STRATEGIC THINKING IN PUBLIC GOODS GAMES WITH TEAMS

CALEB A. COX\textsuperscript{a} AND BROCK STODDARD\textsuperscript{b}

ABSTRACT. We experimentally investigate team behavior in repeated public goods games and use team chat logs to study motives for contribution. Subjects are matched into two-person teams, and each team makes a joint decision in each period. We compare teams with individuals and find similar overall contributions. However, initial contribution is higher and endgame effects are more pronounced for teams. We examine strategic discussions within teams and find strong evidence of concern for repeated game effects and limited backward induction. We also find evidence of confusion and explore its potential sources.

Keywords: public goods, experiment, communication, cooperation, repeated games
JEL Classification: C72 C73 C92 H41

I INTRODUCTION

Team decision making is widespread in social dilemmas in the field. Families contribute to charities, churches, and neighborhood watch efforts. On a larger scale, firms, non-profit organizations, and governments contribute to disaster relief projects and pollution abatement. We study team behavior in repeated public goods games to address two primary questions. First, we examine whether individuals and teams differ in their contribution decision. Second, we use team chat logs to investigate team contribution motives. While recent studies by Kamei (2016) and Auerswald et al. (2016) have begun examining public goods games with teams, most experiments focus on individuals. Furthermore, studying teams allows us to examine strategic thinking through content analysis of discussions between team members making a joint contribution decision, which was not a focus of these related studies. Strategic discussions within teams provide a direct window into the decision making process. Examining teams can thus yield new insight into motives for contribution.

The method of examining team chat logs has recently been used to gain valuable insight into strategic thinking in other contexts such as the prisoner’s dilemma (Kagel and McGee, 2016; Date: March 9, 2018.
\textsuperscript{a}Department of Economics, Virginia Commonwealth University, Snead Hall, 301 W. Main Street, Box 844000, Richmond, VA 23284-4000, USA; cacox@vcu.edu.
\textsuperscript{b}Department of Economics, Appalachian State University, 3095 Peacock Hall, Boone, NC 28608, USA; stoddardbv@appstate.edu.
Cason et al., 2017; Cason and Mui, 2017), signaling games (Cooper and Kagel, 2005), legislative bargaining (Bradfield and Kagel, 2015), ultimatum bargaining (Arkes et al., 2015), coordinated resistance games (Cason and Mui, 2015), and beauty contest games (Burchardi and Penczynski, 2014; Penczynski, 2016). In our setup, groups are composed of multiple decision makers, and each decision maker in a group is a team of two subjects who make a single joint decision in each period. Team members communicate with one another via text chat. Importantly, communication with other teams is not permitted, and team members have identical payoffs so their incentives are aligned, motivating them to work together to form a profitable strategy. By examining chat logs among team members, it is possible to better understand how subjects reason about the public goods game, and what concerns motivate their contribution decisions.

This “two heads” team chat method may be thought of as an elicitation procedure, similar in principle to various methods of eliciting beliefs, preferences, or strategies in economic experiments. However, the data being elicited are qualitative chat messages that reveal thought processes leading to contribution decisions. Chat messages are coded by research assistants into one or more of several categories such as discussing pro-social preferences, payoff maximization, and repeated game effects. These chat codes are used to learn about the underlying motives and strategic thinking of subjects in the experiment.

To examine the effect of playing in teams on contribution, we compare team behavior in public goods games to a baseline individual treatment. To investigate strategic thinking about repeated game effects and backward induction, we also compare the cases of random Strangers matching and fixed Partners matching of decision makers into groups. By comparing team chat logs in Partners and Strangers treatments, our experiment yields new insights on how motivations for contribution vary with repeated interaction in fixed groups. In this way, our experiment is related to the literature comparing Partners and Strangers in public goods games (Andreoni, 1988; Croson, 1996; Keser and van Winden, 2000; Andreoni and Croson, 2008) as well as the much broader literature on cooperation and free-riding in social dilemmas (Ledyard, 1995; Ostrom, 2000; Chaudhuri, 2011). To the best of our knowledge, this study is the first to use content analysis of team chat logs to compare strategic thinking with and without repeated interaction in fixed groups in any game.

We find that behavior is largely similar between individuals and teams. However, initial contribution is higher for teams. Furthermore, endgame effects are more pronounced for teams, as free-riding rates in the last period are greater for teams than individuals.

---

1There is no communication between different teams or incentive to free ride within a team. Thus, our study differs from experiments on cheap talk between rivals (e.g. Isaac and Walker, 1988a; Ostrom et al., 1992; Bochet et al., 2006) or inter-group competitions with intra-group free-riding incentives (e.g. Rapoport and Bornstein, 1987; Hargreaves-Heap et al., 2015; Bhattacharya, 2016).
Aggregate contribution is similar for Partners and Strangers. However, we find differences in strategic thinking. Compared to Strangers, Partners more frequently discuss encouraging cooperation in future periods, as well as expectations of others’ future choices. Nonetheless, concern for repeated game effects occurs with Strangers matching as well.

We find evidence of limited backward induction, with discussions of endgame effects mostly contained in the last few periods. Moreover, discussion of higher-order beliefs integral to the backward-induction process is very rare. Team discussions also reveal evidence of confusion, but relatively little direct evidence of pro-social preferences. Finally, we explore sources of possible confusion revealed in the chat logs, and discuss potential methodological implications for the design of future experiments.

II RELATED LITERATURE

Many studies examine motivations for voluntary contribution to public goods and cooperation in social dilemmas more generally (Ledyard, 1995; Chaudhuri, 2011). Various forms of social preferences such as altruism and warm glow have been used to explain voluntary contribution (see e.g. Andreoni, 1990; Goeree et al., 2002; Crumpler and Grossman, 2008). Several studies on repeated game effects compare fixed Partners matching with random Strangers matching, finding mixed results (Andreoni, 1988; Croson, 1996; Keser and van Winden, 2000; Andreoni and Croson, 2008; Cox and Stoddard, 2015).2 Yamakawa et al. (2016) find repeated game effects to be the primary driver of contribution in a two-player game with a detailed payoff table, with confusion taking a smaller role. Several other experiments, including Andreoni (1995), Houser and Kurzban (2002), and Shapiro (2009) examine the extent to which confusion rather than social preferences drives contribution in public goods games, finding that half or more of all contribution may be explained by confusion. While social preferences, confusion, and repeated game effects may all drive contributions for some subjects, we use team chat logs as a window into subjects’ strategic motivations to examine the relative prevalence of these motives.

Furthermore, most studies of contribution motives focus on individuals. Two recent studies, closely related to ours, compare team vs. individual decision making in public goods games (Kamei, 2016; Auerswald et al., 2016). Unlike these studies, we focus on the analysis of team chat logs to study strategic thinking and motives for contribution. Kamei (2016) compares individuals and two-person teams in public goods games with groups of two decision makers (i.e., two individuals or two teams of two persons each). He finds greater cooperation among

---

2Public goods games with Partners matching typically involve a finitely-repeated game, which in the standard theoretical context should not create repeated game effects. However, related studies on the finitely-repeated prisoner’s dilemma such as Selten and Stoecker (1986), Andreoni and Miller (1993), Cox et al. (2015), and Kagel and McGee (2016) suggest repeated game effects arise nonetheless due to a failure of backward induction.
teams when team members know each others’ identities. In treatments where team members do not know each other’s identities (as in our study), he finds teams are more cooperative than individuals with Partners matching, but not with Strangers matching. Kamei’s results are consistent with our finding of more pronounced endgame effects among teams than individuals. However, we do not find a similar increase in overall cooperation among anonymous teams with Partners matching. This difference in results is interesting, and may be due to some key design differences between the two studies. Kamei’s design uses groups of two decision makers (individuals or teams), and this group size necessitates individual-level feedback. In contrast, our design uses groups of three decision makers and aggregate-level feedback, where each decision maker is informed of the total contribution by all other decision makers, but not the individual contributions of each other decision maker. While results on the role of feedback in social dilemma experiments are mixed, a number of studies suggest that this design feature could be important (e.g. Sell and Wilson, 1991; Kreitmair, 2015; Van der Heijden and Moxnes, 1999; Carpenter, 2004; Cox and Stoddard, 2015).

