Review of Economic Studies (2018) **85**, 964–998 doi:10.1093/restud/rdx031

Advance access publication 9 May 2017

The Demand for Bad Policy when Voters Underappreciate Equilibrium Effects

ERNESTO DAL BÓ
UC Berkeley
PEDRO DAL BÓ
Brown University
and
ERIK EYSTER
LSE

First version received July 2015*; Editorial decision January* 2017*; Accepted May* 2017 (*Eds.*)

Most of the political economy literature blames inefficient policies on institutions or politicians' motives to supply bad policy, but voters may themselves be partially responsible by demanding bad policy. In this article, we posit that voters may systematically err when assessing potential changes in policy by underappreciating how new policies lead to new equilibrium behaviour. This biases voters towards policy changes that create direct benefits—welfare would rise if behaviour were held constant—even if those reforms ultimately reduce welfare because people adjust behaviour. Conversely, voters are biased against policies that impose direct costs even if they induce larger indirect benefits. Using a lab experiment, we find that a majority of subjects vote against policies that, while inflicting direct costs, would help them to overcome social dilemmas and thereby increase welfare. Subjects also support policies that, while producing direct benefits, create social dilemmas and ultimately hurt welfare. Both mistakes arise because subjects fail to fully anticipate the equilibrium effects of new policies. More precisely, we establish that subjects systematically underappreciate the extent to which policy changes will affect the behaviour of other people, and that these mistaken beliefs exert a causal effect on the demand for bad policy.

Key words: Voting, Reform, Political failure, Endogenous policy, Experiment

JEL Codes: C9, D7

1. INTRODUCTION

Theories of voting, and more broadly of democratic politics, rest on the premise that citizens tend to assess correctly the relative merits of the policy options they face. Not surprisingly, existing explanations for political failure (*i.e.* the selection of inefficient policies) typically blame

bad institutions, special interests, or the quality and motives of politicians.[1] In this article, we articulate an account of political failure that assigns some of the blame to voters.

Our basic hypothesis is that when evaluating a new policy, voters tend to underappreciate equilibrium effects, which creates a systematic tendency for them to incorrectly rank policies in welfare terms. Prior claims have been made that the average citizen may not fully understand the implications of equilibrium and consequently misjudge policy. Smith (1776) emphasized the limited grasp of the implications of market equilibrium by the general public, and North (1990) surmised that voters might misperceive the relative merits of different policies and institutions and, hence, demand suboptimal ones. More recently, Beilharz and Gersbach (2004) argued that unawareness of general equilibrium effects may lead citizens to support minimum wages above market ones (see also Romer, 2003). Caplan (2007) documented systematic divergences in the views of voters relative to those of experts on the benefits of, for instance, markets and foreign trade. In political science, the field of political behaviour has long studied patterns of voter opinion with conflicting results on the quality of those opinions (see Bartels, 2012 for a review).

Several challenges make it difficult to prove that voters may demand bad policies because they underestimate equilibrium effects. First, actual government policy occurs in environments so complex that it is almost impossible to conclude that voters' expressed policy preferences rely upon mistaken assessments of equilibrium. We overcome this challenge through an experiment in which the welfare ranking of policies is certain, which allows us to identify mistaken votes. Secondly, documenting that voters err does not establish that they do so systematically, nor that they do so because they underappreciate equilibrium effects. We overcome this second challenge by developing a conceptual framework that isolates conditions for the underappreciation of equilibrium effects to result in a demand for bad policy, and by then investigating experimentally whether perceived equilibrium effects influence voting as predicted by the framework.

In our framework, a policy change takes the form of a change in the game played by citizens; this change preserves the action space but alters payoffs and therefore citizens' incentives to choose each action. The key aspect of the framework is to decompose the effects of the change in policy on a player's payoffs into two main parts: (1) the change in payoff attributable purely to the change in policy, holding actions constant; (2) the change in payoff attributable to players adjusting their actions. The first effect is the "direct" effect of the policy change, while the second is the "indirect" effect due to the change in equilibrium behaviour. To arrive at their policy preferences, voters must assess the sum of these two effects. Our main hypothesis is that voters will tend to focus on the direct effects of the policy change and underappreciate the indirect effects. As a result, voters will favour reforms with positive direct effects, even when undone by negative indirect effects, and reject reforms with negative direct effects, even when more than compensated by positive indirect effects.

To empirically show that the underappreciation of equilibrium effects can result in a demand for bad policy, we focus on the simplest setting in which meaningful equilibrium effects can arise. We consider players who must choose between two 2×2 complete information games, each featuring a unique equilibrium in strictly dominant strategies. Specifically, in our main treatments, subjects begin by playing a Prisoners' Dilemma, and after a few rounds vote on

1. Agency problems include discretion under limited electoral accountability (*e.g.* Barro, 1973; Ferejohn, 1986) and capture (*e.g.* Stigler, 1971; Peltzman, 1976; Coate and Morris, 1995). For accounts of why inept people may self-select into policymaking, see Caselli and Morelli (2004), Messner and Polborn (2004), Besley (2005), and Dal Bó *et al.* (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; Fernandez and Rodrik, 1991), and dynamic inefficiency due to the threat of losing political control (*e.g.* Alesina and Tabellini, 1990, De Figueiredo, 2002, Besley and Coate, 1998).

whether to levy a tax that would create a new game, which we call the Harmony Game. The new game has lower payoffs across all action profiles due to the tax (the direct effect is negative). But because the tax disproportionately lowers the payoffs of defection, the Nash equilibrium in the new game involves cooperation rather than defection, and yields higher equilibrium payoffs than the Prisoners' Dilemma (the indirect effects are positive). As a result, rational subjects who correctly predict behaviour should prefer the Harmony Game. However, subjects who expect behaviour in the Harmony Game not to differ too much from that in the Prisoners' Dilemma will prefer the Prisoners' Dilemma.

Our experiment yields three main findings. First, even though subjects cooperate and earn more in the Harmony Game—which is expected in equilibrium—a majority of them vote to play the Prisoners' Dilemma over the Harmony Game. This finding illustrates political failure in a precisely characterized situation, namely one where the indirect effects of equilibrium adjustment outweigh the direct effects, both in theory and in practice. This finding is robust to the voting institution, namely whether the game is chosen by the majority or a randomly selected dictator. It is also robust to which game subjects play first. That is, a majority of subjects who begin by playing the Prisoners' Dilemma rejects a policy change (moving to the Harmony Game) that has negative direct effects despite its larger positive indirect effects, just as a majority of subjects who begin by playing the Harmony Game supports a policy change (moving to the Prisoners' Dilemma) that has positive direct effects despite its larger negative indirect effects. Secondly, since our framework predicts that voter mistakes should be driven by a lack of appreciation of changes in equilibrium behaviour, in our experiment we elicit beliefs about behaviour to ascertain whether subjects do on average underappreciate the extent to which the change in policy will affect behaviour, and find that they do. Moreover, those subjects who most underappreciate equilibrium effects vote for the Prisoners' Dilemma. Thirdly, we run an additional treatment in which we randomly shock subjects' beliefs about how others play the two games and show that these beliefs affect their votes: the more subjects expect cooperation rates to increase by moving to the Harmony Game, the less they vote for the Prisoners' Dilemma. This shows that the underappreciation of equilibrium effects causes the voting mistakes.

While the main treatments present subjects with a choice between one game for which they had experience and another game for which they had none, we also study a treatment in which subjects acquire experience by voting repeatedly. We find that support for the Prisoners' Dilemma decreases as players become familiar with the new game, suggesting that experience can to some degree substitute for the ability to make equilibrium predictions.

From the perspective of any particular voter, policy changes create indirect effects for two conceptually different reasons: first, they influence other people's actions; secondly, they influence the voter's own action. It turns out that in the games which we examine, underappreciating how others react to policy changes is necessary and sufficient to cause people to vote for the Prisoners' Dilemma. Partly due to this feature of the environment, and partly due to a suspicion that people predict others' behaviour less accurately than they predict their own, we focus on demonstrating that people's beliefs about others' behaviour drive their voting. (In Section 6.4, we discuss how our findings connect to models of non-equilibrium reasoning, including those on level-k thinking (Nagel, 1995; Stahl and Wilson, 1994) and cognitive-hierarchy (Camerer *et al*., 2004).) To supplement our main findings, however, we explore the extent to which some subjects may also miss the fact that their own behaviour will adjust to the new policy, and estimate that nearly one-third of subjects make this mistake.

As with all empirical work, it is important to consider external validity. The fact that these experiments involve students from elite universities making choices in very simple environments suggests that, if anything, our results may understate the extent to which the average citizen underappreciates equilibrium effects in the more complex policy environment in the field.

Of course, outside of the laboratory, politicians or the media could mitigate voters' lack of understanding of equilibrium effects. But some politicians and media may instead exacerbate, or at least exploit, voters' biases. Indeed, most formal theories of electoral politics view politicians not as educators but instead as panderers to voters' policy positions, even wrong ones (Harrington, 1993; Canes-Wrone *et al*., 2001; 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). Ultimately, 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". For these reasons, it is important to understand any potential problems in voters' demand for policies. Our experiments demonstrate the existence of problems and help understand their causes.

Our experimental finding that people underappreciate the equilibrium effect of policies and that this affects their demand for policy suggests that, when considering the attractiveness of policies to voters, we must not look only at the total welfare effect of policies, but instead pay particular attention to their direct effects. The underappreciation of equilibrium effects may distort preferences across a broad class of problems. Voters may too frequently oppose the lifting of price controls, including for instance on housing rents, when they focus on the direct upward pressure on prices and neglect the benefits of expansion in supply. Similarly, they may excessively trust in the congestion-reducing effects of building or enlarging roads when failing to consider how enhanced roads would encourage more commute by car.[2] Voters may not adequately oppose fiscal deficits that get monetized by printing money when they neglect the effect that the expanded monetary base will have on prices. Lastly, activities that generate negative externalities typically call for Pigouvian taxation to help internalize externalities. Excessive carbon emissions is perhaps the single most serious, and still unresolved, policy problem of this form. According to our framework, policymakers could be reluctant to adopt Pigouvian taxes because such taxes do not appeal to voters: their costs are direct (they tend to cause higher prices) while their benefits are indirect (they reduce harmful activities).

The remainder of the article is organized as follows. In Section 2, we situate our work in the context of the literature. In Section 3, we introduce a framework for analysing people who underappreciate equilibrium effects. We describe our experiments in Section 4, and our hypotheses in Section 5. In Section 6, we report data from the experiments, explaining how they demonstrate that people underappreciate the extent to which others react to policy changes, and that this causes them to form the wrong policy preferences and cast the wrong vote. We conclude in Section 7.

2. RELATED LITERATURE

This article relates to the emerging political economy literature incorporating behavioural aspects. Examples include the study of the impact of cognitive dissonance on voting (Mullainathan and Washington, 2009), the analysis of collective action with time-inconsistent voters (Lizzeri and Yariv, forthcoming; Bisin *et al*., 2015), and the behaviour of voters who fail to extract the right information from other voters' actions (Eyster and Rabin, 2005; Esponda and Pouzo, 2017; Esponda and Vespa, 2014). Our work adds to this literature by formalizing and empirically demonstrating a different behavioural mechanism for voting anomalies. It is important to emphasize that our results do not imply that voters will always

2. Although possibly less familiar to economists than the other examples, Downs' (1962) "Law of Congestion" that traffic arrives in proportion to road capacity receives empirical support (Duranton and Turner, 2011).

make mistakes. In fact, there are studies where subjects with enough experience with all of their options make adequate choices, even in fairly involved environments (Agranov and Palfrey, 2015; Ertan *et al*., 2009; Kosfeld *et al*., 2009). Our contribution is to isolate a class of situations that pose a challenge to voters, namely cases in which voters lack familiarity with one of their options and must make equilibrium predictions, and where a tension arises between direct and indirect effects when ranking policy.

