTurk environment because it is impractical, or even impossible, to synchronize the choices of all MTurk subjects within the short (less than one minute) decision time frame that is typical in the laboratory. Our solution for having five rounds with feedback is to run a longitudinal experiment over five consecutive days - Monday through Friday. Thus, we have one round on each day, and each subject makes one decision for every game on each of these days. We first describe the experimental procedures and then explain the details of how we implemented the longitudinal experiment.

[32]Experimental laboratories typically have 20-40 computers, and the maximum team size is about 10 in symmetric games if games are played for multiple rounds with random rematching (or 5 in asymmetric games). Moreover, obtaining a large sample of team games could also bump up against capacity constraints with respect to the subject pool.

[33]Provided by Amazon, MTurk is an online labor market platform where a number of experimental tasks have been done. There have been many experiments in economics conducted in MTurk, for instance, see Horton et al. (2011).

[34]We only ran the Majority treatment because the logistics of the Poll treatment would have been too complicated on MTurk, and the Chat treatment is not feasible on MTurk.

7.1 Design and Procedures

An experimental session consists of 5 rounds (i.e., 5 days), with each session using three different team sizes—1:1, 5:5, and 25:25.These five rounds corresponded to the Phase 2 part of the laboratory experiment.[35] At the beginning of the first day of a session, subjects were randomly assigned to one of three team sizes (1, 5, or 25), and these team size assignments remained fixed throughout the session.[36] The games that were played in each round had the same payoff structure as those in Figure 1[37]

After voting for an action in Round (day) 1, instantaneous feedback about the outcomes was not feasible. Instead, immediately before voting for an action in each of the 8 games in Round (day) 2, each subject was informed about the actions chosen by their Round 1 team and the Round 1 team with which they were matched. That is, the feedback about each game played in Round t was presented to subjects immediately before playing that game again in Round $t + 1$. For feedback in the 5:5 and 25:25 treatments, subjects were also told the vote totals of their Round t team and the Round t team with which they were matched, as in the laboratory sessions.

In contrast to the lab experiment in which the experimenter has full control over the number of subjects, it is not possible to guarantee an exact number of subjects in the MTurk environment, which could potentially lead to difficulties in constructing teams.[38] Therefore, we used the following procedure for randomly assigning teams on each day. After closing each day, for each individual, a computer program randomly (1) selected his/her team members (in 5:5 and 25:25 treatments) and (2) randomly formed a second team of equal size from the remaining population of subjects who were assigned the same team size. This process created random teams for each individual in a way that does not require knowing the exact number of subjects on each day.[39] This procedure also helps

[35]The MTurk design did not include the Phase 1 part.

[36]Subjects were informed of their own team size and that they were always matched randomly in each round (day) with another team of the same size.

[37]One detail of the sequencing of games was slightly different in the MTurk experiment, necessitated by the logistics of feedback between plays of each game. In Phase 2 of the lab experiment, subjects played each game for 5 rounds before moving on to the next game of the sequence. In the MTurk experiment, subjects played 5 rounds (days), making a decision for each game in each round.

[38]There are two main reasons for this problem. First, even if we target a specific number of subjects in MTurk, we might not end up with those numbers since some subjects submit an incomplete task. Such subjects were not eligible for the next day. Second, due to attrition between days, it is even more difficult to predict how many subjects will participate in Round 2 and thereafter.

[39]This meant that there was some overlap in the membership of different teams, but this did not seem problematic because of the very large number (hundreds) of subjects in each session and the fact that team assignments were random. In the team treatments, subjects were told that they would be randomly

us to prevent our results from being biased by the formation of specific teams with large team sizes in the sense that it creates as many teams as individuals in each treatment, implicitly bootstrapping the team data based on individual voting decisions.

On Monday, subjects were recruited from our task posting on MTurk, and subjects who met the following conditions were eligible to view our posting: (1) reside in the U.S.; and (2) approval rate is higher than 90%. At the time of recruiting on Monday, it was clearly announced that this experiment would continue for five days. Upon accepting the task posting, subjects received a link to our online survey page implemented by Qualtrics. Then they were assigned a team size, read through the instructions of the experiment, and were asked to vote for an action in each game.[40] On Tuesday and thereafter, subjects who participated in the previous day received an invitation message from the internal MTurk message system and completed their task.[41]

An obvious concern for a longitudinal experiment is how to minimize attrition. We used two incentives for continued participation that were initially implemented successfully in an earlier multi-day MTurk experiment (Kim, 2022). One incentive was a stick, and the other was a carrot. The stick is that if a subject failed to return on the next day of the experiment, then he or she was not allowed to participate in the remaining days of the experiment. The carrot was a completion bonus on top of all earnings from the experiment. Specifically, a subject received a completion bonus of $3.00 (approximately doubling his or her earnings from the experiment) for full participation in all five days of the experiment. These two incentives were explicitly announced at the time of recruitment.

We ran two sessions, which began on two different Mondays. The first session recruited 509 subjects and the second session recruited 407 subjects. Subjects who participated in the first session were excluded from participating in the second session. For each session, subjects were randomly assigned on day 1 to different team sizes: 1, 5 or 25. Table 5 summarizes the number of subjects that were assigned to each team size on each Monday.[42]

assigned to a team and their team would be randomly matched with another team of the same size.

[40]On the first day, subjects received $0.50 for completing a short demographic survey before starting the main task, which is regarded as a participation fee. At the end of each day in Qualtrics, subjects received a verification code that they needed to enter on the MTurk page. This practice is common on MTurk to prevent subjects from submitting unfinished tasks.

[41]We tried to maintain consistency for the procedures over five days. From Tuesday to Friday, we posted our MTurk task and sent messages to subjects at 7:00 AM, and closed the page at 7:00 PM, allowing subjects up to 12 hours to finish their task. The message was sent directly to the subject's email address that is registered with their MTurk account.

[42]Between Monday and Tuesday, the attrition rates were 0.21 and 0.22 for each starting date, respec-

Table 5: Session summary in the MTurk

Date	1:1	5:5	25:25	Total
12/17/18	0	207	302	509
1/7/19	252	155	0	407

For each round (day), one game was randomly selected for payment. Subjects received their earnings after the experiment was finished. All payments were made through MTurk's internal payment system. The average payoff for 5 days of participation was $5.52.

7.2 Large Teams

7.2.1 PD and WPD Games

We first examine the behaviors on MTurk. Given the team size effect in the lab, a natural conjecture is that team behavior in the 25:25 treatment will be further away from individual behavior in the 1:1 than team behavior in the 5:5 treatment. Indeed, we confirm this conjecture.