Auerswald et al. (2016) examine three-person teams in Partners-matching public goods games with and without punishment. They find that teams contribute more and punish less than individuals. Unlike our experiment, team decisions in Auerswald et al. (2016) are made through voting with unanimity or majority rules rather than through free-form chat. While we do not find a significant difference in contribution overall between individuals and teams, our results share some features with theirs. In both our experiment and their no-punishment treatments, initial cooperation is higher and endgame effects stronger for teams than individuals.

Several studies in psychology compare individuals with teams in the prisoner’s dilemma, including Insko et al. (1988), Insko and Schopler (1992), Bornstein and Ben-Yossef (1994), and Morgan and Tindale (2002). As summarized in Wildschut and Insko (2007), these studies tend to find teams are less cooperative than individuals. Kagel and McGee (2016) study teams in a series of finitely-repeated prisoner’s dilemmas and also find that teams are less cooperative than individuals in early supergames, but find the opposite result in later supergames. As in our study, Kagel and McGee analyze team chats in depth, finding early cooperation is primarily motivated by a belief that it will encourage future cooperation, while early defection is motivated by concern that the other player will defect. They also find that subjects are aware, either initially or once mutual cooperation begins, that the opponent is likely to defect near the end of the game. However, a failure to realize that the opponent is thinking the same about them prevents complete unraveling of cooperation. While there are similarities between the finitely-repeated prisoner’s dilemma and the public goods game, the public goods game can lead to more complex social norms when players make non-binary decisions in groups with
more than two players (Isaac and Walker, 1988b). Furthermore, our experiment allows us to compare motives for contribution with and without repeated interaction in fixed groups.

A number of experiments compare individual behavior with team behavior in a variety of other settings, including dictator games (Cason and Mui 1997; Luhan et al. 2009), ultimatum games (Bornstein and Yaniv 1998; Arkes et al. 2015), trust games (Cox 2002; Kugler et al. 2007), monopoly pricing (Davis and Harless, 1996), beauty contest games (Kocher and Sutter, 2005), rent-seeking contests (Sheremeta and Zhang, 2010), and coordination games (Feri et al., 2010). Summaries of these studies generally conclude that team decisions tend to be more rational and self-interested than individual decisions (Charness and Sutter, 2012; Kugler et al., 2012). Sutter et al. (2013) compares individuals with teams in a variety of normal form games, finding that teams are more strategically sophisticated than individuals.

III Experimental Design and Procedures

We will refer to a decision-making unit (an individual or a team) as a decision maker. In Individual treatments, a decision maker is an individual human subject. In Teams treatments, a single decision maker is a team, composed of two human subjects who make a single joint decision in each period. Groups are composed of three decision makers in all treatments. In all treatments, matching was random, anonymous, and computerized. Partitions between lab stations were used to ensure anonymity.

We used a standard linear public goods game with three decision makers per group. Each decision maker began with an endowment of 80 tokens in a private account. Any whole number from 0 to 80 tokens could be moved to a group account. Each token moved to the group account was multiplied by 2.25 and divided equally among the 3 decision makers in the group (for a marginal per capita return of 0.75). At the end of a period, the three decision makers earned equal shares of the group account plus any tokens in their own private accounts.

Tokens were converted to currency at a rate of 100 tokens to 1 US dollar (or 1 token to 1 US cent). In Teams treatments, these numbers represent the payoffs for each team member, and team members always received identical payoffs. For example, if a team earned 100 tokens in a given period, then both team members received 1 US dollar in payment for that period.

The experiment includes two treatment variables with two levels each in a 2-by-2 design. We consider Teams vs. Individuals and Partners vs. Strangers matching across 15 periods. The number of periods was publicly announced, and each subject’s computer screen displayed the current period and total number of periods throughout the experiment. In the Individual treatment with Partners matching, each decision maker made a private decision in each period and the group of three decision makers remained fixed throughout all 15 periods. In the Individual
treatment with Strangers matching, each decision maker made a private decision in each period and the group of three decision makers was randomly re-matched before every period. At the conclusion of each period, all decision makers learned the aggregate contributions in their group and their own earnings for that period.

In Teams treatments, two human subjects were randomly and anonymously matched and made one joint contribution decision each period. In each period, subjects on the same team could send free-form chat messages to each other for up to 180 seconds. Subjects were prohibited from revealing their identities or personal information in chat messages (and indeed refrained from doing so). Both team members were required to agree on and enter the same contribution choice. If a team failed to do so within 180 seconds, the computer assigned a random contribution. Importantly, chat occurred only in Teams treatments, and only within teams, so that subjects could not send or receive messages with subjects outside their own team. Furthermore, teams always remained fixed throughout the experiment in both Partners and Strangers conditions, so that random re-matching in the Strangers condition refers to the matching of decision makers (two-subject teams) into groups of three teams (six subjects) each.

After preliminary analysis of the data from pilot sessions, the chat logs revealed what may be confusion about incentives for some subjects. In an attempt to minimize possible confusion, further Team and Individual sessions were conducted with revised instructions that provide more information about group and individual maximum and minimum earnings (see Isaac et al., 1994; Saijo and Nakamura, 1995; Ferraro and Vossler, 2010). In analyses reported in Appendix subsection AI.2, we find no significant differences in contribution or confusion about incentives, so the pilot sessions are pooled with all other sessions to increase sample size. The main results of this paper are robust to excluding the data from pilot sessions.

Table I lists each treatment and summarizes its design. All experimental sessions were conducted at the University of South Dakota between Spring semester 2014 and Spring semester 2015. Undergraduate subjects from a wide range of disciplines were recruited from classrooms. At the beginning of each session, subjects individually read a set of instructions, which were then summarized publicly by the experimenter. After reading the instructions, subjects took a post-instruction quiz and were not allowed to continue until all answers were correct. Subjects

---

3 We do not implement perfect Strangers matching, in which two distinct decision makers never face one another more than once. Thus, there is a positive probability of being re-matched with a decision maker in future periods.

4 On average, teams took 60 seconds per period to reach a decision. Failure to agree within 180 seconds is very rare (5 instances in 1,260 observations) and, based on accompanying chat, appears to be primarily accidental.

5 In Strangers treatments, we attempted to keep the matching group as large as possible subject to constraints of lab space and show-up rates. Due to the larger number of subjects needed for TS sessions, the average matching group size in TS is smaller than in IS. In analysis reported in Appendix subsection AI.8, we find no effect of matching group size on contribution decisions.
made all decisions on computers in private. All sessions were programmed and conducted in z-Tree (Fischbacher, 2007). Full instructions are included in Appendix AII. Subjects earned approximately $18.50 on average in sessions lasting 45 minutes (Individual) to 1 hour (Team).

