The Demand for Bad Policy when Voters Underappreciate Equilibrium Effects

Ernesto Dal Bó — UC Berkeley
Pedro Dal Bó — Brown University
Erik Eyster — LSE*

August 26, 2014

Abstract

We study errors in equilibrium thinking that may lead people to choose inefficient policies and institutions. More precisely, we show in an experiment that a majority of subjects vote against policies that would help them overcome social dilemmas. We explain this behavior through subjects' failure to fully anticipate the equilibrium effects of new policy. By eliciting their beliefs about how others will behave under different policies, we show that subjects systematically underappreciate the extent to which policy changes alter other people's behavior, and that those who underappreciate equilibrium effects more are also more likely to demand bad policy. In addition we estimate that one-third of subjects do not appreciate how their own behavior will adapt to the new policy. To the extent that voter opinion affects policy, the underappreciation of equilibrium effects by voters can be a cause of political failure.

JEL codes: C9, D7.

Keywords: reform, policy failure, endogenous policy, cooperation, experiment.

*We thank Daniel Prinz and Santiago Truffa for excellent research assistance as well as Berkeley's XLab and Brown's BUSSEL for support. We are grateful to Ned Augenblick, Eric Dickson, Alessandro Lizzeri, Matthew Rabin and Francesco Trebbi for helpful discussions. We thank participants at various conferences and seminars for their comments and suggestions.

1 Introduction

The political economy field has developed several explanations for why bad policies are implemented when good policies are available. Standard explanations blame aspects of the policy-production process including agency problems, incompetent policy-makers, and institutional failure to efficiently resolve competitive tensions.[1] In this paper, we shift the focus away from policy-makers and institutions, by providing experimental evidence to suggests that some of the blame for bad policies may lie with voters. Although it is standard in the literature to assume that voters correctly assess the merits of the options they face, we show with a simple experiment that people demand bad policies because they fail to correctly predict the equilibrium impact that new policies will have on behavior and welfare. In our experimental setting, this prevents groups from resolving social dilemmas through democratic means.

Writers dating back to Adam Smith have emphasized the limitations of the general public to grasp the implications of market equilibrium considerations (see Smith 1776). In modern times, North (1990) surmised that voters might misperceive the relative merits of different policies and institutions, and hence express suboptimal demand for policy. Caplan (2007) surveys and adds to a literature in political science by documenting a gap on several policy issues between popular opinion and the consensus of professional economists. If professional economists have views closer to the truth, that evidence suggests that voters demand bad policies the impacts of which they do not fully understand.

[1] Agency problems include discretion under limited electoral accountability (e.g., Barro 1973, Ferejohn 1986) and capture (e.g., Stigler 1971, Peltzman 1976, and Coate and Morris 1995). For accounts of why inept people may self-select into policymaking, see Dal Bó and Di Tella (2003), Caselli and Morelli (2004), Besley (2005), Dal Bó, Dal Bó and Di Tella (2006), and Polborn (2006), among others. Institutional failures to efficiently resolve collective disagreements may take the form of status quo bias (e.g., Romer and Rosenthal 1978), delay to reform (e.g., Alesina and Drazen 1991, and Fernandez and Rodrik 1991), and dynamic inefficiency due to the threat of losing political control (e.g., Alesina and Tabellini 1990, De Figueiredo 2002, and Besley and Coate 2007).

Yet it is hard to conclude, based on disagreements with experts, that people demand bad policies. These disagreements may result from voters considering or caring about dimensions not considered or valued by experts. In addition, experts often disagree with each other, and on occasion the minority of experts who agree with the average citizen may in fact be right. Given the complexity of real-world policy debate, it is difficult to determine conclusively that voters demand bad policy on the basis of real-world data. Fortunately, the methods from experimental economics allow us to create environments in which we can evaluate the welfare effects of different policies, design away the standard explanations for bad policies, and study whether voters choose bad policies and why.

We present data from an experiment in which subjects choose whether to participate in a Prisoners' Dilemma game or an alternative game, which we call the Harmony Game, where both cooperation and defection are taxed, but the latter is taxed more. As a result of the asymmetric taxes, cooperation is a dominant strategy in the Harmony Game, and therefore this game leads to higher payoffs than the Prisoners' Dilemma game, both in theory and in practice. In our experiment subjects choose the game following several rounds of experience playing one of the two games. Although subjects do on average earn higher payoffs under the Harmony Game than under the Prisoner's Dilemma, a majority choose to play the Prisoner's dilemma. This result is robust to different voting institutions and orders of play. The fact that most subjects choose to play the Prisoners' Dilemma game is evidence that, under certain circumstances, voters may demand bad policies.

What are those circumstances, and why do voters choose the wrong game? We study situations where voters are inexperienced with one of the options they face, so that a strong enough underappreciation of how other players will change their behavior in the new game is a necessary and sufficient condition for bad choices. Our experiment demonstrates empirically that indeed the key driver for bad votes is a difficulty to predict the extent to which the

new game will alter the behavior of other players. Specifically, we find that on average subjects underestimate the effect that the change in game will have on the behavior of other subjects. Moreover, the subjects who most underappreciate the extent to which cooperation rates differ across games vote most frequently for the Prisoners' Dilemma. This evidence is consistent with the demand for bad policy arising from a lack of understanding of the equilibrium effect of the game change.

Of course, voters' lack of understanding of equilibrium effects would not matter if politicians and the media educated them whenever appropriate. But as Blinder and Krueger (2004, p.328) emphasize, even on matters admitting a technical answer "*the decisions of elected politicians are heavily influenced by public opinion.*" Consistent with this view, most formal theories of electoral politics, by focusing on the rigors of electoral discipline, view politicians not as educators but instead as panderers to voters' policy positions. This is the case even when voters are likely wrong, as in the literature on pandering (Canes-Wrone, Herron and Shotts 2001, and Maskin and Tirole 2004). For-profit media may also pander rather than educate, since they have incentives to bias reporting to match consumers' priors (see Gentzkow and Shapiro 2006 for a theoretical argument, and Gentzkow and Shapiro 2010 for evidence). The concern with pandering by politicians and the media is rooted in the possibility that voters may hold incorrect priors to start with. While the typical Condorcetian assumption in voting models is that voters are on average correct about the merit of the options they face, we lack frameworks to highlight when and why voters may be wrong. To the best of our knowledge, this is the first paper to offer a conceptual framework paired with experimental evidence to investigate these questions.

Our framework predicts voting failures when people tend to focus on the direct impact of policies–the impact on payoffs before changes in behavior are considered–over their indirect impact through equilibrium effects. Our evidence corroborates the underappreciation of

equilibrium effects and the emergence of collective voting mistakes. A consequence is that, all else equal, we can expect the demand side in the market for policies to be biased. A policy with net negative consequences may be preferred to one with net positive consequences if the costs of the bad policy are indirect relative to its benefits, and the benefits of the good policy are indirect relative to its costs. As an example consider the establishment of regulations in the US to reduce carbon consumption, either due to concerns with dependence on foreign oil or with climate change. Experts recommend a Pigouvian tax on CO_2 emissions, but this policy is widely dismissed as politically infeasible, since there is a "reluctance of policymakers to adopt Pigouvian taxes that would affect petroleum consumption" (Knittel 2012, p. 111). This does not mean there is policy inaction to curb CO_2 emissions: less efficient alternatives are enacted, such as miles-per-gallon (MPG) regulations. The studies reviewed by Knittel (2012) suggest that a Pigouvian tax could achieve the same goals of MPG regulations at a fraction of the social cost (from 1/6 to 1/10 depending on the study). MPG regulations are seen as suboptimal compared to a carbon tax for at least two reasons: they impose significant development costs, and are partially undermined by "rebound effects", where the decreased fuel-cost of driving encourages people to drive more.

Our framework and results on the underappreciation of equilibrium effects offer an explanation for why the inefficient policy is chosen. The Pigouvian taxes look unappealing because its costs are direct (higher prices for gas) while the benefits are indirect (lower carbon consumption resulting from the response of others to the tax). With MPG regulations, the equation is inverted. The benefit (more efficiency) is direct, while the costs (e.g., rebound effects) are indirect.

This paper relates to the emerging political economy liteature emphasizing behavioral aspects. Examples are the study of the impact of cognitive dissonance on voting (Mullainathan and Washington 2009), the analysis of collective action with time-inconsistent voters (Bisin,

Lizzeri and Yariv 2011, and Lizzeri and Yariv 2012), and the behavior of voters who fail to extract the right information from other voters' strategies (Eyster and Rabin 2005, Esponda and Pouzo 2010, and Esponda and Vespa 2012).

This paper also relates to a growing experimental literature studying the choice of self-regulatory institutions (see Dal Bó 2011 for a survey). A few findings in that literature are worth highlighting here. Walker, Gardner, Herr and Ostrom (2000) study common-pool problems where players would do better by reducing extraction rates, and where they can put extraction rules to a vote. They find that not all voters propose efficient extraction rules, and that the voting protocol affects the ability to reach a welfare-increasing decision. This failure is enhanced in contexts where subjects are heterogeneous (Magreiter, Sutter, and Dittrich 2005). Sausgruber and Tyran (2005, 2011), test whether people understand that tax incidence does not depend upon who pays the tax. In a setting where the entire incidence falls on buyers, they show that a majority of buyers oppose a tax on buyers whose proceeds are divided equally amongst buyers and sellers but that a majority buyers support a tax on sellers whose proceeds are divided equally amongst buyers and sellers. In addition, Dal Bó (2011) offers evidence that the way in which subjects comprehend their strategic situation could affect their ability to select institutions. In this paper we specifically focus on the underappreciation of equilibrium effects and show that it impairs groups' ability to resolve social dilemmas through democratic means. Like most of these papers, we focus on choices involving an unfamiliar option, which matches institutional reforms or policy choices that are not frequently available. Examples are constitutional changes, privatizations, sweeping health-care reform, or decisive action against global warming.

At an abstract level, our paper relates to the experimental literature documenting failures of backwards induction (e.g., McKelvey and Palfrey 1992, Bone, Hey and Suckling 2009, Levitt, List and Sadoff 2011, and Moinas and Pouget 2013). Our work adds to this

literature by establishing a systematic direction in which people deviate from subgame perfection in a specific class of games: players underestimate how differently their opponents play across subgames with shared action spaces but different payoffs. This error resembles that embodied in Jehiel's (2005) Analogy-Based-Expectations Equilibrium (ABEE), where players mistakenly believe that their opponents' play does not vary across certain decision nodes where in fact it does. Despite the resemblance, Section 6 describes why ABEE cannot account for subjects' behavior in our experiment.

In the next section, we present a simple conceptual framework to explain our ideas and derive our hypotheses. In section 3 we present the experimental design. Section 4 presents benchmark results when games are assigned exogenously. In Section 5 we present the results of our experiment and investigate mechanisms. In Section 6 we discuss alternative explanations, and in Section 7 we conclude.

2 Conceptual framework

In this section, we lay out a simple framework where individual perceptions of the effects of policy changes matter for collective choice, and use it to generate the core hypotheses tested by our experiments. For simplicity, we utilize the same setting here as we do in the laboratory, although the framework extends naturally to any normal-form game. The setup developed here involves a key simplifying assumption to facilitate exposition, and that is that the action taken by a player does not alter the payoff consequences of changes in the other player's action. The appendix includes a generalization of our framework to show that this assumption permits the linear decomposition of the total effects of policy in terms of direct and indirect effects, as characterized in this section.

2.1 Setting

Consider the games $\Gamma = PD, HG$ (for Prisoners' Dilemma and Harmony Game) as displayed in Figure 1, with actions C and D (for cooperate and defect). Focus first on the top panel and the Prisoner's Dilemma game. In the Prisoner's Dilemma, D is a strictly dominant strategy and, hence, (D,D) is the unique Nash equilibrium. Suppose that the Prisoner's Dilemma is currently being played with (D,D) as its outcome, yielding payoffs of $(5, 5)$.