Figure 9 compares the frequency of actions for teams with different sizes. It is clear to note that cooperation increases in team sizes, resulting in the action choices of 25-member teams being almost a pure strategy of cooperation. In all cases, 25-member teams cooperate significantly more often than 5-member teams ($p < 0.01$), with the rate of cooperation in the former case being at least 95% in 4 of 6 cases, reaching almost full convergence to cooperation. Note also that different incentives for cooperation affect the amount cooperation. For instance, in PD games, 25-member teams cooperate more in PD2 where the incentive for defection is weaker than that in PD1. Hence, in the MTurk experiment, teams are consistently *further from Nash equilibrium* than individuals in these games, which is a sharp contrast to all past findings in PD games and also contrasts with some of our lab experiment results.

Result 7. *For PD and WPD games, larger teams cooperate significantly more often than smaller teams, with 25-member teams cooperating nearly all the time. Higher incentives for cooperation lead to higher rates of cooperation.*

tively. However, after Tuesday, more than 93% of subjects returned on each subsequent day. See the Appendix for details.

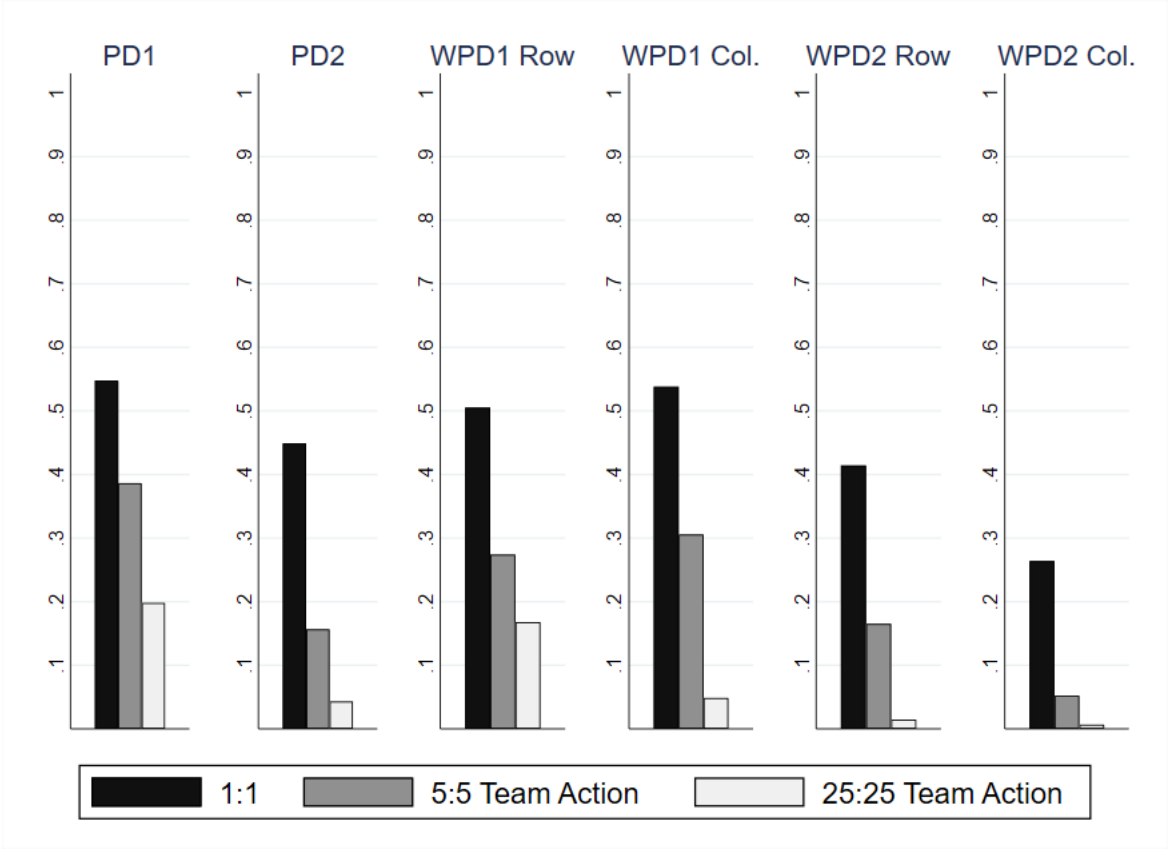


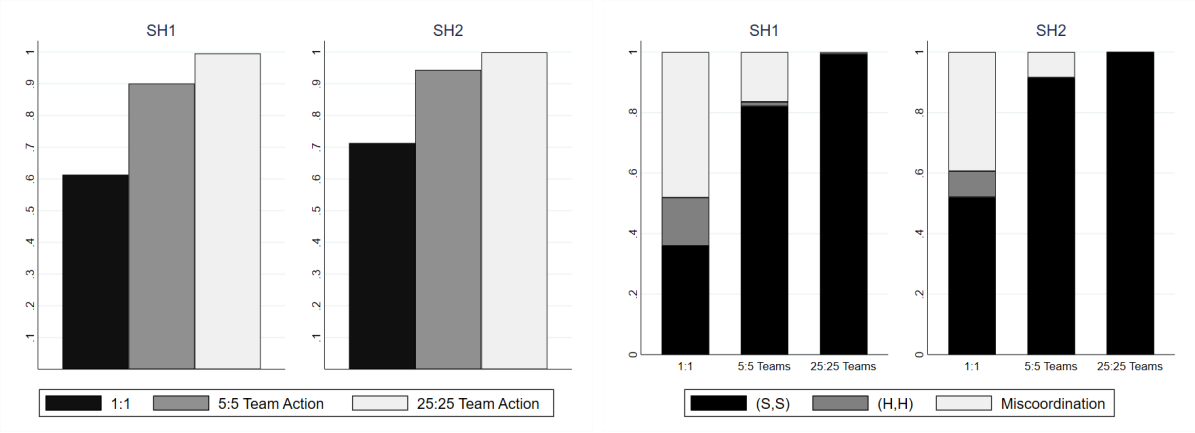
Figure 9: Comparison of 1:1 and team defection frequencies in MTurk.

7.2.2 Coordination Games

We also conjecture the team behavior of 25-member teams based on the comparison between individuals and 5-member teasm on MTurk. If the findings from the comparison between 1:1 and 5:5 on MTurk are qualitatively similar to the lab experiment (i.e., 5:5 teams coordinate more often than 1:1 and, conditional on coordination, coordinate on (S,S) more often), then we also project these differences between 5:5 and 25:25 member teams to go in the same direction.

Figure 10 displays the action choices and coordination rates in SH games. In the left panel, teams in 25:25 choose the riskier strategy significantly more often than individuals in 1:1 and teams in 5:5 in both SH games ($p < 0.01$).[43] Such differences in the choice of actions result in differences in coordination. In the right panel, teams in 25:25 coordinate

[43] As in the laboratory environment, 5:5 teams choose the riskier strategy more frequently than in the 1:1 games ($p < 0.01$).



