Games and Economic Behavior 130 (2021) 179–195

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Pledge-and-review in the laboratory ☆

Steffen Lippert *,[1], James Tremewan [2]

University of Auckland, New Zealand

ARTICLE INFO

Article history:
Received 19 June 2020
Available online 16 August 2021

JEL classification:
C78
C90
D02
H41
Q54

Keywords:
Pledge and review
Public goods
Voluntary contributions
Conditional cooperation

ABSTRACT

We perform a laboratory test of Pledge-and-Review bargaining, implementing a simplified version of the model analysed in Harstad (2021a). In theory, this institution should increase contributions to a public good only if there is uncertainty over the value of possible future payoffs. In contrast, we find that Pledge-and-Review increases efficiency in all the settings we investigate, and that the improvement is most persistent in our setting without uncertainty. Our results suggest that the Pledge-and-Review institution may be useful, even without uncertainty, as it allows conditional cooperators to test, risk free, the cooperativeness of their partners.

1. Introduction

Inspired by the Paris Agreement on climate change, but with additional applications to domestic politics and business negotiations, Harstad (2021a) introduces a novel model of "pledge and review" bargaining. In the model, agents make pledges to a public good which are only implemented after approval by unanimity voting. The main insight of the model is that this institution can increase contributions, but only when there is uncertainty over how agents value the future. We test these predictions in a laboratory experiment which implements a simplified version of the model.

Harstad (2021a) models a repeated two-step bargaining procedure in which parties first simultaneously propose their own individual contributions to a public good and then decide whether they find the overall distribution of pledges acceptable. If one or more parties find the pledges unacceptable, the procedure starts again. Delay between the review and the new pledges makes this a costly outcome. With complete information, the trivial equilibrium in this game coincides with the non-cooperative business-as-usual outcome; but with sufficient uncertainty over parties' cost of delay, *the set of equilibria includes equilibria in which higher than business-as-usual pledges are sustained*.

☆ We are grateful to two anonymous reviewers, the Associate Editor and the Editor, Gary Charness, for their many comments and suggestions that improved the paper considerably. We thank Bård Harstad for helpful discussions, Francis Bloch, Simona Fabrizi, Eberhard Feess, Kai Konrad, Tore Nilssen, Ronald Peeters, Marco Serena and Basil Sharp for their comments, and Carline Bentley for superb research assistance. The experiment was approved by the University of Auckland Human Participants Ethics Committee on 12 November 2018 for 3 years with the Reference Number 022284.

* Corresponding author.
E-mail addresses: s.lippert@auckland.ac.nz (S. Lippert), james.tremewan@auckland.ac.nz (J. Tremewan).
[1] Department of Economics, Centre for Mathematical Social Science, and Te Pūnaha Matatini Centre for Research Excellence, University of Auckland, 12 Grafton Rd, 1010 Auckland, New Zealand.
[2] Department of Economics and Centre for Mathematical Social Science, University of Auckland, 12 Grafton Rd, 1010 Auckland, New Zealand.

https://doi.org/10.1016/j.geb.2021.08.003
0899-8256/

We use laboratory experiments to take a stab at testing the performance of the pledge-and-review procedure in various environments, including one with uncertainty akin to that described in Harstad (2021a). To focus on the role of uncertainty we implement a substantially simplified model, but one which captures the same basic intuition. Our four treatments consist of a standard two-player public goods game, and three *pledge-and-review* treatments which add a unanimity voting stage: if at least one of the two partners votes not to carry out the production of the public good in the second stage, then, instead of restarting the procedure as in Harstad (2021a), both receive a pre-determined disagreement payoff. These three treatments differ in the value of the disagreement payoff for each participant, and the availability of information on these values for the two players.

We determine the disagreement payoffs in three different ways. In the *fixed disagreements payoffs treatment*, we fix them at the players' per-period endowment. In the uncertainty treatment, the disagreement payoffs can be high or low, randomly determined with an expected value equal to the players' per-period endowment, and communicated to both players after the pledge but before the review stage. This introduces uncertainty in a way similar to Harstad (2021a), by increasing the variance of outcomes in case of disagreement. However, this treatment adds not just uncertainty, but also heterogeneity, which can also affect choices for both strategic and behavioural reasons (see Section 2). To disentangle the impact of these two factors, we implement an intermediate treatment: in the *certainty treatment*, we randomly determine the disagreement payoffs in an identical manner to the uncertainty treatment, but communicate them to both players before the pledge stage.

Assigning a high disagreement payoff to a subject in the *uncertainty treatment* is akin to a negotiation partner in Harstad (2021a) who is unexpectedly patient. Such a partner is more inclined to reject a given set of pledges than one who is unexpectedly impatient, which we implement by assigning low disagreement payoffs to subjects in the experiment. Our theoretical predictions mirror Harstad's, with zero (or low) equilibrium pledges when disagreement payoffs are fixed, and pledges equal to the entire endowment when disagreement payoffs are uncertain.

Our empirical findings deviate substantially from our theoretical predictions. First, we find that adding the review stage significantly increases the efficiency of outcomes even without uncertainty over the partners' disagreement payoffs. Second, we find that uncertainty *reduces* the effectiveness of the institution, although it is still beneficial. This cannot be attributed simply to normative conflict resulting from heterogeneity, as removing the uncertainty while maintaining heterogeneity, if anything, improves outcomes.

Pointing towards a possible explanation of our results, we find that subjects are kinder in the treatment with uncertainty than in the treatments without: subjects in the former are more prone to voting against their own financial interests to implement an agreement that benefits their partner, and less prone to sacrificing a beneficial agreement in order to punish a low-contributing partner. This tendency towards more kindness in the treatment with uncertainty suggests that subjects not only hide behind the excuse of uncertainty not to contribute to a public good,[3] but grant this excuse also to others. In the context of pledge-and-review, this kindness backfires: It leads to less than subgame-perfect equilibrium contributions in the case of uncertainty and counteracts the mechanism described in Harstad (2021a).

A noticeable feature of our data is that the positive effects of P&R in the treatment without uncertainty or asymmetries in disagreement payoffs persist for the full twenty repetitions of the game. The usual decline in contributions in public goods games, attributed by Fischbacher and Gächter (2010) to "conditional cooperators" reacting to lower than hoped for contributions by others, is thus avoided. P&R effectively allows for costly second-party punishment, an institution often found to maintain cooperation, through voting down agreements. An advantage of P&R over the usual implementation of costly punishment is that it discourages "antisocial punishment" of cooperators (Herrmann et al., 2008) by endogenously increasing the cost of punishment for those who make low pledges and decreasing the cost for those who make high pledges.[4] Efficiency declines in the other two P&R treatments, but at a slower rate than in the treatment without voting.

In summary, we find that a pledge-and-review institution can increase efficiency in a public goods environment, and that this is robust to heterogeneity and information conditions. Our results suggest that the pledge-and-review institution may be useful, even without uncertainty. It allows for behaviourally reasonable strategies with higher contributions than are typically sustained in public goods games by enabling conditional cooperators to test the cooperativeness of their partners before committing to a particular contribution level.

2. Related literature

Harstad (2021a,b) model the pledge-and-review bargaining procedure associated with the climate negotiations in the Paris Agreement, December 2015.[5] Harstad writes, "... the negotiations leading up to the 2015 Paris Agreement on climate change have been characterized as "pledge and review" (P&R). Before the agreement was signed, each party was asked to submit an intended nationally determined contribution... Any individual country can always decide to not ratify the treaty, after observing the vector of pledges. Thus, in the absence of a world government, the set of contributions must be

[3] This parallels insights in Exley (2015), which finds that subjects use uncertainty as an excuse not to give charitable contributions.