[Figure 1 about here]

Suppose that the players receive the opportunity to switch to playing the Harmony Game, displayed to the right, which can be obtained from the Prisoner's Dilemma by imposing a tax of 1 on C and 4 on D. In a very stylized way, this proposed reform represents a Pigouvian tax. For example, if C corresponds to "light driving" and D "heavy driving," the Prisoner's Dilemma captures a situation where individuals generate excessive carbon emissions because they do not internalize the external costs associated with carbon usage. The new game Harmony Game captures the effects of a carbon tax, in that it imposes a higher tax on heavy driving. How might the players vote on a proposal to switch from the Prisoner's Dilemma to the Harmony Game?

The payoffs in every cell of the Harmony Game lie below those in the corresponding cell of the Prisoner's Dilemma. Intuitively, a carbon tax increases the monetary cost of every level of driving.[2] If players maintained their *status quo* outcome of (D,D), then switching to the new game would reduce payoffs from 5 to 1. This effect, which we call the "ceteris paribus" or "direct effect" of policy, is negative. Voters who ranked the two policies represented by the Prisoner's Dilemma and the Harmony Game games based solely upon these direct effects (or the direct effects as measured from any cell, for that matter) would oppose the carbon-tax

[2]For simplicity, we abstract from the distribution of tax revenue.

proposal.

Of course, adopting the Harmony Game would likely change behavior. Like the Prisoner's Dilemma, the Harmony Game has a unique equilibrium in strictly dominant strategies, but in the Harmony Game the equilibrium is (C,C) rather than (D,D). A switch to the Harmony Game, leads, in equilibrium, to the higher payoffs $(8, 8)$. This gain in payoffs is denoted by $g = 3$ and can be decomposed as $g = d + s + o = -4 + 1 + 6 = 3$. Specifically,

$$\begin{aligned}
\text{overall gain:} \qquad & g = 8 - 5 = 3 \\
\text{ceteris paribus; direct effect (cost):} \qquad & d = 1 - 5 = -4 \\
\text{adjustment by self; indirect effect (benefit):} \qquad & s = 2 - 1 = 8 - 7 = 1 \\
\text{adjustment by others; indirect effect (benefit):} \qquad & o = 8 - 2 = 7 - 1 = 6,
\end{aligned}$$

where the indirect effects stemming from changes in behavior correspond to the dashed arrows in the figure.[3] The ceteris paribus effect represents a direct cost to players before any behavior adjustments occur, yet changes in behavior are what make the Harmony Game good policy. Voters who always expected Nash outcomes would prefer the Harmony Game to the Prisoner's Dilemma. However, individuals probably vary in the extent to which they perceive the indirect effects relative to the direct ones. While we expect players to perceive the direct effects d fully, we hypothesize that they imperfectly appreciate the effects s and o. At this point, the following remark is in order.

Remark 1 *In our setting, failure to recognize the potential for self adjustment is neither necessary nor sufficient to drive a preference for the Prisoner's Dilemma. In contrast, a*

[3]Note that given our choice of payoffs the effect s is the same regardless of whether the other player adjusts his behavior $(8 - 7)$ or not $(2 - 1)$. Similarly, the effect o does not change regardless of whether a player adjusts himself or not. This invariance allows for the linear decomposition of gains, as shown in the appendix.

large enough failure to recognize the potential for others' adjustment is a necessary and sufficient condition to drive a preference for the Prisoner's Dilemma.

To see this, note that given the (D,D) outcome in the Prisoner's Dilemma, if the column player defects in the Harmony Game, then the row player prefers the Prisoner's Dilemma regardless of how he expects to play himself in the Harmony Game; if the column player cooperates in the Harmony Game, then the row player prefers the Harmony Game regardless of how he expects to play himself in the Harmony Game.

To further clarify this point, consider the case where the Harmony Game is currently being played with (C,C) as its outcome, yielding payoffs of $(8, 8)$. The lower panel of Figure 1 depicts this situation. Payoffs are $(8, 8)$ to start with, and the gain from a change to the Prisoner's Dilemma can be decomposed as $g = d + s + o = -3 = 1 + 2 - 6$. Note that when moving from the Harmony Game to the Prisoner's Dilemma the ceteris paribus effect is complemented by an additional benefit from adjustment by self. So it is clear that the only factor that could get players to choose the Prisoner's Dilemma is an underappreciation of the adjustment by others to the new equilibrium. A large enough underappreciation of that adjustment is both necessary and sufficient to rank the Prisoner's Dilemma above the Harmony Game.

Of course, voters' preferring to play the Prisoner's Dilemma does not constitute an error unless the Prisoner's Dilemma actually yields lower payoffs than the Harmony Game, as predicted by Nash equilibrium. We expect this to hold, with the qualification that, as is well known, play in the Prisoner's Dilemma game approaches Nash equilibrium only after a few rounds of play. We then postulate the following,

Hypothesis 1 *Experienced players earn higher payoffs in the Harmony Game than in the Prisoner's Dilemma.*

Given this background, we can now state the first of our two main hypotheses.

Hypothesis 2 *A majority of voters prefer the Prisoner's Dilemma to the Harmony Game, regardless of which game is played initially.*

When voters choose the Prisoner's Dilemma over the Harmony Game, despite earning higher payoffs in the Harmony Game than PG, there is a political failure.

2.2 Making explicit the role of beliefs

The argument above rests on the expositionally convenient assumption that players initially play a Nash equilibrium in whichever game they start from and know that they will continue to play a Nash equilibrium in that game when policy remains unchanged. However, when choosing between games, players may predict non-Nash outcomes in both the proposed game as well as the initial one. To generalize our argument, as well as to substantiate the point that the underappreciation of behavior adjustment by others plays a key role, it is convenient to specify the beliefs that players assign to the actions C and D being played in each game, by themselves and by others.

Denote with β the subjective probability held by a player that the other player cooperates in the Prisoner's Dilemma, and with β' the corresponding probability in the Harmony Game. Let α denote the probability that a player assigns to his own cooperation in the Prisoner's Dilemma and α' the corresponding probability in the Harmony Game. Given these beliefs, and assuming risk neutrality throughout, a player perceives an expected utility from playing each game given by,

$$\begin{aligned} EU_{PD} &= 6\beta - 2\alpha + 5, \\ EU_{HG} &= 6\beta' + \alpha' + 1. \end{aligned}$$

A person with beliefs $[\alpha, \alpha', \beta, \beta']$ prefers the Harmony Game iff,

$$\Delta EU \equiv EU_{HG} - EU_{PD} > 0 \Leftrightarrow \beta' - \beta \equiv \Delta\beta \geq \frac{4 - \alpha' - 2\alpha}{6}. \tag{1}$$

Voters who estimate a lower difference $\Delta\beta$ in others' cooperation rates across games, i.e., those with lower estimates of the indirect effects, tend to prefer the Prisoner's Dilemma to the Harmony Game.

Someone who behaves according to the predictions of classical game theory, namely someone who plays dominant strategies and expects others to do so, will have parameters $[\alpha = 0, \alpha' = 1, \beta = 0, \beta' = 1]$. This person expects a difference in cooperation rates of 100 percentage points across games (i.e., $\Delta\beta = 1 \geq 1/2 = \frac{4-\alpha'-2\alpha}{6}$) and always prefers the Harmony Game. By contrast, consider a person who despite knowing he will always respond to a game by playing his dominant strategy does not necessarily expect others to play dominant strategies. This person will have parameters $[\alpha = 0, \alpha' = 1, \beta, \beta']$ and prefer the Harmony Game iff $\beta' - \beta \geq \frac{1}{2}$, that is, if he expects cooperation in the Harmony Game to be at least 50 percentage points higher than in the Prisoner's Dilemma.[4] Those who estimate the indirect effects to be less than 50 percentage points will support the Prisoner's Dilemma over the Harmony Game.

We do not expect the majority of players to miss the possibility of their own behavior adjusting. Instead, we hypothesize the following:

Hypothesis 3 *(a) A majority of subjects will underappreciate the indirect effects associated with the adjustment of behavior by others; specifically, the average belief differential about cooperation rates $\Delta\beta$ will be smaller than the equilibrium prediction ($\Delta\beta = 1$), and also*

[4] This statement would not change much by introducing plausible risk aversion. For example, with CRRA utility function $u(x) = \frac{x^\rho}{\rho}$, taking the limit case where $\rho \to 0$ gives the condition that $\beta' - \beta$ must exceed approximately 0.6.

smaller than the empirical difference in cooperation rates between games.

(b) Individuals who appreciate the indirect effects of others less will be more likely to support the Prisoner's Dilemma over the Harmony Game.

The core of our investigation concerns Hypotheses 1, 2, and 3. A secondary aim was to identify the cognitive bases for the failures we aim to document. We present two secondary hypotheses. To motivate the first, note that individuals who vote for the Prisoner's Dilemma because they do not expect the behavior of others to change fail to make predictions based on equilibrium considerations; even more, those predictions fail to recognize that others will follow dominant strategies. We conjecture that this failure may be related to a lack of strategic sophistication. We obtained one measure of strategic sophistication by having subjects play a p-beauty contest. We then hold the following:

Hypothesis 4 *Subjects who vote for the prisonners' dilemma are measured to be less sophisticated in the beauty contest.*

To motivate our other secondary hypothesis, note that although the beliefs $\Delta\beta$ we expect the Prisoner's Dilemma supporters to hold conflict with both theory and reality, it may be that conditional on those beliefs, a vote for the Prisoner's Dilemma is optimal. That may be the case for some subjects, yet we conjecture that,

Hypothesis 5 *A significant fraction of subjects vote in a way that is optimal given their reported beliefs only if they fail to appreciate that they themselves will behave differently in the new game.*

The import of the last hypothesis is that we expect some people to miss not only the fact that the behavior of others will change under a different game, but to also miss the fact that

their own behavior will change. As discussed above, this is neither necessary nor sufficient to cause a democratic failure in our setting, and hence not central to our main argument. But it highlights that subjects may display very basic departures from equilibrium thinking.

3 The Experiment

3.1 Design

We begin by explaining the basic structure of experimental sessions in all six treatments and then describe the differences across the treatments. Figure 2 summarizes the design.

[Figure 2 about here]

In Part 1 of the experiment we divided subjects into groups of six. Each subject played against every other one in the group exactly once, resulting in five periods of (one-shot) play in this part of the experiment. The game played varied by group. Groups were randomly assigned to play the Prisoners' Dilemma or the Harmony Game, detailed in Table 1 in the same format in which they were presented to subjects (to maintain neutrality in the labeling of actins, in the experiment we substituted the labels 1 and 2 for the labels C and D used in the table for the respective actions).[5] Subjects knew the game was symmetric, so this representation carries the same information as the normal-form representation. We strove to present all of the relevant information in a way that was both concise and complete, but also not so different from the way in which information might be gleaned in a real-life policy-choice scenario.[6] The exchange rate was $1 per 3 experimental points.

[5] During play, subjects were of course shown only the table corresponding to the game they were playing. At the time of voting, they were shown both tables side by side, as in Table 1.

[6] Even the representation in Table 1, precisely detailing payoffs for each combination of actions, may offer information in a clearer and more structured way than available in real life situation, which should mean the problems we identify in the lab can be compounded by additional mistakes outside of it.

After Part 1, new groups of six were formed randomly for Part 2, which included another five periods of play (6 to 10). At the beginning of Part 2, the game to be played in the next five periods was chosen. One of the main treatment variables is the way in which this choice was made as described below. After the choice of game for Part 2, but before Period 6, subjects reported their beliefs about how a randomly selected opponent in a similar experiment would act in each of the two games. As in periods 1 to 5 in Part 1, in Periods 6 to 10 every subject faced each other subject in his group exactly once.

The two treatment variables are the game that subjects played in Part 1 and the mechanism used to choose the game for Part 2. The treatments arms labeled Control, Random Dictator, Majority and Majority Once had the subjects play the Prisoner's Dilemma game in Part 1, while Reverse Control and Reverse Random Dictator had the subjects play the Harmony Game in Part 1.