(a) Rates of Stag in SH

(b) Rates of Coordination in SH

Figure 10: Team behavior and coordination in the SH games in MTurk

significantly more often than individuals in 1:1 and teams in 5:5. These differences are highly significant ($p < 0.01$). More striking is the fact that 25-member teams *always* coordinate on the payoff-dominant equilibrium in both SH games. The tendency for larger teams to better achieve efficient coordination is at least partly attributable to differences individual team member votes, since individual team members in larger groups are more likely to vote for Stag. For 25-member teams, most team members vote for Stag (81% for SH1 and 84% for SH2). Although equilibrium selection in team games is still an open question theoretically, our experiment empirically lends support to the hypothesis that coordination games played between larger-sized teams are less likely to result in miscoordination and are also more likely to result in the payoff dominant outcome.

Result 8. *For both SH games, teams coordinate better than individuals in 1:1, and 25-member teams coordinate more often on the payoff-dominant equilibrium than 5-member teams.*

8 Conclusion

This paper reports the results from team game experiments conducted both in the laboratory and with MTurk subjects using a design motivated by an equilibrium theory of team games (Kim et al. 2021). Our experiment has several features that distinguish it from previous team game experiments. First, we compare team and individual behavior in games using four variations of prisoners' dilemma games and two variations of

stag hunt games. Second, in contrast to previous experiments that used consensus rule through face-to-face or chat communication, we use three variations on majority voting as the collective decision rule, with the variations designed to provide insights into the roles of communication and consensus formation in teams. By comparing individuals in one-on-one games with team decision probabilities and individual team member vote choices, we can differentiate two channels of team game effects, the consensus effect and the equilibrium effect. Third, while most previous experiments used teams with two or three members, we employ 5-member teams in the laboratory environment and both 5- and 25-member teams in the MTurk environment.

The laboratory experiment produced several findings. First, there are systematic differences between games played by teams and individuals for most of the payoff configurations. For PD and WPD games, while in games with high incentives for defection, teams defect more than individuals, the opposite holds true in games with low incentives for defection. We estimate the team equilibrium model using a one-parameter logit specification. The estimation not only shows a good fit with the prisoners' dilemma data across all treatments and games but also proves to be robust when altruism is included. In SH games, teams miscoordinate less frequently than individuals, and teams achieve the efficient outcome more frequently than individuals, conditional on coordination.

Second, the different collective choice procedures affect vote margins but have little effect on the frequency of team decisions. While the difference in team choice probabilities across the three collective choice procedures are mostly small and insignificant, the effects of procedure on voting behavior are unambiguous: a straw poll and pre-vote communication both make voting more one-sided. This points to the information aggregation role of communication in the group. We also observe significant bandwagon effect in the poll treatment, with team members voting with the minority in the straw poll being more likely to switch their vote than members voting with the majority in the straw poll.

Third, we find supporting evidence for the effect of increasing team size from the large-team MTurk data. The qualitative differences we observed between 1:1 and 5:5 treatments are further reinforced with comparisons between the 5:5 and 25:25 treatments, for all six games. For PD and WPD, teams in the 25:25 treatment cooperate significantly more than teams in the 5:5 treatment and individuals in the 1:1 treatment. In coordination games, teams in the 25:25 treatment coordinate more successfully than teams in 5:5 and individuals in 1:1.

The experiment suggests a number of interesting avenues for future research. First,

as an initial study of the effect of collective choice procedures, we focused on games with binary strategies, and all three collective choice procedures were variations on majority rule. In many teams there is a leader with veto power, and possibly unequal voting weights for the individual members. An investigation of games with non-majoritarian voting rules, or with different voting rules for the each team would be interesting avenues to pursue. Along these lines, games with more than two actions for each team would allow for comparisons between a an even richer variety of collective choice procedures, such as rank-based voting, scoring rules, or sequential agendas.

Second, the finding that teams coordinate more frequently than individuals warrants further investigation since the present study considered only two very simple versions of the stag hunt game. The additional finding that teams tend to coordinate more on the payoff dominant equilibrium raises interesting theoretical questions that deserve further study.

Third, while behavior on the MTurk platform exhibited the same qualitative effects of team size as the laboratory environment, we found that MTurk subjects made more cooperative choices than subjects in the laboratory in all games and in all treatments. This raises some interesting questions about obtaining data for strategic games in such an uncontrolled setting with very low stakes. While the advantages are obvious (low cost, unlimited pool, etc.), more evidence about behavior in strategic games on the MTurk platform is needed. A larger-scale systematic study comparing MTurk and lab behavior in strategic environments could shed valuable light on these issues.

Fourth, our theoretical model underlying the experiment assumes common payoffs among team members. Introducing explicit heterogeneity in team members and whether such heterogeneity helps/hampers teams' efficient behavior would be interesting questions. For example, there might be latent social preferences, in addition to the experimentally-induced monetary payoffs. Social preferences would seem to be especially important for games where the non-cooperative outcomes are Pareto dominated, as in the PD and WPD games. Other kinds of latent preferences would also be relevant in field applications, since team members in organizations might not have identical preferences, as they have different career concerns, incentives, positions in hierarchies, etc. Finally, one could apply the experimental design and methodology introduced in this paper to study a broad range of other games. For example, team game experiments of extensive form games, such as signaling games, could be a productive direction of further research, and theoretical results in Kim et al. (2021) offer some guidance for designing such experiments.

References

[1] Agranov, Marina and Chloe Tergiman (2014). Communication in Multilateral Bargaining. *Journal of Public Economics*, 118: 75-85.

[2] Agranov, Marina and Chloe Tergiman (2019). Communication in bargaining games with unanimity. *Experimental Economics*, 22(2): 350-68.

[3] Ambrus, Attila, Ben Greiner, and Parag A. Pathak (2015). How individual preferences are aggregated in groups: An experimental study. *Journal of public economics*, 129: 1-13.

[4] Austen-Smith, David and Jeffrey Banks (1996). Information Aggregation, Rationality, and the Condorcet Jury Theorem. *American Political Science Review.* 90(1): 34-45.

[5] Bauer, Michal, Jana Cahliková, Dagmara Celik Katreniak, Julie Chytilová, Lubomir Cingl, and Tomáš Želinský (2018). Anti-social behavior in groups, *Working paper.*

[6] Benjamini, Yoav, and Yosef Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1): 289-300.

[7] Blinder, Alan S., and John Morgan (2005). Are two heads better than one? Monetary policy by committee. *Journal of Money, Credit and Banking*, 789-811.

[8] Bornstein, Gary, Ilan Yaniv (1998). Individual and Group Behavior in the Ultimatum Game: Are Groups More "Rational" Players? *Experimental Economics* 1:101-108.

[9] Callander, Steven (2007). Bandwagons and Momentum in Sequential Voting. *Review of Economic Studies.* 74(3): 653-684.

