The international platform of Ludwig-Maximilians University's Center for Economic Studies and the Ifo Institute

CESifo Working Papers
www.cesifo.org/wp

Tipping versus Cooperating to Supply a Public Good

Scott Barrett
Astrid Dannenberg

CESIFO WORKING PAPER NO. 5274
CATEGORY 13: BEHAVIOURAL ECONOMICS
MARCH 2015

ISSN 2364-1428

CESifo
Center for Economic Studies & Ifo Institute

CESifo *Working Paper No. 5274*

Tipping versus Cooperating to Supply a Public Good

Abstract

In some important multi-player situations, such as efforts to supply a global public good, players can choose the game they want to play. In this paper we conduct an experimental test of the decision to choose between a "tipping" game, in which every player wants to contribute to the public good provided enough other players contribute, and a prisoners' dilemma, the classic cooperation game. In the prisoners' dilemma, the first best outcome is attainable, but cannot be sustained as a Nash equilibrium. In the tipping game, only a second best outcome may be attainable, but there exists a Nash equilibrium that is strictly preferred to the one in the prisoners' dilemma. We show that groups do significantly better when they choose the tipping game, and yet many groups repeatedly choose the prisoners' dilemma, indicating a mistaken and persistent tendency to prefer a game with potentially higher payoffs to one having a strategic advantage.

JEL-Code: C720, C920, F530, H410.

Keywords: prisoners' dilemma, tipping game, experiment, public good, agreements, voting, environment, global public goods.

*Scott Barrett**
School of Internnational and Public Affairs
Columbia University
420 W. 118th Street
USA – New York, NY 10027
sb3116@columbia.edu

Astrid Dannenberg
University of Kassel / Germany &
University of Gothenburg / Sweden
ad2901@columbia.edu

*corresponding author

In "tipping" games, players behave differently either side of a "tipping point." In this paper, tipping represents a way of supplying a public good.[1] On one side of the tipping point, no player wants to supply the public good; on the other side, every player wants to supply it. The tipping point thus represents a critical number of providers of the public good. Tipping is to be contrasted with the usual approach to supplying a public good, represented by the prisoners' dilemma, in which every player has a dominant strategy not to contribute.[2] We report the results of an experiment in which members of a group vote to choose which of these games to play, knowing that a majority decides. The choice they face is difficult. The prisoners' dilemma can potentially achieve the overall first best outcome, but it cannot support this outcome as a Nash equilibrium. The tipping game may be able to support a Nash equilibrium that is Pareto-superior to the one in the prisoners' dilemma, but choice of this game might also foreclose any chance of attaining the first best outcome.

The prime example of the situation we have in mind is the provision of a *global* public good. Before countries choose how to play (choose their contributions to the public good) they must first agree on the rules of the game. For example, should they impose limits on the emissions of a pollutant or should they mandate a technology standard, the adoption of which would cause emissions to fall? The first approach is direct and leaves the prisoners' dilemma unchanged. The second approach, under the right conditions, is strategic and can turn the prisoners' dilemma into a tipping game. For example, if the adoption of a new technology entailed substantial network externalities, then it would pay every country to adopt the technology as soon as a critical group of other countries adopted it. The problem with technology standards, however, is that they are rarely the most cost-effective way to meet a particular environmental goal. Adoption of emission limits, by contrast, allows parties the flexibility to meet their obligations

[1] The classic reference to tipping is Schelling (1971), though his concern in this paper is with racial segregation, not the provision of a public good. Runge (1984), building on Sen (1967), has suggested that the provision of public goods may more closely resemble an "assurance game" than a prisoners' dilemma, provided people have "fairminded" preferences. Our paper is more consistent with Snidal (1985), who sees prisoners' dilemma and coordination games as pertaining in different situations. However, in our formulation, players do not simply find themselves in one situation rather than another. Instead, they *choose* the situation they want to be in.

[2] Tipping also differs from the much-studied threshold public goods game in which, once a threshold number of players has contributed, none of the other players wants to contribute to the public good (for a review of threshold public goods experiments, see Croson and Marks 2000).

using the most cost-effective means. However, an agreement specifying emission limits leaves the prisoners' dilemma unchanged and so may have difficulties deterring free riding. Both approaches have been tried in the past to address a number of issues, ranging from climate change to ozone depletion to pollution of the seas (see section 5). But which approach is best? This is a difficult question to answer in general because of the lack of a counterfactual: we don't know what would have happened had these issues been addressed differently. By studying behavior in the lab, however, we can observe the outcomes realized by groups that choose differently. We can also observe whether the groups that choose badly see the error in their ways and reverse their decision at a later point in time.

As the tipping game has different strengths and weaknesses than the prisoners' dilemma, it may not be obvious which one will turn out to be the better choice in the end. Tipping games have multiple equilibria, and it may be difficult for the players to coordinate their behavior. As we explain later, coordination is especially difficult in the tipping game that is capable of sustaining only a second best outcome. The prisoners' dilemma, by contrast, has a unique equilibrium, but decades of experimental research have shown that many people do not play the equilibrium strategy, at least not at the beginning of the game (Ledyard, 1995). Forming expectations is thus difficult in both games, but giving the players the opportunity to update their beliefs may enable them to make better decisions over time.

Previous experiments on the endogenous choice of institutions have shown that individuals and groups often choose naïvely at first (for a review of this literature, see Dal Bó 2011). However, when given the opportunity to revise their initial choice, players often move gradually towards the welfare improving institution. In most cases, this welfare superior institution involves the use of punishments or rewards (e.g. Gürerk, Irlenbusch, and Rockenbach 2006; Ertan, Page, and Putterman 2009; Sutter, Haigner, and Kocher 2010; Markussen, Putterman, and Tyran 2014).[3]

[3] Punishments in the form of economic sanctions are rarely used to influence foreign policy, our main concern in this paper, perhaps because they are often ineffective when they are used (Hufbauer et al. 2007).

Our experiment comes closest to two recent experiments conducted by Dal Bó, Foster, and Putterman (2010) and Dal Bó, Dal Bó, and Eyster (2013).[4] In both of these experiments, players can vote to modify the payoffs of a two-person prisoners' dilemma. In Dal Bó et al. (2010), the players can vote for a fine to be imposed on unilateral defection, an off-equilibrium change that makes mutual cooperation another Nash equilibrium of the game while leaving the payoffs to mutual cooperation unchanged. In this experiment, groups that voted for the change earned higher payoffs, but only about half the groups voted this way. However, because the players were not allowed to revise their choice, we don't know if they would have corrected their mistake in a second voting round.[5] In Dal Bó et al. (2013), the players can vote for a fine that reduces the payoff to playing *every* strategy, but with the payoff to defection falling by more than the payoff to cooperation. In this alternative game, mutual cooperation has a lower payoff than in the original two-player prisoners' dilemma but cooperation becomes the dominant strategy for both players. As in the previous paper, groups that voted for the change earned higher payoffs, but only about half the subjects voted this way. However, in a treatment (Majority Repeated) that allowed subjects to vote repeatedly before each of the five rounds of play, the players learned to overcome their bias in favor of the prisoners' dilemma. By the end of this treatment, only two out of twenty groups were still playing the prisoners' dilemma.

In our experiments, choice of a regime is harder than in Dal Bó et al. (2010) because going for the tipping regime may mean foregoing the opportunity to realize a higher payoff in the prisoners' dilemma. Choice of a regime in our experiments is also harder than in Dal Bó et al. (2013) because our alternative game has two Nash equilibria, only one of which offers a higher payoff compared to the Nash equilibrium in the prisoners' dilemma. Another difference is that, in our experiments, five players vote for which game to play and then play the chosen game *as a group*—a context that is particularly suited to understanding negotiations of multilateral international agreements.

[4] We only became aware of the Dal Bó, Dal Bó, and Eyster (2013) paper after we conducted our experiment.

[5] The primary aim of Dal Bó et al. (2010) is to show that a regime imposing the fine has a bigger effect on behavior when it is chosen by the players who will ultimately be subject to the fine than when it is imposed upon these players without their consent.

In the experiments noted above, by contrast, a group of four (in Dal Bó et al. 2010) or six (in Dal Bó et al. 2013) players chooses which game to play with *pairs* of players then playing the chosen game—a context that is more suited to the study of domestic law making. Finally, we also make public the results of each vote, a design feature that can help the players to coordinate. This assumption is consistent with the way multilateral negotiations are conducted,[6] but would of course be inappropriate for the study of a "democracy" in which the final vote tally is public knowledge but not the voting decisions of particular individuals.

We find that every group that chooses to play the tipping game is able to coordinate perfectly, sustaining a 100 percent group contribution level. As a consequence, even when the tipping game can only sustain a second best outcome, the groups that play the tipping game earn higher payoffs than the groups that play the prisoners' dilemma. Similar to Dal Bò et al. (2013) and other experiments on endogenous institutions, we find that players are initially unsure of which game to play but that, over time, they move towards the regime that pays off more handsomely. When the tipping game can sustain the first best outcome, all groups move quickly and decisively to this game. However, and in contrast to earlier findings, when the tipping game can sustain only a second best outcome, only half the groups move to this regime. The other groups remain trapped in the prisoners' dilemma. These trapped groups sustain more cooperation than the other groups when playing the prisoners' dilemma, but this success ultimately works against these groups' interests as it makes them less likely to switch. The groups that remain trapped believe that they have made the better choice, but all the evidence we have suggests that this belief is wrong. The groups that switch to the tipping game immediately change their behavior and perform better than the groups that stick to the prisoners' dilemma.

In the next two sections we present our underlying model and describe our experimental design and treatments. In Sections 3 and 4 we present our main results on the choices made by individuals and groups, and show how these choices are shaped by

[6] For example, the rules of procedure for the Montreal Protocol say that substantive decisions are to be made by a show of hands or a rollcall vote. (See http://ozone.unep.org/Publications/VC_Handbook/Section_3_Rules_of_Procedure/Rules_of_procedure.shtml.)

expectations. In Section 5 we use our results to interpret several real world examples of agreements to supply a global public good. We conclude with some final observations on our main results.