In the control treatments (Control and Reverse Control), the game for Part 2 of the experiment was chosen at random by the computer. This choice was made once at the beginning of Part 2, and applied for all players in a group and all periods (i.e., all subjects in a given group played the same game in all periods in Part 2). The treatments Random Dictator and Reverse Random Dictator differed from the controls by asking all subjects to choose between the two games at the beginning of Part 2 and then implementing for the group the choice of a randomly selected subject. In the Majority treatment, the game chosen by the majority of the group before Period 6 was implemented for all periods in Part 2. Ties were randomly broken by the computer with even odds. In the Majority Repeated treatment, subjects voted for a game before each period of Part 2. In this treatment, beliefs were not elicited so as not to affect voting behavior in future periods. In all the other treatments, the belief elicitation always occurred after voting, so as not to affect voting. Subjects were informed of the implemented game and not the voting distribution.

At the end of the experiment, subjects played a p-beauty contest (Nagel 1995) to assess their strategic sophistication in simultaneous-move games and filled out a questionnaire providing basic demographics (gender, political ideology, class, major and SAT scores).

We recruited 384 student subjects from UC Berkeley and 384 from Brown University to participate in the experiment. Table 2 shows the number of subjects from each university in each of the six treatments. Sessions lasted around half an hour and earnings ranged from \$16.75 to \$37 with an average of \$27.81 (earnings included a \$5 show-up fee). Table 1 in the appendix displays summary statistics of demographics and beliefs.

4 Benchmark: Does the Harmony Game lead to higher payoffs than the Prisoners' Dilemma?

In order to establish whether voters demand efficient policies as captured by the Harmony Game, we first need to establish that the Harmony Game leads to higher payoffs than the Prisoner's Dilemma, as postulated in Hypothesis 1. Clearly, this is the prediction from game theory: defection in the Prisoner's Dilemma and cooperation in the Harmony Game are both dominant strategies under the assumption that subjects care mainly about monetary payoffs. Hence, any solution concept that assumes that players are rational predicts defection in the Prisoner's Dilemma and cooperation in the Harmony Game (e.g., Nash equilibrium or rationalizability).

But do subjects play close enough to the Nash outcome in each game so that payoffs and cooperation are greater in the Harmony Game than in the Prisoner's Dilemma? The answer is yes. Figure 2 shows the evolution of cooperation and payoffs as a function of the randomly chosen game for Part 2 in the Control and Reverse Control Treatments. The figures show that while there are no significant differences in behavior or payoffs in Part

1 by game selected for Part 2 (consistent with the random assignment of game), behavior and payoffs differ significantly by game in Part 2. Cooperation and payoffs go up when moving from the Prisoner's Dilemma to the Harmony Game, and down when moving from the Harmony Game to the Prisoner's Dilemma. In the Control condition, the shift from the Prisoner's Dilemma to the Harmony Game raises cooperation rates in Part 2 from 16% to 92% (see Appendix Table A2). While this increase of 76% is lower than the 100% predicted by Nash equilibrium, it is well above the 50% needed for a rational player to prefer the Harmony Game to the Prisoner's Dilemma. The changes in behavior are large even in the first interaction in Part 2 (Period 6), and are significant at less than 5% if we consider all periods and significant at less than 1% if we consider only Period 6. The consequent changes in payoffs are slightly more muted in period 6 but also strongly significant over all periods[7] Similar comparisons hold for the other treatments, where the choice of payoff matrix is not random (see Figure A1 and Appendix Tables A2 and A3). In conclusion, even in period 6, the change in game changes cooperation enough to warrant a preference for the Harmony Game over the Prisoner's Dilemma.

Another way to see that behavior across games differs in the direction predicted by theory is to compare the cooperation rates across the two games in Period 5, when the players have already gained experience. Pooling across all treatments where the subjects start to play the Prisoner's Dilemma and the Harmony Game, respectively, we find that the cooperation rate in the Prisoner's Dilemma is 15.5% while that in the Harmony Game is 95% (p-value < 0.0001). The corresponding average payoffs are 5.62 and 7.65, respectively (p-value < 0.0001). Again, the Harmony Game lead to higher payoffs.

In conclusion, behavior and payoffs across the two games vary enough in the direction

[7]The p-values for all comparisons reported in this section are obtained from Wald tests. We adopt the most conservative clustering of standard errors, at the session level. Clustering at the 6-person group level brings most p-values for changes in payoffs in period 6 below significance thresholds.

predicted by standard game theory that voting against the Harmony Game results in lower payoffs in practice as well as in theory, as anticipated in Hypothesis 1. Having established that game payoffs are ranked as in theory, the next question is whether subjects choose games accordingly.

5 The demand for bad policy

Although choosing the Harmony Game leads to higher average payoffs for the subjects, a slight majority of subjects (53.60%) across all treatments voted for the Prisoner's Dilemma game. The lowest share of subjects voting for the Prisoner's Dilemma game is 50.00% under Reverse Random Dictator, while the largest is 60.83% under Majority Once; see Table 3. All of these shares differ significantly from the 0% that would be expected if subjects chose games according to theory.

This is the main result of the paper – a majority of subjects demanded the wrong game or policy. As a result of voting, a majority of subjects (54.55%) ended up in a game in Period 6 that led to lower payoffs than they would have achieved by voting for the Harmony Game. Subjects' tendency to support bad policy is remarkably stable across our various treatments varying the decision mechanism and timing; we will compare the voting shares across treatments later in the paper.

Our design is predicated on the notion that once voters know the counterfactual situation to the status quo they already know, democracy should begin to work. Voting mistakes occur because individuals have difficulty predicting equilibrium responses for a situation that must be unfamiliar. Once it is familiar, making predictions is unnecessary. We investigate this possibility through the Majority Repeated condition. Figure 3 shows the evolution of votes under Majority Repeated. The percentage of subjects voting for the Prisoner's Dilemma

decreases from 50.83% in Period 6 to 28.33% in Period 10. It is noteworthy that even after five rounds of experience in which most subjects observed play under both games, more than a quarter of voters continued to choose the wrong game. But here the power of democracy kicks in: as this lower percentage can only rarely yield a majority for the Prisoner's Dilemma, very few groups end up in the wrong game by Period 10. The percentage of subjects playing the Prisoner's Dilemma decreases from 45% in Period 6 to 10% in Period 10.

Note again that none of the usual explanations for the implementation of bad policies (bad institutions, incompetent or corrupt policymakers, etc.) apply to the simple environments of this experiment. Thus, responsibility for the political failure can only be placed on the subjects, the citizens of this environment. But, what explains why most subjects demanded bad policy?

5.1 Mechanism: failure to appreciate equilibrium effects

According to Hypotheses 3 a) and b), many subjects fail to vote for the Harmony Game because they fail to understand or predict the effect that the payoff matrix has on behavior. Figure 4 shows the distribution of the difference in the beliefs of cooperation between the Harmony and the Prisoner's Dilemma games. We find that, on average, subjects grossly underestimate the effect of the game change on behavior. The average difference in belief of cooperation between the Harmony and the Prisoner's Dilemma games is 35% in Random Dictator and Majority Once while in reality cooperation is 76 percentage points higher in the Harmony Game. Similarly, the average belief difference is 30% in Reverse Random Dictator while the true difference in behavior is 63%.

Moreover, those subjects who most underestimate the effect of the game on behavior are most likely to vote for Prisoner's Dilemma. Figure 5 shows the average elicited belief of cooperation in each game broken down by vote of the subject. In all three treatments in

which beliefs were elicited, subjects who voted for the Prisoner's Dilemma expressed a lower belief that the behavior of others will differ across games. That is, subjects who voted for the Prisoner's Dilemma have a lower estimate of the effect of the payoff matrix on behavior. Notice that for someone who expected behavior to be independent of the payoff matrix, voting for the Prisoner's Dilemma would be optimal.

The relationship between the difference in the beliefs of cooperation and voting is highly statistically significant across treatments and robust to including personal characteristics of the subjects—see Table 4. The *Belief Diff* variable denotes the difference in the belief about the probability of cooperation of other subjects under the Harmony Game relative to the Prisoner's Dilemma (i.e., $\Delta\beta$). This OLS regression shows that an increase in the belief of cooperation of 100% (the theoretical prediction) would decrease the probability of voting for Prisoner's Dilemma by around 50%. Theory predicts that someone who expects switching to the Harmony Game will increase cooperation by anything from 50 to a 100 percentage points should change his chance of voting for the Harmony Game by 100%. So the theoretically predicted effect of beliefs lies between 1 and 2. The empirical coefficient is 0.5, or one half to one fourth of theory's prediction. In our experiment, subjects actions respond far less to their stated beliefs than predicted by theory.

Table 4 –columns 4 to 6– also shows that the relationship between voting for Prisoner's Dilemma and the belief difference is robust to controlling for personal characteristics. Most of these personal characteristics do not have a significant and consistent direct impact on voting across treatments, with the exception of ideology.[8]

The correlation of beliefs and voting documented in Table 4 does not necessarily imply

[8] Table A4 in the appendix shows the relationship between personal characteristics and voting for PD without controlling for belief difference, which is potentially endogenous. In the first stage regressions reported below (where we omit reporting the coefficients of control variables), we observe an effect of Economics major in driving a 10 percentage point wider belief differential when pooling treatments. This would be expected if an Economics education helps think about equilibrium effects. However, the effect is not significant for all treatments when taken separately.

that belief differences have a causal effect on voting. The reason is that people with different beliefs could also differ in dimensions that directly affect voting and are not observable.[9]

To show that beliefs have a causal effect on voting, we exploit exogenous variation in the beliefs held by a subject that is due to the behavior of the players encountered in periods 1 and 2. Because the identity and behavior of a subject's opponent in the first two periods is exogenous, it cannot correlate with any personal characteristic or past behavior of the subject (even the partner in period 2 cannot have played with anybody that has played with the subject in question). A subject who observes more defection in the first two periods while playing a Prisoner's Dilemma should have a greater belief that changing the game affects behavior than a subject that observed more cooperation. Based on this idea, we use the observed behavior of the other players in periods 1 and 2 (measured as a cooperation rate of either 0, 0.5, or 1) as an instrument for beliefs. The approach is admittedly demanding, since by the time beliefs are elicited three more periods of play have occurred. We restrict attention to the three treatments where beliefs were elicited (Random Dictator, Reverse Random Dictator and Majority Once). Panel A in Table 5 shows that the cooperation rate observed in the first two periods has the expected effect on beliefs in all three treatments (positive for Random Dictator and Majority Once, and negative for Reverse Random Dictator). However, while the instrument is strong for Majority Once, it is very weak for the other two treatments.

Panel B in Table 5 shows the second-stage results. For the treatment for which we have an instrument, Majority Once, we can see that subjects who expect the change of games to have a greater impact on cooperation are less likely to vote for the Prisoner's Dilemma. Having a valid instrument in this treatment, we can now ask whether the instrumented estimate

[9] Endogeneity may also occur if subjects examine the games further when being asked to report their beliefs, and report beliefs that justify their past voting choice. Costa-Gomes and Weizsäcker (2008) show evidence compatible with the idea that subjects re-examine strategic situations during the elicitation stage.

differs significantly from the OLS estimate. If not, then one cannot reject the hypothesis that beliefs are exogenous. The coefficients are very similar (-0.006 in the IV specification vs -0.005 in OLS) and a Hausman test cannot reject the null of belief exogeneity.

To sum up, the experimental evidence supports our core hypotheses that unfamiliar policy options involving indirect effects on payoffs will cause political failure, and that this failure will be driven by a majoritarian inability to fully appreciate the equilibrium adjustments triggering those indirect effects. In the remainder of this section and in the next one, we examine our secondary hypotheses.