Predictions

We compare teams with individuals to examine the degree to which playing in teams affects contribution. We also use team chat logs to examine strategic thinking and motives for contribution. Team chat logs were coded into categories by research assistants. Such categories include references to issues such as other-regarding preferences, payoff maximization, concern for future cooperation, and anticipation of endgame effects. Importantly, these categories are not mutually exclusive, and multiple categories may be coded simultaneously. Moreover, teams may be heterogeneous in their motives. We use team chat logs to examine the relative prevalence of various motives for contribution and free-riding as well as the degree to which discussions of these motives correlate with contribution decisions.\(^6\)

As previously discussed, a number of social psychology and economics experiments find individuals to be more cooperative than teams in games such as the prisoners’ dilemma (Wildschut and Insko, 2007). In a variety of settings, teams are more likely than individuals to play the dominant strategy in games with a unique, socially-inefficient dominant strategy equilibrium (Charness and Sutter, 2012).\(^7\)

**Prediction 1** (Team Effect). Individuals will contribute more than teams.

By examining both Partners and Strangers, we can compare the dynamics of contribution and motives across multiple periods of play with and without repeated interaction in fixed groups. Perhaps the most important difference between the Partners and Strangers settings is that with Partners, future interactions may create a concern to encourage future cooperation.

---

\(^6\)Content analysis of team chats has some inherent limitations, as RAs’ interpretation of chat is subjective, and a team member may not discuss all of their motives if they think the other team member is unlikely to agree. Nonetheless, the chat logs provide a rich data set that yields insight into teams’ decision making process.

\(^7\)However, some studies find teams behave in a more pro-social or other-regarding way than individuals, such as Cason and Mui (1997), Kamei (2016), and Auerswald et al. (2016).
We are also interested in learning whether subjects use backward induction in the Partners treatment, perhaps in later periods. Our design allows us to investigate these issues directly by examining team chat logs.

**Prediction 2 (Future Concern).** Discussion of encouraging future contribution by others:

1. Will be positively correlated with contribution
2. Will occur more frequently for Partners than Strangers.
3. Will decrease across periods.

**Prediction 3 (Backward Induction).** Discussion of backward induction reasoning:

1. Will be negatively correlated with contribution.
2. Will occur more frequently for Partners than Strangers.
3. Will increase across periods.

A central question is why people contribute at all despite the free-riding incentive. Two simple explanations are that subjects care about others (Andreoni 1990, Goeree et al. 2002, Crumpler and Grossman 2008) or that subjects are confused about the incentives they face (Andreoni, 1995; Houser and Kurzban, 2002; Shapiro, 2009; Ferraro and Vossler, 2010).

**Prediction 4 (Other-Regarding Preferences).** Discussion of pro-social other-regarding preferences will be positively correlated with contribution.

**Prediction 5 (Confusion).** Discussion suggesting misperception of incentives will be positively correlated with contribution.

In addition to unconditional pro-social motivations, subjects may have reciprocal preferences (Charness and Rabin, 2002). In public goods games, reciprocal preferences might drive contribution and free-riding in the Partners setting (Sefton et al., 2007).

**Prediction 6 (Reciprocity).** Discussion of positive (negative) reactions to others’ past contribution decisions will be positively (negatively) correlated with contribution for Partners, but not for Strangers.

There are also distinct reasons why subjects might choose to free ride: monetary gain or avoiding monetary loss (Rapoport, 1967; Dawes et al., 1986; Dawes and Thaler, 1988; Ahn et al., 2001). If subjects perceive the game as a coordination game (either due to confusion or social preferences), they will contribute only if they are confident that others will contribute.

**Prediction 7 (Free Riding for Profit).** Discussion of monetary gains from free riding will be positively correlated with free riding.
Prediction 8 (Free Riding for Safety). Discussion of concern that others might free ride will be positively correlated with free riding.

IV Results

We first examine treatment differences in contribution levels and proportion of free riders. We then analyze the chat logs of team members. The treatments are identified by Individuals or Teams and the matching mechanism, Partners or Strangers. For brevity in the discussion of the results, treatments will be referred to by the acronyms listed in Table I. For example, the Team-Strangers treatment will be referred to as TS.

Contribution Decisions: Individuals vs. Teams

The left panel of Figure I displays the path of average contribution for each treatment. Average contribution is similar across treatments, but initially higher among teams than individuals. Holding the matching mechanism constant, Wilcoxon ranksum tests show the difference between the contributions of teams and individuals in the first period is significant, \( p \)-values = 0.011 for Strangers and 0.009 for Partners. However, differences in teams’ average contribution across all periods are not significant.\(^8\)

There is more decay in average contribution across decision periods for teams than there is for individuals. Holding constant Individual or Teams, there do not appear to be differences between matching mechanisms, except perhaps in the last period. Contribution increases somewhat in the last period of Strangers treatments, especially with teams.

The right panel of Figure I displays the path of the proportion of complete free riders (decision makers choosing zero contribution) for each treatment. The proportions are not different until the final periods, where the proportion of free riders is greater in Team treatments than in Individual treatments, \( p \)-values = 0.038 for Strangers and 0.095 for Partners.\(^9\) In the final period in the TP treatment, over 50% of teams free ride.

---

\(^8\) The unit of analysis for Wilcoxon ranksum tests of first-period contribution is a decision maker. Across all periods, an independent observation for treatments with Partners matching is a group of three decision makers. An independent observation for treatments with Strangers matching is a session. Tests reveal similar results with decision maker as the unit of analysis. In Appendix subsection AL.1, we report power analyses using G*Power (Faul et al., 2007) examining what treatment effect sizes would likely be detected using these Wilcoxon ranksum tests. We also report a simulation-based power analysis for the regression reported in Model 1 of Table II. While the non-parametric tests may be under-powered for some comparisons, the regression approach allows more powerful tests of treatment effects.

\(^9\) We compare free-riding rates using \( \chi^2 \) tests with clustering adjustment at the group level for Partners and at the session level for Strangers (Rao and Scott 1981, 1984; Sribney 1998).

Table II reports multilevel linear panel regressions that allow for analysis of contribution choices controlling for correlation within decision makers and within groups or sessions (Frechette, 2012; Moffatt, 2016). These multilevel models include random effects at two levels. Decision maker is the first level. The second level is split because the models pool all four treatments: a group of three decision makers for Partners data and a session for Strangers data.

Model 1 includes treatment indicators for Teams and Strangers, and an interaction between these two indicators. We find no significant treatment differences in average contribution, Team, Strangers, and Strangers×Team coefficients (p-values > 0.10). Model 2 also includes a period-trend variable, (unreported) demographic controls, and the one-period lag of mean-centered average contribution of other group members to control for feedback effects (Ashley et al., 2010; Frechette, 2012). Team and Stranger treatment variables remain insignificant in Model 2. Lagged average contribution of others and the period trend of contribution are significant and their coefficients have signs consistent with previous linear public goods studies.

As noted above, Figure I shows stronger downward trends for Teams than for Individuals. Model 3 controls for differential trends between Teams and Individuals by interacting Period and Team. The downward trend in Individual contribution is small, but statistically significant, Period coefficient (p-value < 0.001). Teams show a significantly stronger downward trend, Period×Team coefficient (p-value < 0.001). We also find a larger intercept for Teams (p-value=0.063), consistent with the higher initial contribution for Teams shown in Figure I.10

---

10In Appendix subsection AI.3, we further explore differences between Team and Individual contributions in light of social psychology theories of group polarization (Isenberg, 1986; Cason and Mui, 1997) and the truth-wins norm (Lorge and Solomon, 1955; Kagel and McGee, 2016), finding some support for both. We explore gender differences in Appendix subsection AI.4, finding that men contribute more than women as Individuals, but not in Teams.
Table II. Team vs. Individual Regressions: Average Contribution in Tokens

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>32.80***</td>
<td>35.97***</td>
<td>25.41***</td>
</tr>
<tr>
<td>Lagged Other Cont</td>
<td>—</td>
<td>0.22***</td>
<td>0.20***</td>
</tr>
<tr>
<td>Team</td>
<td>4.78</td>
<td>1.27</td>
<td>10.29*</td>
</tr>
<tr>
<td>Strangers</td>
<td>0.39</td>
<td>1.13</td>
<td>1.14</td>
</tr>
<tr>
<td>Strangers × Team</td>
<td>-1.98</td>
<td>-2.23</td>
<td>-2.28</td>
</tr>
<tr>
<td>Period</td>
<td>—</td>
<td>-0.79***</td>
<td>-0.36***</td>
</tr>
<tr>
<td>Period × Team</td>
<td>—</td>
<td>—</td>
<td>-1.05***</td>
</tr>
<tr>
<td>Demographics</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Standard errors are in parentheses. Multilevel linear regression models with Partners (Strangers) data include random effects for decision maker and group (session). Unreported demographic controls include gender, upper-level economics courses, and previous economics experiments.