[4] For example, in our game if one player contributes everything and the other only half their endowment, by voting down the agreement the former sacrifices 1.5 ECU to reduce the latter's payoff by 7.5 ECU, whereas the latter would have to sacrifice 7.5 ECU to destroy 1.5 ECU of the former.

[5] Note, even though the Paris climate negotiations inspired the model in Harstad (2021a), there are numerous other applications that can be modelled in this fashion, as articulated in Harstad (2021a).

acceptable by everyone that contributes. The 2009 negotiations in Copenhagen, for example, failed because of objections from a small set of countries."

Harstad (2021a,b) argue that the switch from the bargaining procedure followed in the Kyoto Protocol to that followed in the Paris Agreement was one from a top-down approach that attempted to pressure governments into cutting greenhouse gas emissions (GHG) to a bottom-up approach in which countries themselves determine their GHG cuts nationally. Theoretically, Harstad (2021a) shows this means going from an outcome associated with the Nash Bargaining Solution to one where each country's contribution maximizes an asymmetric Nash product with weights on other countries' pay-offs that are smaller than in the Nash Bargaining Solution. As a result, the GHG cuts pledged are typically smaller than those achieved in the top-down approach. However, because pledged cuts are smaller, Harstad (2021b) shows that more countries will endogenously participate, leading to higher investments in green technology and higher aggregate emissions. The crucial ingredient for this prediction is that countries pledge to cut below business-as-usual emissions. While Harstad (2021a,b) provide a theory of how these cuts come about and, thereby, provides a theoretical justification for the use of the pledge-and-review procedure, we complement this approach by testing the performance of stylised versions of the pledge-and-review procedure in the experimental laboratory.

The dynamic game presented in Harstad (2021b), which rationalizes the Paris Agreement, builds on the framework developed in Dutta and Radner (2004, 2006), Harstad (2012, 2016), and Battaglini and Harstad (2016), and complements Dutta and Radner (2019). Dutta and Radner (2019), in particular, show that the Green Climate Fund can lead to efficiency and, thereby, make the Paris Agreement successful. Our paper is, thus, not only complementary to Harstad (2021a,b), but also to Dutta and Radner (2019).

Our main innovation over the existing experimental literature on public goods contributions games is the introduction of the *review stage*. The review allows participants to reject partner contributions that are too low. Ex-post, it thereby gives an opportunity to punish low partner contributions whereas, ex-ante, it provides insurance against the event in which a player makes high contributions to a public good while their partners contribute little. We thereby contribute to the literature that tests various institutional innovations with the aim to overcome the under-provision of voluntary public goods contributions. The experimental literature on public goods games is vast and we focus on the two most relevant strands: voting and heterogeneity.

The lion's share of experiments introducing voting into public goods games relate to the endogenous choice of institutions, with voting taking place prior to contribution decisions.[6] These experiments have no direct relevance to our study. Of the experiments where voting takes place after contribution decisions, most democratize punishment decisions. They test the impact of different voting rules on who gets punished (Ambrus and Greiner, 2019; Casari and Luini, 2009; Cinyabuguma et al., 2005; Van Miltenburg et al., 2014), and by how much (Decker et al., 2003). In our two-player setting, punishment decisions can only be taken by an individual, so these studies have no direct bearing on our own. A more related study is le Sage and van der Heijden (2015), which also involves voting after contributions are made known. However, in contrast to our study, here the public good is always provided according to the contribution decisions. Voting determines whether or not players retain the portion of their endowment they did not contribute, that is, whether they punish low contributions. Like us, they find that this additional stage ameliorates the typical decline in contributions over time. In contrast to our study, the initial contributions in the treatments with voting are not higher than in those without.

Heterogeneity in public goods games can inhibit cooperation by creating "normative conflict" where multiple norms reasonably apply (e.g., Nikiforakis et al., 2012). In the presence of normative conflict, cooperation can deteriorate either because of failure to coordinate on a norm, or because different norms are favoured by different individuals. In contrast to homogeneous public goods games, heterogeneity implies that normative rules such as equality of wealth or income no longer always coincide with equality of absolute or relative contributions. Heterogeneity in wealth levels (Anderson et al., 2008; Cardenas, 2003), endowments (e.g., Cherry et al., 2005; Heap et al., 2016; Zelmer, 2003), and marginal per capita returns (MPCR) on contributions (Fischbacher et al., 2014; Nikiforakis et al., 2012) typically results in lower contributions. However, this has been shown to depend on the linearity of the production function (Chan et al., 1996, 1999), the institutions (Kingsley, 2016), the information setting (Fellner-Röhling et al., 2020), and whether the heterogeneity in the MPCR is rooted in heterogeneous individual costs of contributing or in heterogeneous benefits of the public good (Kölle, 2015). While clearly related, it is not obvious *ex-ante* what the findings of this literature imply in our setting, where *conditional on the public good being provided*, the game is symmetric.

More apparent is the relationship of our study with bargaining experiments where disagreement payoffs vary across players. Here normative conflict can arise from the tension between equalizing income and equalizing gains from trade. Ultimatum game experiments consistently find that heterogeneity in disagreement payoffs leads to more rejections (e.g., Hennig-Schmidt et al., 2008; Knez and Camerer, 1995). Hennig-Schmidt et al. (2018) keep either proposer or responder's disagreement payoff fixed at a low level, then gradually increase the other's: again greater heterogeneity increases disagreement rates, but there is no effect on efficiency as the higher disagreement payoffs compensate, similar to our findings. In a three-player multilateral bargaining experiment, Miller et al. (2018) find that under unanimity rule, immediate agreement is more common with homogenous than heterogenous disagreement payoffs, but no difference under majority rule.

[6] For a survey of this literature, see Dannenberg and Gallier (2019).

Table 1
Payoff table for implemented public goods contributions.

Your own investment in the joint project (ECU)	Other participant's investment in the joint project (ECU)						
	0	**2**	**4**	**6**	**8**	**10**	**12**
0	12 / 12	11.5 / 13.5	11 / 15	10.5 / 16.5	10 / 18	9.5 / 19.5	9 / 21
2	13.5 / 11.5	13 / 13	12.5 / 14.5	12 / 16	11.5 / 17.5	11 / 19	10.5 / 20.5
4	15 / 11	14.5 / 12.5	14 / 14	13.5 / 15.5	13 / 17	12.5 / 18.5	12 / 20
6	16.5 / 10.5	16 / 12	15.5 / 13.5	15 / 15	14.5 / 16.5	14 / 18	13.5 / 19.5
8	18 / 10	17.5 / 11.5	17 / 13	16.5 / 14.5	16 / 16	15.5 / 17.5	15 / 19
10	19.5 / 9.5	19 / 11	18.5 / 12.5	18 / 14	17.5 / 15.5	17 / 17	16.5 / 18.5
12	21 / 9	20.5 / 10.5	20 / 12	19.5 / 13.5	19 / 15	18.5 / 16.5	18 / 18

Closely related to our paper is Reischmann and Oechssler (2018). They introduce the Binary Conditional Contribution Mechanism (BCCM) and explain its merits relative to other public good provision mechanisms (Voluntary Contribution Mechanism (VCM), the Provision Point Mechanism (PPM), Vickrey-Clarke-Groves Mechanisms (VCG), Contractive Mechanisms, and Auctions and Lotteries that reward (stochastically) individual contributions to a public good.) In BCCM, agents simultaneously make conditional contribution offers that take the form "I am willing to contribute to the public good if at least k agents contribute in total." P&R is, theoretically, a generalization of this mechanism, moving from binary to continuous, and allowing contributions to be conditioned on not only the number of other contributors, but also the level of other contributions. A second difference is that the decision on whether to make good on a pledge is made after observing others' proposed contributions. While this makes no difference in theory, it may do so behaviourally. Indeed, it is in exactly this kind of situation (where costly punishment is possible) that strategies specified in advance (i.e., using "the strategy method" in experiments) tend to differ from decisions made after observing other players' actions in otherwise identical environments (Brandts and Charness, 2011). Both these differences bring the model closer to the applications discussed above.