1. Model

There are N symmetric players. In the meta-game, the players first choose which game to play, the A Game or the B Game. They then play the game they have chosen. We begin by describing these individual games.

In the A Game, the players have a binary choice; every player i $(i=1,...,N)$ must choose $q_i \in \{0,1\}$, taking as given the choices made by the other players. Letting k denote the number of *other* players that play $q_j = 1$, i's payoff is assumed to be given by

$$\pi_i^A(1;k) = b(k+1) - c,\ \pi_i^A(0;k) = bk, \tag{1}$$

with $bN > c > b > 0$. In this game, play $q_i^* = 0$ is the unique Nash equilibrium, but full cooperation requires that every player i play $q_i^{FC} = 1$. This is a prisoners' dilemma.

In the B Game, the players have another binary choice; every player i must choose $y_i \in \{0,1\}$. Letting m denote the number of *other* players that choose $y_j = 1$, player i's payoff is assumed to be given by

$$\pi_i^B(1;m) = b(m+1) - c - d,\ \pi_i^B(0;m) = \alpha m. \tag{2}$$

The parameter d represents the cost-penalty to playing the tipping game as compared to the prisoners' dilemma. Assume $d \geq 0$, $N > (c+d-\alpha)/(b-\alpha)$, and $c+d > b > \alpha \geq 0$. It is then easy to show that $\pi_i^B(1;m) > \pi_i^B(0;m)$ for $m > \tau$ and $\pi_i^B(1;m) < \pi_i^B(0;m)$ for $m < \tau$, where $\tau = (c+d-b)/(b-\alpha)$. τ thus represents the "tipping point" for the B game. Our assumptions about the parameters imply $\tau \in (0,N)$.

In the B game, there are two Nash equilibria in pure strategies. In one, every player plays $y_i^* = 0$. In the other, every player plays $y_i^{**} = 1$. All players earn strictly

higher payoffs in this second pure-strategy Nash equilibrium compared to the first one.[7] However, as explained in the next section, it is not obvious that the players will be able to coordinate on this second equilibrium. Moreover, partly for this reason, it is not obvious that the players will choose B over A in the metagame.

2. Experimental Design

Our experiment is played by groups of $N = 5$ players. In the metagame, each group must choose between playing the A game, a prisoners' dilemma, and the B game, a tipping game. The A game is the same in all of our treatments. The B game, however, varies with the treatment. In the treatment *Vote-First-B-10*, groups choose between A and B-10. In *Vote-First-B-8*, they choose between A and B-8. The difference between these treatments is that the Pareto-superior (pure strategy) Nash equilibrium in B-10 yields the same payoff as the full cooperative outcome in the A game, whereas the Pareto-superior Nash equilibrium in B-8 yields a lower payoff compared to the full cooperative outcome in the A game.

In both treatments, the experiment is played in four phases; see Figure 1. At the start of each phase, the players vote to choose the game they want to play, with a simple majority deciding.[8] Afterwards, they play the chosen game in five consecutive contribution rounds, with all the players choosing (simultaneously) whether to contribute to the public good in each round. Since there are four phases, there are 20 contribution rounds in total. It is common knowledge that individual votes are made public to all the

[7] There also exists a mixed strategy equilibrium in which every player earns an expected payoff somewhere in between the payoffs corresponding to these pure strategy equilibria. Letting p denote the probability, from every player i's perspective, that each player $j, j \neq i$, will play $y_j = 1$, the mixed strategy equilibrium involves each player choosing to contribute with probability $p^* = \tau/(N-1)$, yielding each player an expected payoff $E(\pi_i^B) = \alpha\tau$. It is easy to confirm that $\pi_i^B(1;N-1) > E\left(\pi_i^B\left(p^*;p^*(N-1)\right)\right) \geq \pi_i^B(0;0)$.

[8] The voting stage can be thought of as a game for choosing a "frame" for the contributions game that is to be played subsequently. Decisions about framing are routinely made by a vote. For example, the rules of procedure for meetings of the parties to the ozone agreements say that "decisions…on all matters of substance shall be taken by a two-thirds majority vote. . . ." (see http://ozone.unep.org/Publications/VC_Handbook/Section_3_Rules_of_Procedure/Rules_of_procedure.shtml).

players after each voting round and that individual contribution decisions are made public after each contribution round.

In every contribution round, each player is given two playing cards, one red and one black, and must decide which card to return. If player *i* is playing the A game, returning the red (black) card is equivalent to choosing $q_i = 1$ ($q_i = 0$). If playing the B game, returning the red (black) card is equivalent to choosing $y_i = 1$ ($y_i = 0$). In both cases, handing back the red card supplies the public good.

Every player's payoff, relative to the theoretical model, is increased by an amount *s*. This scaling has no effect on the theory, but is needed to ensure that players cannot be left out of pocket when playing the experiment. In the A game, players get $s - c$ if they hand in their red card and *s* if they hand in their black card. Either way, they get *b* for every red card handed in by anyone in the group.

In both versions of the B game, players who hand in their black card get a payoff of *s* plus an amount α for every red card handed in, whereas players who hand in their red card get a payoff $s - c - d$ plus an amount *b* for every red card handed in. The difference between B-10 and B-8 is reflected in the value of *d*. Our experiments assume $\alpha = 0$, $b = 2$, $c = 5$, and $s = 5$ throughout, with $d = 0$ for B-10, and $d = 2$ for B-8. The A and B (that is, B-10 and B-8) games are shown in Figure 2.[9] Here it can be seen that the "10" in B-10 and the "8" in B-8 represent, respectively, the full cooperative payoffs in these two games (the full cooperative payoff in the A game is 10). Note as well that the closed dots in Figure 2 represent Nash equilibria (the mixed strategy equilibria of the B games are "interior"), and the open circles represent the efficient outcomes for the different games. The payoffs are shown in Figure 3.

For which game will people cast their vote? The Pareto-inefficient Nash equilibria in the B games are neither better nor worse than the unique Nash equilibrium in the A game, whereas the Pareto-efficient pure strategy equilibrium in both B games is strictly preferred by all players to the Nash equilibrium of the A game. It might thus seem that

[9] This kind of figure was first developed by Schelling (1978).

the players should vote for B. However, closer inspection reveals a more complex picture.

Some players might form a "first impression" of which game they should choose by looking at the payoffs. From this perspective, there are many reasons for players to prefer A to B. In the *Vote-First-B-10* treatment, for example, the lowest individual payoff is the same in the A and B games, whereas A pays out the highest individual payoff. Similarly, in *Vote-First B-8*, the lowest and the highest individual payoffs are both strictly higher in the A game than in the B game.[10]

Some players might look more deeply into these games, trying to reason through how their co-players will play. For example, in B-10, they might see that coordination on the welfare superior equilibrium in the B game seems likely given that playing Red in this game is both payoff dominant and risk dominant (if each player believes that the other players are equally likely to play Red or Black, then each player can expect that two other players will play Red, in which case each player can expect to get a payoff of 6 by playing Red and a payoff of 5 by playing Black). In B-8, reasoning through how others will play is more difficult. The tipping point is higher for B-8 than for B-10. Moreover, the Pareto inefficient Nash equilibrium is risk dominant, making B-8 a Stag-Hunt-type game. For both reasons, coordination on the Pareto-efficient equilibrium seems less sure in this game.

However, and as noted before, in our experiment individual votes are public knowledge. Votes not only determine the game that is chosen; they also serve as a signal for subsequent contribution decisions. This signalling should be particularly useful in the B games, where the simple majority (at least 3 out of 5) is equal to or greater than the tipping point. In both treatments, it makes the most sense for players to vote for B if they think coordination on the mutually preferred equilibrium will succeed. But players who believe coordination will succeed should then play Red when B is chosen. Hence, *all* the

[10] Note also that in *Vote-First-B-10,* given the choice by each player to play Red or Black, the aggregate payoff is never lower and is often higher in the A game. Also, a person who intends to play Black does at least as well opting for A, whereas a person who intends to play Red is indifferent between A and B. Similarly, in *Vote-First-B-8*, players might be drawn to A because it offers the highest collective payoff. They might also notice that a person intending to play Red is strictly worse off when playing the B game than the A game, given the choices by the other players to play Red or Black, and that a person intending to play Black does at least as well choosing A as B.

B-voters should play Red when B is chosen. But then the A-voters should expect that all the B-voters will play Red, making it in *their* interests to play Red, too. In other words, with vote signalling, players should expect that coordination on the efficient equilibrium will succeed, even for the B-8 treatment. However, this reasoning demands an unusual degree of sophistication. Some players might reason through their decision problem in this way, but others might go with their "first impression" or simply make a guess for how to play.

3. Experimental Results

The experimental sessions were held in a computer lab at the University of Magdeburg, Germany, using undergraduate students recruited from the general student population. In total, 300 students participated in the experiment, each student taking part in one treatment only. There were three treatments (the two *Vote-First* treatments discussed previously and one *Play-First* treatment discussed in Section IV.C) with 20 groups per treatment and five players per group.

The experimental instructions handed out to the students included several numerical examples and control questions.[11] The control questions tested subjects' understanding of the game to ensure that they were aware of the available strategies and the implications of making different choices. After reading the instructions and answering the control questions correctly, subjects began playing the game. In each session, 25 subjects were seated at linked computers (game software Ztree; Fischbacher 2007) and randomly assigned to one of five five-person groups. The subjects did not know the identities of their co-players, but they did know that the membership of their group remained unchanged throughout the session. To ensure anonymity, each individual within a group was identified by a different number, from 1 to 5. During the game, earnings were displayed in tokens. It was public knowledge that payments would be calculated by summing up the number of tokens earned over all 20 contribution rounds and by then applying an exchange rate of €.10 per token. Before and after the game, the subjects were

[11] The experimental instructions are provided in Appendix A.

asked to complete questionnaires.[12] After the final questionnaire was completed, the subjects were paid their earnings in cash.

Our main results for the *Vote-First* treatments are shown in Figures 4 and 5 and summarized in Table 1. Figures 4 and 5 show the average payoff per contribution round for each group (of which there are 20 in total), depending on the game chosen by the group (A in blue, B in orange) over the four phases.[13] A quick look at the figures shows that behavior differs dramatically between *Vote-First-B-10* and *Vote-First-B-8*. We discuss these differences in detail below.