Using 2-tailed tests, ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Model 1: N=2925; Model 2: N=2925; Model 3: N=2730

**Result 1.** Contrary to Prediction 1 (Team Effect), we find that teams are initially more cooperative than individuals. However, contribution decisions across all periods are not significantly different between individuals and teams.

**Chat Log Contents**

Table III describes chat categories for chat messages between members of the same team.11 Most chat categories were pre-specified and related to one or more of the predictions from Section III.12 For instance, PrivOpt and Coord are meant to capture how subjects view the stage game incentives, while categories such as Future and Endgame are intended to address repeated game effects and backward induction. However, the chat code GoBig was added after

11In addition to the categories reported here, we also included several placeholders for categories such as idle or absent chat (17.07% of periods for Partners, 24.18% for Strangers). However, we do not analyze the chat codes for these placeholder categories as they are unrelated to our research questions.

12The categories for Discussing Others are included to explore the degree to which teams are backward looking, forward looking, and engaged in higher-order reasoning about others’ expectations. Such discussions yield insight in related studies such as Kagel and McGee (2016). These categories are indirectly related to several predictions, but are not used to test these predictions directly.
<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Prediction(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Social and Emotional Motives</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ProSocial</td>
<td>Explicit desire to be kind/fair/helpful</td>
<td>4</td>
</tr>
<tr>
<td>Guilt</td>
<td>Explicitly feeling bad for contributing little or contributing more to avoid feeling bad</td>
<td>4</td>
</tr>
<tr>
<td>Anger</td>
<td>Explicit negative emotions in response to others’ contributions</td>
<td>6</td>
</tr>
<tr>
<td>GoBig</td>
<td>Choosing a big number for fun or curiosity</td>
<td></td>
</tr>
<tr>
<td><strong>Perception</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PrivOpt</td>
<td>Perceiving that positive contributions are privately optimal (regardless of others’ contributions)</td>
<td>4, 5</td>
</tr>
<tr>
<td>Coord</td>
<td>Perceiving that contributing is optimal conditional on others contributing (a coordination game)</td>
<td>4, 5</td>
</tr>
<tr>
<td><strong>Strategic and Self-Interested Motives</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Future</td>
<td>Trying to encourage others to contribute in the future</td>
<td>2</td>
</tr>
<tr>
<td>Endgame</td>
<td>Anticipating a change in play in (or near) the last period</td>
<td>3</td>
</tr>
<tr>
<td>Selfish</td>
<td>Intentional free-riding in the hope that others will contribute</td>
<td>7</td>
</tr>
<tr>
<td>Safety</td>
<td>Concern that others will not contribute in the current period as a motive not to contribute</td>
<td>8</td>
</tr>
<tr>
<td><strong>Changing Behavior</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AdjustUp</td>
<td>Contributing more because of past outcomes</td>
<td>6</td>
</tr>
<tr>
<td>AdjustDown</td>
<td>Contributing less because of past outcomes</td>
<td>6</td>
</tr>
<tr>
<td><strong>Discussing Others</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OthersPast</td>
<td>Talking about what others have done in the past</td>
<td></td>
</tr>
<tr>
<td>OthersFuture</td>
<td>Trying to anticipate what others will do</td>
<td></td>
</tr>
<tr>
<td>OthersExpect</td>
<td>Talking about what others expect us to do</td>
<td></td>
</tr>
</tbody>
</table>

reading chat logs from pilot sessions and finding several teams discussing large increases in contribution apparently for fun or out of curiosity.

Note that coding is at the level of a team in a period. Importantly, most chat categories are not mutually exclusive, and multiple categories could be coded for a given team in a given period. For instance, OthersPast and OthersFuture would be coded in the same period if a team discusses the past behavior of other teams and tries to anticipate what other teams will do in the current period or later.

Following the preliminary analysis of pilot chat logs by the co-authors, two research assistants (RAs) were involved in coding all team sessions. The RAs were an undergraduate student from VCU majoring in economics and a graduate student from USD majoring in Accounting. After being instructed how to code the chat messages, the RAs coded the chat for one Partners session and one Strangers session. The RAs then each met with a co-author to review his/her codes. Afterwards, the RAs coded the remaining sessions.
It is useful to look at examples of some chat categories. ProSocial is coded for explicit mentions of motives such as fairness, kindness, or altruism. For example:

\[
\text{we can stick with 20 so we help the group out a little}
\]

Anger is coded for negative emotional reactions to past outcomes. For example:

\[
\text{they are being buttheads. i wouldn’t say thats profanity!}
\]

GoBig represents simply wanting to choose a big number for fun. We did not originally anticipate this category, but added it after reading chats from some of the early sessions. Phrases such as “go big or go home” occur on multiple occasions, as do gambling references such as the following example:

\[
\text{you sure you don’t just wanna go las vegas on this and go all 80?}
\]

PrivOpt represents a perception of the incentives such that contribution is privately profitable, conflating the socially optimal and privately optimal actions. In the following example, a subject thinks that it is more profitable for their team to unilaterally contribute the maximum than to free ride, assuming the other teams’ contributions are fixed at 40 tokens.

\[
\text{We still benefit more if we put all in then none if they only put in their 40}
\]

Coord represents a perception (either due to confusion or social preferences) that contribution is privately profitable conditional on others contributing. For example, the following subjects do not understand why others choose to free ride rather than conditionally cooperate:

\[
\begin{align*}
A: & \text{Do people not get the fact that the more they put in the more everyone gets?} \\
B: & \text{i know. it doesn’t make sense. if everyone did 80 we’d be rolling in it. but [I guess] not}
\end{align*}
\]

As another example, the following subject expresses frustration with being unable to communicate with other teams and a wish for an explicit coordination device:

\[
\text{i just want to say to the whole room and be like EVERYONE CHOOSE 80!!!}
\]

Future represents discussions of encouraging future cooperation and concern that present free-riding might reduce others’ future contributions.\textsuperscript{13} The following is a clear example:

\textsuperscript{13}In Appendix subsection AI.5, we provide examples of discussion regarding potential exploitation of cooperative reputation, defined by a combination of Selfish and Future.
the more we put in the more others might catch on and generate more in the pool, thus making us more

Endgame is coded for discussion of an anticipated change in own contribution or others’ contributions in or near the last period. Teams sometimes discussed such changes in advance, as in the following example:

A: I have an idea for the last round
B: what’s that?
A: Not put anything in
B: Ya, that way it’s all profit?
A: Yep! They probably won’t expect it

Endgame discussions were more frequent near the last period, as in the following example:

last two rounds. [Expletive] these other groups lets just keep all our tokens

Selfish represents explicitly free-riding for monetary gain, as in the following example:

The way I understand it is we want to put in low and hope that the other two groups put in high

In contrast to the previous category, Safety is coded when low contributions are motivated by concern that others will not contribute. For example:

Dude we should probably play it safe the first time just in case the other teams are stingy

AdjustUp and AdjustDown by themselves could be interpreted in a number of ways, including reciprocity, imitation learning, or reinforcement learning. Combining either of these codes with OthersPast suggests reciprocity or imitation learning, but it is very difficult to distinguish between these two interpretations. Chats discussing changing the contribution choice toward others’ contributions could be motivated in either way, for example:

Nice. I feel like one of the teams did all 80. Want to move it up to around 60?