3. Theory and hypotheses

Setup We implement in the laboratory a much simplified two-player, two-stage game version of the model in Harstad (2021a), which captures the same basic mechanism and main result. We begin with a standard two-player public goods game, then add, one at a time, voting, heterogeneity in disagreement payoffs, and finally uncertainty over disagreement payoffs.

The *public goods game treatment*, implements a standard two-player public goods game. Two players $i = 1, 2$ simultaneously choose contributions $x_i \in X_i = \{0, 2, 4, 6, 8, 10, 12\}$ to a public good. The players' payoffs are given by

$$U_{i,A} = 12 - x_i + \frac{3}{4}(x_i + x_j),\ i, j = 1, 2,\ i \neq j.$$

Table 1 illustrates the payoffs from an implemented public goods provision.

In the remaining treatments, these choices are *pledges*, which are only implemented if both players agree in a second voting stage. In this second stage, players observe both pledges and then simultaneously vote whether to accept or to reject the set of pledges. We denote player i's voting decision by $v_i \in \{0, 1\}$, where $v_i = 0$ corresponds to rejecting and $v_i = 1$ to accepting the pledges. If both vote to accept the pledges, the contributions are made and they receive agreement payoffs, $U_{i,A}$ as in the public goods game. If at least one player votes against, they receive disagreement payoffs of

$$U_{i,D} = d_i,\ i = 1, 2.$$

In the *fixed disagreement payoffs treatment*, the disagreement payoffs are given by $d_i = 12$. In the final two treatments, the disagreement payoff are independently and identically distributed random variables, which take the values

$$d_i = \begin{cases} 6 & \text{with probability } \frac{1}{2}, \\ 18 & \text{with probability } \frac{1}{2}. \end{cases}$$

In one of these two treatments, the *certainty treatment*, players learn the realizations of their and their partner's disagreement payoffs before they pledge their contributions. In the other one, the *uncertainty treatment*, they only learn these realizations only after they pledged their contributions but before they vote whether to accept or to reject the pledges. We chose the parametrization of the disagreement payoffs to maximize the predicted difference in outcomes between the treatments.

Equilibria and hypotheses In the *public goods game treatment*, the unique Nash equilibrium is, as usual, one with zero contributions, i.e., $x_i = 0,\ i = 1, 2$.

In the two-stage games, we solve for subgame perfect Nash equilibria in which subjects vote as if they were pivotal.[7] In the review stage the players have the same information irrespective of whether they are in the certainty or the uncertainty treatment. In the review stage, it is in the players' interest to accept a set of pledges if $U_{i,A} \geq U_{i,D}$, and to reject a set of pledges if $U_{i,A} \leq U_{i,D}$.

In the *fixed disagreements payoffs treatment*, $U_{i,D} = d_i = 12$. Hence, it is optimal for players to vote $v_i = 1$ if $U_{i,A} \geq 12$ and $v_i = 0$ if $U_{i,A} \leq 12$. Taking this voting rule into account, there are two subgame perfect Nash equilibria. In the first, both players pledge contributions of zero and are indifferent between accepting and rejecting, i.e., $x_i = 0,\ i = 1, 2$ and $v_i \in \{0, 1\},\ i = 1, 2$. In the second, both players pledge to contribute two units and both vote to accept, i.e., $x_i = 2,\ i = 1, 2$ and $v_i = 1,\ i = 1, 2$.

In the *certainty and the uncertainty treatments*, player i either has a disagreement payoff of $U_{i,D} = d_i = 6$ or one of $U_{i,D} = d_i = 18$. With $x_i \in X_i, i = 1, 2$, any agreement payoff satisfies $U_{i,A} \geq 9$. Players with $d_i = 6$, therefore, strictly prefer to accept any set of pledged contributions to the public good. Players i with $d_i = 18$, on the other hand, strictly prefer to accept sets of pledges (x_i, x_j) such that $12 - x_i + 3(x_i + x_j)/4 > 18 \Leftrightarrow x_j > 8 + x_i/3$, they strictly prefer to reject sets of pledges (x_i, x_j) such that $x_j < 8 + x_i/3$, and they are indifferent between accepting and rejecting if $x_j = 8 + x_i/3$.

Hence, in the *certainty treatment*, when $d_1 = d_2 = 18$, the subgame perfect Nash equilibria are such that either $x_i = 12,\ i = 1, 2$ and $v_i \in \{0, 1\},\ i = 1, 2$; or such that at least one player contributes $x_i < 12$ and the other player votes $v_j = 0$. When $d_1 = 6$ and $d_2 = 18$, there are two reasonable equilibria: one where $(x_1, x_2) = (8, 0)$ or $(x_1, x_2) = (10, 0)$ and $v_i = 1,\ i = 1, 2$. Finally, when $d_1 = d_2 = 6$, the only subgame perfect Nash equilibrium involves pledges $x_i = 0,\ i = 1, 2$ and votes $v_i = 1,\ i = 1, 2$.

In the *uncertainty treatment*, when players pledge their contributions, they have not yet learned the disagreement payoffs and, therefore, have to reason in expectations. Then, as long as it is sufficiently likely (probability greater than $1/4$) that the other player has a high disagreement payoff, in the subgame perfect Nash equilibrium, players pledge $x_i = 12$. They vote $v_i = 1$ for $i = 1, 2$.[8] High pledges serve as an insurance against the case in which a player has a low and their partner has a high disagreement payoff.

We are now ready to state our hypotheses. We begin with choices in the pledge and review stages, which clearly only apply to P&R treatments. We then look at contributions and final payoffs.

In developing our hypotheses, when there are multiple equilibria we predict average data will fall between the minimum and maximum predicted by the different equilibria (the precise number depending on how often each equilibrium is selected). In the treatments with heterogenous disagreement payoffs, we take averages of the values each of the four, equally likely combinations of disagreement payoffs.

Let superscripts denote the public goods (PG), the fixed disagreement payoffs (F), the certainty (C), and the uncertainty (U) treatments; and let subscripts $i, j \in \{L, H\}$ in the certainty treatment denote the disagreement payoffs of an individual i and their partner j where L refers to the low disagreement payoff of 6, and H the high disagreement payoff of 18. The predicted intervals for average pledges are illustrated in Fig. 1.

[7] Formally, in the *fixed* and *certainty treatments* with given disagreement payoffs, a strategy is a pair (x_i, v_i), where v_i maps $X_i \times X_j \mapsto \{0, 1\}$. In the *uncertainty treatment*, strategies must specify voting decisions for each of the four possible combinations of disagreement payoffs, so here v_i instead maps $X_i \times X_j \times D_i \times D_j \mapsto \{0, 1\}$. When we describe the subgame-perfect Nash equilibria, we only present the equilibrium path. The off-path equilibrium actions in the voting stage are readily computable. The restriction that subjects vote as if they were pivotal is behaviourally reasonable. It rules out trivial equilibria in which both subjects vote no irrespective of the pledges. That includes equilibria, in which both players pledge full contributions and if pledges are different from $(12, 12)$, both players vote no, i.e., $v_i = 0$. Alternatively, we could require a version of trembling-hand perfection or that players do not play weakly dominated strategies.