Our article also relates to a growing experimental literature studying the choice of self-regulatory institutions (see Dal Bó (2014) for a survey). A few findings in that literature are worth highlighting here. Walker *et al*. (2000) study common pool problems where players who would benefit from reduced extraction vote on extraction rules. They find that some voters propose inefficient extraction rules, and that the voting protocol affects efficiency. Margreiter *et al*. (2005) show that subject heterogeneity exacerbates inefficiency. Dal Bó (2014) offers evidence consistent with the idea that a better understanding of the strategic situation affects people's ability to select efficient institutions. Kallbekken *et al*. (2011) use laboratory experiments to study people's attitudes towards Pigouvian taxes. Not only does their work focus on different issues than ours (the effects of using the term "tax" versus "fees", of education, and of the distribution of tax revenue), but, more importantly, their experiments cannot shed light on whether the non-imposition of taxes derives from a failure to understand equilibrium effects: all five types of voters predicted to benefit from imposing the tax *in equilibrium* would also (weakly) benefit from imposing the taxes *fixing others' behaviour* (and strictly benefit in four of those five cases). Closer to this paper, Sausgruber and Tyran (2005, 2011) provide experimental evidence which suggests that people may not understand tax incidence: in a market where buyers bear all the burden of a tax, they prefer to impose a larger tax on sellers than a smaller tax on themselves. Because sellers' behaviour is mechanized in the experiments, which buyers understand, it is unclear whether buyers choose the wrong tax because they misunderstand how sellers react to taxes or due to some simple aversion to paying tax of the form studied by Kallbekken *et al*. (2011). Sausgruber and Tyran (2005) elicit subjects' beliefs on the effect of taxes on prices and find that a majority of subjects fail to understand the effect of taxes. However, they do not study how beliefs correlate with voting nor assess the existence of a causal effect. Finally, Dal Bó *et al*. (2010), in their study of the direct effect of democracy, find that 46% of subjects prefer to play a Prisoners' Dilemma game over a coordination game derived from the Prisoners' Dilemma by taxing unilateral defections. This is not evidence that subjects do not understand equilibrium effects as the coordination game has two pure-strategy equilibria, only one of which results in higher equilibrium payoffs than under the Prisoners' Dilemma. In this article, we focus on demonstrating the connection between the underappreciation of equilibrium effects and the demand for bad policy, both by grounding the experiment in a conceptual framework and by choosing a design that can eliminate potential confounds challenging the identification of the effects of interest. One advantage of our design is it minimizes computational complications by presenting subjects with the simplest possible equilibrium change, involving 2×2 games with dominant strategies. In addition, as explained after our main results, we rely on a variety of treatments to eliminate alternative explanations.

At an abstract level, our article relates to the experimental literature documenting failures of backward induction (*e.g*. McKelvey and Palfrey, 1992; Bone *et al*., 2009; Palacios-Huerta and Volij, 2009; Levitt *et al*., 2011; Moinas and Pouget, 2013). We establish a systematic direction of departure 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 think their opponents' play is constant across certain decision nodes. Despite the resemblance, Section 6.7 explains how subjects in our

TABLE 1
The games

Prisoners' Dilemma

	C	D
C	$b-c, b-c$	$-c, b$
D	$b, -c$	$0, 0$

Harmony Game

	C	D
C	$b-c-t_C, b-c-t_C$	$-c-t_C, b-t_D$
D	$b-t_D, -c-t_C$	$-t_D, -t_D$

experiment exhibit a degree of partial sophistication about the relationship between the game and their opponents' actions inconsistent with ABEE.

3. UNDERAPPRECIATION OF EQUILIBRIUM EFFECTS AND PREFERENCES OVER POLICIES: CONCEPTUAL FRAMEWORK

In this section, we present a simple conceptual framework to define the meaning and types of "underappreciation of equilibrium effects" and discuss how they affect voters' preferences over policies.

Consider an agent who will participate in a two-player Prisoners' Dilemma game—see left panel in Table 1. Participants in this game must choose between cooperate (C) and defect (D). Cooperation results in a cost to the agent of c and a benefit b to the other participant, with $b > c > 0$.[3] Given that $c > 0$, it is a dominant strategy to defect, and the Nash equilibrium of this game is that both participants defect leading to (D, D) with payoffs $(0, 0)$; since (C, C) gives both players $b - c > 0$, the equilibrium outcome (D, D) is inefficient.

Now consider a policy proposal that would impose, on each player, taxes t_C on cooperation and t_D on defection. These taxes would transform the game into the one in the right panel in Table 1. Assume that these taxes satisfy the following condition: $b > t_D > t_C + c$. Given the values of the taxes, to cooperate becomes the dominant strategy, leading to an efficient equilibrium: (C, C) with payoffs $(b-c-t_C, b-c-t_C)$. Since individual incentives do not conflict with group objectives in this game, we call it the "Harmony Game".

An agent who anticipates equilibrium behaviour in both games will prefer to impose the taxes and play the Harmony Game, given that the equilibrium payoff in the Harmony Game exceeds the equilibrium payoff in the Prisoners' Dilemma game: $b - c - t_C > 0$. Therefore, if all agents think this way, they will vote for the imposition of taxes, resolving the social dilemma.

However, voters may not correctly appreciate how changes in policy will affect behaviour and, as a result, may not demand the right policies: they may prefer the Prisoners' Dilemma to the Harmony Game. In this section, we offer a taxonomy of the types of mistakes that voters may make when thinking about the effect of policies, and show how these mistakes can affect voters' demand for policy.

Assume that instead of holding equilibrium beliefs about behaviour, voters may hold any belief about their own behaviour and the behaviour of others in the two games. More precisely, assume that a voter believes that she will cooperate with probability α in the Prisoners' Dilemma and probability α' in the Harmony Game while believing that the other player will cooperate with probability β in the Prisoners' Dilemma and probability β' in the Harmony Game. Note that we

3. This specification of a Prisoners' Dilemma game has the property that a player's gain from defection does not depend upon the action of the other player. As discussed in Fudenberg *et al*. (2012), not every Prisoners' Dilemma game can be described this way. Despite the lack of generality, we follow this description as it allows for a simpler analysis and it includes the Prisoners' Dilemma game used in the experiment.

do not attempt to explain here the origin of these beliefs. The goal is to understand how these beliefs affect preferences for the two games. Given her beliefs, and under the assumption that the voter is risk neutral and cares only about her material payoffs, the voter's preferences over the two games will depend on the difference in expected payoff between the two games. The expected gain from moving to the Harmony Game is $G(\alpha,\alpha',\beta,\beta')=EU(HG|\alpha',\beta')-EU(PD|\alpha,\beta)$, where $EU(HG|\alpha',\beta')$ is the expected payoff under the Harmony Game and $EU(PD|\alpha,\beta)$ is the expected payoff under the Prisoners' Dilemma game. The voter would prefer the Prisoners' Dilemma if her beliefs are such that the expected gain of imposing taxes is negative.

The expected gain can be decomposed into three terms. The first term is the direct effect of the game change: $DE=EU(HG|\alpha,\beta)-EU(PD|\alpha,\beta)$. The direct effect DE captures the change in expected payoff from going from the Prisoners' Dilemma to the Harmony Game assuming that the behaviour of both players is not affected by the game change. The second term is the indirect effect due to the adjustment in behaviour by self once in the Harmony Game: $IS=EU(HG|\alpha',\beta)-EU(HG|\alpha,\beta)$. This indirect effect expresses the change in expected payoffs due to the adjustment in one's own behaviour while leaving the behaviour of the other player unchanged. The third effect is the indirect effect due to the adjustment in behaviour by the other player once in the Harmony Game: $IO=EU(HG|\alpha',\beta')-EU(HG|\alpha',\beta)$. This indirect effect describes the change in expected payoffs due to the change in the behaviour of others while leaving the behaviour of the self constant at the new level (α'). It is easy to check that these three effects add up to the total expected gain from a change in game: $G=DE+IS+IO$.[4]

For our two games, this decomposition can be simply expressed in terms of payoff parameters, taxes, and beliefs. The direct effect is the expected payment of taxes under the belief that behaviour will be as in the Prisoners' Dilemma: $DE=-(\alpha t_C+(1-\alpha)t_D)$. This effect is negative, and its absolute value is decreasing in α. The indirect effect from the adjustment by self (IS) equals the change in the probability of cooperation by self times the amount saved by cooperating (the tax to defection minus the cost of cooperation and the tax to cooperation): $IS=(\alpha'-\alpha)(t_D-c-t_C)$. This effect is increasing in the believed change in cooperation by self ($\alpha'-\alpha$). Finally, the indirect effect from the adjustment by other (IO) equals the change in the probability of cooperation by other ($\beta'-\beta$) times the benefit from the other's cooperation (b): $IO=(\beta'-\beta)b$.

As an important benchmark, we can easily calculate the value of these effects if the player predicts Nash equilibria in both games: $\alpha=\beta=0$ and $\alpha'=\beta'=1$: $DE^{NE}=-t_D$, $IS^{NE}=t_D-c-t_C$, and $IO^{NE}=b$. The total gain in equilibrium is $G^{NE}=b-c-t_C$ which is greater than zero given the assumptions on payoffs and taxes, so a voter predicting equilibrium behaviour would prefer the Harmony Game.

We can compare these equilibrium effects on payoffs with those perceived by a voter who does not predict equilibrium behaviour. For such a voter, the underappreciation of the indirect effect on payoffs due to the adjustment by self is proportional to the underappreciation of how much self will adjust behaviour in equilibrium: $IS=(\alpha'-\alpha)IS^{NE}\leq IS^{NE}$. Similarly, a voter who does not hold equilibrium beliefs underappreciates the indirect effect on payoffs due to the adjustment by others in proportion to the underappreciation of how much the other player will adjust behaviour in equilibrium: $IO=(\beta'-\beta)IO^{NE}\leq IO^{NE}$.[5]

4. An alternative decomposition would consider first a change in the behaviour of others and then a change in own behaviour. As will be clear later, such a decomposition is equivalent to the one defined above for the class of games considered in this section.

5. To be clear, a person who mistakes the sign of equilibrium effects (*e.g.* $\alpha=\beta=1$ and $\alpha'=\beta'=0$) is also said to underappreciate equilibrium effects.

The following proposition establishes a particular relationship between the underappreciation of equilibrium effects and a preference for the bad policy, namely for the Prisoners' Dilemma game.

Proposition 1. *A voter has a preference for the Prisoners' Dilemma over the Harmony Game if and only if she sufficiently underappreciates the indirect effect due to the adjustment in behaviour by the other player (IO) relative to Nash equilibrium predictions.*

Proof. A voter has a preference for the Prisoners' Dilemma over the Harmony Game if and only if her perceived gains from moving to the Harmony Game are negative:

$$G=D+IS+IO=-(\alpha t_C+(1-\alpha)t_D)+(\alpha'-\alpha)(t_D-c-t_C)+(\beta'-\beta)b<0.$$

This condition holds if and only if $\beta'-\beta<\alpha\frac{t_C}{b}+(1-\alpha)\frac{t_D}{b}-(\alpha'-\alpha)\frac{t_D-c-t_C}{b}$. Given that $b>t_D>t_C+c$, the right-hand side of the previous inequality attains a maximum of $\frac{t_D}{b}<1$ with $\alpha=\alpha'=0$ and a minimum of $\frac{t_C}{b}<1$ with $\alpha=\alpha'=1$. It follows that if the voter has expectations about others driven by Nash equilibrium predictions (*i.e.* $\beta'-\beta=1$) she can never have a preference of the Prisoners' Dilemma. It also follows that a voter with a preference for the Prisoners' Dilemma must have $\beta'-\beta<\frac{t_D}{b}$ and that a voter with $\beta'-\beta<\frac{t_C}{b}$ must have a preference for the Prisoners' Dilemma. Therefore, a voter has a preference for the Prisoners' Dilemma if and only if $\beta'-\beta$ is sufficiently small relative to its value of 1 under Nash equilibrium predictions (how small depending on the direction of the implication); As $IO=(\beta'-\beta)b$ the proposition follows. ‖

This proposition is developed for the case when the voter contemplates a move from the Prisoners' Dilemma to the Harmony Game, but it holds also for the reverse move from the Harmony Game to the Prisoners' Dilemma. In sum, the voter will have a preference for the Prisoners' Dilemma game if and only if she sufficiently underappreciates how the change in game will affect the behaviour of the other player relative to equilibrium. Of course, for voting for the Prisoners' Dilemma to be a mistake, actual behaviour in the two games must resemble equilibrium behaviour. In the next section, we explain how we bring this environment to the laboratory to study whether actual behaviour matches equilibrium behaviour, whether subjects underappreciate the indirect effect due to the adjustment of others, and whether this affects subjects' preferences over games.

4. THE EXPERIMENT

The experiment brings to the laboratory the environment studied in the previous section with a particular choice of parameters: $b=6$, $c=2$, $t_C=1$, and $t_D=4$ over a baseline payoff of 5.[6] This combination of parameters results in the payoff matrices in Table 2. The actions C and D were respectively labelled "1" and "2" in the experiment to ensure a neutral presentation. The exchange rate was \$1 per three experimental points.

Our main experiment involved six treatments. We begin by explaining here the basic structure of the experimental sessions in all six treatments of the main experiment, and then describe the differences across the treatments. In Section 6.4, we describe an additional experiment.

In Part 1 of the main experiment, we divided subjects in each session into groups of six. Each subject played against every other one in the group exactly once, resulting in five periods of

6. There are two reasons why we consider taxes on both actions. One is realism: Pigouvian taxes under the socially optimal action are not necessarily zero in a general set-up with externalities. For example, that would be the case if CO_2 emissions were taxed and the socially optimal amount of emissions were positive. By restricting attention to the socially

TABLE 2
The Prisoners' Dilemma and Harmony Games

Prisoners' Dilemma (PD)

	C	D
C	9,9	3,11
D	11,3	5,5

Harmony Game (HG)

	C	D
C	8,8	2,7
D	7,2	1,1

(one-shot) play in this part of the experiment. This was done to minimize reputational concerns arising from repeated interaction, which could lead to cooperation in the Prisoners' Dilemma without the need for taxation. The game played in Part 1 varied by treatment. In some treatments, all groups played the Prisoners' Dilemma in Table 2; in other treatments, all groups played the Harmony Game in Table 2. All groups in a session belonged to the same treatment.