We find that the strategic sophistication of subjects in simultaneous move games, as proxied by the number chosen in the p-beauty contest game, is not related to the voting decision in any of the treatments. This refutes our secondary Hypothesis 4. It could also indicate that the p-beauty contest number is not a good measure of strategic sophistication (for example, it is not the case that smaller numbers are necessarily a better choice given that others do not play Nash) or that what is crucial in the voting decision is the capacity or inclination to think about the behavior of others in future stages and not in a simultaneous-move game as in the beauty-contest game.[10]

5.2 Do subjects understand how changing the game will affect their own behavior?

We have shown that subjects that vote for the bad policy greatly underestimate the effect that a policy change will have on the behavior of others. But it is also possible that subjects do not appreciate that their own behavior will change as well, as postulated in our secondary Hypothesis 5.

[10] We exclude self-reported SAT scores for two reasons: first, because not all subjects provided this information, including it would reduce the number of observations in the analysis; second, SAT scores do not significantly predict voting, and excluding them does not change our results.

To study the share of subjects who fail to appreciate that their own behavior depends on the policy, we postulate a simple mixture model where individuals have one of two types $t \in (R, I)$ (for Responsive, and Inertial, respectively) depending on the way they think about their actions in each game.[11] The Responsive type is one who holds beliefs (β, β') about the cooperation rates by others in the Prisoner's Dilemma and the Harmony Game, respectively, but always recognizes he can respond in any game by playing his dominant strategy. The Inertial type does not realize he will adjust his own actions. This type considers that if he played action D(C) in the last round, he will continue to play it in the next, even if the game changes.

If we compute the expected payoff differential between the Prisoner's Dilemma and the Harmony Game for beliefs $[\alpha, \alpha', \beta, \beta']$ we obtain

$$\Delta u^t(\Delta\beta) = -6\Delta\beta + 4 - \alpha' - 2\alpha.$$

The key aspect differentiating the types is that the term $-\alpha' - 2\alpha$ is 1 for Responsive types and is either 0 or -3 for Inertial types who defected or cooperated in period 5, respectively. Thus, $\Delta u^R(\Delta\beta) = -6\Delta\beta + 3$ and $\Delta u^I(\Delta\beta, c) = -6\Delta\beta + 4 - 3c$, where c is an indicator variable for whether the subject cooperated in period 5.

We postulate the existence of a share s of Responsive types, and $1 - s$ of Inertial types. For the purposes of empirical identification of the share s, we assume that a Responsive (Inertial) type votes for the Prisoner's Dilemma game with a probability that depends on the payoff differential $\Delta u^R(\Delta\beta)$ $(\Delta u^I(\Delta\beta, c))$. To account for empirical errors, we will

[11] One way to study this would be to elicit beliefs about players' own actions. We did not do this in order not to disturb the elicitation of beliefs about others, which are key to our core hypotheses. An alternative was to add another condition, but given the large size of the experiment (768 subjects), we opted to investigate this secondary hypothesis via the structural approach presented in this section. While our Hypothesis 5 was formulated ex ante, the precise assumptions on types presented here were developed ex post.

assume that such probability is given by a logistic cdf with parameters (μ, σ). Thus, a player i with type t votes for Prisoner's Dilemma with a probability $F\left(\Delta u^t\left(\Delta\beta, c\right), \mu, \sigma\right)$, where F denotes the logistic distribution. It follows that the probability of a Prisoner's Dilemma vote by a player i, given a share s of Responsive types is,

$$P\left(v = PD|\Delta\beta, s\right) = sF\left(\Delta u^R\left(\Delta\beta\right), \mu, \sigma\right) + \left(1 - s\right) F\left(\Delta u^I\left(\Delta\beta, c\right), \mu, \sigma\right),$$

Given a profile of votes $\mathbf{v} = [v_1, ..., v_N]$ where $v_j = 1$ denotes a vote for Prisoner's Dilemma by subject j, and $v_j = 0$ a vote for Harmony Game, we have that the overall probability of such a profile is,

$$\prod_{i=1}^{N} P\left(v = PD|\Delta\beta, s, \mu, \sigma\right)^{v_i} \left(1 - P\left(v = PD|\Delta\beta, s, \mu, \sigma\right)\right)^{1-v_i},$$

which yields the log-likelihood,

$$L\left(s, \mu, \sigma|\mathbf{v}, \boldsymbol{\Delta\beta}\right) = \sum_{i=1}^{N} \left\{ \begin{array}{c} v_i \ln\left[sF\left(\Delta u^R\left(\Delta\beta\right), \mu, \sigma\right) + \left(1 - s\right) F\left(\Delta u^I\left(\Delta\beta, c\right), \mu, \sigma\right)\right] \\ + \left(1 - v_i\right) \ln\left[s\left(1 - F\left(\Delta u^R\left(\Delta\beta\right), \mu, \sigma\right)\right) + \left(1 - s\right)\left(1 - F\left(\Delta u^I\left(\Delta\beta, c\right), \mu, \sigma\right)\right)\right] \end{array} \right.$$

We estimate the parameter s, by maximizing $L\left(s, \mu, \sigma|\mathbf{v}, \boldsymbol{\Delta\beta}\right)$ given the voting data $\mathbf{v}$ and the vector of elicited beliefs $\boldsymbol{\Delta\beta}$. Clearly, in this estimation we take beliefs to be exogenous—this is a maintained assumption with some support from the exogeneity test performed earlier in relation with the instrumental-variables findings. We pool the data for the three conditions where beliefs were elicited, namely Random Dictator, Reverse Random Dictator, and Majority Once, for a total of 408 observations.

The estimate of the share of Responsive types s, presented in Table 6, is 67%. The Wald test for the share of Responsive types being less than 100% yields p-values of 0.042 and

0.057 depending on whether the standard errors are clustered respectively at the individual or group level. The estimates for the parameters (μ, σ) of the logistic distribution are 0.8 and 1.33, which is reasonable given the size of the payoff differentials in the games. Taken together, these findings support the notion that a fraction of the players vote without appreciating how their own play will adjust following a policy change. The point estimate suggests that a full third of the players make this error. This is striking, given that all that is required is to forecast that one will play a different, dominant, strategy in a 2x2 game following the change in policy.

6 Ruling out alternative mechanisms

The variation of treatments allows us to rule out some alternative mechanisms. One possibility is that the Prisoner's Dilemma obtains a majority in the Random Dictator and Majority treatments because a status quo bias causes some people to choose their initial game even if it is a suboptimal one. When the initial game is the Harmony Game instead, a reluctance to try new, risky things, should reinforce a preference for the initial, and also optimal, game and secure virtually unanimous support for the Harmony Game. However, in the Reverse Random Dictator condition the Prisoner's Dilemma garnered 50% of the vote. This vote share is only 3 points smaller than (and statistically indistinguishable from) that under Random Dictator. This evidence rules out the status quo bias possibility.

The Random Dictator treatment allows us to rule out forms of pivotal thinking as a source of the demand for bad policy. Under majority rule, the vote of a citizen matters only if pivotal. Two types of reasoning may dilute a subject's incentive to vote for the Harmony Game in that situation. First, the subject may have expectations of a large majority (in either direction), leaving her with negligible chances of being pivotal and therefore with no

incentive to carefully consider how to vote. Second, even if the chance of pivotality is not negligible, a subject may interpret the event of being pivotal as a sign that a large share of the subjects are not rational (if they were rational, they would vote for the Harmony Game). If many subjects are not rational, they may not respond to the change in game as theory predicts, and voting for the Harmony Game may be a bad idea. Under Random Dictator, pivotality has a clear, and non-negligible chance of 1/6. Moreover, the event of being pivotal does not depend on the votes of others and hence it cannot constitute a signal of how others may play in the Harmony Game. Therefore, the majoritarian 52.98% of votes for the Prisoner's Dilemma under Random Dictator cannot be explained by the previous pivotality concerns. The share of votes in favor of the Prisoner's Dilemma is greater under Majority Once (60.83%), but the difference is not statistically significant.

We derived our hypotheses from a framework where subjects may have difficulty appreciating how the behavior of others will adjust in a new game, and our main hypotheses meet with support in the data. Nevertheless, our framework would warrant less attention if the majority vote for Prisoner's Dilemma could be explained by strictly rational motives in an equilibrium framework. For example, a majority of players could be rational types who recognize their dominant strategies in each game, and they may fear that a non-trivial (yet minoritarian) subset of players are not rational. In the eyes of the rational types, those non-rational types intend to support the Prisoner's Dilemma and not play dominant strategies in each game. Then the rational types may optimally choose to support the Prisoner's Dilemma as well, creating a majority for Prisoner's Dilemma even when the majority of players are fully rational. This account is untenable, however. As shown in the conceptual framework section, a change in cooperation rates as low as 50% will still yield incentives to vote for the Prisoner's Dilemma under risk neutrality. So rational voters must expect a *majority* of others to fail to play dominant strategies to find it advantageous to vote for the Prisoner's Dilemma.

In other words, a majority of irrational types is needed to explain a majority for Prisoner's Dilemma. But what if subjects are risk averse? Could this open the door to a majoritarian vote for Prisoner's Dilemma while the proportion of irrational subjects are a minority? As we showed in footnote 4, even under extreme risk aversion the size of this minority would have to be very substantial (around 35%) and in our experiments virtually everyone plays dominant strategies in both Harmony Game and Prisoner's Dilemma, violating the assumption of a non-trivial share of irrational players. Thus, rationalizing the Prisoner's Dilemma majority in a model where the majority of players are both rational and correct about their environment appears difficult. This supports our interpretation, behavioral in nature, that individuals underappreciate equilibrium effects.

Several models of strategic naivety make predictions other than SPNE in our setting, and one may wonder whether the underappreciation of equilibrium effects we identify is a particular case of the phenomena explained by those theories. Here we discuss three such theories, namely the level-k model of strategic thinking (Stahl and Wilson 1994, Nagel 1995), Camerer, Ho and Chong's (2004) related "cognitive hierarchy model," and Jehiel's (2005) analogy-based-expectation equilibrium (ABEE). We then explain why they do not provide a fully satisfactory account of the patterns in our data.

The level-k model of strategic thinking summarizes players' strategic sophistication by the parameter k, where a level-k type of player best responds to beliefs that her opponent is a level-$k-1$ type of player (for $k \geq 1$), and a level-0 type randomizes uniformly over actions. Experimental work has estimated levels one and two to be the most frequent types across the universe of laboratory games where the model has been estimated (see, e.g., Crawford, Costa-Gomes and Iriberri 2013). In our setting, a level-0 type would cooperate and defect with equal probability in both games, leading a level-1 type to vote for Prisoner's Dilemma and play the dominant strategy in both Prisoner's Dilemma and Harmony Game. Because

level-1 types play dominant strategies, all higher levels vote for the Harmony Game. Hence, the level-1 type, better than any other type, fits the behavior of the majority of our subjects who vote for Prisoner's Dilemma. However, a theory predicated on level-1 makes at least one prediction that is contradicted by the data, namely that subjects voting for Prisoner's Dilemma predict that cooperation rates do not vary across Prisoner's Dilemma and Harmony Game. This does not match the fact that most individuals voting for Prisoner's Dilemma in our experiments, even if they underestimate the difference in cooperation across the two games, still predict more cooperation in the Harmony Game than in Prisoner's Dilemma. Those voting for the Prisoner's Dilemma estimate on average an increase in cooperation of about 20% (which is signficantly different from zero with a p-value < 0.0001).

A similar prediction is made by Jehiel's (2005) ABEE model. In this approach each player is modeled as bundling her opponents' decision nodes into partitional "analogy classes"; each player holds correct beliefs about her opponents' distribution of actions across each class, yet mistakenly believes that the frequency of each action played is constant across every node in an analogy class. There are two analogy classes of interest in our setting. Players with the finest analogy classes put Prisoner's Dilemma and Harmony Game in different analogy classes; each game has a different dominant strategy, and because players correctly predict actions in each class, the ABEE outcome coincides with subgame-perfect equilibrium. Alternatively, players who bundle Prisoner's Dilemma and Harmony Game into a single, coarser analogy class would predict that their opponents play the same way in both games. As with level-k models, this prediction is not supported by the data.