[8] There are also equilibria where $x_i = 12,\ v_i = 1$ if $d_i = 6$, and $v_i \in \{0, 1\}$ if $d_i = 18$. These equilibria involve the player with a high disagreement payoff punishing the player with a low disagreement payoff with probability $p_i \in (0, \frac{1}{2})$ for pledging their full endowment (see Online Appendix B). We view these equilibria as behaviourally unlikely and will disregard them for our hypotheses. Indeed, in our experiment this only occurred three times out of 414 observations with asymmetric disagreement payoffs in the certainty treatment and once out of 395 observations with asymmetric disagreement payoffs in the uncertainty treatment.

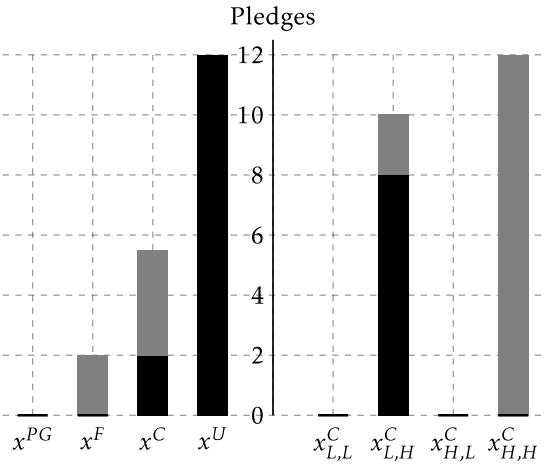

Fig. 1. Predictions of average pledges. Height of black columns: Minimum average equilibrium pledges. Height of grey columns: Maximum average equilibrium pledges. Left panel: Average equilibrium pledges by treatment: $x^{PG} = 0$, $x^F \in [0, 2]$, $x^C \in [2, 5.5]$, and $x^U = 12$. Right panel: Average equilibrium pledges in the certainty treatment by disagreement pay-off pairs: $x^C_{L,L} = x^C_{H,L} = 0$, $x^C_{L,H} \in [8, 10]$, and $x^C_{H,H} \in [0, 12]$.

Hypothesis 1. *Pledges:* $x^F \in [0, 2]$, $x^C \in [2, 5.5]$, $x^U = 12$. Average pledges in the four treatments can be ordered as follows: $x^F \leq x^C < x^U$.

Let a be proportion of agreements. Then the theory presented above yields $a^F \in [0, 1]$, $a^C \in [0.75, 1]$, $a^U \in [0.75, 1]$, that is, no useful treatment comparisons. In the asymmetric treatments however, voting yes is always strictly preferred when both players have low disagreement payoffs, sometimes strictly preferred when disagreement payoffs are different, and never strictly preferred when both have high disagreement payoffs. We therefore predict that $a^t_{LL} \geq a^t_{LH} = a^t_{HL} \geq a^t_{HH}$ for treatment $t \in \{C, U\}$. Note that this implies that agreements are more likely the greater are the possible efficiency gains. We can also test the assumption of self-interested voting – based on many laboratory experiments, one would expect social preferences to play a role. Finally, for given pledges, in the absence of non-standard preferences, voting behaviour should be the same in C and U.

Hypothesis 2. *Review:*

(a) $a^F = a^C = a^U$ (no difference in agreement rates across treatments).
(b) $a^t_{LL} \geq a^t_{LH} = a^t_{HL} \geq a^t_{HH}$ for $t \in \{C, U\}$ (agreements are more common when greater efficiency gains are possible).
(c) $v_i = 1$ if $U_{i,A}(x_i, x_j) \geq U_{i,D}(x_i, x_j)$ and $v_i = 0$ otherwise (subjects vote to maximize payoff).
(d) $v^C_i(x_i, x_j, d_i, d_j) = v^U_i(x_i, x_j, d_i, d_j)$ (conditional on disagreement payoffs and pledges, voting behaviour is the same in C and U).

The theory also allows us to rank the contributions that are implemented in the pledge-and-review procedure. Denote these implemented contributions by $\hat{x}$.

Hypothesis 3. *Contributions:* $\hat{x}^{PG} = 0$, $\hat{x}^F \in [0, 2]$, $\hat{x}^C \in [2, 5.5]$, $\hat{x}^U \in [9, 12]$. Average contributions in the four treatments can be ordered as follows: $\hat{x}^{PG} \leq \hat{x}^F \leq \hat{x}^C < \hat{x}^U$.

Given that we are ultimately interested not in the size of contributions themselves, but in efficiency, we also make predictions about the expected payoffs. In PG there are never positive contributions, so payoffs equal endowments, whereas in F payoffs vary from 12 to 13, depending on how often each equilibrium is played. In C, efficiency gains come from both agreements between pairs of players with low disagreement payoffs, and the contributions of players with a low disagreement payoff who are matched with a high disagreement payoff player. Hence, overall, expected payoffs vary from 14.5 to 14.75. Expected payoff in U, however, is 18: in all matchings, players with low disagreement payoffs gain 12 while those with high disagreement payoffs gain nothing. These results are show in Fig. 2.

Hypothesis 4. *Payoffs:* Average payoffs in the four treatments can be ordered as follows: $G^{PG} \leq G^F < G^C < G^U$.

As in Harstad (2021a), pledges, contributions, and efficiency are higher when there is uncertainty. Note also that our C treatment does not correspond to the setting with certainty in Harstad (2021a); rather our F treatment does. As in that paper, in the F treatment, pledges, contributions and efficiency are predicted to be (near) zero. The fact that contributions in F

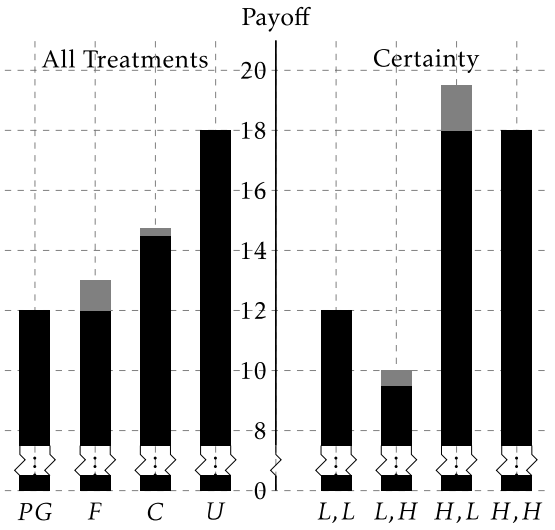

Fig. 2. Predictions of expected equilibrium payoffs. Height of black columns: Minimum (average) equilibrium payoffs. Height of grey columns: Maximum (average) equilibrium payoffs. Left panel: Equilibrium payoffs by treatment. Right panel: Equilibrium payoffs in the certainty treatment by disagreement pay-off pairs.

are predicted to be weakly higher than in PG as an artefact of the discrete nature of the choice set. In a continuous setting, voting would not increase contributions in this treatment. In Online Appendix A, we show that the model predictions do not change substantially when we account for preferences with inequity aversion *á la* Fehr and Schmidt (1999).

4. Experimental design

The experiment consisted of precise implementations of each of the four models described in the previous section. Instructions, which can be found in Online Appendix G along with screenshots, were written in neutral language.

We implemented a between subject design with 16 sessions, four for each treatment. In each session, with one exception, there were 20 subjects; each assigned to one of two matching groups of ten participants. In one of the sessions, in which we ran the fixed disagreement payoffs treatment, we had one matching group of ten subjects. We, therefore, had eight independent observations for the certainty, uncertainty, and public goods treatments and seven independent observations for the fixed disagreement payoffs treatment. Subjects played 20 rounds in each session with random re-matching within their matching group in each round. In the uncertainty and certainty treatment, the disagreement payoffs were randomly determined after each re-matching, so subjects' disagreement payoffs, and those of their partner, varied from round to round. After each round, subjects received full feedback on their own and their partner's choices and payoffs.