After Part 1, new groups of six were formed randomly for Part 2, which included another five periods of play (6–10). 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 subjects learned that choice, subjects reported their beliefs about how a randomly selected opponent in a similar experiment would act in each of the two games.[7] The belief elicitation occurred after voting, so as not to affect voting, and before subjects learned about the voting outcome, so as not to have the outcome affect reported beliefs. Subjects were informed of the implemented game and not the voting distribution. As in periods 1–5 in Part 1, in periods 6–10 every subject faced each other subject in the group exactly once. Subjects were paid for their earnings in all ten periods in Parts 1 and 2.

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 treatment arms labelled Control, Random Dictator, Majority and Majority Once had the subjects play the Prisoners' Dilemma game in Part 1, while Reverse Control and Reverse Random Dictator had the subjects play the Harmony Game in Part 1. We included treatments with different games in Part 1 to ensure that any demand for the Prisoners' Dilemma did not derive from status quo bias. The instructions for the conditions in which the Prisoners' Dilemma was played first presented the Harmony Game as derived from the Prisoners' Dilemma by applying a tax of 1 experimental point on action 1 (cooperation) and of 4 points on action 2 (defection). By contrast, the treatments in which the Harmony Game was played first did not mention a tax, but rather a subsidy. This difference should mitigate any concern that visceral reactions against the notion of taxation drive voting for the Prisoners' Dilemma. We employed the tax-subsidy terminology because it conveys clearly the relation between the games and brings the exercise closer to the way in which policy debates are characterized. Table 3 summarizes the experimental design.

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,

optimal and individually optimal actions, the Prisoners' Dilemma simplifies a more general environment with a large number of actions, where typically most (if not all) of the actions will be taxed to varying degrees. Secondly, we wanted to eliminate any unnecessary differences between the two actions that would arise from taxing only *D*.

7. Beliefs were elicited through an incentivized mechanism that is robust to variation in risk attitudes, as in Karni (2009). See also Grether (1992), Holt (2007, pp. 384–385) and Möbius *et al*. (2014). Tying incentives to the behaviour of subjects in a different session may create uncertainty in subjects' minds about whether all details remain constant across experiments. Nevertheless, we wanted to avoid subjects entertaining the notion that their own behaviour in Part 2 of the experiment could affect their belief-elicitation payout. Details of the elicitation screens are available in the Online Appendix.

TABLE 3
Experimental design—treatments

	Part 1	Part 2		
Treatment	Game	Game	Game choice institution	Game choice before
Control	PD	PD or HG	Random	Period 6
Reverse Control	HG	PD or HG	Random	Period 6
Random Dictator	PD	PD or HG	Random Dictator	Period 6
Reverse Random Dictator	HG	PD or HG	Random Dictator	Period 6
Majority Once	PD	PD or HG	Simple Majority	Period 6
Majority Repeated	PD	PD or HG	Simple Majority	Periods 6–10

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). These treatments allow us to incentivize the belief elicitation in the other treatments. Together with the other treatments, the control treatments also allow us to corroborate the theoretical prediction that subjects would be better off in the Harmony Game than in the Prisoners' Dilemma game, regardless of whether the change in game happens randomly or by choice.

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 the choice of a randomly selected subject for the group. In the Majority Once 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 equal probability. 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 behaviour in future periods.

Both random dictator and majority institutions were considered to make sure that the choice of subjects was robust to the voting institution. The Majority Repeated treatment was included to study the evolution of game choices by subjects.

During play, subjects were shown the payoff function corresponding to the game they were playing. This information was displayed as a table with a row for each possible outcome of the game, as shown in the Online Appendix. Subjects knew the game was symmetric, so this representation carried the same information as the normal-form representation shown in Table 2.[8] At the time of voting, subjects were shown both tables side by side to facilitate comparison between the games. Moreover, since they faced no time limit to vote, participants had ample time to think about the two different games.

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).

The experiment was programmed and run using z-Tree (Fischbacher, 2007). We recruited 384 student subjects from UC Berkeley and 384 from Brown University to participate in 43 sessions of the experiment. Table 4 shows the number of subjects and sessions from each university in each of the six treatments. Sessions lasted around half an hour and earnings ranged from $16.67 to $37 with an average of $27.86 (earnings included a $5 show-up fee). Appendix Table 11 displays summary statistics of demographics and beliefs.

8. We find no evidence that this representation affected behaviour as the levels and evolution of cooperation was consistent with those found in the literature.

TABLE 4
Number of subjects and sessions by treatment and place

	Subjects			Sessions		
	Berkeley	Brown	Total	Berkeley	Brown	Total
Control	60	60	120	3	4	7
Reverse Control	60	60	120	3	4	7
Random Dictator	84	84	168	3	5	8
Reverse Random Dictator	60	60	120	4	4	8
Majority Once	60	60	120	3	4	7
Majority Repeated	60	60	120	3	3	6
Total	384	384	768	19	24	43

5. HYPOTHESES

With the particular values of the parameters used in the experiment, we can now revisit the conceptual framework so as to provide the precise hypotheses that we test with the experimental data.

A voter who expects equilibrium behaviour in both games will expect a gain of $G^{NE}=8-5=3$ from moving to the Harmony Game. This gain divides into the three effects discussed in Section 3. The direct effect in equilibrium is just the reduction in payoff due to the move from PD to HG leaving the outcome (D,D) constant: $DE^{NE}=-4$. The indirect effect due to the adjustment by self is the increase in payoff due to the move from (D,D) to (C,D) in HG: $IS^{NE}=2-1=1$. The indirect effect due to the adjustment by other is the increase in payoff due to the move from (C,D) to (C,C) in HG: $IO^{NE}=8-2=6$.[9]

However, voters may have beliefs about their own behaviour and the behaviour of others in the two games that disagree with equilibrium beliefs. As in Section 3, assume that a voter believes that she will cooperate with probability α in the Prisoners' Dilemma and probability α' in the Harmony Game, while believing that the other player will cooperate with probability β in the Prisoners' Dilemma and probability β' in the Harmony Game. Given these beliefs, the advantage of moving from the Prisoners' Dilemma to the Harmony Game is:

$$G=EU(HG|\alpha',\beta')-EU(PD|\alpha,\beta)=-4+2\alpha+\alpha'+6(\beta'-\beta). \tag{1}$$

Proposition 1 establishes that a voter has a preference for the Prisoners' Dilemma if and only if she sufficiently underappreciates the indirect effect associated with the adjustment of others. This means that $\beta'-\beta$ must be sufficiently smaller than the Nash equilibrium value of 1, as $IO=6(\beta'-\beta)$ in this case. We can calculate how small $\beta'-\beta$ must be by finding the condition on $\beta'-\beta$ such that $G<0$. This condition is:

$$\beta'-\beta<\frac{4-\alpha'-2\alpha}{6}.$$

Note that the right-hand side of this inequality defines a threshold that can reach values between $\frac{1}{6}$ and $\frac{2}{3}$, depending on the values of α and α'. Voters who estimate a difference in

9. Similarly, a voter in a "reverse" treatment who expects equilibrium behaviour in both games will expect a (negative) gain from moving from HG to PD equal to $G^{NE}=5-8=-3$. This total gain can again be decomposed in the three effects: $DE^{NE}=9-8=1$, $IS^{NE}=11-9=2$, and $IO^{NE}=5-11=-6$. Note that the absolute value of the indirect effect due to the adjustment by others is the same regardless of which game is played first.

others' cooperation rates across games below the threshold (*i.e.* those with lower estimates of the indirect effect due to the adjustment of others) should prefer the Prisoners' Dilemma to the Harmony Game. Since the maximum value of this threshold lies below one, a player with a preference for the Prisoners' Dilemma must have beliefs such that $\beta' - \beta < 1$. In other words, a voter with a preference for the Prisoners' Dilemma must underappreciate the adjustment by others. Consider for example a person who knows that she will herself always play the dominant strategy. This person's parameters are $(\alpha = 0, \alpha' = 1, \beta, \beta')$ and will prefer the Prisoners' Dilemma if $\beta' - \beta < \frac{1}{2}$. That is, she will prefer the Prisoners' Dilemma if she expects cooperation in the Harmony Game to be at most 50 percentage points higher than in the Prisoners' Dilemma.[10]

The experiment was designed to test whether subjects underappreciate the response of others to a change in game leading them to form the wrong preferences over games. For a preference for the Prisoners' Dilemma to be wrong, it must be that the Harmony Game actually results in higher average payoffs than the Prisoners' Dilemma. In other words, it is necessary that the actual behaviour in the two games resemble equilibrium behaviour sufficienly well that data and theory rank payoffs across the two games consistently. We expect this to hold, leading to the following hypothesis.

Hypothesis 1. *The Harmony Game results in higher average payoffs than the Prisoners' Dilemma.*

We expect that even when the Harmony Game results in higher average payoffs than the Prisoners' Dilemma, a majority of subjects may underappreciate equilibrium effects leading to the following hypothesis.

Hypothesis 2. *A majority of subjects prefers the Prisoners' Dilemma to the Harmony Game.*

As discussed before, the preference for the Prisoners' Dilemma can only arise from an underappreciation of the indirect effect due to the adjustment by others. This leads to the next two hypotheses.

Hypothesis 3. *A majority of subjects underappreciates the indirect effects associated with the adjustment of behaviour by others. The average belief differential about cooperation rates $\beta' - \beta$ is smaller than the equilibrium prediction $\beta' - \beta = 1$, and smaller than the empirical difference in cooperation rates between games.*

Hypothesis 4. *Subjects who appreciate less the indirect effect due to the adjustment of others are more likely to support the Prisoners' Dilemma over the Harmony Game.*

The core of our investigation concerns Hypotheses 1 to 4. We present next two secondary hypotheses. To motivate the first, note that subjects who vote for the Prisoners' Dilemma because

10. This statement would not change much by introducing plausible risk aversion. For example, a subject with a quadratic utility function with no other income and who believes there is 10% cooperation in the Prisoners' Dilemma would prefer the Harmony Game if $\beta' - \beta$ is above $\sim$0.56. If this subject instead has a baseline income of as little as \$10, then the critical value is below 0.51, and it becomes 0.5007 when baseline income is \$100. Other utility functions yield a similar picture even for subjects who are arbitrarily risk averse. Consider a subject with CRRA utility function $u(x) = \frac{x^\rho}{\rho}$ and who believes there is 10% cooperation in the Prisoners' Dilemma. The critical value of $\beta' - \beta$ converges to 0.61 when the subject becomes arbitrarily risk averse ($\rho \to 0$) and he has no income outside of the experiment. For baseline incomes of \$10 and \$100, the limit critical values become as low as 0.514 and 0.501, respectively.

they do not expect the behaviour 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 poor strategic reasoning. We obtained one measure of strategic sophistication in simultaneous move games by having subjects play a *p*-beauty contest. We then hold the following:

Hypothesis 5. *Subjects who vote for the Prisoners' Dilemma are measured to be less sophisticated in the p-beauty contest.*

To motivate our other secondary hypothesis, note that some subjects may miss not only the fact that others' behaviour depends upon the game, but also that their own behaviour does too. As discussed above, this is neither necessary nor sufficient to cause a preference for PD in our setting, and hence not central to our main argument. However, it highlights that subjects may display very basic departures from equilibrium thinking. Our last hypothesis then is that:

Hypothesis 6. *Some subjects underappreciate the indirect effect through the adjustment by self.*

6. RESULTS

6.1. *Benchmark: the Harmony Game leads to higher payoffs than the Prisoners' Dilemma*

Do subjects play close enough to the Nash outcome in each game that cooperation and payoffs in the Harmony Game exceed those in the Prisoners' Dilemma? The answer is yes, supporting Hypothesis 1. Consider first the two treatments that assigned games for Part 2 exogenously. The game played in Part 2 strongly affects behaviour and payoffs. On average, across all Part 2 periods, in the Control treatment there is 92% cooperation when playing the Harmony game, versus only 16% cooperation when playing the Prisoners' Dilemma. In the Reverse Control treatment, the respective cooperation rates are 93% in the Harmony Game versus 30% in the Prisoners' Dilemma. Although below the 100 percentage points predicted by Nash equilibrium, both cooperation differentials (76% and 63%) are significantly different from zero, and are also significantly larger than the 50 percentage points needed for a rational player to prefer the Harmony Game to the Prisoners' Dilemma. The higher cooperation under the Harmony Game leads to higher earnings than in the Prisoners' Dilemma–this difference is significant at the 1% level in both Control and Reverse Control.[11] Figure 1 shows the period-by-period evolution of cooperation and payoffs as a function of the game played in Part 2 in the control treatments. One noteworthy pattern is that the change of game affects behaviour immediately, already by period 6.