Since the level-k model constrains level-k players to believe their opponents are level-$k-1$, one may think it is not flexible enough to match our data, but that a more flexible framework in the same spirit could. Camerer, Ho and Chong's (2004) "cognitive hierarchy model" allows for more flexible beliefs: for instance, level-2 types believe that they face a

distribution of level-0 and level-1 types that coincides with the population distribution in the experiment. This account faces the same hurdle that denied the rational explanation we proposed above: virtually no subjects play the dominated action in either Harmony Game or Prisoner's Dilemma, so level-2 types must assign close to probability one to level-1 types in the cognitive-hierarchy model too. Thus, if there are no level-0 players, higher level players have no reason to vote for Prisoner's Dilemma in this model. Yet, we observe majoritarian support for Prisoner's Dilemma in the data.

7 Conclusion

We have experimentally identified a collective failure to democratically resolve social dilemmas driven by an underappreciation of how behavior changes when games change. Players underestimate, on average, how much others' behavior changes following a game change, or policy reform. In addition, a non-trivial share of players appear to fail to appreciate that their own behavior will differ across games. Our evidence suggests that unfamiliar policy options that contain "hidden" costs or benefits that will accrue once behavior adjusts, can be a challenge for direct democracy. To the extent that candidates compete by associating themselves with policy proposals they expect to be popular, the challenge may extend to representative democracy as well.

Of course, identifying collective failures and a tendency to underestimate equilibrium effects in the laboratory does not necessarily mean that we have identified an important source of bad policy in real life. One could hope that public discourse and political competition results in voters learning about the total effect of policies, thus bridging the gap between public opinion and reliable evidence. However, a vast literature in economics and political science—both theoretical and empirical—has considered politicians as reflecting, more than

shaping, the positions of voters, while the field of political behavior continues to document persistent discrepancies between public opinion and the consensus of policy experts. To the extent that public opinion and voter preferences matter in democracies, understanding how people think about policy choice appears relevant for our knowledge of how societies choose to regulate themselves.

8 Appendix

8.1 A more general framework

Here we present a slightly more general framework to link beliefs about cooperation $[\alpha, \alpha', \beta, \beta']$ to the direct and indirect effects d, s, o. In particular, we show that under some assumptions on payoffs, it is possible to linearly decompose the gain from playing one game over the other as $g = d + s + o$, where $s = \Delta\alpha X$, $\omega = \Delta\beta Y$, where $\Delta\alpha, \Delta\beta$ are the beliefs about differential cooperation across games by self and others respectively, and where X, Y are functions of the game payoffs.

Assume general Prisoner's Dilemma and Harmony Game games as follows:

Prisoner's Dilemma

	C	D
C	R	S
D	T	P

Harmony Game

	C	D
C	$R-x$	$S-x$
D	$T-y$	$P-y$

with the usual restriction that $T > R > P > S$ and that $y - x > max\{T - R, P - S\}$. In addition we assume that the effect of the other player's action on own payoffs is the same regardless of a player's own action as in the experiment: $R - S = T - P$.

Define the direct effect as $d = EU(HG, \alpha, \beta) - EU(PD, \alpha, \beta)$. It can be easily shown

that, regardless of T, R, P and S: $d = -(\alpha x + (1-\alpha) y)$, where x and y are the Pigouvian taxes.

Define the indirect effect due to the adjustement of others as $o = EU(HG, \alpha, \beta') - EU(HG, \alpha, \beta)$. Under the assumption that $R - S = T - P$: $o = (\beta' - \beta)(T - P)$.

Define the indirect effect due to the adjustement of self as $s = EU(HG, \alpha', \beta) - EU(HG, \alpha, \beta)$. Under the assumption that $R - S = T - P$: $s = (\alpha' - \alpha)(S - P - x + y)$.

Note that the change in expected payoffs from the game change is $\Delta EU = \alpha'(y - x) - y + (\beta' - \beta)(T - P) + (\alpha' - \alpha)(S - P)$. This difference in expected utilities can be written as: $\Delta EU = d + o + s$, where $d = \alpha'(y - x) - y$, $o = (\beta' - \beta)(T - P)$ and $s = (\alpha' - \alpha)(S - P)$.

This linear decomposition would not hold if $R - S \neq T - P$.

References

[1] Alesina, A. and G. Tabellini (1990). "Voting on the Budget Deficit," *American Economic Review* 80(1), 37-49.

[2] Alesina, A. and A. Drazen (1991). "Why are Stabilizations Delayed?" *American Economic Review* 81, 1170-1188.

[3] Andreoni, J., and J.H. Miller (1993). "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence," *Economic Journal* 103(4), 570-85.

[4] Bartels, L. (1996). "Uninformed Votes: Information Effects in Presidential Elections," *American Journal of Political Science* 40(1), 194-230.

[5] Bartels, L. (2012). "The Study of Electoral Behavior," in Jan Leighley (ed.) The Oxford Handbook of American Elections and Political Behavior. Oxford University Press.

[6] Besley, T. (2005). "Political Selection," *Journal of Economic Perspectives* 19(3), 43-60.

[7] Besley, T. and S. Coate (1998). "Sources of Inefficiency in a Representative Democracy: A Dynamic Analysis," *American Economic Review* 88(1), 139-56.

[8] Bisin, A., A. Lizzeri and L. Yariv (2011). "Government Policy with Time Inconsistent Voters," Unpublished manuscript. http://www.econ.nyu.edu/user/lizzeria/Deficit_Nov2_11AL.pdf

[9] Blinder, A. and A. Krueger (2004), "What Does the Public Know about Economic Policy, and How Does It Know It?" *Brookings Papers on Economic Activity* 2004(1), 327-397.

[10] Bone, J., Hey, J.D., and J. Suckling (2009), "Do People Plan?" Experimental Economics 12, 12-25.

[11] Camerer, Colin F., Ho, Teck-Hua, and Juin Kuan Chong "A cognitive hierarchy model of games," *Quarterly Journal of Economics* 119(3): 861-898.

[12] Canes-Wrone, B., M.C. Herron, and K.W. Shotts (2001). "Leadership and Pandering: A Theory of Executive Policymaking," *American Journal of Political Science* 45, 532-550.

[13] Caselli, F., and M. Morelli (2004). "Bad Politicians," *Journal of Public Economics* 88(2), 759-82.

[14] Charness, G. and M. Rabin (2002). "Understanding Social Preferences With Simple Tests," *Quarterly Journal of Economics* 117(3), 817-869.

[15] Coate, S. and S. Morris (1995). "On the Form of Transfers to Special Interests," *Journal of Political Economy* 103(6), 1210-35.

[16] Costa-Gomes, Miguel and Georg Weizsäcker (2008). "Stated Beliefs and Play in Normal-Form Games," *Review of Economic Studies* 75: 729-762.

[17] Crawford, V. and N. Iriberri (2007). "Level-k Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?" *Econometrica* 75(6), 1721–70.

[18] Dal Bó, E., and R. Di Tella. (2003). "Capture by Threat," *Journal of Political Economy* 111 (October), 1123-54.

[19] Dal Bó, E., Dal Bó, P., and R. Di Tella (2006). "Plata o Plomo?: Bribe and Punishment in a Theory of Political Influence," *American Political Science Review* 100(1), 41-53.

[20] 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. http://www.econ.brown.edu/fac/Pedro_Dal_Bo/institutionschapter.pdf

[21] De Figueiredo, R. (2002). "Electoral Competition, Political Uncertainty, and Policy Insulation," *American Political Science Review* 96(2), 321-333.

[22] Ertan, A., T. Page and L. Putterman (2005). "Who to Punish? Individual Decisions and Majority Rule in Mitigating the Free-Rider Problem?" *European Economic Review* 53(5), 495-511.

[23] Esponda, I. (2008). "Behavioral Equilibrium in Economies With Adverse Selection" *American Economic Review* 98(4), 1269-91.

[24] Esponda, I. and D. Pouzo (2010). "Conditional Retrospective Voting in Large Elections," Unpublished manuscript. http://people.stern.nyu.edu/iesponda/Ignacio_Esponda/Research_files/ep-CRV-jan-12.pdf

[25] Esponda, I. and E. Vespa (2012). "Hypothetical Thinking and Information Extraction: Strategic Voting in the Laboratory," Unpublished manuscript. http://people.stern.nyu.edu/iesponda/Ignacio_Esponda/Research_files/ev-pivvotelab-jun07-12.pdf

[26] Eyster, E. and M. Rabin (2005). "Cursed Equilibrium,"' *Econometrica* 73(5), 1623-1672.

[27] Fehr, E. and K.M. Schmidt. (1999). "A theory of fairness, competition, and cooperation," *Quarterly Journal of Economics* 114(3), 817–868.

[28] Gentzkow, M. and J. Shapiro. (2010). "What drives media slant? Evidence from U.S. newspapers," *Econometrica* 78(1), 35-71.

[29] Isaac, R., K. McCue, and C. Plott (1985). "Public goods provision in an experimental environment," *Journal of Public Economics* 26, 51-74.

[30] Jehiel, P. (2005). "Analogy-based expectation equilibrium," *Journal of Economic Theory* 123 (2), 81–104.

[31] Karni, E. (2009). "A Mechanism for Eliciting Probabilities," *Econometrica*, 77, 603–606.

[32] Kim, O., and M. Walker (1984). "The free rider problem: experimental evidence," *Public Choice* 43: 3-24.

[33] Knittel, C. (2012). "Reducing Petroleum Consumption from Transportation," *Journal of Economic Perspectives* 26(1), 93-118.

[34] Levitt, S., List, J. and S. Sadoff (2011). "Checkmate: Exploring Backward Induction among Chess Players," *American Economic Review* 101(2): 975-90.

[35] Lizzeri, A. and L. Yariv (2011). "Collective Self-Control," Unpublished manuscript. http://www.hss.caltech.edu/~lyariv/Papers/Trees.pdf

[36] Margreiter, M., M. Sutter, and D. Dittrich (2005). "Individual and Collective Choice and Voting in Common Pool Resource Problem with Heterogeneous Actors," *Environmental & Resource Economics* 32, 241-71.

[37] Maskin, E. and J. Tirole (2004). "The Politician and the Judge: Accountability in Government," *American Economic Review* 94(4).

[38] McKelvey, R. and T. Palfrey (1992). "An Experimental Study of the Centipede Game", *Econometrica* 60: 803-836.

[39] Moinas, S. and S. Pouget (2013). "The Bubble Game: An Experimental Study of Speculation", *Econometrica* 81(4): 1507-1540.

[40] Mullainathan, S. and E. Washington (2009). "Sticking with Your Vote: Cognitive Dissonance and Political Attitudes," *American Economic Journal: Applied Economics* 1(1), 86-111.

[41] North, D. C. (1990). *Institutions, Institutional Change and Economic Performance*, Cambridge University Press: Cambridge.

[42] Peltzman, S. (1976). "Toward a More General Theory of Regulation," *Journal of Law and Economics* 19, 211-240.

[43] Rabin, M. (1993). "Incorporating Fairness Into Game Theory and Economics," *American Economic Review* 83, 1281-1302.

[44] Romer, T. and H. Rosenthal (1978). "Political Resource Allocation, Controlled Agendas and the Status Quo," *Public Choice* 33, 27-43.

[45] Sausgruber, R. and J-R Tyran (2005). "Testing the Mill Hypothesis of Fiscal Illusion", *Public Choice* 122: 39-68.

[46] Sausgruber, R. and J-R Tyran (2011). "Are we taxing ourselves? How deliberation and experience shape voting on taxes," *Journal of Public Economics* 95, 124-176.

[47] Smith, Adam (1776). *An Enquiry into the Nature and Causes of the Wealth of Nations* W. Strahan and T. Cadell: London.

[48] Stigler, G. (1971). "The Regulation of Industry" *The Bell Journal of Economics and Management Science* 2, 3-21.