After reading the instructions, subjects had to correctly answer a series of control questions before proceeding. After playing the 20 rounds, the experiment concluded with a questionnaire including demographic information, questions relating to the motivation for the subjects' choices and their perception of the motivation of other subjects, self-evaluation questions pertaining to the subjects' risk, time, and other-regarding preferences (Falk et al., 2016), as well as the cognitive reflection test (Frederick, 2005).

The experiment was conducted at the University of Auckland Business School's DECIDE computer lab. Participants were recruited via ORSEE (Greiner, 2015) and the procedures were computerized using z-Tree (Fischbacher, 2007). Subjects were paid according to their performance in a randomly picked round of the session. Experimental currency units (ECU) denote the payoffs in the experiment and, with an exchange rate of 1.5, they were paid out to subjects in New Zealand dollars (NZD). Sessions lasted around 70 minutes, and subjects earned on average NZ$23. The experiment was approved by the University of Auckland Human Participants Ethics Committee on 12 November 2018 for 3 years with the Reference Number 022284.

5. Results

We are mainly interested in testing Hypotheses 3 and 4, which regard contributions and efficiency. However, to gain some insight into what underlies our main results, we first report results on pledges and voting in the three voting treatments.

All non-parametric tests use matching-group averages as independent observations. We use Mann-Whitney (MW) tests to test for differences in distributions, and stochastic inequality tests (SIT: Schlag, 2008) when we wish to make a directional

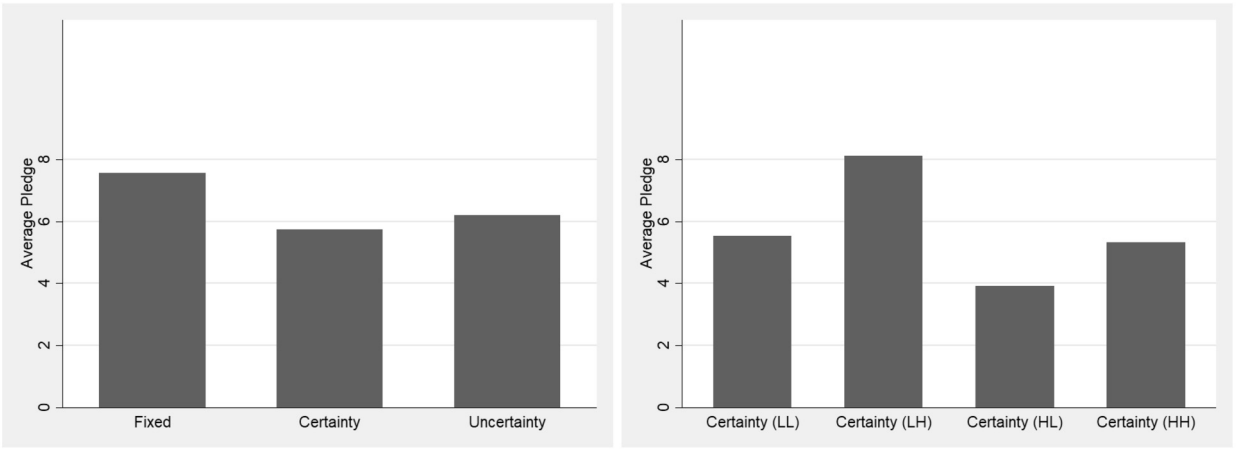

Fig. 3. Average pledges.

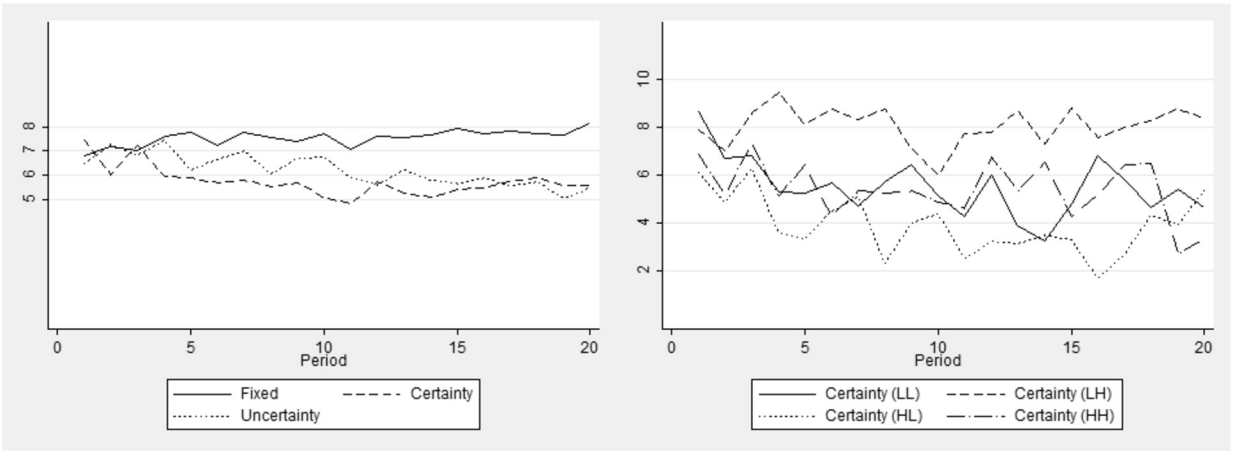

Fig. 4. Average pledges by period.

claim.[9] When comparing pledges to theoretical predictions we use two-tailed binomial tests (BT), i.e. average pledges are significantly greater/less than a number ($p < 0.01$) if the relationship holds in 8/8 matching groups, and weakly different ($p = 0.070$) if the relationship holds in 7/8 matching groups. In F we only have a significant result when 7/7 matching groups are in line ($p = 0.016$).

Standard errors in all regressions are clustered at the matching-group level and include subject random effects. Results on time trends are based on one of the two regressions in Online Appendix D, depending on whether or not the data is broken down by player-type in the heterogenous treatments.

5.1. Pledges

Fig. 3 plots the average pledges in each of the voting treatments.[10] Two departures from equilibrium predictions are immediately apparent. First, the average pledge in F of 7.6 is substantially higher than the highest equilibrium pledge of 2. Second, the average pledge in U is only 6.2 in contrast to the equilibrium prediction of 12. These results are robust, obtaining in all matching groups: average pledges are statistically greater than 5.3 in F (BT: $p = 0.016$), and less than 8 in U (BT: $p < 0.01$). They are also not diminished by learning, as can be seen in Fig. 4 which plots average pledges over the 20 periods: in fact, if anything, pledges in each treatment move *away* from equilibrium predictions. Linear time trends are not statistically significant in F, but negative in U ($p < 0.01$).

[9] Unless the two distributions being compared have identical shapes and differ only in location, the Mann-Whitney test is only an exact test of differences in distributions (Wilcox, 2001, pages 231-232). Rejection can result from, for example, a difference in variance, skewness, or kurtosis, even where there is no difference in mean or median (e.g. Fagerland and Sandvik, 2009). The stochastic inequality is an exact test that allows for directional inference, assuming only independence of observations (Schlag, 2015).

[10] The full distribution of pledges for each treatment can be found in Online Appendix C.