Table IV reports summary statistics for the chat codes. First, we report the Cohen’s $\kappa$ measure of inter-rater reliability (Cohen, 1960). A score of $\kappa = 1$ indicates perfect agreement, while $\kappa = 0$ indicates no more than chance agreement. We focus our analysis mainly on chat categories with at least moderate agreement ($\kappa \geq 0.4$) based on the benchmark scale of Landis and Koch (1977). The categories with the lowest agreement, such as discussions of guilt and others’ expectations, were coded very infrequently.
Table IV. Chat Code Statistics

<table>
<thead>
<tr>
<th>Code Frequencies by Period</th>
<th>Code Frequencies by Team</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cohen’s Kappa</td>
</tr>
<tr>
<td>ProSocial</td>
<td>0.31</td>
</tr>
<tr>
<td>Guilt</td>
<td>0.00</td>
</tr>
<tr>
<td>Anger</td>
<td>0.52</td>
</tr>
<tr>
<td>GoBig</td>
<td>0.28</td>
</tr>
<tr>
<td>PrivOpt</td>
<td>0.44</td>
</tr>
<tr>
<td>Coord</td>
<td>0.17</td>
</tr>
<tr>
<td>Future</td>
<td>0.49</td>
</tr>
<tr>
<td>Endgame</td>
<td>0.58</td>
</tr>
<tr>
<td>Selfish</td>
<td>0.51</td>
</tr>
<tr>
<td>Safety</td>
<td>0.29</td>
</tr>
<tr>
<td>AdjustUp</td>
<td>0.46</td>
</tr>
<tr>
<td>AdjustDown</td>
<td>0.64</td>
</tr>
<tr>
<td>OthersPast</td>
<td>0.78</td>
</tr>
<tr>
<td>OthersFuture</td>
<td>0.63</td>
</tr>
<tr>
<td>OthersExpect</td>
<td>0.00</td>
</tr>
</tbody>
</table>

$\chi^2$ tests with clustering adjustment at the group level for Partners and at the session level for Strangers (Rao and Scott 1981, 1984; Sribney 1998). Using 2-tailed tests, ***, **, and * indicate significant treatment differences at the 1%, 5%, and 10% levels.

Table IV also reports chat code frequencies by period and team. Frequency by period is the percentage of periods in which a chat category was coded, averaging over the two RAs.14 Frequency by team is the percentage of teams for which a chat category was ever coded, averaging over the two RAs. ProSocial was coded for a higher percentage of periods and for more teams with Partners than Strangers, $p$-values = 0.004 and 0.009, respectively. Interestingly, ProSocial was coded fairly infrequently among both Partners and Strangers. However, agreement between the RAs on this category was somewhat weak, in part because it was rarely coded. Furthermore, based on our own reading, ProSocial chats largely did not focus on simple forms of other-regarding preferences such as altruism or warm glow. Instead, the most common theme was building a cooperative norm, overlapping with the Future category, such as in the following example:

*Let’s move up again. if we can build mutual trust with the other teams, we can keep it high*

Future was coded more frequently for Partners than Strangers, both by period and by team, $p$-values = 0.001 and 0.002, respectively. AdjustUp was also coded more frequently by period

---

14In a previous version of this paper, our analysis involved a third RA, an undergraduate student from VCU majoring in economics. The third RA was the tie-breaker for cases of disagreement by the first two RAs, entering codes only in such cases (see Gangadharan et al., 2017). In the chat code data used in that previous version of the analysis, a chat category was coded if (a) both of the first two RAs coded it, or (b) one of the first two RAs coded it and the third RA agreed. The main results remain largely similar using this approach.
and team for Partners, \( p \)-values = 0.051 and 0.035, respectively. However, we do not find evidence of such a difference for AdjustDown, \( p \)-values = 0.368 by period and 0.512 by team. Furthermore, while Endgame was coded somewhat more frequently by period and team for Partners, the difference is not statistically significant, \( p \)-values = 0.366 by period and 0.480 by team. While the majority of teams in the Partners treatment anticipate endgame effects, Strangers mention this category almost as often. Large majorities of teams were coded as discussing OthersPast and OthersFuture. OthersFuture was discussed more frequently by period for Partners, \( p \)-value=0.003. Few teams were coded as discussing OthersExpect representing a more advanced level of reasoning. However, in part because of how rarely OthersExpect was coded, agreement is no more than chance, \( \kappa = 0.00 \). Still, the small percentage of teams discussing this category is consistent with the results of Kagel and McGee (2016), who found subjects did not frequently discuss higher-level reasoning in a repeated prisoner’s dilemma with teams. GoBig was coded more frequently by period and by team for Strangers, \( p \)-values = 0.029 and 0.059, respectively. However, agreement was somewhat weak for GoBig, \( \kappa = 0.28 \).

**Result 2.** We find significantly more discussion of the Future category for Partners than Strangers, consistent with Prediction 2 (Future Concern).

**Result 3.** In partial support of Prediction 3 (Backward Induction), the majority of Partners teams anticipate endgame effects, while Strangers teams mention endgame effects somewhat less frequently. However, the difference is not significant.

**Result 4.** We do not find strong, direct evidence of pro-social motives (Prediction 4). However, agreement between RAs on the ProSocial category is somewhat weak.

**Result 5.** We find some evidence that free-riding motives include both profit (Prediction 7) and safety (Prediction 8). However, agreement between RAs on the Safety category is somewhat weak.

*Strategic Thinking Across Periods*

It is useful to examine the dynamics of some chat codes over multiple periods to see when teams discuss specific categories. Figure II shows the frequency of discussion of concern for future cooperation (Future) across all 15 periods of the experiment. For Partners, chat about future concern remains around 10% for most periods, with higher frequency in periods 3 and 8, and dropping off only in the last two periods. Among Strangers, discussion of future concern is less frequent, and 83.0\% of mentions occur in the first seven periods.
The left panel of Figure III shows the frequency of discussion of anticipated endgame effects across periods. For both Partners and Strangers, most chats about endgame effects occur late in the game, with the most frequent discussions in period 13 for Partners and 15 for Strangers.

The right panel of Figure III shows the frequency of first mentions of anticipated endgame effects across periods. Focusing on first mentions allows us to see at what point in the game teams first begin to realize that behavior may change in the final period. As with all mentions of endgame effects, first mentions of endgame effects occur primarily late in the game. For Partners, 94.7% of all first mentions occur in period 10 or later, and for Strangers, 76.0%. It is interesting that first mentions of endgame effects occur in period 10 or later less frequently for Strangers than Partners. However, first mentions of endgame effects occur in the final period more frequently for Strangers (28.0% of all first mentions) than for Partners (13.2%).

**Result 6.** Consistent with Prediction 2 (Future Concern), mentions of concern for future cooperation persist until the last two periods for Partners. For Strangers, most mentions of future concern occur in the first seven periods.

**Result 7.** In partial support of Prediction 3 (Backward Induction), for both Partners and Strangers, the vast majority of mentions and first mentions of endgame effects occur in period 10 or later.
Figure III. Frequency of endgame chat across periods

Relating Chat Logs to Choices

In this subsection, we examine correlations between chat codes and contribution choices. We interpret the following regression results as correlations and not as causal effects because chat codes may be endogenous.\(^\text{15}\) However, analyzing these correlations is still useful to understand the relationship between team chats and contribution.