[10] Cason, Timothy N., Sau-Him Paul Lau, and Vai-Lam Mui (2019). Prior Interaction, Identity, and Cooperation in the Inter-Group Prisoner's Dilemma. *Journal of Economic Behavior and Organization*, 166: 613-629.

[11] Cason, Timothy N. and Vai-Lam Mui (1997). A laboratory study of group polarisation in the team dictator game. *Economic Journal*, 107(444): 1465-1483.

[12] Charness, Gary and Matthew Jackson (2007). Group play in games and the role of consent in network formation. *Journal of Economic Theory*, 136:417-445.

[13] Charness, Gary and Matthias Sutter (2012). Groups Make Better Self-interested Decisions. *Journal of Economic Perspectives*, 26(3):157-176.

[14] Cooper, David J., and John H. Kagel (2005). Are Two Heads Better Than One? Team versus Individual Play in Signaling Games. *American Economic Review*, 95(3):477-509.

[15] Cox, James C (2002). Trust, Reciprocity, and Other-Regarding Preferences: Groups vs. Individuals and Males vs. Females. In *Advances in Experimental Business Research*, edited by Rami Zwick and Amnon Rapoport, 331-50. Dordrecht: Kluwer Academic Publishers.

[16] Elbittar, Alexander, Andrei Gomberg, and Laura Sour (2011). Group Decision-Making and Voting in Ultimatum Bargaining: An experimental Study. *B.E. Journal of Economic Analysis and Policy (Contributions)*, 11(1): Article 53.

[17] Fischbacher, Urs. z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, 10(2): 171-178.

[18] Feri, Francesco, Bernd Irlenbusch, and Matthias Sutter (2010). Efficiency gains from team-based coordination: large-scale experimental evidence, *American Economic Review*, 100(4): 1892-1912.

[19] Gillet, Joris, Arthur Schram, and Joep Sonnemans (2009). The tragedy of the commons revisited: The importance of group decision-making, *Journal of Public Economics* 93(5-6): 785-797.

[20] Goeree, Jacob K., Charles A. Holt, and Thomas R. Palfrey (2016). Quantal Response Equilibrium, *Princeton University Press.*

[21] Guarnaschelli, Serena, Richard D. McKelvey, and Thomas R. Palfrey (2000) An Experimental Study of Jury Decision Rules. *American Political Science Review* 94(2): 407-423.

[22] Horton, John J., David G. Rand, and Richard J. Zeckhauser (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, 14(3): 399-425.

[23] Insko C.A., Hoyle R.H., Pinkley R.L., Hong G.Y., Slim R.M., Dalton B, Lin Y-H.W., Ruffin P.P., Dardis G.J., Bernthal P.R. (1988). Individual-group discontinuity: the role of a consensus rule. *Journal of Experimental Social Psychology*, 24:505-519.

[24] Kim, Jeongbin (2022). Discounting, dynamic consistency, and cooperation in an infinitely repeated game experiment. *Working paper*.

[25] Kim, Jeongbin, Thomas R. Palfrey, and Jeffrey R. Zeidel (2022). Games Played by Teams of Players. *American Economic Journal: Microeconomics*, 14(4): 122-157.

[26] Kocher, Martin, Sabine Straub, and Matthias Sutter (2006). Individual or team decision-making—causes and consequences of self-selection. *Games and Economic Behavior* 56(2): 259-270.

[27] Kocher, Martin G. and Matthias Sutter (2005). The Decision Maker Matters: Individual versus Group Behavior in Experimental Beauty-Contest Games. *Economic Journal* 115(500): 200-223.

[28] Kugler, Tamar, Gary Bornstein, Martin G. Kocher, and Matthias Sutter (2007). Trust between Individuals and Groups: Groups are Less Trusting Than Individuals But Just as Trustworthy. *Journal of Economic Psychology*, 28(6): 646-57.

[29] Kugler, Tamar, Edgar E. Kausel, and Martin G. Kocher (2012). Are Groups More Rational Than Individuals? A Review of Interactive Decision Making in Groups. *Wiley Interdisciplinary Reviews - Congnitive Science*, 3(4):471-82.

[30] Luhan, Wolfgang J., Martin G. Kocher, and Matthias Sutter (2009). Group polarization in the team dictator game reconsidered. *Experimental Economics*, 12(1): 26-41.

[31] Martinelli, Cesar and Thomas Palfrey (2020). Communication and Information in Games of Collective Decision (with C. Martinelli). In M. Capra, R. Croson, T. Rosenblatt, and M. Rigdon eds. *The Handbook of Experimental Game Theory*, Edward Elgar: Cheltenham, 348-375.

[32] McKelvey, Richard D. and Thomas R. Palfrey (1995). "Quantal Response Equilibria for Normal Form Games." *Games and Economic Behavior*, 10:6-38.

[33] McKelvey, Richard D. and Thomas R. Palfrey (1998). "Quantal Response Equilibria for Extensive Form Games." *Experimental Economics*, 1:9-41.

[34] Palfrey, Thomas R. and Pogorelskiiy (2019), Communication Among Voters Benefits the Majority Party. *Economic Journal* 129(618): 961-990.

[35] Palfrey, Thomas R., Howard Rosenthal, and Nilanjan Roy (2017), How Cheap Talk Enhances Efficiency in Threshold Public Goods Games. *Games and Economic Behavior*, 101: 234-59.

[36] Sutter, Matthias (2005). Are four heads better than two? An experimental beauty-contest game with teams of different size. *Economics letters*, 88(1):41-46.

[37] Wildschut, Tim and Chester A. Insko (2007). Explanations of interindividual-intergroup discontinuity: A review of the evidence. *European Review of Social Psychology*, 18:175-211.

[38] Wildschut, Tim, Hein FM Lodewijkx, and Chester A. Insko (2001). Toward a reconciliation of diverging perspectives on interindividual-intergroup discontinuity: The role of procedural interdependence. *Journal of Experimental Social Psychology*, 37(4): 273-285.