[49] Walker, J., R. Gardner, A. Herr and E. Ostrom (2000). "Collective Choices in the Commons: Experimental Results on Proposed Allocation Rules and Votes," *The Economic Journal* 110(1), 212-34.

Figure 1: Direct and Indirect Effects of Policy Change

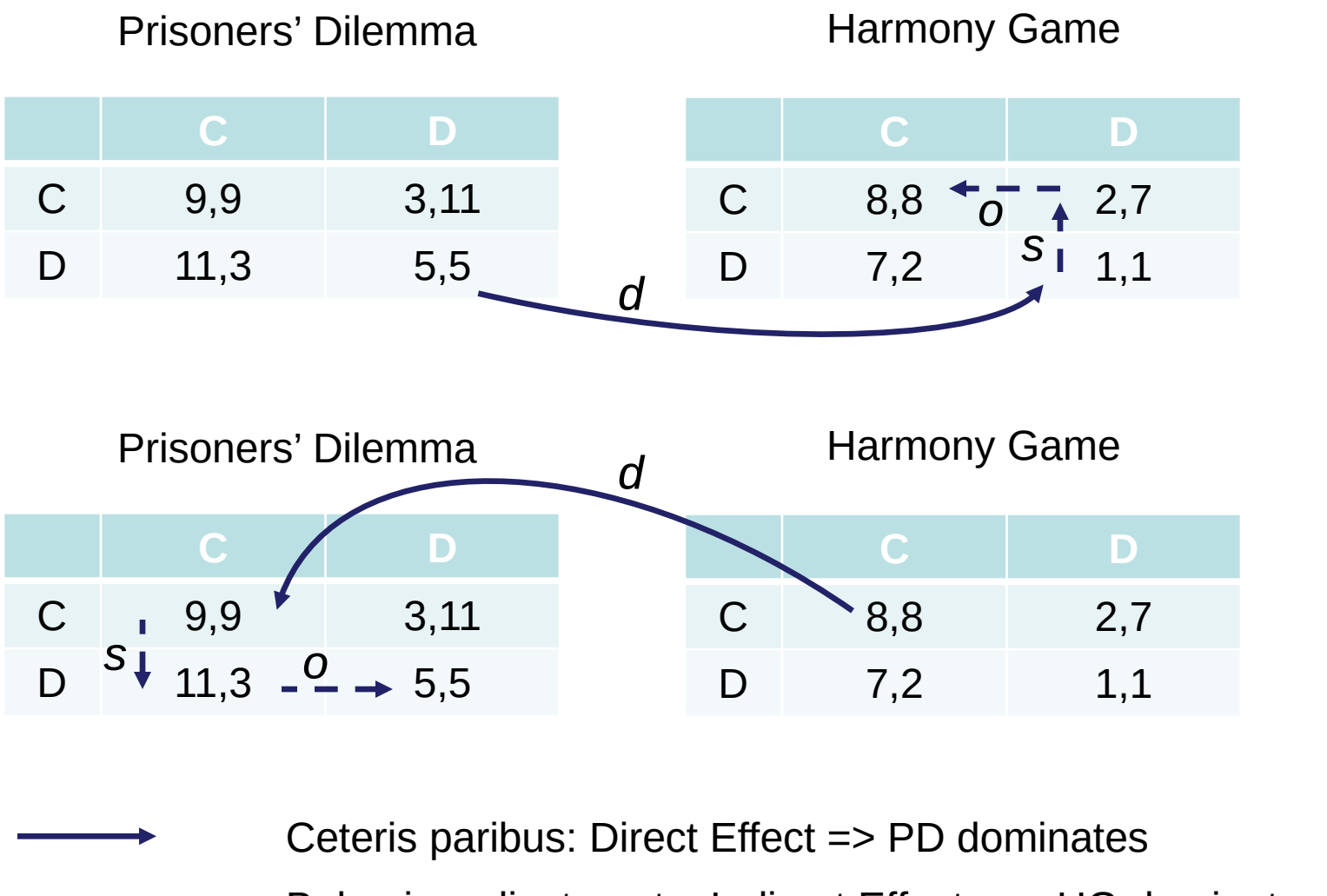
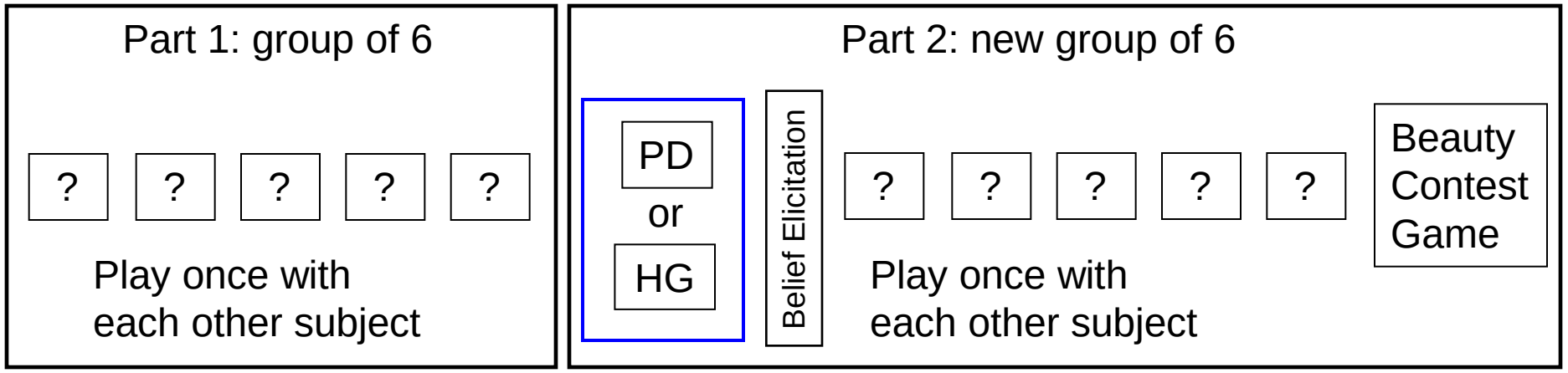


Figure 2: Experimental Design:



Questionnaire

Gender
Ideology
Class
Major
SATs

Treatments:

1. Control: PD first, game chosen randomly
2. Reverse Control: ALT first, game chosen randomly
3. Random Dictator: PD first, random dictator
4. Reverse RD: ALT first, random dictator
5. Majority Once: PD first, majority voting
6. Majority Repeated: PD first, majority voting before each game in part 2

Figure 3: The Power of Nash Equilibrium:

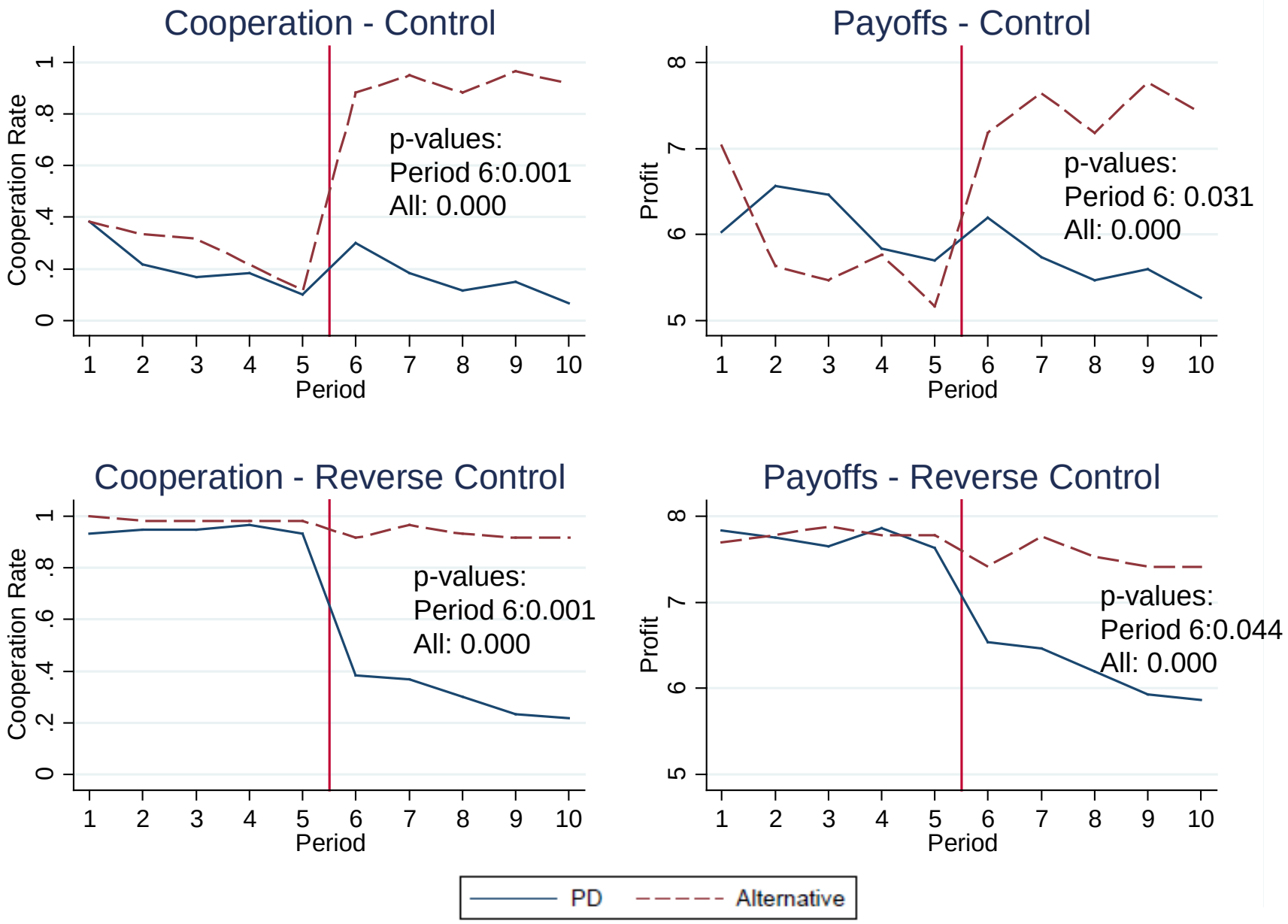


Figure 4: Evolution of Voting and Chosen Game in Majority Repeated

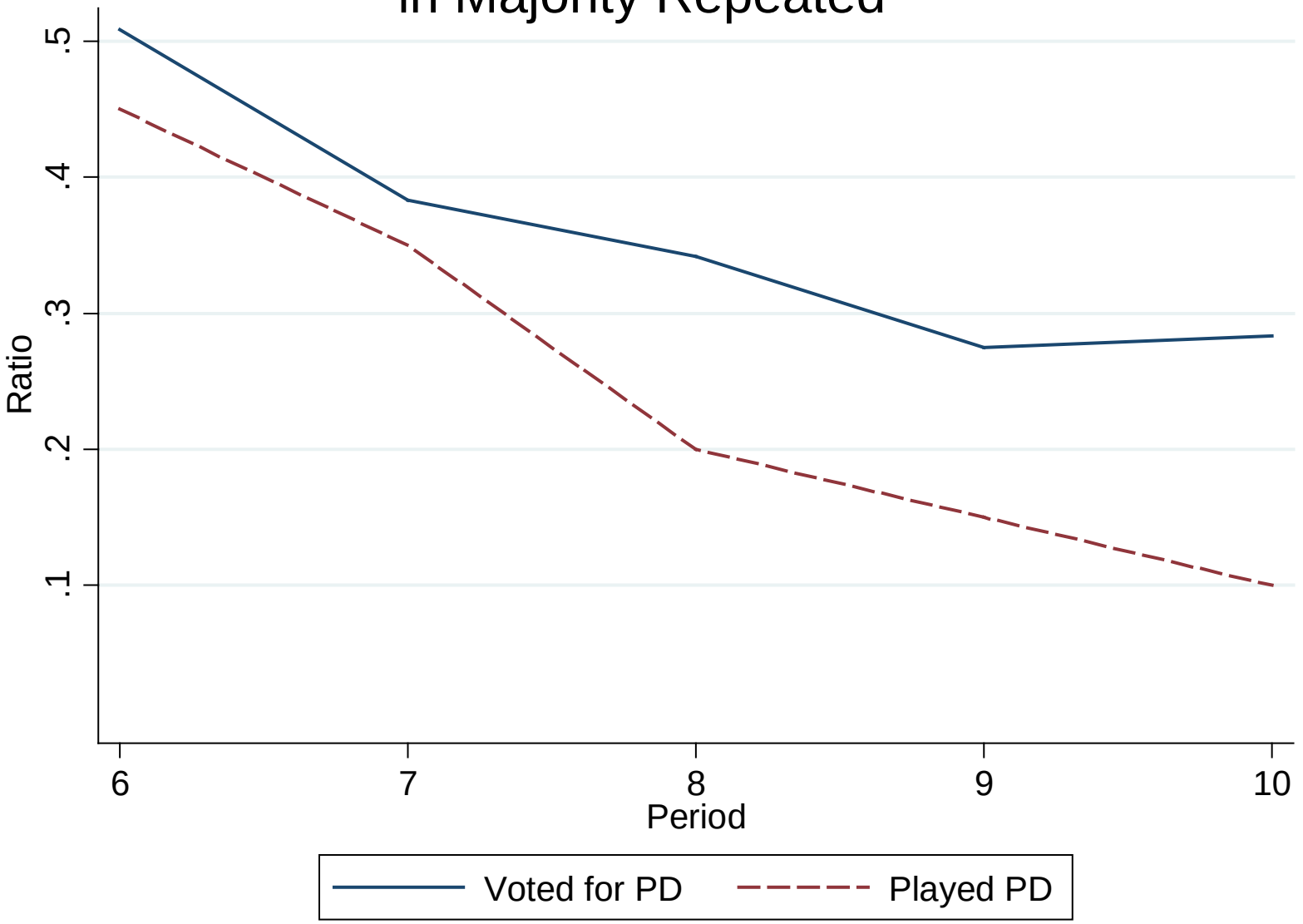


Figure 5: Distribution of Difference in Beliefs of Cooperation Between Games (HG – PD)

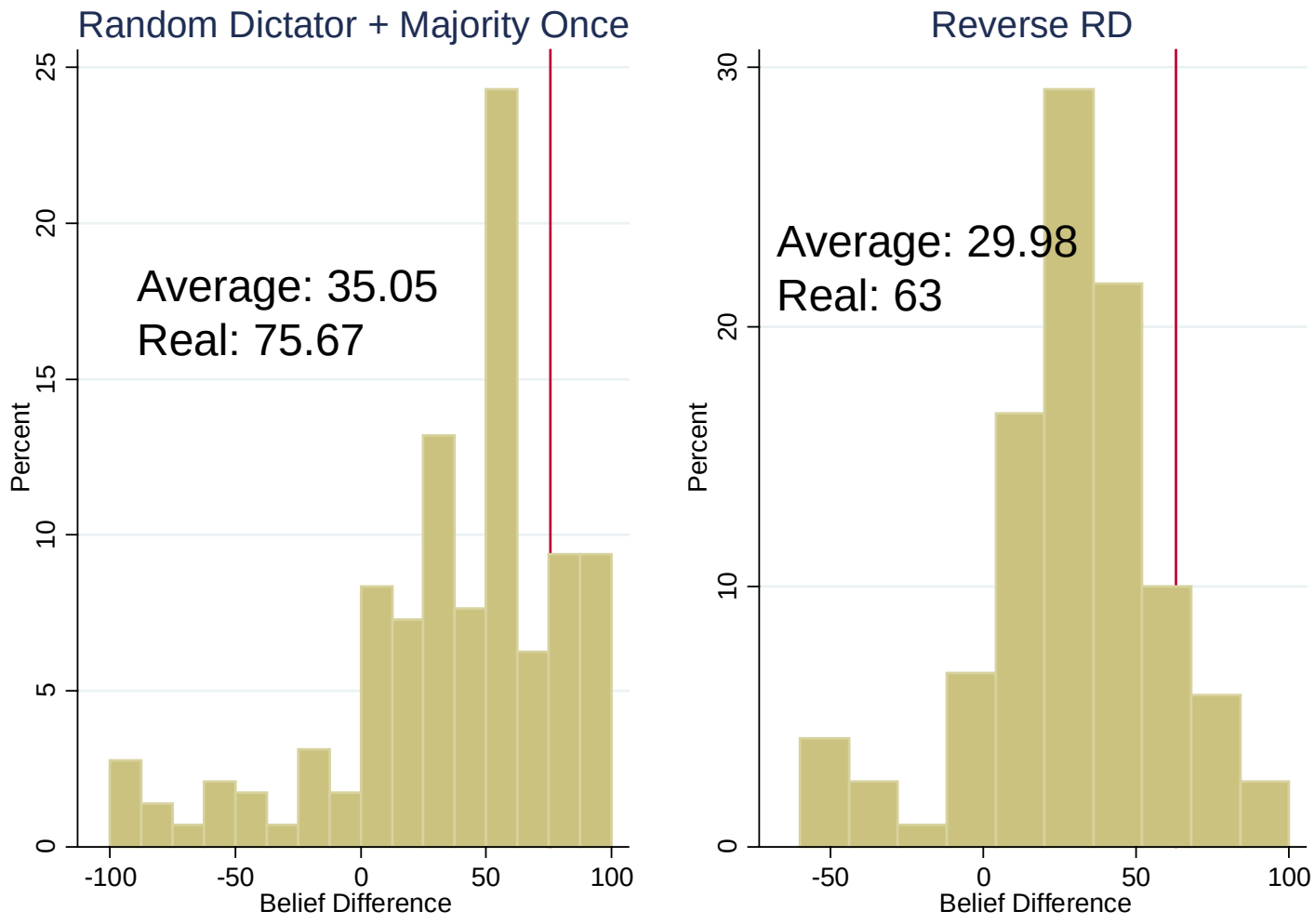


Figure 6: Beliefs of Cooperation and Voting

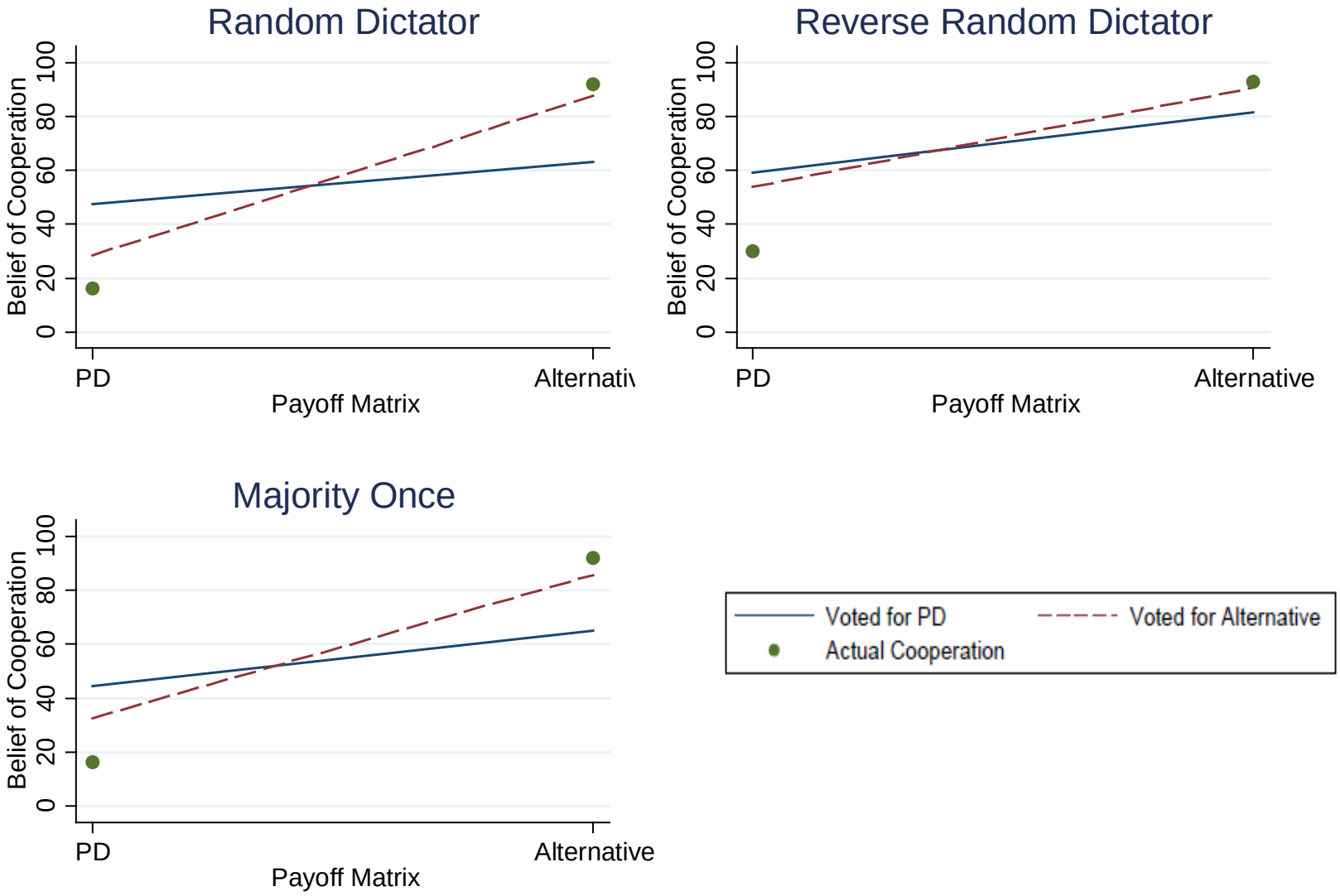
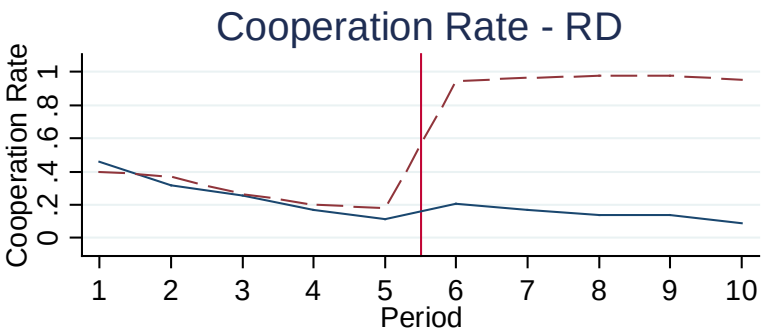
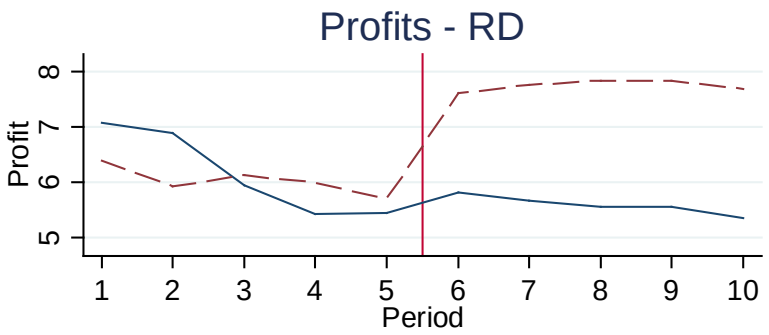
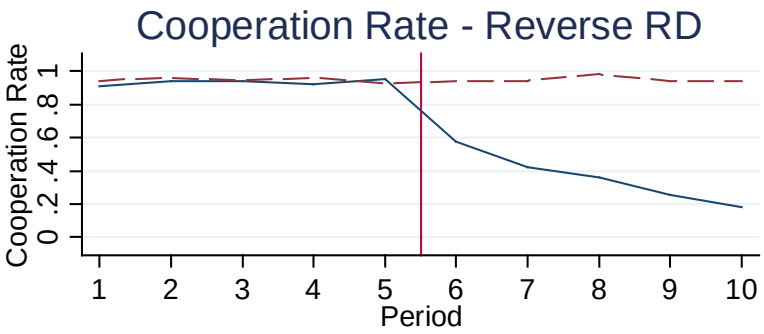
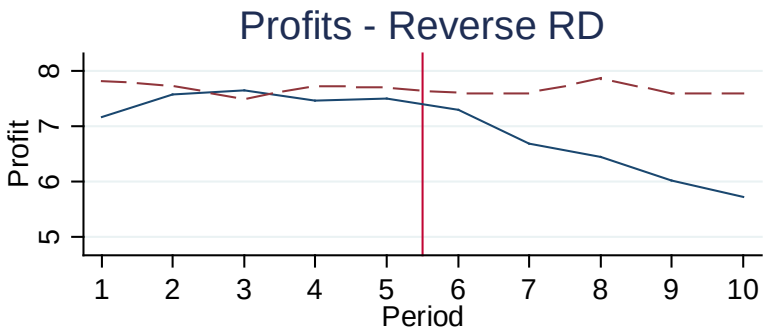