While the Control and Reverse Control treatments help to determine the effects of exogenous game assignment, one may wonder whether these effects extend to the treatments in which subjects themselves choose which game to play in Part 2. The answer is again a strong yes. If we take all voting treatments together, instead of the control treatments, in Part 2 there is 96% of cooperation in the Harmony Game versus 23% under the Prisoners' Dilemma. Consequently, subjects playing

11. The *p*-values for all comparisons reported in this section are obtained from Wald tests. For these tests we run ordinary least squares (OLS) regressions of cooperation or payoffs on dummy variables for each game. We adopt the most conservative clustering of standard errors, at the session level for overall Part 2 outcomes and period 6 payoffs, and at the level of group assignment in Part 1 for outcomes in Part 1 and behaviour in period 6. The significant results on cooperation and payoff comparisons when pooling treatments are robust to using matching pairs sign-rank tests at the session level. This applies to cooperation and payoff comparisons in Part 2 for the Control and Reverse Control treatments, but these tests are ill-suited for analysing treatments separately because voting may result in no variation within session in terms of what game subjects end up playing, causing much data and power loss.

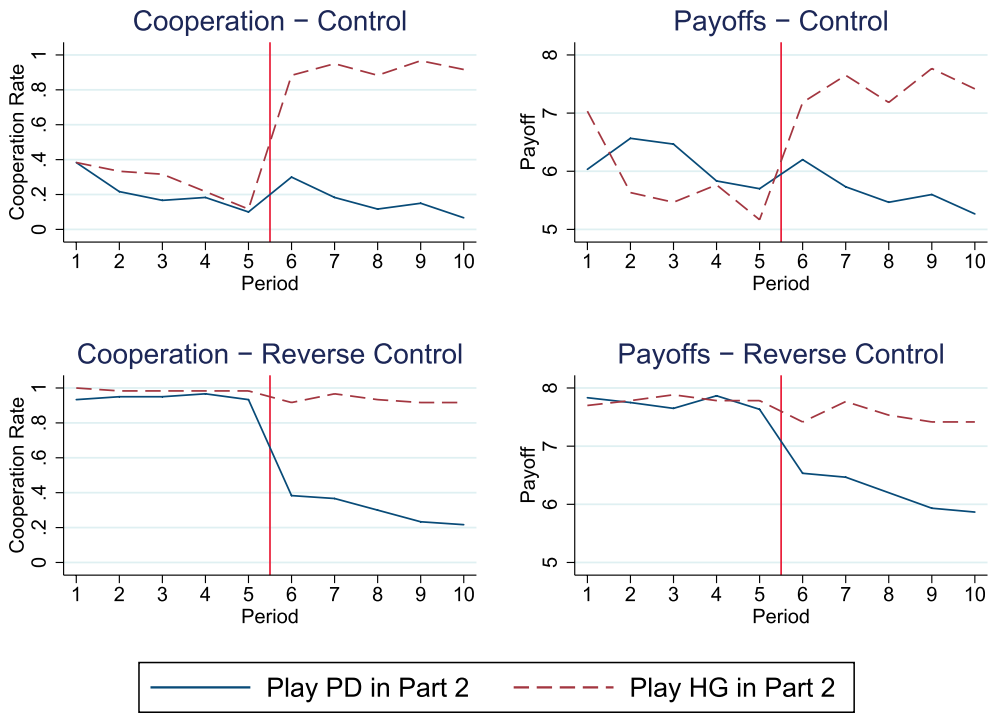

FIGURE 1
Comparing Prisoners' Dilemma and Harmony Game

the Harmony Game obtain significantly higher payoffs than those playing the Prisoners' Dilemma (7.74 versus 5.90 points, a 30% increase, which amounts to $3.07 in Part 2 of the experiment). The differences in cooperation rates and payoffs in Part 2 between the two games are statistically significant at the 1% level and are robust to considering each treatment separately. Appendix Figure 5 displays the evolution of cooperation and payoffs for all treatments. Appendix Figures 6 and 7 display the evolution of cooperation for each separate treatment. Appendix Table 12 offers the associated quantitative information, namely the cooperation rates and payoffs depending on the game that is played in Part 2 for each treatment.

Another way to see that behaviour across games differs in the direction predicted by theory is to compare the cooperation rates across the two games in period 5, once the players have already gained experience. Pooling across all treatments, we find that the cooperation rate in the Prisoners' Dilemma is 15.5% while that in the Harmony Game is 95% ($p < 0.0001$). The corresponding average payoffs are 5.62 and 7.65, respectively ($p < 0.0001$). Again, the Harmony Game leads to higher payoffs.

In conclusion, behaviour and payoffs across the two games vary enough in the direction 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.[12] Having established that the games'

12. This ranking of games is unlikely to be affected by social preferences. For example, subjects with inequity aversion as in Fehr and Schmidt (1999) will have a stronger taste for the Harmony Game, which produces less inequality in practice than the Prisoners' Dilemma. See the Online Appendix for substantiation of this claim.

TABLE 5
Prisoners' Dilemma vote shares by treatment at beginning of part 2

Treatment	Vote PD (%)
Random Dictator	52.98
Reverse Random Dictator	50.00
Majority Once	60.83
Majority Repeated	50.83
Total	53.60

payoffs rank empirically as they do theoretically, we turn our attention to whether subjects choose games accordingly.

6.2. *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 Prisoners' Dilemma game at the beginning of Part 2, supporting Hypothesis 2—see Table 5. The lowest share of subjects voting for the Prisoners' Dilemma game is 50.00% under Reverse Random Dictator, while the largest is 60.83% under Majority Once. All of these shares differ significantly from the 0% that would be expected if subjects chose games according to theory.

This is an important result of the article—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. The tendency of subjects 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 article. We also study later the evolution of votes as subjects gain experience in Majority Repeated.

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. Responsibility for the implementation of bad policies falls entirely on the subjects, the citizens of this environment. But, what explains why a majority of subjects demanded bad policy?

6.3. *Mechanism: failure to appreciate equilibrium effects*

The conceptual framework presented in Section 3 showed that subjects in the environment studied in the experiment can only have a preference for the Prisoners' Dilemma if they underappreciate the indirect effect due to the adjustment in the behaviour of others. This led to the hypotheses that a majority of subjects will underappreciate the indirect effect due to the adjustment of others (Hypothesis 3) and that those who underappreciate this effect more will be more likely to vote for the Prisoners' Dilemma (Hypothesis 4). As we describe next, we find evidence that strongly supports both hypotheses.

First, we find that, on average, subjects grossly underestimate the effect of the game change on the behaviour of others, consistent with Hypothesis 3. As Figure 2 shows, the distribution of the difference in the beliefs of cooperation between the Harmony and the Prisoners' Dilemma games ($\beta' - \beta$) is far from both the observed difference in behaviour and the equilibrium one. The equilibrium difference is 100 percentage points, and the observed difference is marked by a

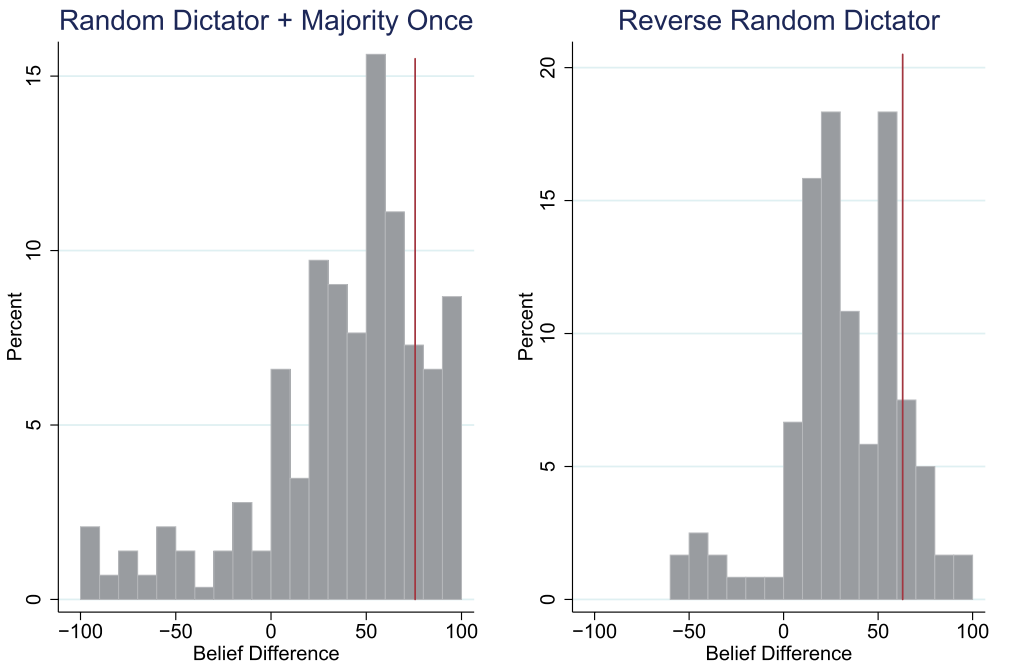

FIGURE 2

Distribution of difference in beliefs of cooperation between games (Harmony Game Minus Prisoners' Dilemma)

Notes: Vertical lines denote observed cooperation difference in Control (left) and Reverse Control (right).

vertical line in Figure 2. The average difference in belief of cooperation between the Harmony and the Prisoners' Dilemma games is 35 percentage points in the pooled Random Dictator and Majority Once treatments (with a median of 45) while in reality cooperation is 76 percentage points higher in the Harmony Game (as measured in the Control treatment, in which games are randomly assigned).[13] Similarly, the average belief difference is 30 percentage points in Reverse Random Dictator (with a median of also 30) while the true difference in behaviour is 63 percentage points (as measured in the Reverse Control treatment, in which games are randomly assigned). On average, subjects predict an effect of the game change half the size of its true effect.[14]

Secondly, consistent with Hypothesis 4, subjects who more strongly underestimate the effect of the game on behaviour are more likely to vote for the Prisoners' Dilemma. Figure 3 shows the average elicited belief of cooperation in each game broken down by vote of the subject and treatment. The triangles show the belief of cooperation in each game held by subjects who voted for the Prisoners' Dilemma, while the squares show the beliefs of subjects who voted for the Harmony Game. The dots represent the observed cooperation rates in each game. The solid line connects across games the beliefs of subjects who voted for the Prisoners' Dilemma while the broken line connects across games the beliefs of subjects who voted for the Harmony Game. In

13. The differences in cooperation by the very subjects who vote are not far from those in the Control treatment, as can be seen in Table 12 in the Appendix.

14. The underappreciation of equilibrium effects is driven both by an overestimation of cooperation in the Prisoners' Dilemma and an underestimation of cooperation in the Harmony Game in all treatments. This suggests that subjects have difficulty making equilibrium predictions even for games with which they have some experience. We return to the issue of experience when discussing learning in Section 6.8.

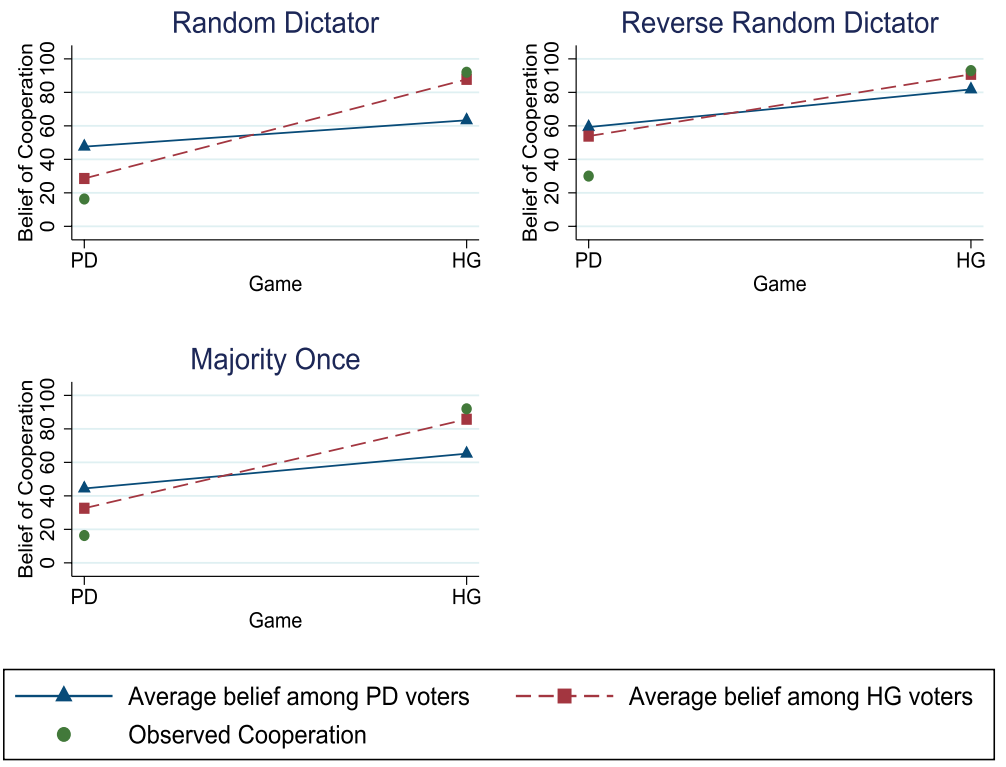

FIGURE 3
Belief of cooperation and voting

all three treatments in which beliefs were elicited, subjects who voted for the Prisoners' Dilemma expressed a lower belief that the behaviour of others will differ across games. That is, subjects who voted for the Prisoners' Dilemma have a lower estimate of the effect of the game change on behaviour.