Appendix

Tables for Figures in the main text

Table 6: Figures in Section 6.1 PD and WPD games

	Figure 5: Frequency of one-sided votes					
Treatments	PD1	PD2	WPD1 Row	WPD1 Col.	WPD2 Row	WPD2 Col.
Majority	0.73	0.43	0.67	0.55	0.43	0.43
Poll	0.83	0.60	0.70	0.73	0.57	0.78
Chat	0.78	0.64	0.88	0.88	0.87	0.88

	Figure 6: Comparison of 1:1 and 5:5 defection frequencies					
Treatments	PD1	PD2	WPD1 Row	WPD1 Col.	WPD2 Row	WPD2 Col.
1:1	0.84	0.72	0.77	0.79	0.69	0.53
5:5 Team action	0.91	0.66	0.88	0.84	0.51	0.36

Table 7: Figures in Section 6.2 Coordination games

	Figure 7: Team behavior and coordination in the SH games	
	Figure 7(a): Rates of Stag in SH	
	SH1	SH2
1:1	0.30	0.52
5:5 Team action	0.41	0.71
	Figure 7(b): Rates of Coordination in SH	
	SH1 1:1	SH2 1:1
(S,S)	0.09	0.27
(H,H)	0.50	0.23
Miscoordination	0.41	0.50
	SH1 5:5 Teams	SH2 5:5 Teams
(S,S)	0.23	0.58
(H,H)	0.41	0.17
Miscoordination	0.36	0.25

Table 8: Figures in Section 7.2.1 PD and WPD games

	Figure 8: Comparison of 1:1 and team defection frequencies in MTurk					
Treatments	PD1	PD2	WPD1 Row	WPD1 Col.	WPD2 Row	WPD2 Col.
1:1	0.55	0.45	0.51	0.54	0.41	0.26
5:5 Team action	0.39	0.16	0.27	0.31	0.17	0.05
25:25 Team action	0.20	0.04	0.17	0.05	0.02	0.01

Table 9: Figures in Section 7.2.2 Coordination games

	Figure 9: Team behavior and coordination in the SH games in MTurk	
	Figure 9(a): Rates of Stag in SH	
	SH1	SH2
1:1	0.61	0.71
5:5 Team action	0.90	0.94
25:25 Team action	1.00	1.00
	Figure 9(b): Rates of Coordination in SH	
	SH1 1:1	SH2 1:1
(S,S)	0.36	0.52
(H,H)	0.16	0.09
Miscoordination	0.48	0.39
	SH1 5:5 Teams	SH2 5:5 Teams
(S,S)	0.82	0.91
(H,H)	0.01	0.00
Miscoordination	0.17	0.09
	SH1 25:25 Teams	SH2 25:25 Teams
(S,S)	0.99	1.00
(H,H)	0.00	0.00
Miscoordination	0.01	0.00

Returning rates in MTurk

Table 10: The rate of returning in each session of the MTurk experiment

	12/17/18		1/7/19	
Round	# of subjects	Returning rate	# of subjects	Returning rate
1	509		407	
2	402	0.79	316	0.78
3	372	0.93	296	0.94
4	348	0.94	283	0.96
5	338	0.97	278	0.98

Team choice frequencies over rounds in the lab and the MTurk

Although the direct comparison of behaviors across the different environments (lab and MTurk) is not intended in our research design, it is worth mentioning for future research how different those behaviors are. Figures from 11 to 14 show the moving averages over three rounds for rates of defection and stag. For PD and WPD games, both defection frequencies in the 1:1 treatment and team defection frequencies in the 5:5 treatment are always significantly lower in the MTurk environments than in the lab environment ($p < 0.01$). Such significant differences become more prominent as subjects gain experience and in Round 5, the differences are highly significant ($p < 0.01$) except for column players for the 1:1 treatment comparison in WPD1 and WPD2 ($p = 0.019$ and $p = 0.016$, respectively).

A similar pattern is observed in SH games. The rates of choosing a riskier action both in the 1:1 treatment and in the 5:5 treatment in the MTurk environment are always significantly higher than those in the lab environment ($p < 0.01$). Learning reinforces such differences in SH games that all comparisons between the lab environment and the MTurk environment are also highly significant in Round 5 ($p < 0.01$). Taken together, we find that action choices in the MTurk environment are more cooperative and willing to take more risk than those in the lab environment.

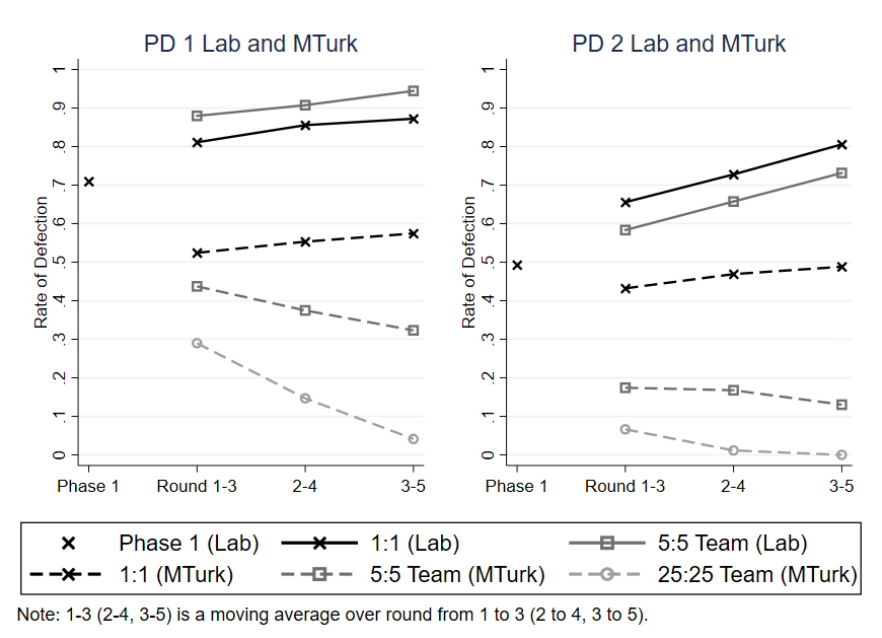


Figure 11: Comparison of 1:1 and team defection frequencies in PD games.