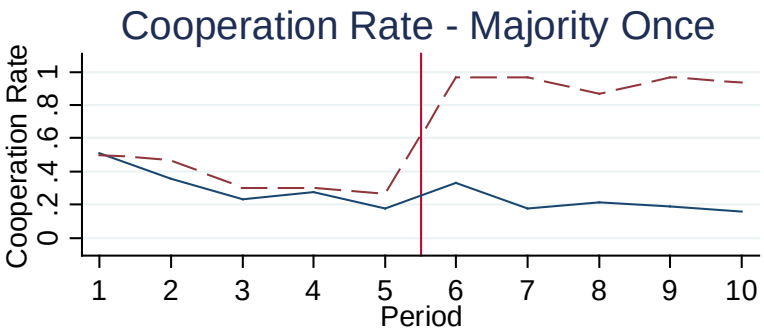
Figure A1: Comparing PD and HG











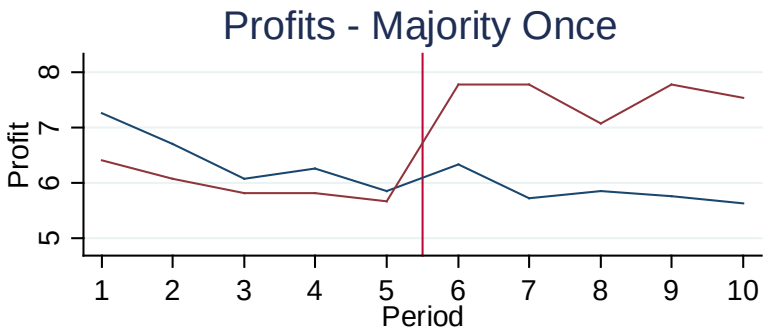


Table 1: The Two Games

Prisoners' Dilemma			Harmony Game		
Your choice	Other's choice	Earnings	Your choice	Other's choice	Earnings
C	C	9	C	C	8
D	C	11	D	C	7
C	D	3	C	D	2
D	D	5	D	D	1

Note: Earnings are in experimental points (3 experimental points equal $1).

Table 2: Number of Subjects by Treatment and Place

	Berkeley	Brown	Total
Control	60	60	120
Reverse Control	60	60	120
Random Dictator	84	84	168
Reverse Random Dictator	60	60	120
Majority Once	60	60	120
Majority Repeated	60	60	120
Total	384	384	768

Table 3: Prisoner's Dilemma vote shares by Treatment

Treatment	Vote PD	Play PD
Random Dictator	52.98%	46.43%
Reverse Random Dictator	50.00%	55.00%
Majority Once	60.83%	75.00%
Majority Repeated	50.83%	45.00%
Total	53.60%	54.55%

Note: Play PD reports the share of people playing the Prisoner's Dilemma in Period 6. This differs from the vote share since vote shares differ by group.

Table 4: Beliefs and Voting for PD (Dependent Variable: Vote PD)

	Treatment					
	RD	Reverse RD	Majority Once	RD	Reverse RD	Majority Once
	(1)	(2)	(3)	(4)	(5)	(6)
Belief Difference	-0.005	-0.004	-0.005	-0.005	-0.004	-0.005
	[0.001]***	[0.001]**	[0.001]***	[0.000]***	[0.001]***	[0.001]***
Male				-0.126	0.091	-0.207
				[0.068]*	[0.102]	[0.083]**
Year				-0.007	-0.003	0.081
				[0.032]	[0.037]	[0.039]**
Ideology				0.035	0.038	0.045
				[0.018]*	[0.017]**	[0.022]*
Economics				0.051	0.175	-0.052
				[0.062]	[0.174]	[0.151]
Political Science				-0.194	0.24	0.093
				[0.222]	[0.170]	[0.119]
Brown University				0.07	0.059	0.149
				[0.080]	[0.111]	[0.073]*
Beauty Number				-0.001	0	0
				[0.002]	[0.002]	[0.002]
Constant	0.698	0.619	0.774	0.622	0.343	0.43
	[0.037]***	[0.078]***	[0.044]***	[0.122]***	[0.165]*	[0.185]**
Observations	168	120	120	168	120	120
R-squared	0.2	0.06	0.16	0.24	0.12	0.24

Note: Year denotes year in college. Ideology from 0 to 10 from most liberal to most conservative. Belief Difference denotes the difference between beliefs of cooperation under Harmony and Prisoner's Dilemma games. Robust standard errors in brackets clustered by part 1 group: * significant at 10%; ** significant at 5%; *** significant at 1%

Table 5: Instrumenting for Beliefs

Panel A: First Stage (dependent variable: Belief Difference)

	Treatment					
	RD	Reverse RD	Majority Once	RD	Reverse RD	Majority Once
	(1)	(2)	(3)	(4)	(5)	(6)
Average other coop in periods 1 & 2	-7.122	15.008	-42.89	-8.915	21.166	-47.546
	[10.489]	[21.063]	[7.552]***	[10.032]	[19.966]	[6.701]***
Personal Characteristics Included	N	N	N	Y	Y	Y
Constant	38.944	15.63	52.572	23.745	0.439	35.767
	[5.609]***	[20.935]	[4.740]***	[21.358]	[19.130]	[16.855]**
Observations	168	120	120	168	120	120
R-squared	0	0.01	0.14	0.07	0.06	0.17
F of excluded instrument	0.46	0.51	32.25	0.79	1.12	50.35

Panel B: Second Stage (dependent variable: Vote for PD)

	Treatment					
	RD	Reverse RD	Majority Once	RD	Reverse RD	Majority Once
	(1)	(2)	(3)	(4)	(5)	(6)
Belief Difference	-0.031	-0.022	-0.007	-0.026	-0.021	-0.006
	[0.040]	[0.020]	[0.003]**	[0.026]	[0.017]	[0.003]**
Personal Characteristics Included	N	N	N	Y	Y	Y
Constant	1.656	1.154	0.83	1.112	0.669	0.449
	[1.432]	[0.589]*	[0.110]***	[0.650]*	[0.318]**	[0.186]**
Observations	168	120	120	168	120	120

Note: Personal characteristics include gender, year of studies, ideology, Economics and Political Science concentrations and number from beauty contest game. Belief Difference denotes the difference between beliefs of cooperation under Harmony and Prisoner's Dilemma games. Robust standard errors in brackets clustered by part 1 group: * significant at 10%; ** significant at 5%; *** significant at 1%

Table 6: Structural estimates

s (share of Rational types	0.67
	[0.17]
Wald test $s \neq 1$; p-value:	0.06
Logistic distribution parameters	
μ	0.80
	[0.15]
σ	1.34
	[0.35]

Note: Pooled sample from Random dictator, Reverse random dictator and Majority once treatments. Robust standard errors clustered by part 1 group in brackets.

Appendix Table A1: Summary Statistics

	Obs.	Mean	Std. Dev.	Min	Max
Male	768	0.43	0.50	0	1
Year	768	2.70	1.21	1	5
Ideology	768	3.54	2.14	0	10
Economics	768	0.15	0.36	0	1
Political Science	768	0.05	0.21	0	1
Brown U.	768	0.50	0.50	0	1
Beauty Contest Number	768	36.67	21.35	0	100
Math SAT	662	723.95	71.77	400	800
Verbal SAT	644	700.19	77.45	400	800
Belief of C in PD	408	44.26	25.79	0	100
Belief of C in HG	408	77.74	26.02	0	100
Belief Difference	408	33.47	-41.11	-100	100
Earnings	768	27.81	3.27	16.75	37

Appendix Table A2: Cooperation comparison between games by treatment

	Control			Reverse Control			Random Dictator			Reverse RD			Majority Once		
	Part 1	Part 2		Part 1	Part 2		Part 1	Part 2		Part 1	Part 2		Part 1	Part 2	
Periods	All	6	All	All	6	All	All	6	All	All	6	All	All	6	All
Part 2 Game															
HG	27%	88%	92%	99%	92%	93%	28%	94%	96%	95%	94%	95%	37%	97%	94%
PD	21%	30%	16%	95%	38%	30%	26%	21%	15%	93%	58%	36%	31%	33%	21%
Diff. p-value	0.288	0.001	0.000	0.319	0.001	0.000	0.786	0.000	0.000	0.661	0.001	0.000	0.421	0.000	0.000

Note: p-values calculated using Wald tests with s.e. clustered at session level.

Appendix Table A3: Payoff comparison between games by treatment

	Control			Reverse Control			Random Dictator			Reverse RD			Majority Once		
	Part 1	Part 2		Part 1	Part 2		Part 1	Part 2		Part 1	Part 2		Part 1	Part 2	
Periods	All	6	All	All	6	All	All	6	All	All	6	All	All	6	All
Part 2 Game															
HG	5.81	7.18	7.44	7.79	7.42	7.51	6.04	7.61	7.75	7.70	7.61	7.66	5.95	7.77	7.58
PD	6.12	6.20	5.65	7.75	6.53	6.20	6.16	5.82	5.59	7.48	7.30	6.44	6.42	6.33	5.85
Diff. p-value	0.186	0.115	0.001	0.641	0.117	0.004	0.652	0.000	0.000	0.259	0.392	0.001	0.019	0.008	0.005

Note: p-values calculated using Wald tests with s.e. clustered at session level.

Appendix Table A4: Personal Characteristics and Voting for PD (Dependent Variable: Vote PD)

	Treatment			
	RD	Reverse RD	Majority Once	Majority Repeated
	(1)	(2)	(3)	(4)
Male	-0.208	0.061	-0.219	-0.272
	[0.083]**	[0.116]	[0.097]**	[0.100]**
Year	0.006	-0.012	0.063	-0.023
	[0.031]	[0.040]	[0.041]	[0.043]
Ideology	0.027	0.04	0.049	-0.004
	[0.018]	[0.018]**	[0.025]*	[0.016]
Economics	-0.034	0.164	-0.081	-0.086
	[0.083]	[0.171]	[0.175]	[0.149]
Political Science	-0.147	0.263	0.07	-0.102
	[0.215]	[0.141]*	[0.180]	[0.164]
Brown University	0.066	0.044	0.136	0.043
	[0.087]	[0.121]	[0.084]	[0.116]
Beauty Number	-0.001	0.001	0	0.003
	[0.002]	[0.003]	[0.002]	[0.002]
Constant	0.519	0.26	0.321	0.589
	[0.153]***	[0.175]	[0.188]	[0.185]***
Observations	168	120	120	120
R-squared	0.06	0.06	0.08	0.11

Note: Year denotes year in college. Ideology from 0 to 10 from most liberal to most conservative. Belief Difference denotes the difference between beliefs of cooperation under Harmony and Prisoner's Dilemma games. Robust standard errors in brackets clustered by part 1 group: * significant at 10%; ** significant at 5%; *** significant at 1%