The relationship between the difference in the beliefs of cooperation and voting is highly statistically significant—see Table 6. The "Belief Difference" variable denotes the difference in the belief of cooperation of other subjects under the Harmony Game relative to the Prisoners' Dilemma (*i.e.* $\beta' - \beta$) which displays a significant correlation with voting.

Table 6—column 2—shows that the relationship between voting for the Prisoners' Dilemma and the belief difference is robust to controlling for personal characteristics.[15] Most of these personal characteristics do not have a significant association with voting, with the exception of ideology.[16]

15. 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; secondly, SAT scores do not significantly predict voting, and excluding them does not change our results.

16. Since personal characteristics could affect voting via beliefs, in Appendix Table 13 we examine the association between personal characteristics and voting after dropping the beliefs variable. We find no robust relationship between personal characteristics and voting.

TABLE 6
Beliefs and Voting for Prisoners' Dilemma (dependent variable: vote for PD)

	Random Dictator, Reverse RD, Majority Once	
	(1)	(2)
Belief Difference	−0.005***	−0.005***
	(0.000)	(0.000)
Male		−0.085*
		(0.050)
Year		0.020
		(0.020)
Ideology		0.036***
		(0.011)
Economics		0.032
		(0.062)
Political Science		0.047
		(0.101)
Brown University		0.074
		(0.052)
Beauty Number		−0.000
		(0.001)
Constant	0.698***	0.527***
	(0.028)	(0.082)
Observations	408	408
Clusters	68	68
R^2	0.143	0.175

Notes: OLS specification. Belief Difference denotes the difference in beliefs of cooperation percentage under HG and PD. Year denotes year in college. Ideology from 0 (most liberal) to 10 (most conservative). Economics and Political Science denote subjects' major. Robust standard errors clustered by Part 1 group: *** significant at 1%, * at 10%, respectively.

6.4. *Underappreciation of equilibrium effects causes voting for the Prisoners' Dilemma*

The correlation between beliefs and voting documented in Table 6, while predicted by our framework, does not imply that subjects' beliefs about how others will play the two games exert a causal effect on how they vote. Indeed, an endogeneity problem may arise if people with different beliefs also differ in unobservable dimensions that directly affect voting, or if people introspect further about the games at the time of reporting their beliefs and report beliefs to justify their past voting choice.[17]

Establishing a causal link between subjects' beliefs and their votes requires a source of exogenous variation in subjects' beliefs. To create such a source, we conducted an additional experiment very much analogous to the Random Dictator treatment presented before, except that in the new experiment subjects received information about how past subjects played the two different games in prior treatments. Providing different subjects with different (but truthful) information generated random variation in their beliefs about how others played the two games. Of course, information about how past subjects played might alter subjects' beliefs about their own future behaviour. To neutralize any such effect, we constrained subjects to defect in the Prisoners' Dilemma and cooperate in the Harmony Game.

17. Costa-Gomes and Weizsäcker (2008) show evidence compatible with the idea that subjects re-examine strategic situations during the belief-elicitation stage.

TABLE 7
Beliefs and voting for Prisoners' Dilemma in the additional experiment (dependent variable: vote for PD)

	(1)	(2)
Belief Difference	−0.003***	−0.002**
	(0.001)	(0.001)
Personal characteristics	N	Y
Constant	0.660***	0.618***
	(0.044)	(0.114)
Observations	192	192
Clusters	32	32
R^2	0.037	0.049

Notes: OLS specification. Robust standard errors clustered by Part 1 group: *** significant at 1%, ** at 5%.

To ensure that subjects perceived their information about others' past behaviour as relevant, we changed Part 2 of the experiment. Rather than interact with other human subjects in Part 2, each subject interacted with a computerized counterpart that chose between cooperate and defect according to the actual choice rates of some groups in one of the treatments of the main experiment.[18] The main treatment of this new experiment was to show different subjects different data about past subjects' behaviour, and thereby induce different beliefs about the likely actions of their computerized counterpart. These data comprised the behaviour of four groups of actual subjects in the Control treatment, two of which had played the Prisoners' Dilemma and two of which had played the Harmony Game in Part 2 of the experiment. Denoting these groups PD-1, PD-2, HG-1, and HG-2, the computerized counterpart was programmed to cooperate in the Prisoners' Dilemma (Harmony Game) according to the average cooperation rate across both PD-1 and PD-2 (HG-1 and HG-2). Subjects knew this, and that they would get some information about behaviour in each game. Half of the subjects learned cooperation rates in period 6 for groups PD-1 and HG-1, while the other half learned cooperation rates in period 6 for groups PD-2 and HG-2. The first half of subjects saw cooperation rates of 50% in period 6 of both games, while the second half saw a cooperation rate of 17% in period 6 of the Prisoners' Dilemma and 83% in period 6 of the Harmony Game. These different observed cooperation rates constitute the treatment variable. All aspects of the information treatment were administered between Part 1 and Part 2 of the experiment (namely, after Part 1 games, but before voting, belief elicitation, and Part 2 games).

We recruited ninety-six student subjects from UC Berkeley and ninety-six from Brown University to participate in this additional experiment. Sessions lasted around half an hour, and earnings ranged from $17.75 to $34.50, with an average of $26.45 (including a $5 show-up fee). Appendix Table 14 displays summary statistics of demographics and beliefs.

This additional experiment replicates the results obtained in the main treatments. A majority of subjects vote for the Prisoners' Dilemma (59%). Subjects, on average, underappreciate the equilibrium effect that changing the game has on behaviour: the average belief difference between games is 28% (median of 30%), less than half of the difference in cooperation observed between games. Finally, as seen in Table 7, subjects' reported belief differences are negatively correlated with voting for the Prisoners' Dilemma.

As was the case with the correlation showed in Table 6, the correlation in Table 7 does not establish causation. In this case, however, subjects' belief differences depend in part on the information provided to them about past behaviour in both games. Since this information was

18. The computerized counterpart's choices were independent across periods.

randomly assigned, and subjects' own actions were fixed, the exogenous part of the variation in subjects' beliefs about the behaviour of others can be used to identify the effect of beliefs on voting. Equation (1) in Section 5 isolates the exclusion restriction assumption underlying our exercise. With (α, α') fixed by design, the only way in which information about the behaviour of past subjects affects voting is by altering a subject's beliefs (β, β') about others' likely behaviour in each game, which in turn has implications for the expected utility of voting for each game.[19] Thus, the information treatment can be used as an instrument for subjects' beliefs about others' behaviour.

Panel A in Table 8 shows the first-stage result of the instrumental variable (IV) analysis. The instrument is the variable *Saw High Difference*, which equals one if the subject saw the cooperation rates from the previous two groups with high difference in behaviour (17% cooperation in the Prisoners' Dilemma and 83% cooperation in the Harmony Game) and equals zero if the subject saw the two groups with no difference in behaviour (50% cooperation in both games). The instrumental variable *Saw High Difference* is highly statistically significant, and it affected beliefs on average by increasing the belief difference by around 16 percentage points. Those who received information displaying no changes in behaviour across games updated their beliefs, on average, to a belief differential of 20.4% (these subjects expected on average 47.4% probability of cooperation in the Prisoners' Dilemma and 68% in the Harmony Game). Subjects who received information indicating a large difference updated to a larger belief difference in cooperation: 36% (with 38.4% in the Prisoners' Dilemma and 74% in the Harmony Game).[20] As expected given random assignment, controlling for personal characteristics of the subjects does not affect the results, as seen in column (2).

Panel B in Table 8 shows the second-stage results. We find that belief difference affects the voting decisions of the subjects. An increase in the belief that behaviour will change across games results in a significant decrease in the probability of voting for the Prisoners' Dilemma. A 1 percentage point change in a subject's belief difference leads to a fall in the propensity to vote for the Prisoners' Dilemma of roughly 2 percentage points.[21]

Note that the IV estimates are larger in absolute value than the OLS ones presented in Tables 6 and 7, suggesting that the OLS estimates underestimate the true effect of beliefs on voting. This could be due to both omitted variables that biased the OLS estimate against our hypothesis, or to measurement error in beliefs. It is also possible that the information shock had heterogeneous effects, and that the IV estimate reflects the response of a subset of subjects whose voting is more responsive to their beliefs. It is worth noting that our results are robust to corrections for weak instruments.[22]

19. This is plausible, as it is hard to hypothesize ways in which the information treatment could affect voting directly, if not through its effects on the beliefs (β, β'). Priming or emulation effects may arise, but these pertain to the way one may behave in each game–which our design rules out by fixing own actions–while the behaviour of interest, voting, is of a different nature to the behaviour on which information is given.

20. The effect is stronger if measured in terms of medians. The first group had a median expectation of cooperation in the Prisoners' Dilemma of 50% and 70% in the Harmony Game. Those treated with information indicating a strong change in behaviour had a median expectation of cooperation of 30% in the Prisoners' Dilemma and 80% in the Harmony Game.

21. The reduced form effect of the *Saw High Difference* treatment is to lower the vote share for the Prisoners' Dilemma by 34.4 percentage points (from a vote share of 76% to slightly below 42%). This was attained through an average change in the belief differential of about 16 percentage points, yielding the IV point estimate of $-0.022 (\approx -0.344/16)$.

22. We construct weak-instrument-robust confidence intervals following the Hansen and Chernozhukov (2008) approach, which allows for clustering, for both the estimates with and without controls. Working with a 5% confidence level for rejecting the null that beliefs do not affect voting, the intervals exclude zero and are $[-0.07, -0.012]$ without controls and $[-0.06, -0.009]$ with controls.

TABLE 8
Instrumenting for beliefs

Panel A: First Stage (dependent variable: belief Difference)		
	(1)	(2)
Saw high difference	15.549***	16.232***
	(5.517)	(5.540)
Personal characteristics	N	Y
Constant	20.396***	7.946
	(4.190)	(9.798)
Observations	192	192
R^2	0.043	0.115
F-test of saw high difference = 0	7.940	8.585
Panel B: Second stage (dependent variable: vote for PD)		
Belief difference	−0.022***	−0.022***
	(0.008)	(0.008)
Personal characteristics	N	Y
Constant	1.211***	0.933***
	(0.246)	(0.213)
Observations	192	192
Clusters	32	32

Notes: IV specification. Belief Difference denotes the difference in beliefs of cooperation under HG and PD. Robust standard errors clustered by Part 1 group: *** significant at 1%.

To sum up, the additional experimental evidence supports our core hypotheses that voters may demand bad policies, and that this demand follows from their inability to fully appreciate the equilibrium adjustments of others.

6.5. *Voting for the Prisoners' Dilemma and strategic sophistication*

We now return to our main experiment and investigate our secondary hypotheses. Going back to Table 6 and Appendix Table 13, 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. Given that the *p*-beauty contest number may not be a perfect measure of strategic sophistication (*e.g.* it is not the case that smaller numbers are necessarily a better choice given that others do not play Nash), we also studied whether there may be a nonlinear relationship. We find that including a quadratic term does not change the lack of relationship. There is also no difference in voting between those with a number below and above the median, or below and above 66.66 (numbers above 66.66 are dominated). This lack of relationship between the the *p*-beauty contest number and voting refutes our secondary Hypothesis 5. This is not so surprising in the light of recent research showing that subjects' strategic sophistication is not persistent across games (see Georganas *et al*., 2015). Moreover, this result could indicate that what is crucial in the voting decision is the capacity or inclination to think about the behaviour of others in future stages and not in a simultaneous move game as in the *p*-beauty contest game.

6.6. *Do subjects understand how changing the game will affect their own behaviour?*

We have shown that subjects who vote for the bad policy greatly underestimate the effect that a policy change has on others' behaviour. Some subjects may not only fail to anticipate how others adjust to policy changes, but also how they themselves will respond. Our secondary Hypothesis 6 postulates that there is a non-trivial share of such subjects.