Table V reports multilevel panel regressions using the same random-effects structure as the previous regressions in Table II. Models 4–5 report multilevel linear panel regressions with contribution as the dependent variable. Models 6–7 report multilevel linear panel regressions with the first difference in contribution decisions as the dependent variable. The first difference in contribution decisions is the difference between contribution levels in period \(t\) and period \(t-1\). Models 8–9 report multilevel logistic regressions. The dependent variable is an indicator of complete free-riding (zero contribution). Odds ratios are reported in Models 8–9, so that estimates less than 1 indicate negative correlations, while estimates greater than 1 indicate positive correlations. Relevant chat categories are included as independent variables.\(^\text{16}\) Chat categories with lower reliability scores (\(\kappa < 0.4\)) are excluded. In addition to chat categories, we include lagged average contribution of others (Models 4–5 & 8–9) or lagged first difference of average

\(^{15}\)Such issues of potential endogeneity of elicited explanatory variables are not uncommon in related experimental studies. For example, in studies using elicited beliefs to explain strategic choices in games, the elicited beliefs may be endogenous in the same way as our chat codes.

\(^{16}\)When AdjustUp is coded, complete free-riding never occurs. In Models 8–9, these observations are dropped from the regression. The results are robust to omitting AdjustUp completely without dropping these observations.
<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Contribution Partners Model 4</th>
<th>Contribution Strangers Model 5</th>
<th>First-Difference Partners Model 6</th>
<th>First-Difference Strangers Model 7</th>
<th>Free-Riding Partners Model 8</th>
<th>Free-Riding Strangers Model 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>46.46***</td>
<td>51.56***</td>
<td>-1.35</td>
<td>1.48</td>
<td>0.05**</td>
<td>0.01***</td>
</tr>
<tr>
<td></td>
<td>(5.88)</td>
<td>(6.68)</td>
<td>(3.66)</td>
<td>(3.16)</td>
<td>(0.05)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Lagged Other Cont</td>
<td>0.25***</td>
<td>0.09**</td>
<td>—</td>
<td>—</td>
<td>0.97**</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.04)</td>
<td></td>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Lagged Diff Other Cont</td>
<td>—</td>
<td>—</td>
<td>0.12*</td>
<td>0.01</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.03)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Last</td>
<td>-0.37</td>
<td>0.24</td>
<td>-1.82</td>
<td>10.74**</td>
<td>1.09</td>
<td>2.50</td>
</tr>
<tr>
<td></td>
<td>(5.56)</td>
<td>(4.64)</td>
<td>(6.41)</td>
<td>(5.31)</td>
<td>(0.96)</td>
<td>(1.82)</td>
</tr>
<tr>
<td>Period</td>
<td>-1.01***</td>
<td>-1.23***</td>
<td>0.07</td>
<td>-0.07</td>
<td>1.11*</td>
<td>1.13***</td>
</tr>
<tr>
<td></td>
<td>(0.24)</td>
<td>(0.21)</td>
<td>(0.30)</td>
<td>(0.26)</td>
<td>(0.06)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>Selfish_Ever</td>
<td>-15.23***</td>
<td>-33.68***</td>
<td>-0.93</td>
<td>-4.27**</td>
<td>2.20</td>
<td>36.61***</td>
</tr>
<tr>
<td></td>
<td>(4.45)</td>
<td>(5.88)</td>
<td>(2.61)</td>
<td>(2.31)</td>
<td>(1.83)</td>
<td>(31.43)</td>
</tr>
<tr>
<td>PrivOpt_Ever</td>
<td>23.35***</td>
<td>21.46***</td>
<td>2.11</td>
<td>2.32</td>
<td>0.05*</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>(8.88)</td>
<td>(7.47)</td>
<td>(5.00)</td>
<td>(2.88)</td>
<td>(0.09)</td>
<td>(0.98)</td>
</tr>
<tr>
<td>AdjustUp</td>
<td>0.90</td>
<td>13.18***</td>
<td>19.37***</td>
<td>29.61***</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>(3.59)</td>
<td>(3.73)</td>
<td>(4.28)</td>
<td>(4.21)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lagged AdjustUp</td>
<td>-0.75</td>
<td>7.95**</td>
<td>1.89</td>
<td>-1.83</td>
<td>0.67</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>(3.48)</td>
<td>(3.79)</td>
<td>(3.97)</td>
<td>(4.15)</td>
<td>(0.65)</td>
<td>(0.37)</td>
</tr>
<tr>
<td>AdjustDown</td>
<td>-16.44***</td>
<td>-18.41***</td>
<td>-22.77***</td>
<td>-30.16***</td>
<td>1.10</td>
<td>7.80***</td>
</tr>
<tr>
<td></td>
<td>(2.87)</td>
<td>(2.60)</td>
<td>(3.34)</td>
<td>(2.81)</td>
<td>(0.66)</td>
<td>(3.59)</td>
</tr>
<tr>
<td>Lagged AdjustDown</td>
<td>-7.77***</td>
<td>-6.99***</td>
<td>6.45**</td>
<td>9.57***</td>
<td>1.36</td>
<td>1.85</td>
</tr>
<tr>
<td></td>
<td>(2.85)</td>
<td>(2.48)</td>
<td>(3.24)</td>
<td>(2.72)</td>
<td>(0.87)</td>
<td>(0.88)</td>
</tr>
<tr>
<td></td>
<td>(6.06)</td>
<td>(5.18)</td>
<td>(6.66)</td>
<td>(5.38)</td>
<td>(13.57)</td>
<td>(3.76)</td>
</tr>
<tr>
<td>Lagged Anger</td>
<td>-0.14</td>
<td>-9.76*</td>
<td>1.96</td>
<td>1.22</td>
<td>0.27</td>
<td>2.94</td>
</tr>
<tr>
<td></td>
<td>(6.37)</td>
<td>(5.38)</td>
<td>(7.02)</td>
<td>(5.74)</td>
<td>(0.95)</td>
<td>(2.76)</td>
</tr>
<tr>
<td>Future</td>
<td>17.50***</td>
<td>15.89***</td>
<td>7.82**</td>
<td>7.06</td>
<td>0.01*</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>(3.45)</td>
<td>(5.28)</td>
<td>(3.88)</td>
<td>(6.17)</td>
<td>(0.03)</td>
<td>(0.18)</td>
</tr>
<tr>
<td>Lagged Future</td>
<td>14.83***</td>
<td>5.36</td>
<td>-2.36</td>
<td>-10.59*</td>
<td>0.47</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>(3.35)</td>
<td>(3.19)</td>
<td>(3.79)</td>
<td>(5.91)</td>
<td>(0.41)</td>
<td>(0.89)</td>
</tr>
<tr>
<td>Endgame_Ever</td>
<td>12.84***</td>
<td>6.84</td>
<td>0.36</td>
<td>2.51</td>
<td>0.27</td>
<td>0.13**</td>
</tr>
<tr>
<td></td>
<td>(5.15)</td>
<td>(7.09)</td>
<td>(2.85)</td>
<td>(2.56)</td>
<td>(0.23)</td>
<td>(0.12)</td>
</tr>
<tr>
<td>Endgame_Ever × Last</td>
<td>-10.98</td>
<td>0.13</td>
<td>-7.13</td>
<td>-8.91</td>
<td>41.08***</td>
<td>3.31</td>
</tr>
<tr>
<td></td>
<td>(7.44)</td>
<td>(7.06)</td>
<td>(8.55)</td>
<td>(8.07)</td>
<td>(57.56)</td>
<td>(3.78)</td>
</tr>
</tbody>
</table>

Demographics

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Yes</th>
<th>Yes</th>
<th>Yes</th>
<th>Yes</th>
</tr>
</thead>
</table>

Standard errors are in parentheses. Multilevel models with Partners (Strangers) data have random effects for teams/individuals and group (session). Models for contribution and first-difference of contribution (4-7) are multilevel linear regressions. Models for free-riding (8 & 9) are multilevel logistic regressions. Unreported demographic controls include gender, upper-level economics courses, and previous economics experiments.