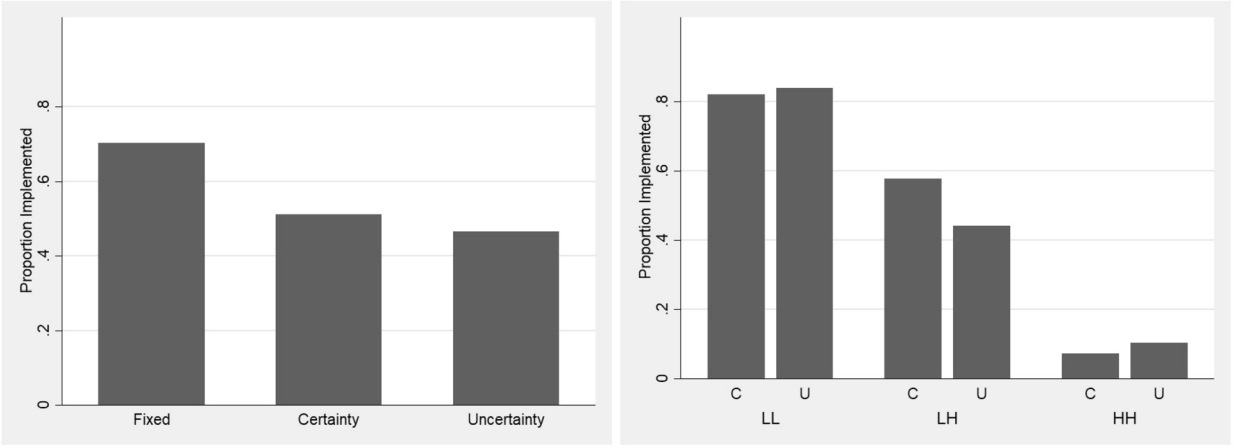

Fig. 5. Proportion of agreements.

The average pledge in C of 5.7 is marginally above the maximum equilibrium prediction of 5.5, but not statistically different according to a binomial test (five sessions are above, and three below). Breaking this down into the four types of games within this treatment we can see that the higher than expected pledges are a result of subjects in C_{LL} and C_{HL} pledging substantially more than the predicted zero: average pledges are greater than 5.5 in C_{LL} (BT: $p < 0.01$), and greater than 3.9 in C_{HL} (BT: $p < 0.01$). Average pledges in C_{LH} at 8.1 were within the predicted range of 8-10. In C_{HH}, any choice could be consistent with some equilibrium, and were 5.3 on average. There is a negative time trend overall ($p = 0.030$), with the only significant trend for C_{LL} ($p = 0.042$).

Mann-Whitney tests find statistically significant differences in distributions of pledges between F and both C ($p = 0.021$) and U ($p = 0.083$), but not between C and U. Stochastic inequality tests show that average pledges tend to be higher in F than C ($p = 0.070$), but find no evidence of a directional difference at conventional levels between F and U.

Result 1. *Pledges:* Pledges are higher in F and lower in U than predicted. Heterogeneity in disagreement payoffs alters the distribution of pledges, and reduces pledges in the presence of uncertainty.

5.2. Review

Fig. 5 shows the proportion of games where there is an agreement, that is, both subjects vote yes.[11] In F, 70% of games end in agreement, significantly more according to stochastic inequality tests than the 53% in C ($p = 0.016$) or the 46% in U ($p = 0.016$). The difference between the latter two treatments is not significant. It should be noted, however, that the probability of an agreement in the treatments with heterogenous disagreement payoffs increases dramatically with potential efficiency gains, from around 10% when both are high, to 40-60% when mixed, to over 80% when both are low. The signs of these differences hold within *every* matching group in both C and U, and are therefore all statistically significant (sign test: $p < 0.01$).

Fig. 6 shows the proportion of successful agreements over time. There is a weakly significant increase in F ($p = 0.054$) and decrease in U ($p = 0.054$). The latter is driven by the decline in games where both subjects have high disagreement payoffs ($p < 0.01$).

Voting, by and large, reflected self-interest, with 90% of votes across the treatments consistent with payoff-maximization. Of the remainder, 5.9% were consistent with costly punishment (reducing both own and partner's payoff), while 3.8% were altruistic (reducing own payoff but increasing partner's). Only 10 of the 4600 votes were clearly irrational (reducing own payoff while not affecting partner's). Details of the proportion of subjects voting yes for each combination of pledges by treatment can be found in Online Appendix E.[12]

Fig. 7 shows costly punishment rates across treatments, which are 2-3 p.p. higher in F than the other treatments, but there is only statistical evidence of a difference with respect to C (MW: $p = 0.071$). The only combination of disagreement payoffs where there is a statistical difference between C and U is when both subjects have high disagreement payoffs (MW: $p = 0.035$), with subjects in the latter punishing twice as often. These results are robust to adjusting for one's own and one's partners contributions (see Online Appendix D). The regressions find, in addition, significantly ($p < 0.01$) less costly punishment among subjects in U_{HL} compared to C_{HL}.

[11] These proportions are shown separately for each combination of pledges in Online Appendix C.

[12] What we refer to as costly punishment and altruism could also be driven by social norms, or by other social preferences, such as inequity aversion or reciprocity. We use the terms costly punishment and altruism for convenience as the data do not allow us to distinguish between these different explanations.

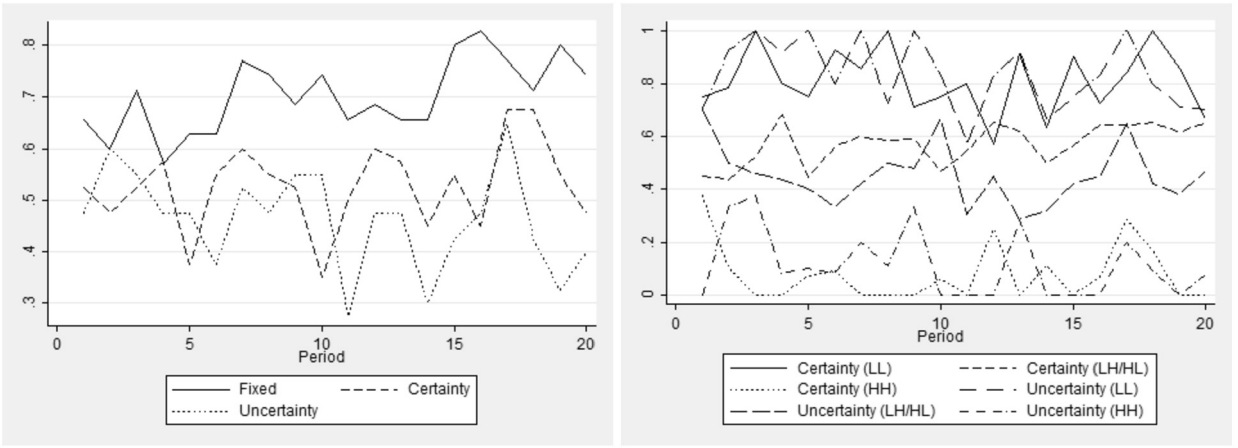

Fig. 6. Proportion of agreements by period.

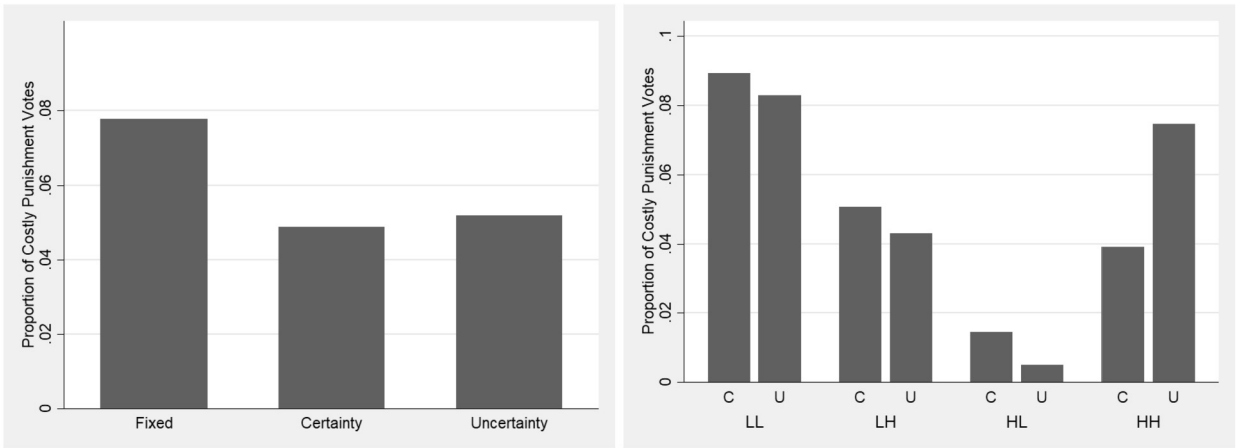

Fig. 7. Proportion of votes consistent with costly punishment.