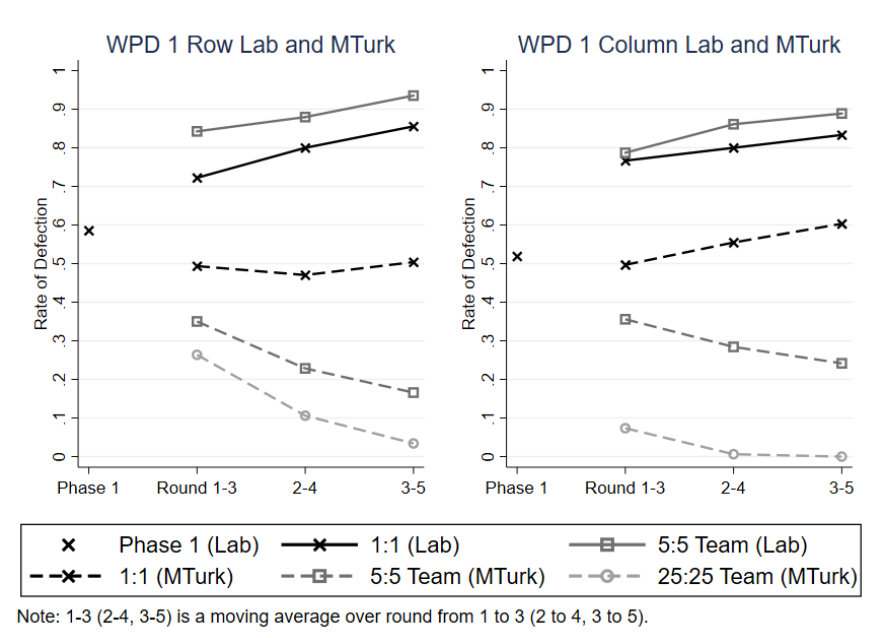
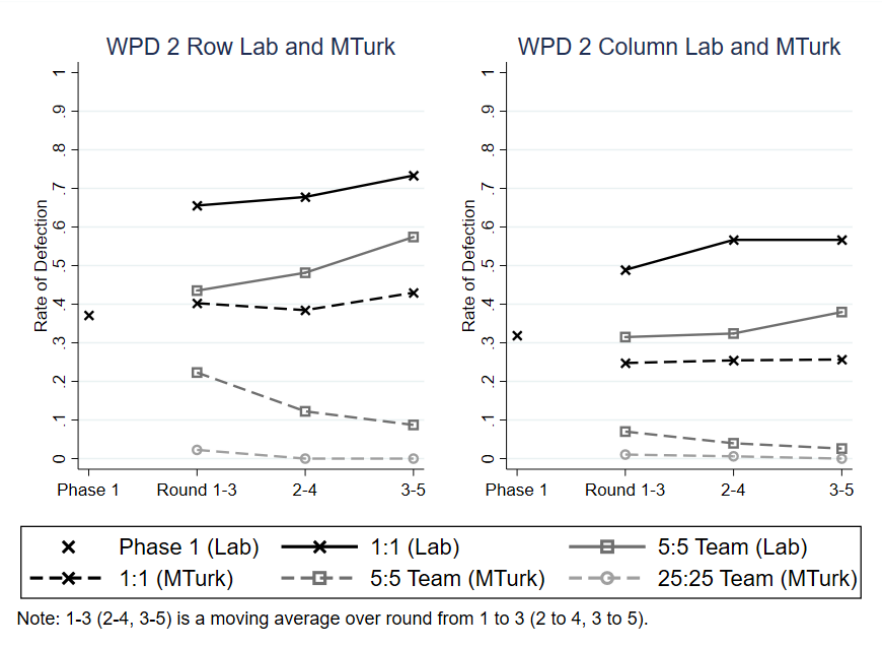
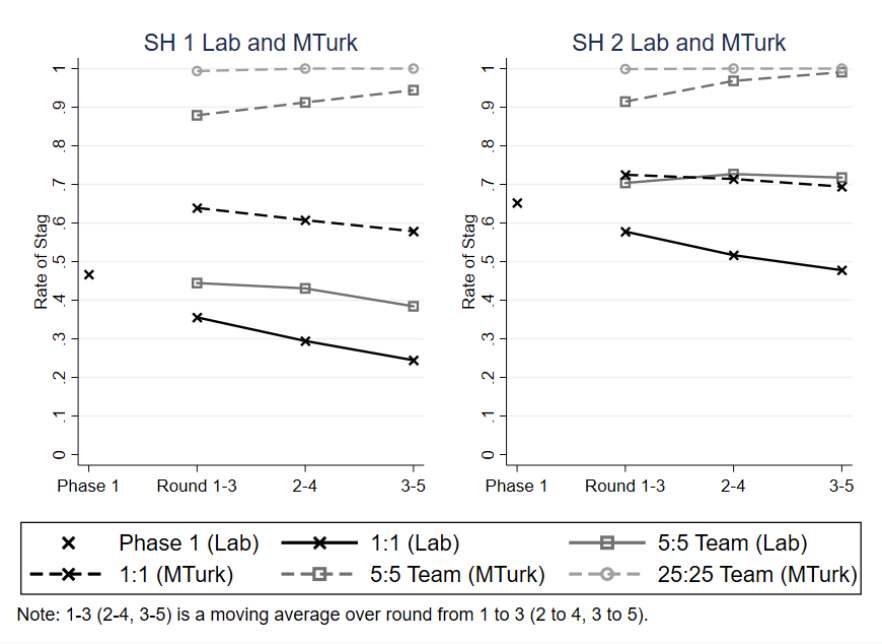


Figure 12: Comparison of 1:1 and team defection frequencies in WPD1 game.



Note: 1-3 (2-4, 3-5) is a moving average over round from 1 to 3 (2 to 4, 3 to 5).

Figure 13: Comparison of 1:1 and team defection frequencies in WPD2 game.



Note: 1-3 (2-4, 3-5) is a moving average over round from 1 to 3 (2 to 4, 3 to 5).

Figure 14: Comparison of 1:1 and team stag frequencies in SH games.

The logit/altruism model and estimates

The altruism model we use introduces a single altruism parameter, A, which is the weight each player places on the other team's payoff. Thus, for example, in WPD1, the payoff matrix with an altruism weight of A is:

	C	D
C	35+35A,35+35A	77+16A,16+77A
D	16+49A,49+16A	58+58A,58+58A

This results in MLE estimates of $A = 0.056$ and $\lambda = 0.069$. The scatter plot of estimated vs. observed defection probabilities is shown in Figure 15, and there is not much quantitative difference in the estimates, compared to the model without altruism. The estimated defection probabilities in the altruism model are shown in Table 11.

Table 11: Team Equilibrium Estimation Results: Altruism Model

Game and Role	Team Size	Observed	Estimated
PD1	1×1	0.84	0.83
	5×5	0.81	0.82
PD2	1×1	0.72	0.69
	5×5	0.61	0.70
WPD1Row	1×1	0.77	0.77
	5×5	0.79	0.78
WPD1Col	1×1	0.79	0.67
	5×5	0.75	0.73
WPD2Row	1×1	0.69	0.61
	5×5	0.49	0.61
WPD2Col	1×1	0.53	0.41
	5×5	0.38	0.45

The parameter A in the altruism model is significantly different from 0 at the 1% level, using a likelihood ratio test ($\chi^2 = 6.74$), and the fit improves slightly relative to the one-parameter model ($R^2_{A,\lambda} = 0.73$ vs. $R^2_\lambda = 0.69$).