Using 2-tailed tests, ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Model 4: N=462; Model 5: N=714; Model 6: N=429; Model 7: N=663; Model 8: N=386; Model 9: N=634;

The variables for chat codes are calculated by averaging across RAs in each period. For instance, if only one RA coded Future in that period, the Future variable would take a value of 0.5. For chat codes that suggest persistent preferences or perceptions, for all periods we use a variable representing the proportion of RAs who ever coded the category for that team contribution of others (Models 6-7).17 We also include a period-trend variable and an indicator variable for the last period.18

The Model 6-7 results are robust to using the level rather than difference of others’ lagged contribution. Variation Inflation Factors (VIFs) are examined for potential multicollinearity between the independent variables. The VIFs are generally small. The highest VIF we find is 3.08 for the Last period dummy with Partners data, which is correlated by construction with Period and the interaction term Endgame_Ever × Last.
(Selfish_Ever, PrivOpt_Ever, and Endgame_Ever). For example, if only one RA coded a team as anticipating endgame effects in any period, the Endgame_Ever variable would take a value of 0.5 for all periods. For chat codes of a more transitory nature, contemporaneous and one-period lagged variables were included. An interaction between Endgame and the last-period indicator is included to capture a potential endgame effect.\footnote{As robustness checks, we also estimated models with fixed effects at the group level (Partners) or session level (Strangers) and random effects at the team level. The results of these models are similar to the results reported in Table V. Models with period fixed effects in place of the period trend variable also give similar results.}

Concern for future cooperation positively correlates with contribution and is highly significant for both Partners and Strangers, Future Models 4-5 \( p \)-values = 0.001 and 0.003, respectively. For Strangers, this result suggests teams were trying to boost contribution in their session, realizing they may be randomly re-matched with teams in later periods and/or matched with teams who were matched with their previous group members. However, the one-period lagged coefficient for discussing this category is significant only in models with Partners data, Lagged Future Models 4-5 \( p \)-values = 0.001 and 0.302, respectively. Moreover, Partners coded as concerned about future cooperation tended to free-ride less frequently, Future Model 8 \( p \)-value = 0.062.

**Result 8.** Supporting Prediction 2 (Future Concern), there is strong evidence for the encouragement of future cooperation by others for Partners. There is some evidence for future concern for Strangers as well.

In Model 4, anticipating endgame effects correlates positively with contribution for Partners, Endgame_Ever Model 4 \( p \)-value = 0.013. The coefficient for the interaction of Endgame_Ever and Last is negative and insignificant, Models 4-5 \( p \)-values = 0.140 and 0.985, respectively. However, for Partners, Endgame chat is positively correlated with free-riding in the last period, Endgame_Ever \( \times \) Last Model 8 \( p \)-value = 0.008. Thus, for Partners, discussion of endgame effects suggests limited backward induction, with higher overall contribution before the final period, but more complete free-riding in the final period. For Strangers, however, anticipating endgame effects is associated with significantly lower free-riding, Endgame_Ever Model 9 \( p \)-value = 0.027.\footnote{The stronger interaction Endgame \( \times \) Last for free riding in TP compared to TS is consistent with the pattern in the right panel of Figure I. Free riding rises sharply in the last period in TP, but rises gradually in TS.}

Taken together with the previous evidence on future concern, these results suggest that many teams actively try to encourage future cooperation, but do not always attempt to exploit cooperative reputations in later periods.
Result 9. In partial support of Prediction 3 (Backward Induction), there is an endgame effect in the frequency of free-riding for Partners. However, there is little evidence of an endgame effect on overall contribution.

The coefficient for a team ever being coded as perceiving contribution as privately optimal is positively correlated with contribution, PrivOpt_Ever Models 4-5 $p$-values = 0.009 and 0.004, respectively. Such a perception of the game might be due to fairly strong unconditional prosocial preferences or due to misunderstanding the incentive structure of the game.

Result 10. Perceiving contribution as privately optimal correlates positively with contribution. This evidence is consistent with both Prediction 4 (Other-Regarding Preferences) and Prediction 5 (Confusion).

In Models 4-5 and 8-9, lagged average contribution of others and the period trend of contribution are significant and their coefficients have signs consistent with previous public goods experiments. In Models 4 and 5, the coefficient for negative emotions in response to low contributions by others is negative and marginally significant for Strangers, Anger Model 5 $p$-value = 0.086. Negative emotion in response to low contribution by others is correlated with more free-riding for Partners, Anger Model 8 $p$-value = 0.046.

In Models 6-7, the coefficient for increasing contribution because of past outcomes is positive and highly significant, AdjustUp $p$-value < 0.001 in each model. Changing contributions in response to past outcomes could be interpreted as reciprocity, reinforcement learning, or imitation learning. However, we find similar signs, sizes, and significance levels for this chat category for Partners and Strangers, despite reciprocity appearing less salient as a motivation for Strangers. Thus, while teams do adjust behavior in response to previous outcomes, we do not find strong support for reciprocity as an explanation for this behavior.

In each of the first-difference panel Models 6-7, the coefficient for decreasing contribution because of past outcomes is negative and highly significant, AdjustDown $p$-values < 0.001 in each model. Once again, we find coefficients of similar signs, sizes, and significance levels for Partners and Strangers, suggesting that reinforcement or imitation learning is a more plausible explanation than direct reciprocity.

Result 11. In contrast with Prediction 6 (Reciprocity), anger in response to others’ past behavior is negatively correlated with contribution for Strangers, but not for Partners. While teams do adjust their contribution in response to past outcomes, the effect is similar for Partners and Strangers. Moreover, anger is positively correlated with free-riding for Partners.

In Models 4-5, the coefficient for a team ever being coded as selfish (intentional free-riding) negatively correlates with contribution, with a larger effect for Strangers, Selfish_Ever Models
4-5 $p$-values $\leq 0.001$ in each model. For Strangers, discussing intentional free-riding is associated with significantly more free-riding, Selfish_Ever Model 9 $p$-value $<0.001$.

**Result 12.** There is evidence in favor of Prediction 7 (Free-Riding for Profit) for Strangers. For Partners, chat in this category is correlated with lower average contribution, but not more frequent complete free-riding.

**V Confusion in Team Chat Logs**

Previous studies such as Andreoni (1995), Houser and Kurzban (2002), and Shapiro (2009) show evidence that confusion plays a substantial role in public goods games. Confusion is typically identified as a residual source of contribution after other explanations such as social preferences are removed. By exploring our team chat logs, it is possible to gain a more direct understanding of the potential sources of confusion and how they might be mitigated. The chat codes assigned by research assistants suggest some subjects may misunderstand the game in a variety of important ways. Subjects may misperceive the incentives, believing that contribution is privately optimal, either unconditionally or conditional on the contribution of others. Subjects may also perceive substantial repeated game effects in the Strangers treatment, either due to a rational concern about being randomly rematched with the same group members, or due to misunderstanding the matching protocol.

To better understand the sources of confusion, we first personally read the chat logs in their entirety. We then defined several categories of confusion, including misunderstanding incentives, naive reinforcement learning, use of gambling terminology, and misunderstanding the matching protocol. We then had two RAs independently code the chat messages. These RAs were an MA student in Economics at VCU and an MBA student at USD. The confusion categories are not mutually exclusive, and multiple categories could be coded simultaneously.