3.1. Voting

Consider, to begin, the voting behavior of individuals, summarized in Figure 6. In *Vote-First-B-10*, 57 per cent of players voted for game B in the initial phase, rising to 91 per cent by the fourth and final phase.[14] In *Vote-First-B-8*, 11 per cent of the players voted for B initially, rising to 51 per cent by the final phase. The switching behavior of all individuals taken together (in favor of B) is similar for the two treatments, but the initial support given to A rather than B differs greatly.

The behavior of individuals is consistent with these aggregate observations. In *Vote-First-B-10*, 36 per cent of the players started by voting for A, and then switched to B at some point without ever switching back, whereas in *Vote-First-B-8*, 37 per cent of the players voted this way. In *Vote-First-B-10*, 4 per cent switched from B to A before switching back to B, whereas in *Vote-First-B-8*, 7 per cent did this. Finally, in *Vote-First-B-10*, 6 per cent switched from B to A without ever switching back, whereas in *Vote-First-B-8*, 11 per cent behaved in this same way. Again, the main difference in behavior is reflected in the "core" support for A rather than B. In *Vote-First-B-10*, 51 per cent of the players voted for B every time, whereas just 3 per cent voted for A every time. In

[12] The post-play questionnaire results are discussed in Section 4.2; the pre-play questionnaire results are discussed in footnote 16.
[13] We only show payoffs as the figures for contributions reveal a nearly identical pattern.
[14] Individual voting behavior in the initial phase of *Vote-First-B-10* is surprisingly similar to the results observed by Dal Bó, Foster, and Putterman (2010). In their experiment, 54 percent of players voted to play the coordination game. However, as we show here, support for the coordination game quickly increases when players have the chance to revise their choice.

Vote-First-B-8, this behavior was almost reversed, with 38 per cent voting for A every time and just 7 per cent voting for B every time. To summarize:

Result 1. *In Vote-First-B-10 there is core support for B, whereas in Vote-First-B-8 there is core support for A. Vote switching behavior is very similar for the two treatments, with the vast majority of switchers moving from A to B.*

This voting behavior is reflected in the choices made at the group level. In *Vote-First-B-10*, 11 out of 20 groups started by playing B and never switched. The other nine groups initially gave their support to A, but all of these groups switched to B at the next opportunity, never to look back again. Support for B was thus prompt and decisive. In *Vote-First-B-8*, *all* groups started by playing A. In the second phase, two groups switched to B, but these groups subsequently switched back to A before returning to B again in the last phase. These groups' support for B was tentative. Four groups persisted in playing A until the last phase, when they finally switched to B. These groups' support for B was reluctant. Ten groups *never* chose B. These groups were strongly attracted to A and/or repelled by B (we discuss these effects later). Overall, the difference in group-behavior between the two treatments is highly significant. The proportion of groups choosing to play B is significantly higher in *Vote-First-B-10* than in *Vote-First-B-8* (Fisher's exact test, $p < 0.01$ for each phase).[15]

Result 2. *In Vote-First-B-10, groups were initially divided in their support for A and B, but support quickly shifted to B; ultimately, group support for B was universal. In Vote-First-B-8, all groups started out supporting A, but over time about half the groups hesitantly and reluctantly switched to B; the other groups never chose B.*

3.2. Contributions and Payoffs

In *Vote-First-B-10*, the groups that chose A in the first phase contributed 31 per cent of their red cards in the first contributions round, declining to 9 per cent by the fifth round, for an average of 21 per cent (see Table 1). The groups that chose B in the first phase of *Vote-First-B-10*, by contrast, started out making high contributions and then increased

[15] Unless stated otherwise, all statistical tests reported in this paper are two-sided and take the group as unit of observation—a conservative approach.

these, quickly reaching the maximum level. Taking the group average for the first phase as the unit of observation, a Mann-Whitney-Wilcoxon (MWW) test shows that the difference in contributions between the groups that played A and the groups that played B is highly significant (p = 0.00). The players in *Vote-First-B-10* also received a higher average payoff when they played B than when they played A (p = 0.00).

Over all four phases of *Vote-First-B-8*, contributions in the A game (averaged over all groups playing A) generally declined (see Table 1).[16] Contributions in the first phase of the A game averaged 39 per cent, dropping to 10 per cent by the last phase. Contributions started at 62 per cent in the first contributions round, declining to 5 per cent by the 20th round. As in *Vote-First-B-10*, contributions in the B-8 game settled at the optimal level by the end of every phase in which B was played. Also, following each vote, average contributions for the groups playing B are always significantly higher than for the groups playing A (MWW test, $p < 0.05$ for each phase). Average payoffs are also higher for the groups that chose B rather than A. Here, the differences are weakly significant for the second and third phases ($p < 0.10$), and highly significant for the last phase (p = 0.00).

Result 3. *For both of the Vote-First treatments, contributions and payoffs are significantly higher when groups play B than when they play A.*

3.3. The Prisoners' Dilemma Trap

The behavior of Group 25 (see Figure 5) demonstrates the allure of the A game in *Vote-First-B-8*. The players in this group are initially drawn to A, probably because playing A has the potential of yielding a higher payoff. The problem is that this potential can only be realized if all the players hand in their red cards when playing A, and the group is unable to sustain much cooperation for long. By contrast, these same players coordinate flawlessly when playing B. Being unable to sustain a first best, Group 25 eventually settles for the second best. Many other groups behave similarly. However, about half do

[16] As shown in Figure 5, contributions by individual groups reflect a similar pattern.

not—and these groups, the ones that persist in playing A, earn a lower payoff than the groups that switch. Together, Results 2 and 3 imply:

Result 4. *In Vote-First-B-10, all groups converge quickly to the B game, and then coordinate flawlessly, sustaining the first best outcome. In Vote-First-B-8, some groups move hesitantly and reluctantly to the B game, eventually coordinating flawlessly and sustaining the second best outcome. The other groups remain "trapped" in the A game. These groups cling stubbornly to A even though they would almost certainly do better by switching to B.*

The last conclusion follows from the observation that every group that switched to B changed its behavior and did better. The reason we have to qualify our conclusion with the words "almost certainly" is that we cannot rule out the possibility that the groups that switched did better because of the characteristics of their members and that the groups that didn't switch, having a different membership, might not have done better by switching. There is no way to test this hypothesis directly, but our experiment does offer supporting evidence.

First, we can observe how well the players who voted for A did when they were forced to play B. Table 2 (left side) compares the A-voters who played A with the A-voters who were forced to play B (because of the way their fellow group members voted) for each phase. It shows that A-voters always earned more when playing the B game than when playing the A game. The differences in between-group behavior within the same phase are not statistically significant in the second and the third phases but they are highly significant in the fourth phase (MWW test, $p = 0.00$).[17]

Second, we can also do within-group comparisons. Ten groups switched from A to B at some point (we ignore here the second switch from A to B by groups 21 and 25; see Figure 5). Comparing the payoffs of the players who voted for A in both phases, before and after their group switched to B, we find that 75 per cent of these (same) A-voters earned strictly more when they were forced to play B because of how their group

[17] A more conservative comparison would include only the groups that have either two or three A-voters (see right side of Table 2). These groups differ by only one A-vote, and so may be less likely to differ in some unobservable ways. The results remain essentially the same. In all phases, A-voters earned more when playing B than when playing A. The difference in the fourth phase remains significant (MWW test, $p = 0.01$).

voted. A Wilcoxon signed-rank test that takes the group as the unit of observation shows that the A-voters who got their way and were able to play A earned a lower payoff (a result that holds with weak significance; $p = 0.06$) than when they were in a minority and were forced to play B.

Both sets of results show that the A-voters who got their way, and were able to play A because of how their co-players voted, earned less on average than the A-voters who found themselves in a minority and were thus forced to play B. Although we cannot prove that the groups that got stuck in A would have done better had they played B, the evidence just presented certainly points in this direction. We provide additional evidence for this claim in the next section.

4. Analysis of Expectations

What are the reasons some groups remain trapped in A and some switch to B in *Vote-First-B-8*? The analysis developed below draws from varying kinds of evidence, and yields a strong result:

Result 5. *In Vote-First-B-8, whether groups switch to B or persist in playing A depends on two different forces. Low expectations for successful cooperation "push" groups away from A. High expectations for successful coordination "pull" groups towards B. Both effects are necessary; neither is sufficient.*

Our evidence for this result is of three types. First, we are able to infer expectations from the choices observed in the games. Second, we asked the players in an ex post questionnaire what they expected and what motivated their choices. Finally, we conducted two additional treatments, called *Play-First*, in which we manipulated expectations by requiring that groups gain experience playing both games before choosing the game they would prefer to play.

4.1. Inferring Expectations From "Actual" Behavior

In *Vote-First-B-8*, *all* groups initially chose to play A, a group behavior that reflects an expectation by a majority that payoffs will be higher when playing A than when playing

B. Plainly, groups must have a disappointing experience playing A before being willing to try B. The push effect is thus necessary for getting players to move to B.[18]

We can also show that a stronger push effect increases the likelihood of any group moving to B. Table 3 presents results from a series of probit regressions. Columns 1 and 2 show regressions of the decision by individuals to vote for A in any phase, conditional on this individual having played A in the previous phase.[19] The results reveal voting inertia: individuals tend to vote for A if they voted for A previously. This was to be expected since, as previously noted, once individuals vote for B they rarely switch to A. However, the results also reveal that the decision to vote for A depends strongly on the experience associated with having played A previously. The players that switched their vote to B had a particularly bad experience when playing A in the previous phase. Note that this effect is observed irrespective of whether an individual's experience is measured by his or her payoff when playing A (column 1) or his or her group's aggregate contribution level when playing A (column 2). Finally, column 3 shows that this result also holds at the group level: groups are more likely to stick with A if they experienced a higher contribution level when playing A in the previous phase.

To summarize, in *Vote-First-B-8*, all groups are initially drawn to A. Some are later "pushed" into trying B if and when their experience in playing A proves disappointing. This effect of getting groups to try B is crucial. Of the ten groups that tried B at some point, all but one ended up choosing B in the last voting round. Of the 11 groups that chose A in the last round, only one had ever tried playing B before.