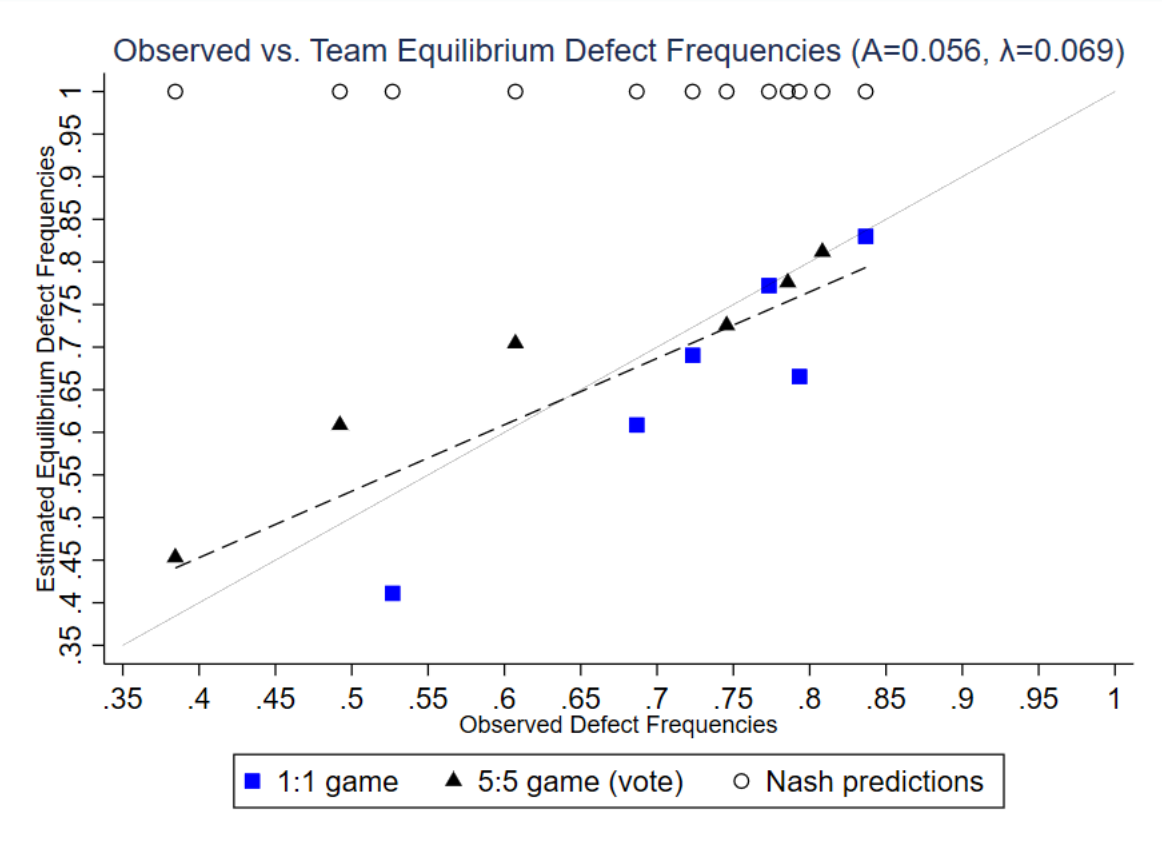


Figure 15: Fitted equilibrium defect frequencies vs. observed defect frequencies. Altruism model.

Online Appendix (not for publication)

Instructions (Poll treatment)

General instructions

Thank you for coming to the experiment. You are about to participate in an experiment on decision-making. Your earnings will depend partly on your decisions, partly on the decisions of others, and partly on chance.

The entire session will take place through computer terminals, and all interaction between participants will take place through the computers. Please do not talk or in any way try to communicate with other participants during the session.

There will be 2 phases in this experiment. Each phase will start with an instruction period

in which you will be given a description of the main features of that phase. At the end of phase 1, you will receive the instructions for phase 2. If you have any questions during this instruction period, raise your hand and your question will be answered so everyone can hear.

For each phase, some of your decisions will be randomly selected for payment. Your earnings are denominated in points, and each point has a value of \$0.05, i.e. 1 point = \$0.05. In other words, every 100 points generates \$5 in earnings for you. In addition to your earnings from decisions, you will receive a show-up fee of \$8. At the end of the experiment, your earnings will be rounded up to the nearest dollar amount. All your earnings will be paid with cash in private at the end of the experiment.

Phase 1 Decision making

1. In phase 1, you will be asked to make decisions in 8 matches. You are randomly matched with another participant for each separate match. This random pairing changes in every match.

2. In each match you will be asked to choose an action. By clicking each action's label (A or B), you can choose your action. An example for choices and payoffs is as follows:

	The other's choice of action	
Your choice of action	A	B
A	100, 200	300, 400
B	500, 600	700, 800

The numbers in the cells of the table represent your earnings in points. These numbers are only for illustration and are different from the numbers in the experiment. The first entry in each cell represents your earnings (in points), while the second entry represents the earnings (in points) of the other participant you are matched with. The earnings numbers in the table will change in every match, and you will be randomly re-matched with another participant in each match. - As you can see, this shows the earnings associated with each pair of action choices, one by you and one by the other participant you are matched with. Once you and the other participant have each selected a choice, those two choices will determine your earnings (in points) for this match.

For example, if:
You select A and the other selects A, your earnings equal 100 points while the other's earnings equal 200 points.
You select A and the other selects B, your earnings equal 300 points while the other's earnings equal 400 points.

3. At the end of each match, you will be reminded of the action you made for that match. However, the action chosen by the other participant will only be revealed to you at the very end of the experiment. In other words, you will be informed of the payoff that is determined by you and the other participant in each match at the end of the experiment.
4. At the end of the experiment, the computer will randomly select exactly one of your 8 matches from Phase I for actual payment.
- Are there any questions?

Phase 2 Instructions

Beginning of phase 2
Phase 2 will also have 8 matches, but now each match has 5 rounds. That is, for each match, you will participate in 5 rounds, one after another, so you will be making a total of 40 decisions in Phase 2.

Team assignment
At the beginning of every round of every match, you will be randomly re-grouped into 5 member teams. That is, including yourself, your team will always have exactly 5 members, but the membership of your team changes randomly every round of every match. Also at the beginning of every round of every match, your 5 member team will be randomly matched with another 5 member team.

Team decision
1. In each round, your team will make a collective decision to choose an action. You will be asked to vote for an action in each round, and your team decision will be determined by majority rule.
2. That is, your team decision will be chosen as an action for which more than a half of your team members voted. For instance, if there are 2 members of your team who vote

for action A and 3 members who vote for action B, then your team's collective decision will be action B.

3. An example for choices and payoffs is as follows:

	The other team's choice of action	
Your team's choice of action	A	B
A	100, 200	300, 400
B	500, 600	700, 800

The first entry in each cell represents the payoff for each of your team members, while the second entry represents the payoff of each of the other team members.

For example:

If your team's collective decision is action A and the other team's collective decision is action A, each member of your team (including you) earns 100 points while each member of the other team earns 200 points.

If your team's collective decision is action A and the other team's collective decision is action B, each member of your team (including you) earns 300 points while each member of the other team earns 400 points.

4. Each round consists of ***two stages***: polling stage and voting stage. In the polling stage, you will be asked to select an action that you think your team should collectively choose. The poll result of your team will be revealed to all of your team members in the voting stage. Your response in the polling stage is not binding for your vote in the next stage.