To study the share of subjects who fail to think through how their own behaviour depends on policy, we postulate a simple mixture model where individuals can be one of two types $t \in (R, I)$ (for Rational and Inertial, respectively) depending on the way they think about their actions in each game.[23] The Rational type is one who holds beliefs (β, β') about the cooperation rates by others in the Prisoners' Dilemma and the Harmony Game, respectively, but recognizes he will play his dominant strategy in each game. The Inertial type does not realize that his behaviour will differ across games. 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 Prisoners' Dilemma and the Harmony Game for beliefs $[\alpha, \alpha', \beta, \beta']$, we obtain

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

where $\Delta\beta = \beta' - \beta$. The key aspect differentiating the types is that the term $-\alpha' - 2\alpha$ is -1 for Rational 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 Rational types, and $1-s$ of Inertial types. For the purposes of empirical identification of the share s, we assume that a Rational (Inertial) type votes for the Prisoners' 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 assume that such probability is given by a logistic cdf with mean μ and standard deviation σ. Thus, a player i with type t votes for Prisoners' Dilemma with a probability $F\left(\Delta u^I(\Delta\beta_i, c_i), \mu, \sigma\right)$, where F denotes the logistic distribution. It follows that the probability of a Prisoners' Dilemma vote by a player i, given a share s of Rational types is,

$$P(v_i = PD | \Delta\beta_i, c_i, s, \mu, \sigma) = sF\left(\Delta u^R(\Delta\beta_i), \mu, \sigma\right) + (1-s)F\left(\Delta u^I(\Delta\beta_i, c_i), \mu, \sigma\right).$$

Given N subjects, we define the profile of period 5 actions as $\mathbf{c} = [c_1, ..., c_N]$ where $c_i = 1$ denotes cooperation by subject i in period 5 and $c_i = 0$ denotes defection. Similarly, we define the profile of votes as $\mathbf{v} = [v_1, ..., v_N]$, where $v_i = 1$ denotes a vote for Prisoners' Dilemma by subject i, and $v_i = 0$ a vote for Harmony Game. We have that the overall probability of such a profile of votes is,

$$\prod_{i=1}^{N} P(v_i = 1 | \Delta\beta_i, c_i, s, \mu, \sigma)^{v_i} (1 - P(v_i = 1 | \Delta\beta_i, c_i, s, \mu, \sigma))^{1-v_i},$$

23. One way to test Hypothesis 6 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 main experiment (768 subjects), we opted to investigate this secondary hypothesis via the structural approach presented in this section. While our Hypothesis 6 was formulated *ex ante*, the precise assumptions on types presented here were developed *ex post*.

which yields the log-likelihood,

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

We estimate the parameter s, by maximizing $L(s,\mu,\sigma|\mathbf{v},\Delta\beta,\mathbf{c})$ given the voting data $\mathbf{v}$, the period 5 behaviour $\mathbf{c}$, and the vector of elicited beliefs $\Delta\beta$. Clearly, in this estimation we take beliefs to be exogenous. We pool the data for the three treatments 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 Rational types s, presented in Table 9, is 67% when we consider all subjects. The Wald test for the share of Rational types being equal to 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, allowing us to reject the null of no Inertial types. 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 commit this error. This is striking, given that all that is required is to forecast that one will play a different dominant strategy in a 2×2 game following the change in policy.

The existence of the logistic disturbance term implies that any voter could vote for or against the Prisoners' Dilemma; but the beliefs $\Delta\beta$ and the types (R,I) affect the likelihood of the vote going one way or another. For voters with $\Delta\beta$ between $\frac{1}{6}$ and $\frac{2}{3}$, however, the type (R,I) alone can affect the vote, even with a zero realization of the disturbance term. One may consider the empirical variation offered by those voters as more central, and wonder about the robustness of the result to restricting the sample to those voters. We report the estimates of our model on this restricted sample in the second column of Table 9. Although we lose a large number of observations (from 408 to 232), the fraction of Rational types remains significantly different from one (point estimate 0.4, p-value 0.084). The point estimate does not significantly differ from the one obtained using the full sample.

We have heretofore assumed that the chance of being Rational versus Inertial does not depend upon a subject's beliefs $\Delta\beta$. But a subject who holds Nash beliefs about others (*i.e.* $\Delta\beta=1$) seems unlikely to ignore how her own action varies across games. To allow for the possibility that those who better predict others' behaviour better predict their own behaviour, we re-estimate our model to allow for our parameter s to depend upon whether the subject's beliefs $\Delta\beta$ lies above or below the median value. This produces estimates of s equal to 95% versus 26%, respectively. Note that this results on an average estimated prevalence of Rational subjects of 60%, which is close to the original estimate. This suggests that our estimate of the overall prevalence of Rational subjects is quite robust to relaxing the assumption of independence of types and beliefs about others.

6.7. *Ruling out alternative mechanisms*

The variation of treatments allows us to rule out some alternative mechanisms. One possibility is that the Prisoners' Dilemma attracts 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. Instead, when the Harmony Game is played first, a reluctance to try new 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 treatment the Prisoners' Dilemma garnered 50% of the vote. This vote share is only 3 points smaller than

TABLE 9
Structural estimates

	Subjects	
	All	$\frac{1}{6} \leq \Delta\beta \leq \frac{2}{3}$
s (Share of rational types)	0.670	0.399
	(0.173)	(0.348)
μ	0.801	1.097
	(0.149)	(0.549)
σ	1.336	0.978
	(0.350)	(0.493)
Observations	408	232
p-value of Wald test $s \neq 1$	0.057	0.084

Notes: Pooled sample from Random Dictator, Reverse Random Dictator and Majority Once. Robust standard errors clustered by Part 1 group in parentheses.

(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, a vote 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 expect a large majority (in either direction), leaving her with negligible chance of being pivotal and therefore with no incentive to carefully consider how to vote. Secondly, 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 subjects do not expect behaviour to correspond to equilibrium (for if they did, they would vote for the Harmony Game). Subjects could take this as a sign that other subjects will themselves 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 Prisoners' Dilemma under Random Dictator cannot be explained by the previous pivotality concerns. The share of votes in favor of the Prisoners' 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 behaviour of others will adjust to a new game, and our main hypotheses are supported by the data. Nevertheless, our framework would warrant less attention if the majority vote for Prisoners' Dilemma could be explained by strictly rational motives. Yet, as shown above, a rational subject (one who knowingly plays the dominant action in each game) cannot prefer the Prisoners' Dilemma unless she expects cooperation in the Harmony Game to be at most 50 percentage point greater than in the Prisoners' Dilemma. Given the actual difference in cooperation rates is much higher (*e.g.* 76 and 63 percentage points in the Control and Reverse Control treatments, respectively), a rational subject who has fairly accurate beliefs about real behaviour cannot prefer the Prisoners' Dilemma. Moreover, as we showed in footnote 10, even sizeable risk aversion makes rational subjects only slightly more willing to vote for the Prisoners' Dilemma. Thus, rationalizing the Prisoners' Dilemma majority in a model where the majority of players are both rational and correct about their environment appears difficult.

We have attributed the demand for bad policy to subjects' failure to appreciate the equilibrium effects of policy changes. Several existing models of strategic naivety make predictions other than subgame-perfect ones in our setting, and one may wonder whether the mistake 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 *et al*.'s (2004) related "cognitive hierarchy model," and Jehiel's (2005) ABEE. Although each theory captures some facet of subjects' non-equilibrium behaviour, each also falls short of providing 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 *et al*., 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 the Prisoners' Dilemma and play the dominant strategy in both the Prisoners' Dilemma and the 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 behaviour of the majority of our subjects who vote for the Prisoners' Dilemma. However, a theory predicated on level-1 makes at least one prediction that is contradicted by the data, namely that subjects voting for the Prisoners' Dilemma predict that cooperation rates do not vary across the Prisoners' Dilemma and the Harmony Game. This does not match the fact that most individuals voting for the Prisoners' 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 the Prisoners' Dilemma. Those voting for the Prisoners' Dilemma estimate on average an increase in cooperation of about 20 percentage points (which is significantly 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 modelled 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 the Prisoners' Dilemma and the 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 the Prisoners' Dilemma and the 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 *et al*.'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 hurdle that few subjects play the dominated action in either the Harmony Game or the Prisoners' Dilemma, so level-2 types must assign a high probability to level-1 types in the cognitive hierarchy model. But, if there are few level-0 players, higher level players have no reason to vote for the Prisoners' Dilemma in this model. Yet, we observe majoritarian support for the Prisoners' Dilemma in the data.

The underappreciation of indirect effects we document reflects a failure of contingent reasoning as each subject compares how others would act across the two games. All three theories just reviewed deal with contingent reasoning, but make too extreme a prediction about the degree to which subjects will fail. While we think the main gist is correct that inadequate contingent reasoning is to blame, more nuanced theorizing seems necessary to fully understand the drivers of the underappreciation of equilibrium effects. We hope our findings will spur further work in the area.

A concern about our main results is that a framing effect, in the form of an aversion to the word "tax"used to introduce the Harmony Game, may create support for the Prisoners' Dilemma.[24] Three considerations mitigate this concern. First, the use of the word "tax" was extremely sparse: it was used once in instructions read aloud, and did not appear in the screens facing subjects during the experiment before voting (it was only present in the feedback screen after the vote). Secondly, if the tax framing drove results under the Random Dictator and majority treatments, then the Reverse Random dictator treatment, in which taxation was not mentioned, should have eliminated support for the Prisoners' Dilemma. Yet it did not. Thirdly, our framework predicts that a manipulation of beliefs about the behaviour of others should affect voting, which we demonstrate in our additional experiment. A tax-framing effect cannot explain that result, since the frame did not change across belief-manipulation treatments.

Finally, one might worry that social preferences transform the experiment's monetary payoffs into utilities in such a way that people may rationally vote for the Prisoners' Dilemma. As we argue in detail in the Online Appendix, social preferences should push voting towards the Harmony Game. But the simplest and most direct evidence against the concern that social preferences may cause voting for the Prisoners' Dilemma comes from our additional experiment. Because each subject in Part 2 of this experiment neither chooses an action nor plays against a human opponent, the subject's voting decision reduces to a single-person decision problem that has no implications for any other subject's payoff. The fact that subjects in this treatment that neutralizes social preferences behave in a similar way to subjects in our main experiment—majorities of both vote for the PD—establishes that social preferences do not drive our findings.

6.8. *Learning under repeated majority voting*

Our basic design presented subjects with a choice between a game with which they had experience and one with which they lacked experience. This sought to capture substantive situations of interest where some policy options are new to the population, as well as to motivate subjects to make conjectures that ought to be informed by equilibrium predictions. In that context, we showed that beliefs about behaviour are not tightly driven by equilibrium considerations and this led a majority of subjects to mistakenly prefer the Prisoners' Dilemma over the Harmony Game. Against such backdrop, one may conjecture that gaining experience with the less familiar game will lead subjects to rank the games correctly, and one may also wonder how quick and complete that learning will be.

The Majority Repeated treatment allows us to study the evolution of voting as subjects gain experience. The percentage of subjects voting for the Prisoners' Dilemma decreases from 50.83% in period 6 to 28.33% in period 10—see Figure 4.

24. There is some evidence from non-incentivized, survey-choice experiments of tax aversion, namely people declare a stronger preference for avoiding costs of a given magnitude when these are portrayed as a tax (Sussman and Olivola, 2011).

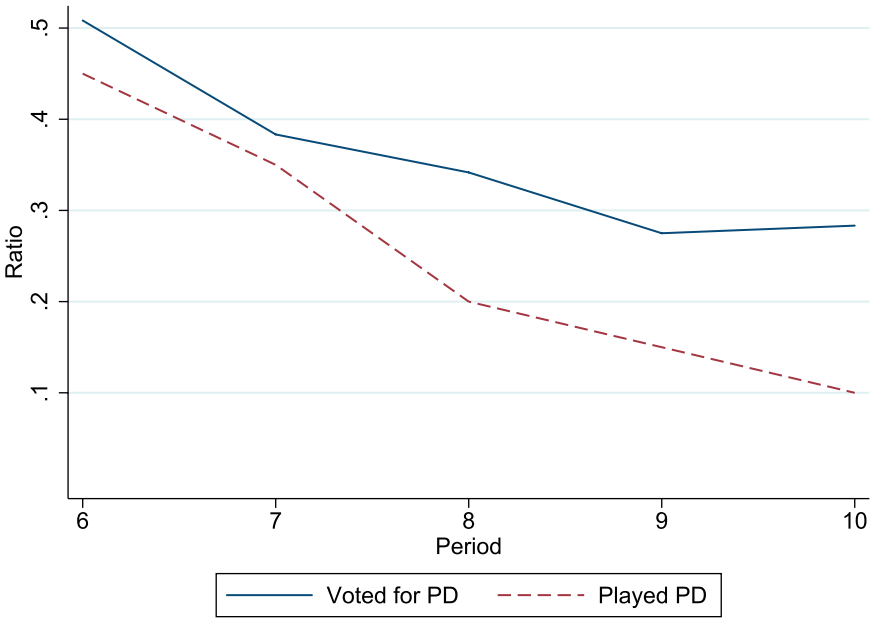

FIGURE 4
Evolution of votes and outcomes in majority repeated

The fact that vote shares for the Prisoners' Dilemma decrease with experience suggests that there is learning and that in our context experience can substitute for the ability to make equilibrium predictions based on theory. However, it is noteworthy that even in the fifth period after the first vote, more than a quarter of subjects continued to choose the wrong game. One possibility is that many subjects still vote for the Prisoners' Dilemma simply because they have not been exposed to the Harmony Game. But this is not the case. The percentage of subjects playing the Prisoners' Dilemma decreases from 45% in period 6 to 10% in period 10.[25] In other words, a non-trivial share of votes for the Prisoners' Dilemma persists until period 10 despite the fact that by then most voters have been exposed to the Harmony Game. More specifically, the second column in Table 10 shows, by period, the percentage of Prisoners' Dilemma voters who had played the Harmony Game before. Of course, none of them had played the Harmony Game before voting in period 6. However, by period 8 more than half of the Prisoners' Dilemma voters had played the Harmony Game before.