Result 6. *Doing better in the A game makes individuals less likely to vote for B; but since payoffs are generally higher in the B game than in the A game, greater success in the A game paradoxically makes players worse off overall.*

[18] There is, of course, a chance that had individuals been sorted differently, at least one group would have comprised a majority of first-time B voters. In our experiment, out of 100 players in the *Vote-First-B-8* treatment, 86 voted for A at the first opportunity and 14 voted for B. The probability that a group of five players drawn at random from this pool of 100 players will contain at least three first-time B-voters can be shown to be less than two percent.

[19] We obviously exclude from this regression all the first-phase votes, which depend only on expectations. This leaves 300 observations (three phases times 100 players voting in each phase). However, we must also exclude the 30 observations corresponding to situations in which groups played B in the previous phase, leaving 270 observations.

This last result, which we revisit in our concluding section, depends again on the groups that are trapped in A doing better were they to switch to B. We provide more support for this claim later in sub-section C.

We now turn to the pull effect. Recall from Section 2 that it can only pay players to vote for B if they believe coordination will succeed. But if players believe that coordination will succeed, then they will want to contribute their red cards when playing B. Hence, players who expect coordination to succeed when playing B should be more inclined to vote for B and to play Red when B is chosen. Players who expect coordination to fail should be less inclined to vote for B and to play Red when B is chosen. As explained in Section 2, sophisticated reasoning suggests that even the A voters should play Red if B is chosen. However, not everyone may reason this way.

Table 4 presents a probit regression of individual contribution decisions in the first round of playing A (column 1) or B (column 2), conditional on this individual having played A in the previous phase.[20] The results reveal remarkable differences between the groups that play A and those that play B. For the groups that play A there is no significant difference between A-voters and B-voters. What drives their contribution decision is their contribution when playing A previously: the lower a player's average contribution in the previous phase the less likely the player is to hand in the red card in the first round of the next phase. In other words, free riders tend to remain free riders and cooperators tend to remain cooperators.

Lagged contributions in the A game do not have a significant effect on contributions in the first round after a group has switched to B. However, we find a significant difference in the contributions of the players who *vote* for A and the players who *vote* for B: B-voters are more likely to hand in their red card than A-voters when playing the B game for the first time. This implies that B-voters must be more optimistic

[20] Only in this first contribution round are expectations determined exclusively by the voting outcome and previous experience playing A. Again, we exclude from this regression all the first-phase observations as well as those corresponding to when B was played in the previous phase. In total, groups chose A after having played A in the preceding phase 42 times, making (since there are five players per group) 210 observations; groups played B after having played A in the previous phase a total of 12 times, giving 60 observations.

about coordination succeeding in the B game—presumably the reason they voted for B in the first place. This is the pull effect.

As discussed in Section 2, ambiguity about the prospects of coordination succeeding in B-8 should be resolved by vote signalling. It thus appears that the A-voters who play Black when B is chosen may have failed to read this signal. This failure can also help to explain why these people voted for A in the first place.

4.2. Ex Post Questionnaire

Table 5 presents responses by the players to a questionnaire given after they had finished playing.[21] In *Vote-First-B-10*, we distinguish between groups that played B every time and those that played A at least once (of course, in this treatment, no group played A more than once). In *Vote-First-B-8*, we distinguish between groups that played A every time and those that played B at least once.

Two observations stand out. First, expectations for successful coordination are very high in *Vote-First-B-10.* They are also high in *Vote-First-B-8* for the groups that played B at least once—a demonstration of the pull effect. However, expectations for successful coordination are noticeably lower for the individuals in groups that never played B in *Vote-First-B-8*. Interestingly, these players' expectations for contribution levels overall are similar for the B game and the A game (compare their responses to the first two questions in the table). All other players have very different expectations for the two games (for the first two questions, compare the responses of the players who played A every time in *Vote-First-B-8* with the responses of the other players).[22]

Second, almost all the players in *Vote-First-B-10* would recommend that a new group of participants play B rather than A. By contrast, individuals who took part in *Vote-First-B-8* were divided. A large majority of those who played A every time would

[21] Responses to an ex post questionnaire are likely to reflect both expectations and experience. However, a pre-play questionnaire might have biased subsequent behavior in the game. Also, much of the dynamics occurred during the game and would not have been captured by a pre-play questionnaire.

[22] Dal Bó et al. (2013) obtain a similar result. In their experiment, players who voted for the prisoners' dilemma were less likely to believe that behavior would be different for the two games.

recommend A, whereas most of the players who played B at least once would recommend B—further confirmation of the pull effect.

We also asked our participants in an open-ended question to give the reason for their recommendation. Many of the players in *Vote-First-B-8* who played A every time and who also recommended that others play A said that, in their view, game A was the better game. A typical answer was, "I would recommend game A and wish them a more cooperative group than the one I had." These players seemed to believe that the level of cooperation was determined by the group and not by the game. One of the players who played B in *Vote-First-B-8* and who also recommended B said this: "Play A with people you know and trust, but play B with people you don't know." This answer reflects a better strategic understanding of the different incentives created by the two games.[23]

4.3. Inferring Expectations From Play-First Treatments

We have so far demonstrated that there exists both a push and a pull effect. We know that the push effect is necessary (and that the pull effect alone is not sufficient) because no group chose B without first trying A. Here we report the results of two new treatments. These show that the pull effect is also necessary in order for players to choose B over A. Analysis of these new treatments also provides further evidence for the push effect.

In both of the new treatments, the players must choose between A and B-8. In treatment *A-First*, the players must play the A game in the first phase and the B game in the second phase. After that, they play the same way as in the *Vote-First* treatments, voting and then playing five contribution rounds in the third phase, and then repeating

[23] Of course, this only begs the higher order question of what determines strategic understanding. Before playing our experiment, we asked the players for their academic major, the number of semesters they had completed at university, and their final secondary school grade (known in Germany as the Abitur). We also asked them to play a "beauty contest game" in order to obtain a measure of their strategic sophistication. In particular, in each session participants were asked to choose a number between zero and 100, knowing that the person who chose the number closest to two-thirds of the session average would receive a prize of €10. Since the unique Nash equilibrium of this game is to choose zero, lower numbers should imply a deeper level of strategic reasoning (Bosch-Domènech *et al.* 2002). In contrast to Dal Bó, Foster, and Putterman (2010), however, we did not find any significant correlations between the personal characteristics of the players or the numbers they chose in the beauty contest game and the way these individuals voted in our *Vote-First* treatments. A plausible interpretation of our results is that voting was determined by expectations, and that expectations could not be predicted from these elicited variables.

this sequence in the fourth and final phase. Treatment *B-First* is the same as *A-First* except that the players play B followed by A before voting in third and fourth phases; see Figure 7.

In *Vote-First*, the players must discover for themselves which game is best to play without the benefit of experience. This game comes closest to how people must play in the real world. However, and as we have seen, expectations can be mistaken. This is the reason for the *Play-First* treatments. These ensure that the players have experience playing both games before voting. By comparing these treatments with *Vote-First* we can thus determine how expectations in both games affect group behavior. By having the players play A first followed by B, or B first followed by A, we can also determine whether the order of experience has a separate effect from the experience itself.

As shown in Table 6, we do not find significant differences between the *A-First* and *B-First* treatments as regards how groups vote beginning in the third phase (Fisher's exact test, $p > 0.10$ for each phase) or the contributions they make following these votes, conditional on their choice of A or B (MWW test, $p > 0.10$ for each phase). We thus pool the data for both treatments and call the combined treatment *Play-First-B-8*. The results for this combined treatment are shown in Figure 8.

Our focus is on whether the outcomes observed in the first two phases of *Play-First*, when all groups are required to play both A and B precisely once, affect the choice of which game to play in the second two phases. We are also interested in knowing how the choices made in these two voting phases compare with the choices made in the first two voting phases of *Vote-First*.

Before turning to these questions, we should note that contributions and payoffs, conditional on the game that has been chosen, reflect a similar pattern as before. As in *Vote-First-B-8*, the groups that chose to play B at the start of the third phase of *Play-First-B-8*, contributed significantly more than the groups that chose to play A (MWW test, $p < 0.01$ for each phase). They also got a significantly higher payoff ($p < 0.01$ for each phase).

The important difference between *Vote-First-B-8* and *Play-First-B-8* lies in the choice of which game to play in the two phases when voting is first allowed. Only two

out of 20 groups chose to play B at least once in the first two phases of *Vote-First-B-8*, whereas, 15 out of 20 groups chose to play B at least once in the two voting phases of *Play-First-B-8* (Fisher's exact test, $p < 0.01$ for each phase). We infer from this evidence that the contrast in behavior between the two treatments reflects a difference in expectations (with this difference being shaped by behavior in the non-voting phases of *Play-First*).

The surprise, perhaps, is that *any* group would choose A in the voting phases of *Play-First-B-8*. However, there were five instances of coordination failure in the non-voting phases of *Play-First* (see Figure 8, groups 42, 44, 48, 54, and 57), an outcome never observed in *Vote-First*. The reason for this failure is probably due to the players being denied any opportunity to signal their intentions by voting.[24] As noted in Section 2, the prospects of players being able to coordinate on the mutually preferred equilibrium in the B game are unclear for treatment B-8 in the absence of vote-signalling. When coordination on this equilibrium failed in the non-voting phases, groups always chose to play A in the voting phases. Chastened by their bad experience playing B, these groups never attempted to play B again. Indeed, failure to coordinate on the mutually preferred equilibrium in the first two phases of *Play-First* is perfectly correlated with whether or not groups try game B *at all* in the last two phases (Spearman's $rho = 1.00$, $p = 0.00$). A bad experience when playing B made these groups pessimistic about the prospects of coordination succeeding, squelching the pull effect.

Of course, Section 2's theoretical argument for using vote signaling as a coordinating device should not be affected by the way the game was played in the absence of voting. Had the players understood that voting could signal intentions, they should have been able to coordinate on the mutually preferred equilibrium in the B game. The fact that they did not coordinate in this way is thus further evidence that people fail to appreciate the value of vote signaling.