Fig. 8 shows how altruistic voting rates vary across treatments. It is almost non-existent in F, but makes up 3.5% of votes in C, and 6.9% in U. Stochastic inequality tests show all treatment differences to be statistically significant at the 5% level or greater. Altruistic voting can never occur when both subjects have a low disagreement payoff, as agreement always leads to both players being better off, and disagreement to both being worse off. It is more common in U than C for all other combinations of disagreement payoffs, but significantly so only for subjects with a high disagreement payoff who are partnered with a subject with a low disagreement payoff (stochastic inequality: $p = 0.041$): in this case, 9% of votes are altruistic in C, and 19% in U. Again regressions adjusting for contribution levels support these results, with the additional finding of a significant increase in altruistic voting in U_{LH} compared to C_{LH}.

Figs. 7 and 8 concern relatively small numbers: none of the entries except for one is above 0.1. So, at first glance, one might caution not to over-interpret these results. However, the fact that, for example, costly punishment is not observed more often does not mean it does not play a substantial role in influencing outcomes by discouraging certain choices. Even just a 1 in 10 chance of receiving costly punishment for under-pledging is non-negligible. Finally, we think the variation of non-payoff-maximizing behaviour across treatments itself can be interesting and we report it also for that reason: In some types of games, e.g., U_{HL}, non-self-interested behaviour reaches almost 20% of all voting decisions.

Result 2. *Review:*

(a) Agreements are more common in F than either C or U.
(b) In C and U, agreements are more common when greater efficiency gains are possible.
(c) Voting largely follows self interest, but there is non-negligible evidence of costly punishment and altruism.
(d) Overall there is tendency for subjects in U to be kinder than in C, with less costly punishment (significant for HL) and more altruistic voting (significant for LH and HL). The only exception in direction of difference is costly punishment for HH.

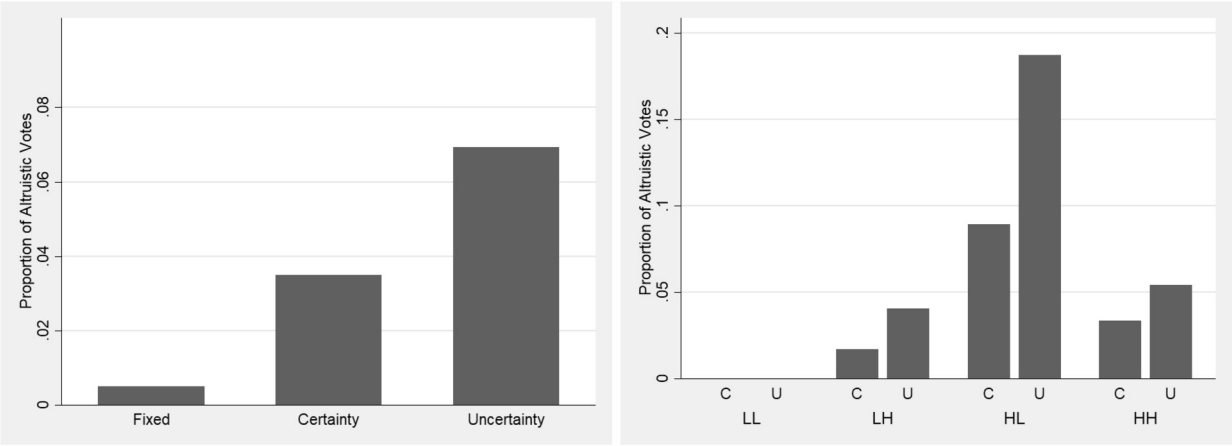

Fig. 8. Proportion of votes consistent with altruism.

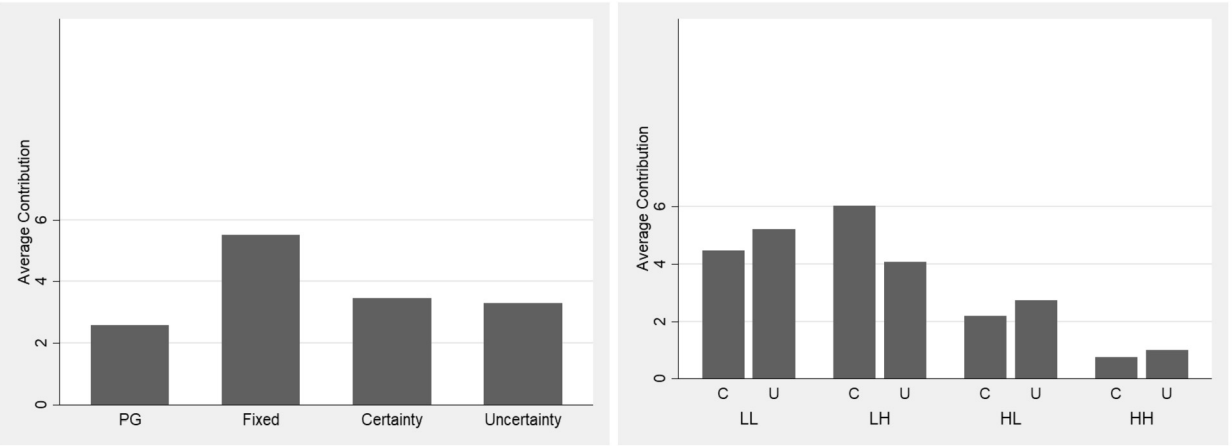

Fig. 9. Average contributions.

5.3. Contributions

With pledge-and-review, *pledges only become contributions* once they have been *ratified* in the review stage. If the pledges are not ratified, actual contributions are zero, so in all pledge-and-review treatments, *unconditional* average contributions will be lower than pledges. Average contributions *conditional on being accepted*, however, will be higher than pledges: it is the pledge vectors with low partner-pledges that will be rejected. This implies that average contributions in successful agreements will also be significantly higher than contributions in the PG treatment.

Fig. 9 shows the unconditional average contributions by treatment. Average contributions are 2.6, 5.5, 3.5, and 3.3 in PG, F, C, and U, respectively. Contributions are significantly higher in F than PG (SIT: $p = 0.021$), C (SIT: $p = 0.031$), and U (SIT: $p = 0.035$). The difference between the unconditional average contributions in F and those C or U is readily explained by subjects with $d_i = 18$, who are only present in C and U, voting down pledges that are similar to the pledges in F. The only other statistically significant treatment difference is a weak difference in distributions between PG and C (MW: $p = 0.074$), but a SIT provides no evidence of a directional effect. The only significant difference in distributions of contributions in the four types of games in C and U occurs for subjects with a low disagreement payoff matched with a subject with a high disagreement payoff, with average contributions higher in the former than the latter (MW: $p = 0.036$, SIT: $p = 0.144$).

Fig. 10 shows how contributions evolve over time. We compare the trends by regressing contributions on period, treatment dummies, and period-treatment interactions. There are significantly negative linear time trends in both PG and U, both significantly more negative than F ($p < 0.01$, $p < 0.01$) and C ($p < 0.01$, $p < 0.032$). No other pairwise comparison is statistically significant at conventional levels. In C, only with two low disagreement payoffs is there a (weakly) significant time trend ($p = 0.060$); in U, contributions fall over time when both subjects have high disagreement payoffs ($p = 0.013$), and even more dramatically when both have low ($p < 0.01$).