Why do some subjects who have experienced both games continue to vote for the wrong one? Is it the case that those subjects have met defectors in the Harmony Game? The third column in Table 10 shows the percentage of Prisoners' Dilemma voters who had played the Harmony Game before and had observed a difference in cooperation rates greater than 50 percentage points, the cutoff to prefer the Harmony Game over the Prisoners' Dilemma. Note that in every period, more than 70% of Prisoners' Dilemma voters who had played the Harmony Game before had observed cooperation rates in the Prisoners' Dilemma and the Harmony Game that warranted voting for

25. The reason is simple: given the simple majority rule, quickly after the vote share for the Prisoners' Dilemma drops below 50%, the power of democracy kicks in: majorities for the Prisoners' Dilemma in each group become less frequent, and only a small share of subjects end up in the wrong game by period 10.

TABLE 10
The information of Prisoners' Dilemma voters

Period	Played HG before (%)	Share of those who played HG before who observed at least 50 percentage points more cooperation in HG than PD (%)
6	0	
7	32.60	80
8	51.22	85.71
9	57.58	78.95
10	73.53	72

the Harmony Game. So, a large fraction of Prisoners' Dilemma voters had information in favour of the Harmony Game but still preferred to vote for the Prisoners' Dilemma by the end of the experiment.

In conclusion, while many subjects who began by voting for the Prisoners' Dilemma switch with experience to the Harmony Game, others fail to learn, even with the benefit of observations that favour voting for the Harmony Game. This evidence yields a nuanced message: inexperience with a policy whose attractiveness depends upon equilibrium effects can generate substantial bias in policy preference; with more balanced experience, the bias in policy preference decreases but does not disappear entirely—some fraction of the biased preferences identified in our main experiments persists. While the particular levels of the demand for bad policy found in the lab should be taken with caution, one interpretation of our experimental results is that when offered an unfamiliar policy, people err frequently enough that groups often select the wrong policy, even when relying on majority voting. With more balanced experience, majoritarian mistakes are rare, but policy distortions could still occur when groups select policy through mechanisms that are less stark than majority voting, and which place positive weight on the opinion of all voters (*e.g.* collective bargaining) or that may end up selecting the loser of a plurality vote (*e.g.* electoral college).

7. CONCLUSION

We have experimentally identified a demand for bad policy driven by voters' underappreciation of how behaviour changes when policies change. Voters in our experiment underestimate, on average, how much other people's behaviour will change following a change in the game played. In addition, a non-trivial share of voters appear to fail to appreciate that their own behaviour will differ across games. Our evidence suggests that unfamiliar policy options can be a challenge for voters when these policies contain "hidden" costs or benefits that will accrue once behaviour adjusts. An example of such a policy is a Pigouvian tax, which generates a direct monetary cost on taxpayers as well as indirect benefits in equilibrium by inducing those same taxpayers to internalize negative externalities. More generally, our results help explain why voters may not always support the policy proposals of economists that are beneficial mainly through indirect equilibrium effects, but may support populist proposals that are costly through indirect equilibrium effects.

Of course, identifying a demand for bad policy in connection with a tendency to underestimate equilibrium effects in the laboratory does not necessarily mean that outside of the lab such demand will dominate forces promoting good policies. One might hope that public discourse and political competition lead voters to learn about the total effect of policies, thus bridging the gap between

public opinion and reliable evidence. However, as discussed in the "Introduction", a vast literature in economics and political science—both theoretical and empirical—has considered politicians as reflecting, more than shaping, the positions of voters. To the extent that public opinion and voter preferences matter for the selection of policies, understanding how people think about policies appears relevant for our knowledge of how societies choose to regulate themselves. This article makes a contribution to that understanding.

APPENDIX

TABLE 11
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	37.25	20.93	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	33.48	41.11	– 100	100
Earnings	768	27.86	3.27	16.67	37

TABLE 12
Comparison between Prisoners' Dilemma and Harmony Game by treatment

Panel A: Cooperation

	Control			Reverse Control			Random Dictator		
	Part 1	Part 2		Part 1	Part 2		Part 1	Part 2	
Periods	All	6	All	All	6	All	All	6	All
HG in Part 2 (%)	27	88	92	99	92	93	28	94	96
PD in Part 2 (%)	21	30	16	95	38	30	26	21	15
Diff. *p*-value	0.186	0.000	0.006	0.209	0.000	0.004	0.72	0.000	0.000
	Reverse RD			Majority Once			Majority Repeated		
	Part 1	Part 2		Part 1	Part 2		Part 1	Part 2	
Periods	All	6	All	All	6	All	All	6	All
HG in Part 2 (%)	95	94	95	37	97	94	28	97	98
PD in Part 2 (%)	93	58	36	31	33	21	24	20	17
Diff. *p*-value	0.656	0.000	0.000	0.431	0.000	0.010	0.352	0.000	0.003

Panel B: Payoffs

	Control			Reverse Control			Random Dictator		
	Part 1	Part 2		Part 1	Part 2		Part 1	Part 2	
Periods	All	6	All	All	6	All	All	6	All
HG in Part 2	5.81	7.18	7.44	7.79	7.42	7.51	6.04	7.61	7.75
PD in Part 2	6.12	6.20	5.65	7.75	6.53	6.20	6.16	5.82	5.59
Diff. *p*-value	0.122	0.125	0.008	0.576	0.134	0.008	0.586	0.007	0.000
	Reverse RD			Majority Once			Majority Repeated		
	Part 1	Part 2		Part 1	Part 2		Part 1	Part 2	
Periods	All	6	All	All	6	All	All	6	All
HG in Part 2	7.70	7.61	7.66	5.95	7.77	7.58	5.82	7.79	7.83
PD in Part 2	7.48	7.30	6.44	6.42	6.33	5.85	6.35	5.81	5.67
Diff. *p*-value	0.211	0.326	0.000	0.048	0.034	0.019	0.052	0.019	0.006

Panel C: Number of observations and number of sessions (in parentheses)

	Control (7)			Reverse Control (7)			Random Dictator (8)		
	Part 1	Part 2		Part 1	Part 2		Part 1	Part 2	
Periods	All	6	All	All	6	All	All	6	All
HG in Part 2	300	60	300	300	60	300	450	90	450
PD in Part 2	300	60	300	300	60	300	390	78	390
	Reverse RD (8)			Majority Once (7)			Majority Repeated (6)		
	Part 1	Part 2		Part 1	Part 2		Part 1	Part 2	
Periods	All	6	All	All	6	All	All	6	All
HG in Part 2	270	54	270	150	30	150	330	66	450
PD in Part 2	330	66	330	450	90	450	270	54	150

Notes: p-values from Wald tests. Standard errors clustered by Part 1 group for Part 1 outcomes and period 6 cooperation, and by session for Part 2 outcomes and period 6 payoffs. Due to few clusters, session-clustered SEs are Webb (2014) wild bootstrapped. For Majority Repeated, behaviour in Part 1 as a function of game played in period 6.

TABLE 13
Personal Characteristics and voting for Prisoners' Dilemma (dependent variable: Vote for PD)

	(1) All	(2) RD	(3) Reverse RD	(4) Majority Once	(5) Majority Repeated
Male	−0.167***	−0.208**	0.061	−0.219**	−0.272**
	(0.049)	(0.083)	(0.116)	(0.097)	(0.100)
Year	0.005	0.006	−0.012	0.063	−0.023
	(0.019)	(0.031)	(0.040)	(0.041)	(0.043)
Ideology	0.027***	0.027	0.040**	0.049*	−0.004
	(0.010)	(0.018)	(0.018)	(0.025)	(0.016)
Economics	−0.030	−0.034	0.164	−0.081	−0.086
	(0.064)	(0.083)	(0.171)	(0.175)	(0.149)
Political Science	0.052	−0.147	0.263*	0.070	−0.102
	(0.096)	(0.215)	(0.141)	(0.180)	(0.164)
Brown University	0.076	0.066	0.044	0.136	0.043
	(0.051)	(0.087)	(0.121)	(0.084)	(0.116)
Beauty Number	0.000	−0.001	0.001	−0.000	0.003
	(0.001)	(0.002)	(0.003)	(0.002)	(0.002)
Constant	0.456***	0.519***	0.260	0.321	0.589***
	(0.082)	(0.153)	(0.175)	(0.188)	(0.185)
Observations	528	168	120	120	120
R^2	0.038	0.058	0.063	0.083	0.107

Notes: OLS specification. Year denotes year in college. Ideology from 0 to 10 from most liberal to most conservative. Economics and Political Science denote subjects' major. Robust standard errors clustered by Part 1 group: *** significant at 1%, ** at 5%, * at 10%, respectively.

TABLE 14
Summary statistics—additional experiment

	Obs.	Mean	Std. Dev.	Min	Max
Male	192	0.42	0.49	0	1
Year	192	2.57	1.27	1	5
Ideology	192	2.98	1.99	0	8
Economics	192	0.19	0.39	0	1
Political Science	192	0.02	0.14	0	1
Brown University	192	0.50	0.50	0	1
Beauty Contest Number	192	38.14	23.20	0	100
Math SAT	161	718.11	85.18	350	800
Verbal SAT	159	706.29	75.77	490	800
Belief of C in PD	192	42.88	25.19	0	100
Belief of C in HG	192	71.05	21.90	0	100
Belief Difference	192	28.17	37.25	−100	100
Earnings	192	26.45	2.92	17.75	34.50

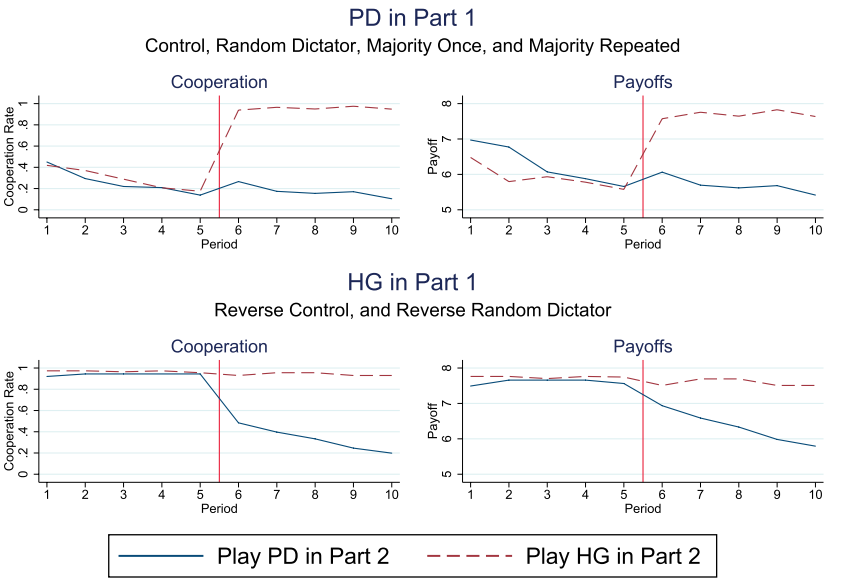

FIGURE 5

Comparing Prisoners' Dilemma and Harmony Game: all treatments

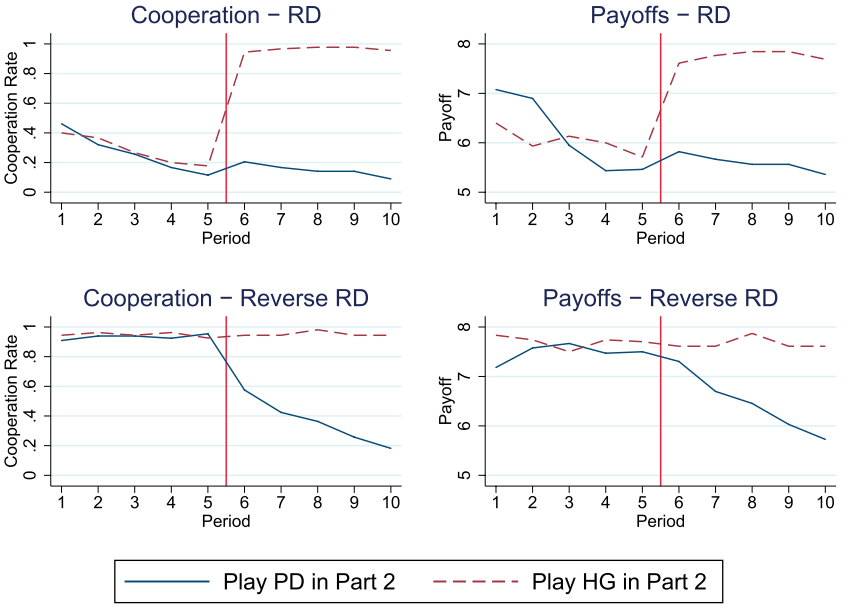

FIGURE 6

Comparing Prisoners' Dilemma and Harmony Game: Random Dictator and Reverse Random Dictator