5. In the voting stage, you will be asked to vote for an action. The poll results for your team will be shown in parentheses in the payoff table. Once everyone votes for an action, you will be informed of the both teams' decisions and the actual votes made in each team. That is, you will be informed of exactly how many members vote for each action on each team. In the payoff table, the number of votes for each action made in each team will be

presented in parentheses. This will also tell you the outcome and point earnings of that round, which will be highlighted on your screen in purple.

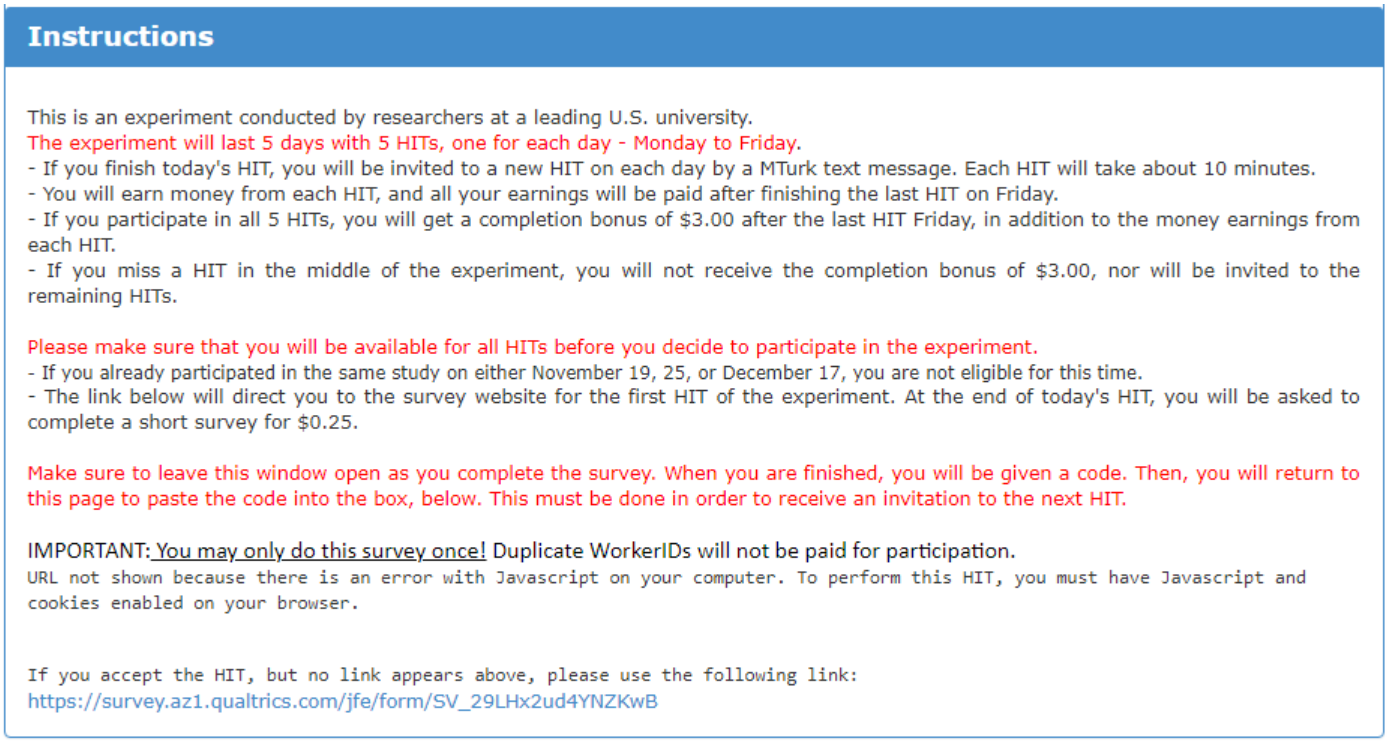
6. After every round of every match, you will be randomly assigned to a new 5 member team, and then your team will be randomly matched with another 5 member team.

7. At the end of the experiment, one round for each match will be randomly selected for your phase 2 earnings. That is, 8 rounds in total will be randomly selected for your phase 2 earnings. Your phase 2 earnings will then be computed as the sum of your earnings in those 8 randomly selected rounds, converted from points to dollars using the exchange rate of 1 point = 5 cents.

Before we start, let me remind you that:

- In phase 2, there are 8 matches each of which has 5 rounds. At the beginning of each round, you will be randomly assigned to a 5-member team, and your team will be randomly matched with another team.
- Each round consists of two stages. **In the polling stage**, you will be asked to select an action that you think your team should collectively choose. Your response in the polling stage is not binding for your vote in the next stage.
- **In the voting stage**, you will be shown the poll result and asked to vote for an action. Your team decision will be determined by majority rule, i.e., your team decision will be the action for which 3 or more members of your team voted.
- Once everyone votes for an action, you will be informed of the team decisions and votes made in each team. That is, you will be informed of how many members vote for each action in each team, and hence, the outcome of the round.
- At the end of the experiment, one round for each of the 8 matches will be randomly selected for your phase 2 earnings. Your phase 2 earnings equal the sum of your earnings in those 8 randomly selected rounds. Your total earnings for the experiment equal your phase I earnings plus your phase 2 earnings plus your $8 show up fee.

Screen shot of the recruiting page in MTurk (Monday)



Instructions

This is an experiment conducted by researchers at a leading U.S. university.
The experiment will last 5 days with 5 HITs, one for each day - Monday to Friday.
- If you finish today's HIT, you will be invited to a new HIT on each day by a MTurk text message. Each HIT will take about 10 minutes.
- You will earn money from each HIT, and all your earnings will be paid after finishing the last HIT on Friday.
- If you participate in all 5 HITs, you will get a completion bonus of $3.00 after the last HIT Friday, in addition to the money earnings from each HIT.
- If you miss a HIT in the middle of the experiment, you will not receive the completion bonus of $3.00, nor will be invited to the remaining HITs.

Please make sure that you will be available for all HITs before you decide to participate in the experiment.
- If you already participated in the same study on either November 19, 25, or December 17, you are not eligible for this time.
- The link below will direct you to the survey website for the first HIT of the experiment. At the end of today's HIT, you will be asked to complete a short survey for $0.25.

Make sure to leave this window open as you complete the survey. When you are finished, you will be given a code. Then, you will return to this page to paste the code into the box, below. This must be done in order to receive an invitation to the next HIT.

IMPORTANT: You may only do this survey once! Duplicate WorkerIDs will not be paid for participation.
URL not shown because there is an error with Javascript on your computer. To perform this HIT, you must have Javascript and cookies enabled on your browser.

If you accept the HIT, but no link appears above, please use the following link:
https://survey.az1.qualtrics.com/jfe/form/SV_29LHx2ud4YNZKwB