We are particularly interested in clear cases of confusion about the incentive structure of the game. Chat logs indicating a belief that contribution is conditionally or unconditionally optimal could be due to social preferences or to misperception of incentives. Thus we instructed the RAs to be conservative in interpreting particular chats as misunderstanding incentives, favoring cases where team chats are focusing on monetary payoffs rather than preferences. The chat code frequency by team of misunderstanding incentives is 32.1%, averaging over the two RAs, with moderate agreement between RAs (κ = 0.49).

Another related form of confusion suggested in the chat logs is naive reinforcement learning. While the previous type of confusion indicates an incorrect understanding of the incentives in the game, this type of confusion suggests that a team views the game as a black box rather than
thinking deeply about incentives. In such chats, teams indicate a decision to contribute because they observed a positive contemporaneous correlation between their own contributions and payoffs within the same period. Reinforcement toward free-riding is harder to distinguish from insight into the game, so we do not interpret such chats as confusion. The frequency by team of naive reinforcement learning is 26.8%. However, agreement between RAs is somewhat weaker for this category ($\kappa = 0.29$).

We also found that some teams used gambling terminology such as referring to contributions as “bets” or “bids.” This finding is unexpected, similar to the prevalence of “Go Big” chat. While gambling-related framing of the public goods game is not necessarily indicative of confusion, it does appear suggestive that subjects may view the group account as similar to a slot machine where higher “bets” also increase potential contemporaneous rewards. Such a mental model of the game could lead subjects to incorrect intuitions about the consequences of their decisions. In interpreting such chats, we instructed RAs not to include idioms such as “best bet,” or the frequently-used phrase “all in,” which could be a gambling reference or simply a literal reference to putting all tokens into the group account. The frequency by team of gambling references is 13.7%, with moderate agreement between RAs ($\kappa = 0.46$).

Another source of potential confusion worth examining is misunderstanding the matching protocol. For Partners, we do not find any indication of such misunderstanding. For Strangers, the frequency by team of misunderstanding the matching protocol is 8.8% (typically suggesting a belief that groups are fixed), with moderate agreement between RAs ($\kappa = 0.40$). Another 14.7% of Strangers teams explicitly indicate concern for the possibility of random rematching with decision makers from previous groups ($\kappa = 0.51$). This second type of repeated game concern for Strangers is not due to confusion. However, it underscores the finding that some Strangers teams do appear to be concerned about repeated game effects, either due to confusion about matching or due to the chance of randomly meeting past decision makers in future periods.

In additional analyses reported in Appendix subsection AI.6, we report regressions relating contribution choices to the confusion categories discussed in this section.\footnote{Other idiosyncratic types of confusion occurred occasionally, such as misunderstanding how many periods remained (4.8% of teams, $\kappa = 0.50$), despite this information being displayed throughout the experiment.} We find that teams ever coded as misunderstanding incentives contribute significantly more, for both Partners ($p$-value=0.009) and Strangers ($p$-value<0.001), consistent with Prediction 5 (Confusion). Moreover, we find suggestive evidence that this correlation is somewhat stronger for Strangers than for Partners, discussed in Appendix subsection AI.7. This result may help explain why overall contribution is similar for Partners and Strangers, despite the greater prevalence of future concern in Partners matching.
Designers of future experiments may benefit from understanding these sources of potential confusion. While we used standard public goods game instructions, including examples and review questions, additional training of subjects may be useful to help them better understand the incentive structure of the game. Review questions, wording of the instructions, and payoff tables might be designed specifically to address the sources of potential confusion identified in this experiment. Increasing incentives by using larger monetary stakes might also motivate subjects to think more deeply about the experiment. Subject pool differences are also possible, with perhaps less confusion at highly-selective elite universities. Group size might be important, as groups of two individuals may be less confused than larger groups (Yamakawa et al., 2016). Experimenters seeking to eliminate repeated game effects, while allowing multiple periods, may need to provide additional explanation of the matching protocol. It may also be necessary to increase the number of subjects in the lab and/or reduce the number of periods, and if possible, use “Perfect Strangers” matching where no subject faces any particular rival more than once. While further study is needed to determine the best solutions, understanding these issues is an important step toward improving the design of public goods experiments.

VI Conclusion

In this experiment, we compare contribution choices of individuals and two-person teams in public goods games and examine the motives of teams by analyzing their chat logs. We find similar overall contribution for teams and individuals. However, initial contribution is higher for teams, and there are stronger endgame effects for teams, with more frequent free-riding in the final period. These features of the data are qualitatively similar to the results found in Auerswald et al. (2016), though they find higher overall contribution among teams, while we do not. Kamei (2016) also finds stronger endgame effects for teams, but finds substantially higher contribution for teams with Partners matching. A number of design differences, such as group size and feedback, might explain these differing results. Further research is needed to better understand how these design features influence team contribution choices.

Our analysis of team chat logs indicates Partners and Strangers differ substantially in the degree to which they are forward-looking in their decisions. Partners discuss concern for future contribution and the future actions of others more frequently than Strangers. Chat logs further suggest that teams are limited in their use of backward induction to anticipate changes in play near the end of the game. Most anticipation of endgame effects occurred in later periods. Such discussion is associated with higher average contribution and less frequent free-riding in early periods, but more frequent free-riding in the last period, particularly for Partners.
We find little direct evidence of pro-social motives for contribution. Discussions of such other-regarding preferences are somewhat rare, especially among Strangers. Nonetheless, subjects may be motivated by pro-social preferences despite the low frequency of direct discussions to this effect. Other categories such as perceiving contribution as privately optimal could be interpreted as (indirectly) indicating pro-social preferences. We find evidence that some subjects perceive incentives in this way. Pro-social other-regarding preferences might transform the monetary payoffs into utility payoffs with such structures. However, we think confusion plays a role, as other features in the chat suggest confusion of various types. Some teams appear to misunderstand the pecuniary incentives. Some use gambling terminology, such as referring to a contribution as a “bet” or “bid.”

Our team chat analysis is most directly relevant to contribution decisions made by teams. However, we find team and individual contribution decisions are largely similar. Thus our chat analysis may also provide suggestive evidence on the motives underlying individual contribution decisions. Nonetheless, future research might compare cooperation of individuals and teams in other environments.

Understanding the motives of contributors is important for policy makers. Contributors’ concern for future cooperation may be particularly useful. Organizations seeking repeated monetary contributions or volunteer work might encourage potential contributors to provide a positive example for others. Similar strategies might be used to discourage anti-social behaviors such as littering and unsafe driving. In the workplace, managers might discourage shirking within work groups by reminding workers of the importance of future cooperation. Studying the effectiveness of such priming strategies in strengthening the future concern motive in social dilemmas is another interesting direction for future research.

ACKNOWLEDGEMENTS

We thank Jared Weber, Ashley Halloway, Clarence Castro, Jonathan Howard, and Reed Muehler for excellent research assistance. We are grateful to Co-Editor Tim Cason and two anonymous referees for very helpful comments and suggestions. We also thank John Kagel, David Cooper, James Walker, Arlington Williams, Lata Gangadharan, Oleg Korenok, Tim Salmon, Peter McGee, Glenn Dutcher, Kenju Kamei, Edward Millner, Javier Portillo, Andrzej Baranski, David Kingsley, Lawrence De Geest, David Bruner, Dave McEvoy and the participants of the 2015 IMEBESS Meeting, CEA Conference, ESA North American Meeting, SEA Annual Meetings, and seminars at UMass Amherst Resource Economics, USD Beacom School of Business,
and VCU Economics. Any remaining errors are the responsibility of the authors. Financial support for was generously provided by the Beacom Opportunity Fund at the University of South Dakota. Support for additional research assistants was provided by VCU School of Business.

REFERENCES