Importantly, we also find that the groups that failed to coordinate on the mutually preferred equilibrium in the first two phases of *Play-First* also performed poorly when playing A (see Figure 8). Their average contribution rate over the last two phases of

[24] Behavior may also have been affected by the players not choosing for themselves which game to play. See Dal Bó, Foster, and Putterman (2010) and Sutter, Haigner, and Kocher (2010).

playing A is just 8 percent. These players surely were under no illusions about cooperation in the A game, but they were pessimistic about the prospects of coordination succeeding in the B game. This demonstrates that the push effect is only a necessary and not a sufficient condition for switching, and that the pull effect is also necessary.

Apart from the five groups that failed to coordinate in the non-voting phases of *Play-First*, only two other groups (49 and 60) played A in the final phase of this treatment. The behavior of these groups resembles that of group 28 in *Vote-First-B-8* (see Figure 5). These groups probably voted for A in the final voting phase believing or hoping that their contributions, which were high when they played B previously, would remain high if they switched to A, yielding them a larger payoff. We'll never know, but it seems that these groups probably regretted this last switch, and that they would have chosen differently had they to do over again.[25] In any event, it's clear that the main difference between *Vote-First* and *Play-First* consists in the cases in which coordination failed. When coordination succeeded in the B game, making the players optimistic about the prospects for coordination, groups chose B over A. When coordination failed in the B game, making the players pessimistic about the prospects for coordination, groups chose A over B.

Note finally that *Play-First* also provides more evidence of the push effect. There is a strong correlation between the average contribution level in the A game when played in the non-voting phase and in the voting phase in which a group chose to play B rather than A (Spearman's $rho = 0.53$, $p = 0.04$). Groups that performed poorly when playing A in the non-voting phase chose to play B at the first opportunity. Groups that performed better when playing A in the non-voting phases needed to play A in another (frustrating) phase before switching to B.

5. Applications

[25] In the ex post questionnaire, nine out of the ten students in these two groups recommended that a new group play B; only one player recommended A. These responses lend support to our hypothesis that these groups would have switched back to B if given one more opportunity.

In this section we show how our experimental results can be helpful for interpreting three real world examples of international agreements adopting different approaches.

We begin with the International Convention for the Prevention of Pollution by Ships, more commonly known as MARPOL. MARPOL establishes a technology standard for oil tankers, ensuring that a tanker's oil cargo is kept physically separate from its ballast water. Previously, most oil pollution in the oceans resulted from tankers flushing out their ballast water mixed with oil. Under MARPOL, however, port states can protect their coasts simply by restricting entry to tankers meeting the new standard—that is, by banning trade involving the old technology. As the global market for ocean shipping is characterized by strong network externalities, this technology-standards approach creates incentives for port states and tanker owners alike to adopt the new standard once assured that a critical mass of others will adopt the new standard. MARPOL thus made protection of the oceans a tipping game.[26]

However, choice of this approach came at a cost. The direct approach of limiting emissions was "cheaper, more economically efficient, and 'in theory…. a good idea'" (Mitchell 1994: 434), but was difficult to monitor. The mandated technology-standards approach, by contrast, "was expensive both in terms of capital and the reduction to cargo-carrying capacity" (Mitchell 1994: 434), but was easy to monitor and so could be enforced. Today, virtually all oil tankers comply with the MARPOL standard. However, as in our *Vote-First-B-8* treatment, negotiators adopted MARPOL's coordination approach very reluctantly. They first sought to reduce discharges directly and they persisted in trying to make this approach work for more than fifty years. It was not until the 1970s that they switched to the technology-standards approach.

The Montreal Protocol on protecting the ozone layer works a little differently than MARPOL, but has had a similarly transformative effect. Montreal restricts both the consumption and production of chlorofluorocarbons (CFCs), while also banning trade in CFCs and products containing CFCs between parties and non-parties. Under Montreal, provided enough countries limit their consumption of CFCs, exporters want to produce the CFC substitutes; and provided enough countries produce the substitutes, importers

[26] For a theoretical model showing this kind of transformation, see Barrett (2006); see also Barrett (2003).

want to limit their consumption of CFCs. Like MARPOL, Montreal's approach makes protection of the ozone layer a tipping game.[27] The important difference is that Montreal sustains an outcome that is indistinguishable from a first best. Rather than mandate a particular substitute (a technology standard), Montreal only mandates reductions in CFCs (a performance standard), leaving it to the parties ("the market") to choose which substitutes to employ. As in our *Vote-First-B-10* treatment, negotiators of the Montreal Protocol adopted the coordination approach right from the start.

Unlike our first two examples, the Kyoto Protocol on climate change typifies the direct approach to the prisoners' dilemma. Kyoto specifies national greenhouse gas emission limits without the support of an agreed enforcement mechanism.[28] When this approach was first put to the test, it crumbled. The United States refused to ratify the agreement, Canada withdrew from it, and Japan decided not to participate in the Protocol's second phase. While other countries, notably members of the European Union, have taken steps to reduce their emissions, overall the agreement has had little if any effect (Aichele and Felbermayr 2011). Interestingly, Kyoto incorporates several flexible implementation mechanisms including a provision allowing emissions trading. The people who negotiated Kyoto thus focused their attention on cost-effectiveness, not enforcement.

There is now widespread recognition that the Kyoto Protocol's approach has failed—a necessary condition, our research shows, for players to be willing to try an alternative approach. Our research also suggests that, to be willing to make this switch, players must be optimistic about the prospects of the alternative succeeding. Here there is also sign of change. For example, in June 2013, the United States and China agreed to promote a phase down of hydrofluorocarbons (HFC, a chemical that does not destroy the ozone layer but that is one of the six greenhouse gases targeted by the Kyoto Protocol) in an amendment to the Montreal Protocol.[29] Such a piecemeal approach to limiting climate

[27] For a theoretical model of this transformation, see Barrett (1997; 2003). See also Heal and Kunreuther (2012).

[28] Article 18 says that any compliance mechanism applying with "binding consequences" must be agreed by amendment, and no such amendment has been adopted.

[29] To be specific, the U.S. and China "agreed to work together and with other countries through multilateral approaches that include using the expertise and institutions of the Montreal Protocol to phase down the production and consumption of HFCs, while continuing to include HFCs within the scope of the [United

change cannot sustain a first best outcome, but our research suggests that negotiators would do well to explore further opportunities for tipping, including second best approaches like technology standards combined with trade restrictions.

6. Conclusions

In many settings players can decide on the rules of the game before they begin playing the game. For example, when negotiators meet to adopt an international agreement to provide a public good, they must decide which game to play. A prisoners' dilemma can potentially achieve the overall first best outcome, but collective action in this game is difficult to enforce. Collective action is easier to enforce in a tipping game, but choice of this game may foreclose the possibility of attaining the first best.

The problem with choosing between these games is that players can't be certain which game will work best. Our experiment shows that players are quick to choose the tipping game when doing so enables them to sustain the overall first best outcome. However, they are reluctant to choose this game when doing so means settling for second best, even if the second best outcome is better than the one that results when the players try, but fail, to cooperate in the prisoners' dilemma. Many groups become trapped in the prisoners' dilemma, believing that they have chosen wisely when they would almost certainly do better by switching. Of course, we cannot exclude the possibility that at least some of the groups that were trapped in the A game would have failed to coordinate in the B game (were they to have played B). After all, if group members are pessimistic about coordination succeeding, failure in the B game will be a self-fulfilling prophecy. However, all the evidence we have points to the conclusion that groups do better by switching.

Our finding that the groups that cooperate more successfully are more likely to stick with the prisoners' dilemma parallels an earlier finding by Orbell and Dawes (1993) that when players are free to choose whether to play a prisoners' dilemma or not to play at all, cooperators are more likely to choose to play than non-cooperators. However, while this tendency is to the advantage of cooperators in the Orbell and Dawes (1993) experiment,

---

Nations Framework Convention on Climate Change] and its Kyoto Protocol provisions for reporting of emissions." See: http://www.whitehouse.gov/the-press-office/2013/06/08/united-states-and-china-agree-work-together-phase-down-hfcs.

we find that it is to their disadvantage when players have to choose between a prisoners' dilemma and a tipping game.

Our results confirm the tendency observed in previous studies for players to misapprehend the consequences of the choice of which game to play (Dal Bó et al. 2013). However, this tendency is unusually striking in our experiment. In the treatment in which coordination can sustain only a second best outcome, *every* group started out by choosing the prisoners' dilemma. This game appears to be the default choice when players are unsure how the two games will be played.

The previous literature has also found that, when given the opportunity to revise their choice of institution, players will gradually move towards the welfare improving institution (Gürerk, Irlenbusch, and Rockenbach 2006; Ertan, Page, and Putterman 2009; Markussen, Putterman, and Tyran 2014).[30] Our results are different and more unsettling. We find that a significant number of groups remain loyal to the prisoners' dilemma even after they have witnessed their repeated failure to sustain much cooperation in this game. Over the course of our experiment, cooperation in the prisoners' dilemma deteriorated significantly, and yet only half of the groups switched to the tipping game. To be willing to switch, groups not only had to become disillusioned with cooperation in the prisoners' dilemma; they also had to be hopeful about the prospects for coordination in the tipping game.

Overestimating the ability of one's group to cooperate and underestimating its ability to coordinate both lead to suboptimal choices. The skill needed to anticipate other players' behavior in the two games is thus crucial. Our research shows that this is a skill that some people and therefore some groups lack. In particular, comparison of the *Vote-First* and *Play-First* treatments shows that awareness of vote-signalling behavior is crucial to success in the B game—and, therefore, to the players' willingness to vote for B—and yet many voters seem oblivious of the signalling effect of voting. It remains for future research to show whether our results are unique to the game choice studied in our experiment or whether these results reflect a more general tendency for players to misapprehend the meaning of signals.

[30] For example, in a recent experiment on endogenous punishment institutions, Markussen, Putterman, and Tyran (2014: 303) found that "voters manage surprisingly well to self-organize for collective action, and . . . provide a remarkable example of efficient endogenous emergence of institutions."

Appendix A. Experimental Instructions

Here we provide the instructions for the *Vote-First-B-10* treatment, translated from German. Instructions for the other treatments are available upon request.