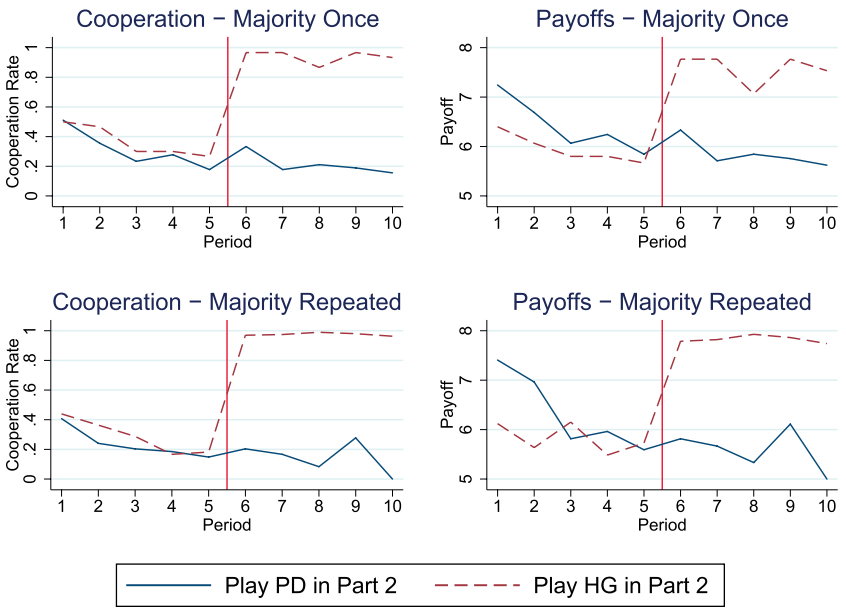

FIGURE 7
Comparing Prisoners' Dilemma and Harmony Game: Majority Once and Majority Repeated

Acknowledgments. We thank Alexander Kiam, Jeongbin Kim, Daniel Prinz, Santiago Truffa and Stephen Yen for excellent research assistance as well as Berkeley's XLab and Brown's BUSSEL for support. We are grateful to Anna Aizer, Ned Augenblick, Eric Dickson, Willie Fuchs, Alessandro Lizzeri, Matthew Rabin, Francesco Trebbi, Reed Walker, Georg Weizsäcker, and Noam Yuchtman for helpful discussions. We thank participants at various conferences and seminars for their comments and suggestions. Eyster thanks the ERC for financial support.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

REFERENCES

AGRANOV, M. and PALFREY, T. R. (2015), "Equilibrium Tax Rates and Income Redistribution: A Laboratory Study", *Journal of Public Economics*, **130**, 45–58.
ALESINA, A. and TABELLINI, G. (1990), "Voting on the Budget Deficit", *American Economic Review*, **80**, 37–49.
ALESINA, A. and DRAZEN, A. (1991), "Why are Stabilizations Delayed?" *American Economic Review*, **81**, 1170–1188.
BARRO, R. J. (1973), "The Control of Politicians: An Economic Model", *Public Choice*, **14**, 19–42.
BARTELS, L. (2012), "The Study of Electoral Behaviour", in Leighley, J., (ed.), *The Oxford Handbook of American Elections and Political Behaviour* (Oxford University Press).
BEILHARZ, H. J. and GERSBACH, H. (2004), "General Equilibrium Effects and Voting into a Crisis" (Discussion Paper No. 4454, CEPR).
BESLEY, T. (2005), "Political Selection", *Journal of Economic Perspectives*, **19**, 43–60.
BESLEY, T. and COATE, S. (1998), "Sources of Inefficiency in a Representative Democracy: A Dynamic Analysis", *American Economic Review*, **88**, 139–156.
BISIN, A., LIZZERI, A. and YARIV, L. (2015), "Government Policy with Time Inconsistent Voters", *American Economic Review*, **105**, 1711–1737.
BLINDER, A. and KRUEGER, A. (2004), "What Does the Public Know about Economic Policy, and How Does It Know It?" *Brookings Papers on Economic Activity*, **2004**, 327–397.
BONE, J., HEY, J. D. and SUCKLING, J. (2009), "Do People Plan?" *Experimental Economics*, **12**, 12–25.
CAMERER, C. F., HO, T.-H. and CHONG, J. K. (2004), "A Cognitive Hierarchy Model of Games", *Quarterly Journal of Economics*, **119**, 861–898.
CANES-WRONE, B., HERRON, M. C. and SHOTTS, K. W. (2001), "Leadership and Pandering: A Theory of Executive Policymaking", *American Journal of Political Science*, **45**, 532–550.

CAPLAN, B. (2007), *The Myth of the Rational Voter: Why Democracies Choose Bad Policies*, (Princeton: Princeton University Press).
CASELLI, F., and MORELLI, M. (2004), "Bad Politicians", *Journal of Public Economics*, **88**, 759–782.
CHERNOZHUKOV, V. and HANSEN, C. (2008), "The Reduced Form: A Simple Approach to Inference with Weak Instruments", *Economics Letters*, **100**, 68–71.
COATE, S. and MORRIS, S. (1995), "On the Form of Transfers to Special Interests", *Journal of Political Economy*, **103**, 1210–1235.
COSTA-GOMES, M. and WeIZSÄCKER, G. (2008), "Stated Beliefs and Play in Normal-Form Games", *Review of Economic Studies*, **75**, 729–762.
CRAWFORD, V., COSTA-GOMES, M. A. and IRIBERRI, N. (2013), "Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications", *Journal of Economic Literature*, **51**, 5–62.
DAL BÓ, E., DAL BÓ, P. and DI TELLA, R. (2006), "Plata o Plomo?: Bribe and Punishment in a Theory of Political Influence", *American Political Science Review*, **100**, 41–53.
DAL BÓ, P. (2014), "Experimental Evidence on the Workings of Democratic Institutions", in *Economic Institutions, Rights, Growth, and Sustainability: the Legacy of Douglass North*, (Cambridge: Cambridge University Press).
DAL BÓ, P., FOSTER, A. and PUTTERMAN, L. (2010), "Institutions and Behaviour: Experimental Evidence on the Effects of Democracy", *American Economic Review*, **100**, 2205–2229.
DE FIGUEIREDO, R. (2002), "Electoral Competition, Political Uncertainty, and Policy Insulation", *American Political Science Review*, **96**, 321–333.
DURANTON, G. and TURNER, M. (2011), "The Fundamental Law of Road Congestion: Evidence from US Cities", *American Economic Review*, **101**, 2616–2652.
DOWNS, A. (1962), "The Law of Peak-Hour Expressway Congestion", *Traffic Quarterly*, **16**, 393–409.
ERTAN, A., PAGE T., and PUTTERMAN, L. (2009), "Who to punish? Individual Decisions and Majority Rule in Mitigating the Free Rider Problem", *European Economic Review*, **53**, 495–511.
ESPONDA, I. and POUZO, D. (2017), "Conditional Retrospective Voting in Large Elections", *American Economic Journal: Microeconomics*, **9**, 54–75.
ESPONDA, I. and VESPA, E. (2014), "Hypothetical Thinking and Information Extraction: Strategic Voting in the Laboratory", *American Economic Journal: Microeconomics*, **6**, 180–202.
EYSTER, E. and RABIN, M. (2005), "Cursed Equilibrium", *Econometrica*, **73**, 1623–1672.
FEREJOHN J. (1986), "Incumbent Performance and Electoral Control", *Public Choice*, **50**, 5–25.
FERNANDEZ, R. and RODRIK, D. (1991), "Resistance to Reform: Status Quo Bias in the Presence of Individual-Specific Uncertainty", *American Economic Review*, **81**, 1146–1155.
FISCHBACHER, U. (2007), "z-Tree: Zurich Toolbox for Ready-made Economic Experiments", *Experimental Economics*, **10**, 171–178.
FUDENBERG, D., RAND, D. G. and DREBER, A. (2012), "Slow to Anger and Fast to Forgive: Cooperation in an Uncertain World", *American Economic Review*, **102**, 720–749.
FEHR, E. and SCHMIDT, K. M. (1999), "A Theory of Fairness, Competition, and Cooperation", *Quarterly Journal of Economics*, **114**, 817–868.
GENTZKOW, M. and SHAPIRO, J. (2006), "Media Bias and Reputation", *Journal of Political Economy*, **114**, 35–71.
GENTZKOW, M. and SHAPIRO, J. (2010), "What Drives Media Slant? Evidence from U.S. Newspapers", *Econometrica*, **78**, 35–71.
GEORGANAS, S., HEALY, P. J. and WEBER, R. A. (2015), "On the Persistence of Strategic Sophistication", *Journal of Economic Theory*, **159**, 369–400.
GRETHER, D. M. (1992), "Testing Bayes Rule and the Representativeness Heuristic: Some Experimental Evidence", *Journal of Economic Behaviour & Organization*, **17**, 31–57.
HARRINGTON, J. E. (1993), "Policy, Economic Performance, and Elections", *American Economic Review*, **83**, 27–42.
HOLT, C. A. (2007), *Markets, Games, and Strategic Behaviour* (Boston: Pearson/Addison Wesley).
JEHIEL, P. (2005), "Analogy-Based Expectation Equilibrium", *Journal of Economic Theory*, **123**, 81–104.
KALLBEKKEN, S., KROLL, S. and CHERRY, T. (2011), "Do You Not Like Pigou, Or Do You Not Understand Him? Tax Aversion and Revenue Recycling in the Lab", *Journal of Environmental Economics and Management*, **62**, 53–63.
KARNI, E. (2009), "A Mechanism for Eliciting Probabilities", *Econometrica*, **77**, 603–606.
KOSFELD, M., OKADA, A. and RIEDL, A. (2009), "Institution Formation in Public Goods Games", *American Economic Review*, **99**, 1335–1355.
LEVITT, S., LIST, J. and SADOFF, S. (2011), "Checkmate: Exploring Backward Induction among Chess Players", *American Economic Review*, **101**, 975–990.
LIZZERI, A. and YARIV, L. (Forthcoming), "Collective Self-Control", *American Economic Journal: Microeconomics*.
MARGREITER, M., SUTTER, M. and DITTRICH, D. (2005), "Individual and Collective Choice and Voting in Common Pool Resource Problem with Heterogeneous Actors", *Environmental & Resource Economics*, **32**, 241–271.
MASKIN, E. and TIROLE, J. (2004), "The Politician and the Judge: Accountability in Government", *American Economic Review*, **94**, 1034–1054.
McKELVEY, R. and PALFREY, T. (1992), "An Experimental Study of the Centipede Game", *Econometrica*, **60**, 803–836.
MESSNER, M. and POLBORN, M. K. (2004), "Paying politicians", *Journal of Public Economics*, **88**, 2423–2445.
MÖBIUS, M. M., NIEDERLE, M., NIEHAUS, P. and ROSENBLAT, T. (2014), "Managing Self-Confidence: Theory and Experimental Evidence", unpublished manuscript.

MOINAS, S. and POUGET, S. (2013), "The Bubble Game: An Experimental Study of Speculation", *Econometrica*, **81**, 1507–1540.
MULLAINATHAN, S. and WASHINGTON, E. (2009), "Sticking with Your Vote: Cognitive Dissonance and Political Attitudes", *American Economic Journal: Applied Economics*, **1**, 86–111.
NAGEL, R. (1995), "Unraveling in Guessing Games: An Experimental Study", *American Economic Review*, **85**, 1313–1326.
NORTH, D. C. (1990), *Institutions, Institutional Change and Economic Performance* (Cambridge: Cambridge University Press).
PALACIOS-HUERTA, I. and VOLIJ, O. (2009), "Field Centipedes", *American Economic Review*, **99**, 1619–1635.
PELTZMAN, S. (1976), "Toward a More General Theory of Regulation", *Journal of Law and Economics*, **19**, 211–240.
ROMER, D. (2003), "Misconceptions and Political Outcomes", *The Economic Journal*, **113**, 1–20.
ROMER, T. and ROSENTHAL, H. (1978). "Political Resource Allocation, Controlled Agendas and the Status Quo", *Public Choice*, **33**, 27–43.
SAUSGRUBER, R. and TYRAN, J.-R. (2005), "Testing the Mill Hypothesis of Fiscal Illusion", *Public Choice*, **122**, 39–68.
SAUSGRUBER, R. and TYRAN, J.-R. (2011), "Are We Taxing Ourselves? How Deliberation and Experience Shape Voting On Taxes", *Journal of Public Economics*, **95**, 124–176.
SMITH, A. (1776), *An Enquiry into the Nature and Causes of the Wealth of Nations* (London: W. Strahan and T. Cadell).
STAHL, D. O. and WILSON, P. W. (1994), "Experimental Evidence on Players' Models of Other Players", *Journal of Economic Behaviour & Organization*, **25**, 309–327.
STIGLER, G. (1971), "The Regulation of Industry" *The Bell Journal of Economics and Management Science*, **2**, 3–21.
SUSSMAN, A. B. and OLIVOLA, C. Y. (2011), "Axe the Tax: Taxes Are Disliked More than Equivalent Costs", *Journal of Marketing Research*, **48**, S91–S101.
WALKER, J., GARDNER, R., HERR, A. and OSTROM E. (2000), "Collective Choices in the Commons: Experimental Results on Proposed Allocation Rules and Votes", *The Economic Journal*, **110**, 212–234.
WEBB, M. (2014). "Reworking Wild Bootstrap Based Inference for Clustered Errors" (Queen's Economics Department Working Paper No. 1315).