Welcome to our experiment!

1. General information

In our experiment you can earn money. How much you earn will depend on the game-play, or more precisely on the decisions you and your fellow co-players make. For a successful run of this experiment, it is essential that you do not talk to other participants. Now read the following rules of the game carefully. If you have any questions, give us a hand signal. We will come to you and answer them.

2. Game rules

There are 5 players in your group, meaning you and 4 other players. Each player is faced with the same decision problem. All decisions are anonymous. For this reason, you will be identified by a number (between 1 and 5), which you will see in the lower left corner of your display.

There are two games, Game A and Game B. At the beginning, every player in your group will vote for one of the two games. After that, and before the game starts, the players' votes will be displayed to everyone. The game that receives the most votes (at least 3 out of 5) will be played by the group. Thus, the group plays either Game A or Game B.

In each game, you will receive two cards, a Red Card and a Black Card. You will be asked to hand in one of the two cards. Your payoff will depend on which game is played (A or B), which card you hand in (Red or Black), and which cards your four co-players hand in. The following two tables show your payoff for all possible outcomes in each game.

Voting

A Game		Number of Red Cards handed in by your co-players				
		0	1	2	3	4
Your choice	Red	2	4	6	8	10
	Black	5	7	9	11	13

B Game		Number of Red Cards handed in by your co-players				
		0	1	2	3	4
Your choice	Red	2	4	6	8	10
	Black	5	5	5	5	5

Here are some examples for how to read the tables:

If the group plays the A Game and two of your co-players hand in their Red Card (and the other two co-players hand in their Black Card), you get 6 tokens if you hand in your Red Card and you get 9 tokens if you hand in your Black Card.

If the group plays the B Game and two of your co-players hand in their Red Card (and the other two co-players hand in their Black Card), you get 6 tokens if you hand in your Red Card and you get 5 tokens if you hand in your Black Card.

The game (A or B) that receives the most votes in the group (at least 3 out of 5) will be played five times consecutively. In each round you will be asked to hand in either the Red Card or the Back Card. After this, your group will vote again, play the chosen game another five times, and so on. In total, your group will vote four times and play the chosen game five times after each vote. Hence, you and your co-players will decide which card to hand in 20 times in total. You will play with the same group of players throughout all rounds. The sum of tokens you earn across all 20 rounds will be paid to you in cash at the end. You will get €0.10 for each token. For example, if you earn 150 tokens in total, you will get €15.00.

3. Control questions

Please answer the following control questions.

a. Right or wrong? At the beginning all players will vote for Game A or Game B. After everyone votes, and before the game starts, you will learn how your co-players voted and

they will learn how you voted. The game that receives the most votes will be played by the group.

○ Right ○ Wrong

b. Right or wrong? The group will vote four times in total. After each vote, the chosen game will be played for five rounds.

○ Right ○ Wrong

c. Assume that the group plays the A Game and one of your co-players hands in the Red Card (and the other three co-players hand in their Black Card). What is your payoff if you hand in your Red Card? _____ What is your payoff if you hand in your Black Card? _____

d. Assume that the group plays the B Game and one of your co-players hands in the Red Card (and the other three co-players hand in their Black Card). What is your payoff if you hand in your Red Card? _____ What is your payoff if you hand in your Black Card? _____

e. Assume that the group plays the A Game and three of your co-players hand in their Red Card (and the other co-player hands in the Black Card). What is your payoff if you hand in your Red Card? _____ What is your payoff if you hand in your Black Card? _____

f. Assume that the group plays the B Game and three of your co-players hand in their Red Card (and the other co-player hands in the Black Card). What is your payoff if you hand in your Red Card? _____ What is your payoff if you hand in your Black Card? _____

g. Assume that the group plays the A Game and all four of your co-players hand in their Red Card (and no one hands in the Black Card). What is your payoff if you hand in your Red Card? _____ What is your payoff if you hand in your Black Card? _____

h. Assume that the group plays the B Game and all four of your co-players hand in their Red Card (and no one hands in the Black Card). What is your payoff if you hand in your

Red Card? _____ What is your payoff if you hand in your Black Card? _____

Please also consider other examples! Give us a hand signal after you have answered all the control questions. We will come to you and check that you have answered all the questions correctly. The game will begin after we have checked the answers of all the participants and answered any questions you may have. Good luck!

Acknowledgements. We are grateful to Geir Asheim, Raphael Calel, Alessandra Casella, Bård Harstad, Robert Keohane, Brad LeVeck, Thomas Schelling, Alessandro Tavoni, and David Victor for comments on a first draft. We particularly want to thank Geir Asheim for unravelling the vote-signalling effect in our model. We are also grateful to the MaXLab team at Magdeburg University for use of their laboratory, and to the Princeton Institute for International and Regional Studies research community on Communicating Uncertainty: Science, Institutions, and Ethics in the Politics of Global Climate Change for financially supporting our experiments.

References

Aichele, R. and Felbermayr, G. (2011). 'Kyoto and the Carbon Footprint of Nations', *Journal of Environmental Economics and Management,* vol. 63, pp. 336-354.

Barrett, S. (1997). 'The Strategy of Trade Sanctions in International Environmental Agreements', *Resource and Energy Economics,* vol. 19, pp. 345-361.

Barrett, S. (2003). *Environment and Statecraft: The Strategy of Environmental Treaty-Making*, (Oxford: Oxford University Press).

Barrett, S. (2006). 'Climate Treaties and 'Breakthrough' Technologies', *American Economic Review* (*Papers and Proceedings*), vol. 96, pp. 22-25.

Bosch-Domènech, A., Montalvo, J. G., Nagel, R., and Satorra, A. (2002). 'One, Two, (Three), Infinity, . . . : Newspaper and Lab Beauty-Contest Experiments', *American Economic Review,* vol. 92, pp. 1687-1701.

Croson, R. T. A., Marks, M. B. (2000). 'Step Returns in Threshold Public Goods: A Meta- and Experimental Analysis', *Experimental Economics*, vol. 2, pp. 239–259.

Dal Bó, P. (2011). 'Experimental Evidence on the Workings of Democratic Institutions,' forthcoming in Economic Institutions, Rights, Growth, and Sustainability: the Legacy of Douglass North, Cambridge University Press: Cambridge.

Dal Bó, E., Dal Bó, P., and E. Eyster (2013). 'The Demand for Bad Policy When Voters Underappreciate Equilibrium Effects,' Working Paper, Haas School of Business, University of California, Berkeley.

Dal Bó, P., Foster, A., and Putterman, L. (2010). 'Institutions and Behavior: Experimental Evidence on the Effects of Democracy', *American Economic Review,* vol. 100, pp. 2205–2229.

Ertan, A., Page, T., and Putterman, L. (2009). 'Who to Punish? Individual Decisions and Majority Rule Mitigating the Free Rider Problem', *European Economic Review,* vol. 53, pp. 495–511.

Fischbacher, U. (2007). 'Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments', *Experimental Economics,* vol. 10, pp. 171-178.

Gürerk, Ö., Irlenbusch, B., and Rockenbach, B. (2006). 'The Competitive Advantage of Sanctioning Institutions', *Science,* vol. 312, pp. 108-111.

Heal, G. and Kunreuther, H. (2012). 'Tipping Climate Negotiations', in *Common Sense and Climate Change: Essays in Honor of Thomas Schelling*, Robert Hahn and Alistair Ulph, eds. (Oxford: Oxford University Press).

Hufbauer, G. C., Schott, J. J., Elliott, K. A., and Oegg, B. (2007). *Economic Sanctions Reconsidered, 3rd Edition* (Washington, DC: Institute for International Economics).

Ledyard, J. O. (1995). 'Public goods: a survey of experimental research', in *Handbook of Experimental Economics*, John Kagel and Alvin Roth, eds. (Princeton: Princeton University Press)

Markussen, T., Putterman, L., and Tyran, J.-R. (2014). 'Self-Organization for Collective Action: An Experimental Study of Voting on Sanction Regimes', *Review of Economic Studies*, vol. 81, pp. 301-324.

Mitchell, R. B. (1994). 'Regime Design Matters; Intentional Oil Pollution and Treaty Compliance', *International Organizat*ion, vol. 48, pp. 425-458.

Orbell, J. M. and Dawes, R. M. (1993). 'Social Welfare, Cooperators' Advantage, and the Option of Not Playing the Game', *American Sociological Review*, vol. 58, pp. 787-800.

Runge, C.F. (1984). 'Institutions and the Free Rider: the Assurance Problem in Collective Action,' *The Journal of Politics*, vol. 46, no. 1, pp. 154-181.

Schelling, T.C. (1971). 'Dynamic Models of Segregation,' *Journal of Mathematical Sociology,* vol. 1, pp. 143-186.

Schelling, T. C. (1978). *Micromotives and Macrobehavior* (New York: W.W. Norton & Co).

Sen, A.K. (1967). 'Isolation, Assurance and the Social Rate of Discount,' *Quarterly Journal of Economics*, vol. 81, no. 1, pp. 112-124.

Snidal, D. (1985). 'Coordination versus Prisoners' Dilemma: Implications for International Cooperation and Regimes,' *American Political Science Review*, vol. 79, no. 4, pp. 923-942.

Sutter, M., Haigner, S., and Kocher, M. G. (2010). 'Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations', *Review of Economic Studies*, vol. 77, pp. 1540–1566.

Table 1

Vote-First Treatments

Phase	Game	*Vote-First-B-10*			*Vote-First-B-8*		
		Percent of groups	Average percent red cards	Average payoff	Percent of groups	Average percent red cards	Average payoff
I	A	45	21	6.1	100	39	7.0
	B	55	99	9.8	0	-	-
II	A	0	-	-	90	24	6.2
	B	100	100	9.9	10	90	7.2
III	A	0		-	80	26	6.3
	B	100	100	10	20	90	7.1
IV	A	0		-	55	10	5.5
	B	100	100	10	45	94	7.5

Table 2

A-Voters in Vote-First-B-8

Phase	Game	All groups			Only groups with two or three A-voters		
		No. of groups	Average no. of A-voters per group	Average A-voter payoff	No. of groups	No. of A-voters per group	Average A-voter payoff
I	A	20	4.5	7.0	1	3	6.7
	B	0	-	-	0	-	-
II	A	18	3.8	6.3	6	3	6.8
	B	2	2	7.4	2	2	7.4
III	A	16	3.5	6.4	9	3	6.5
	B	4	1.5	7.1	2	2	7.0
IV	A	11	3.4	5.6	7	3	5.4
	B	9	1.3	7.3	4	2	7.0

Table 3

Probit Regression on Voting for and Selecting A in Vote-First-B-8

	Individual level		Group level
	(1)	(2)	(3)
Variables	Voting decision (A = 1, B = 0)	Voting decision (A = 1, B = 0)	Game selection (A = 1, B = 0)
Lagged voting decision	1.339***	1.351***	
	(0.221)	(0.219)	
Lagged individual payoff in A	0.228***		
	(0.068)		
Lagged group contribution in A		1.645**	6.966**
		(0.826)	(3.657)
Constant	-2.535***	-1.603***	-0.843
	(0.610)	(0.504)	(0.849)
Observations	270	270	54
Number of subjects	100	100	
Number of groups			20

Random effects probit regression. Standard errors in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Regressions at the individual level include group dummies, which are insignificant and not shown here. Dependent variables: voting decision = 1 if player voted for A in current phase, 0 otherwise. Game selection = 1 if group played A in current phase, 0 otherwise. Independent variables: lagged voting decision = 1 if player voted for A in previous phase, 0 otherwise. Lagged individual payoff in A = player's average payoff in the previous phase of playing A. Lagged group contribution in A = group's average contribution in previous phase of playing A.

Table 4

Probit Regression of Individual Contribution Decision in Vote-First-B-8

	Individual contribution decision (Red = 1, Black = 0)	
	(1)	(2)
Variables	Game A in current phase	Game B in current phase
Voting decision (A = 1, B = 0)	-0.359	-1.684***
	(0.215)	(0.476)
Lagged individual contribution in A	1.628***	0.0199
	(0.364)	(0.972)
Constant	-0.601	1.326**
	(0.403)	(0.614)
Observations	210	60
Number of subjects	90	50

Random effects probit regression. Standard errors in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Regressions include group dummies, which are insignificant and not shown here. Dependent variable: contribution decision = 1 if player played Red, 0 otherwise (only contribution decisions in the first round of the current phase of playing A or B are included). Independent variables: Voting decision = 1 if player voted for A, 0 otherwise. Lagged individual contribution in A: average number of red cards contributed in the previous phase of playing A.

Table 5

Responses to the Ex Post Questionnaire (Percent of Subjects)

Question	Answer	*Vote-First-B-10*		*Vote-First-B-8*	
		Played B every time (n = 55)	Played A at least once (n=45)	Played A every time (n = 50)	Played B at least once (n = 50)
Did you expect your fellow co-players to hand in their red card in Game A?	Very much	7	20	30	26
	Somewhat	15	47	34	50
	Little	47	18	32	18
	Not at all	30	16	4	6
Did you expect your fellow co-players to hand in their red card in Game B?	Very much	98	96	38	84
	Somewhat	2	4	30	4
	Little	0	0	12	4
	Not at all	0	0	20	8
If you could give advice to a new group of participants, which game would you recommend that they play?	Game A	4	0	82	26
	Game B	96	100	18	74

Table 6

Play-First Treatments

Phase	Game	*Play-A-First-B-8*			*Play-B-First-B-8*		
		Percent of groups	Average percent red cards	Average payoff	Percent of groups	Average percent red cards	Average payoff
I	A	100	32	6.6	0	-	-
	B	0	-	-	100	80	6.8
II	A	0	-	-	100	50	7.5
	B	100	73	6.3	0	-	-
III	A	60	24	6.2	60	23	6.2
	B	40	100	8.0	40	100	8.0
IV	A	40	9	5.5	30	16	5.8
	B	60	99	7.9	70	99	7.9

Fig. 1. *Vote-First Treatments*

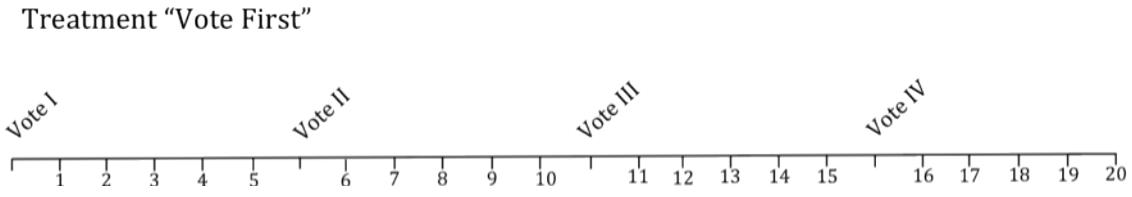


Fig. 2. *The A and B Games*

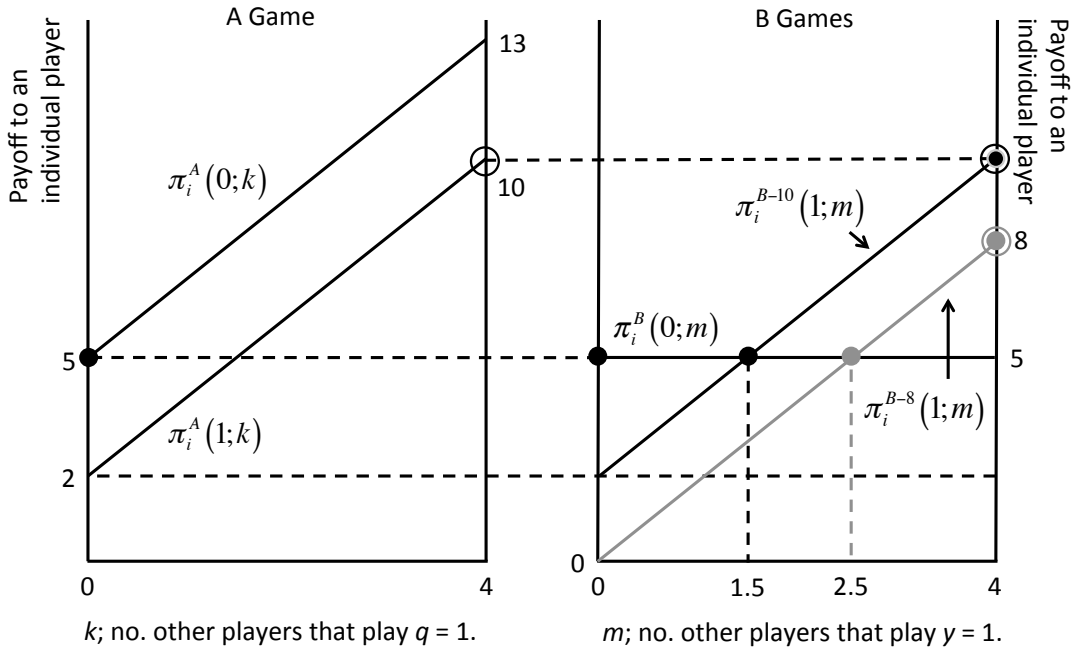
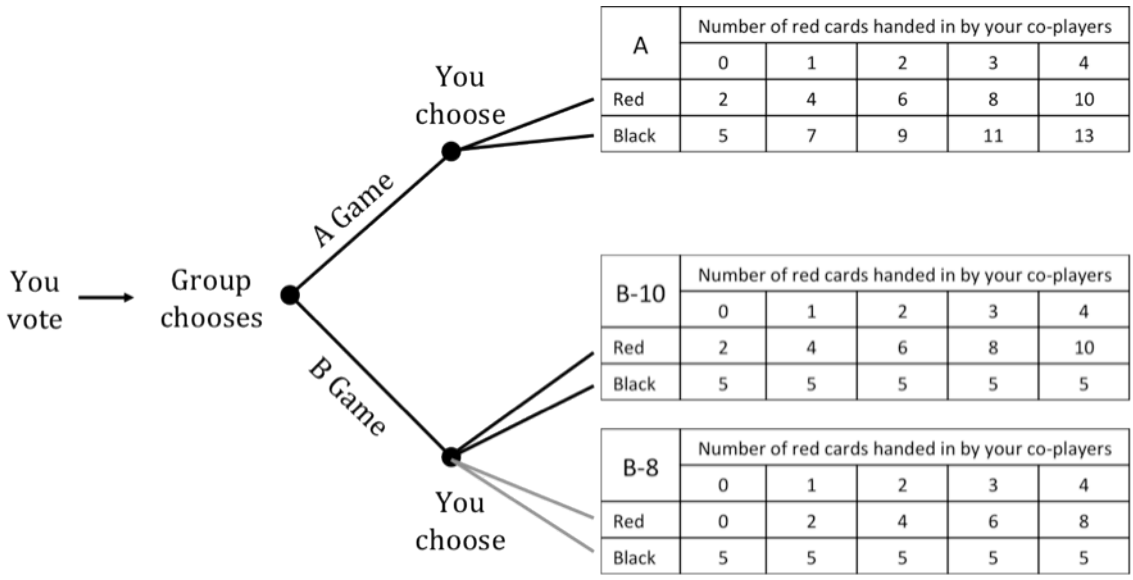


Fig. 3. *Voting Stage and Payoffs*

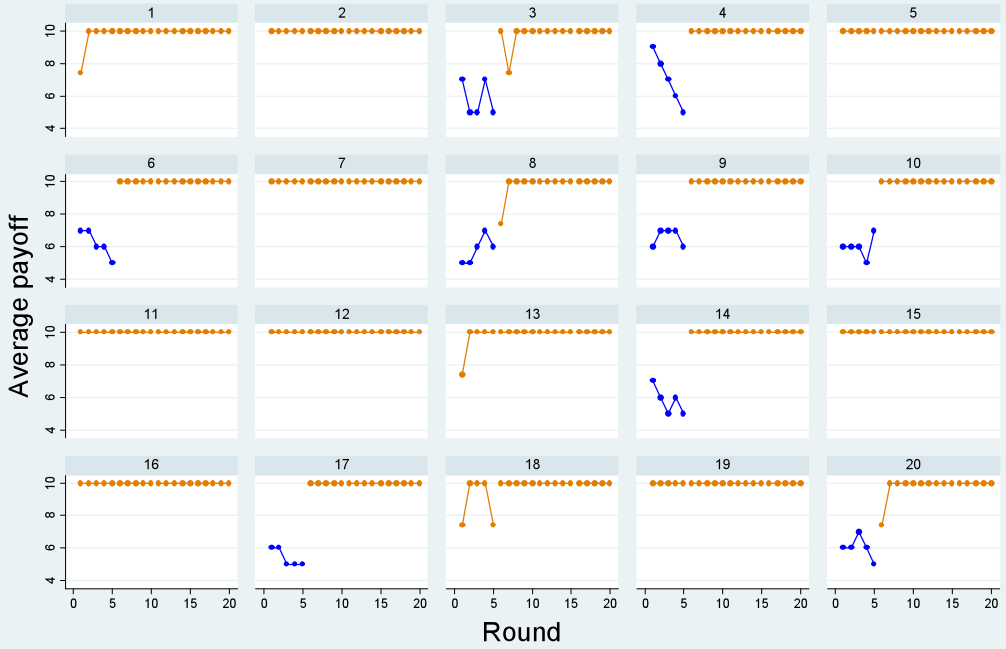


A	Number of red cards handed in by your co-players				
	0	1	2	3	4
Red	2	4	6	8	10
Black	5	7	9	11	13

B-10	Number of red cards handed in by your co-players				
	0	1	2	3	4
Red	2	4	6	8	10
Black	5	5	5	5	5

B-8	Number of red cards handed in by your co-players				
	0	1	2	3	4
Red	0	2	4	6	8
Black	5	5	5	5	5

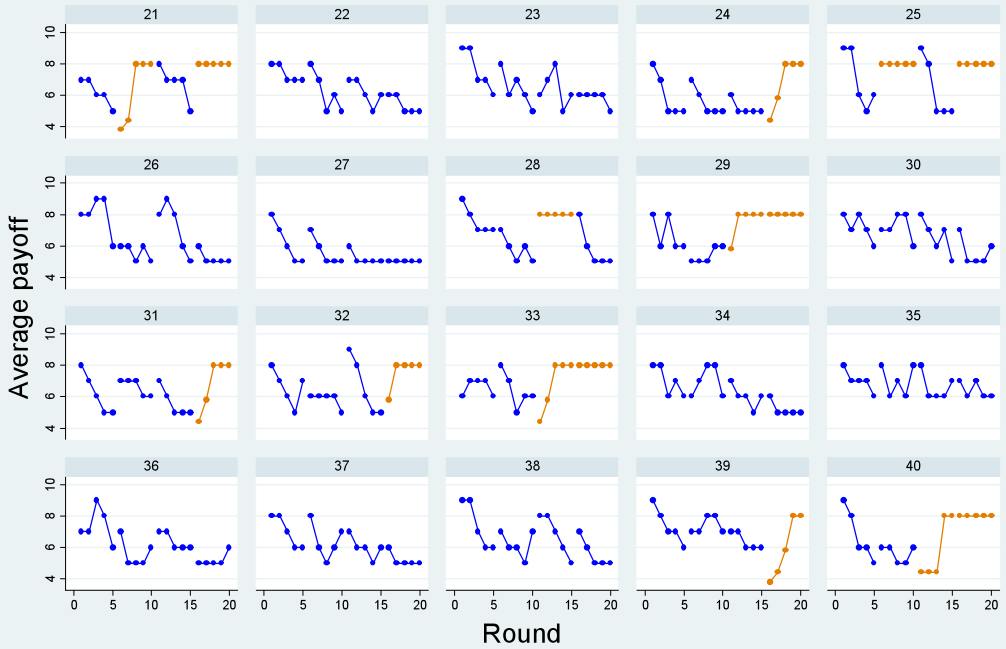
Fig. 4. *Payoffs over Time by Group for Vote-First-B-10*



Blue = A Orange = B

A has one Nash equilibrium with payoff 5. B has two pure strategy Nash equilibria with payoff 5 and 10.

Fig. 5. *Payoffs over Time by Group for Vote-First-B-8*



Blue = A Orange = B

A has one Nash equilibrium with payoff 5. B has two pure strategy Nash equilibria with payoff 5 and 8.

Fig. 6. *Individual Voting Decisions in the Vote-First Treatments*

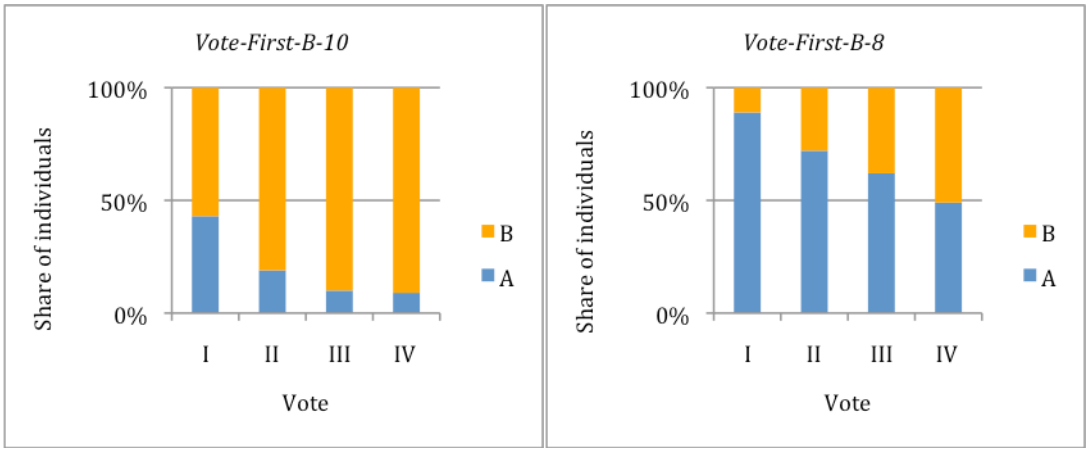


Fig. 7. *Play-First Treatments*

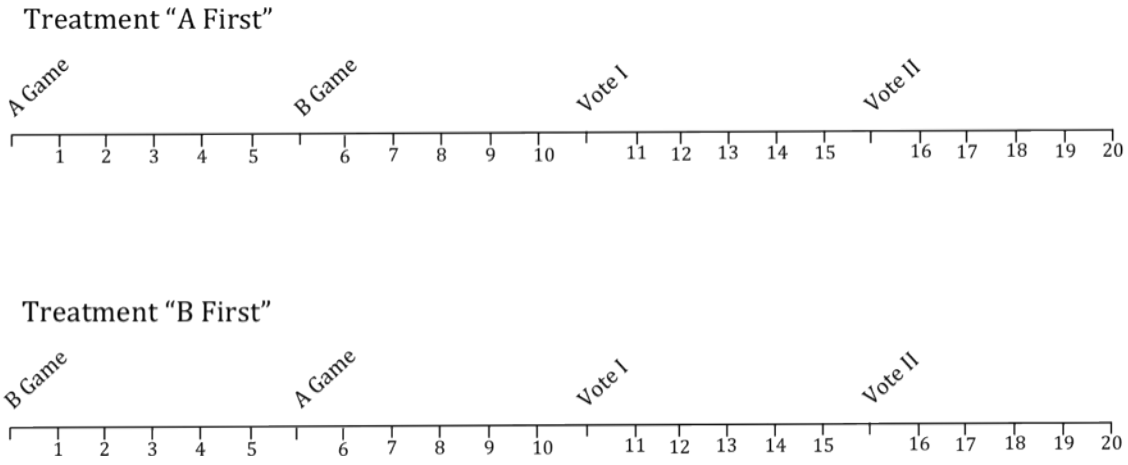
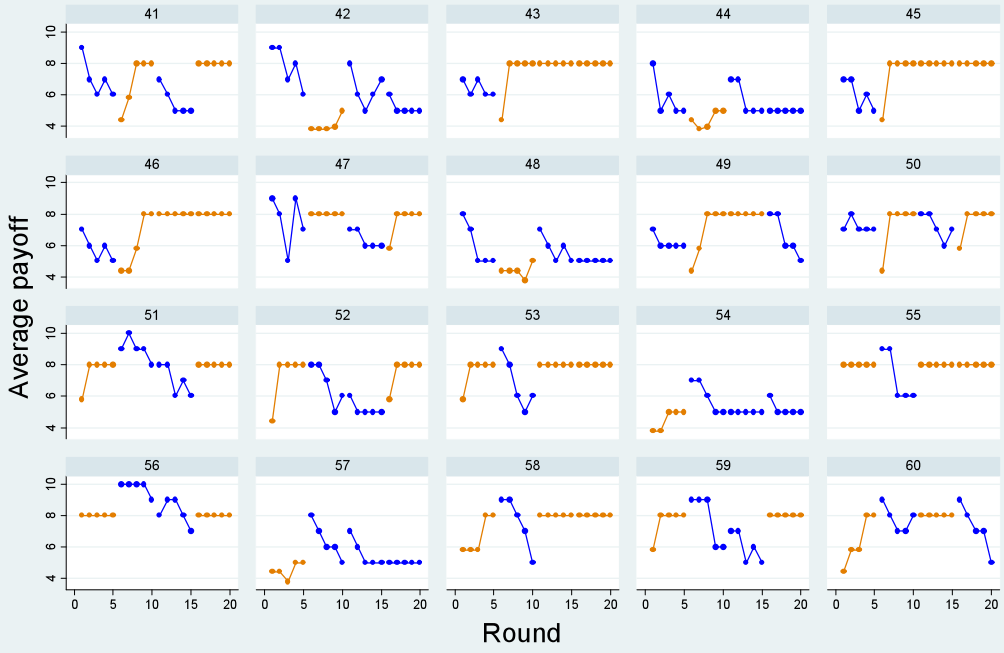


Fig. 8. *Payoffs over Time by Group for Play-First-B-8*



Blue = A Orange = B

Groups 41-50 played A-B in the exogenous part. Groups 51-60 played B-A.