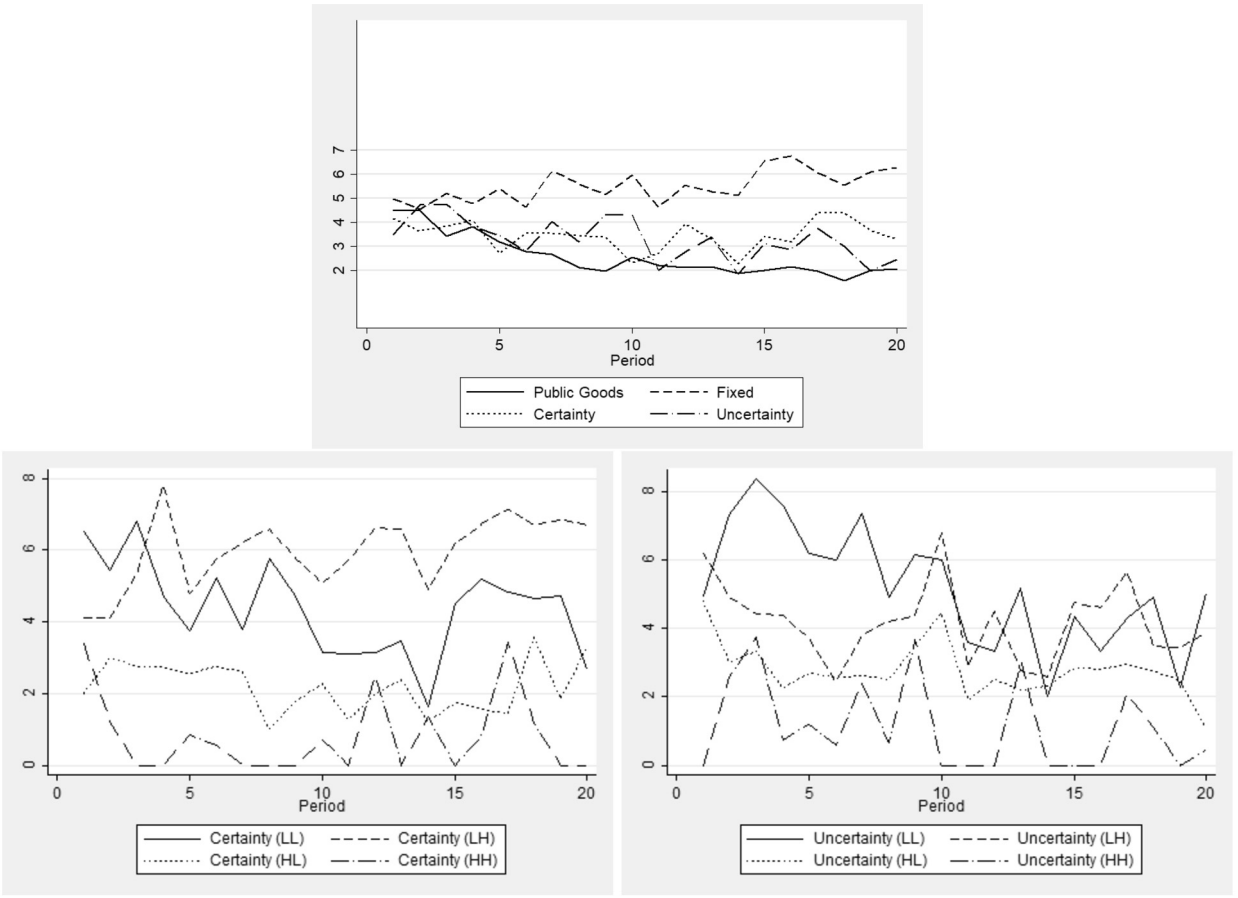

Fig. 10. Average contributions.

Result 3. *Contributions:* Contributions are significantly higher in F than PG, C or U.

Fig. 11 illustrates average contributions conditional on being accepted. Average conditional contributions are significantly lower in PG than the other three treatments (SIT: $p = 0.016$, $p < 0.01$, $p < 0.01$ for F, C, and U). We observe no significant treatment differences overall across the pledge-and-review treatments. When one player has a low and the other a high disagreement payoff, average contributions in the C treatment are higher than those in the U treatment for subjects with the low (MW: $p = 0.036$, SIT: $p = 0.138$), whereas they are lower for subjects with the high disagreement payoff (MW: $p = 0.046$, SIT: $p = 0.164$). This illustrates the relative kindness of the latter in U, where they are willing to sacrifice more by voting to agree.

Fig. 12 depicts how contributions in successful agreements vary over periods. There are negative time trends in C ($p = 0.098$) and U ($p < 0.01$). The time trend in U is significantly more negative than that in F ($p = 0.010$) but statistically indistinguishable from that in C. When we break down the C and U treatments by disagreement payoff pairs, there are negative trends in C_{LL} ($p = 0.058$) and U_{LL} ($p < 0.01$).

5.4. Efficiency

Fig. 13 illustrates the main result of our paper: There is a statistically and economically significant increase in payoffs in all three pledge-and-review treatments over the PG treatment; F (SIT: $p = 0.021$); C (SIT: $p < 0.01$); U (SIT: $p < 0.01$). In the pledge-and-review treatments, payoffs are approximately 10% higher than in the PG treatment, resulting in more than twice the efficiency gain over the no-cooperation benchmark of 12. Between the pledge-and-review treatments, we observe no significant difference in payoffs.[13]

[13] The only significant difference we obtain is for payoffs of subjects with a high disagreement payoff paired with subject with a low one. These subjects are better off in the C than in the U treatment (SIT: $p = 0.010$).

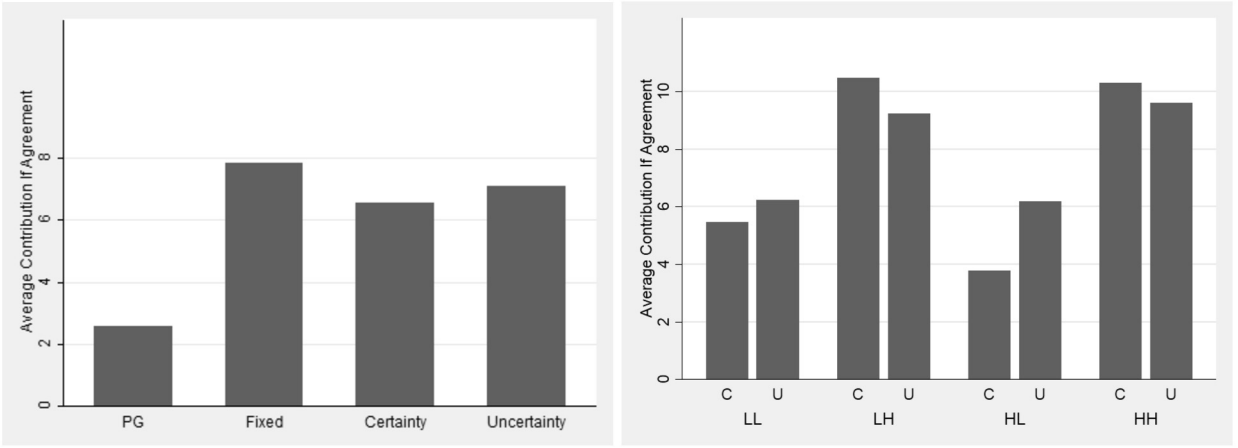

Fig. 11. Average contributions in successful agreements.

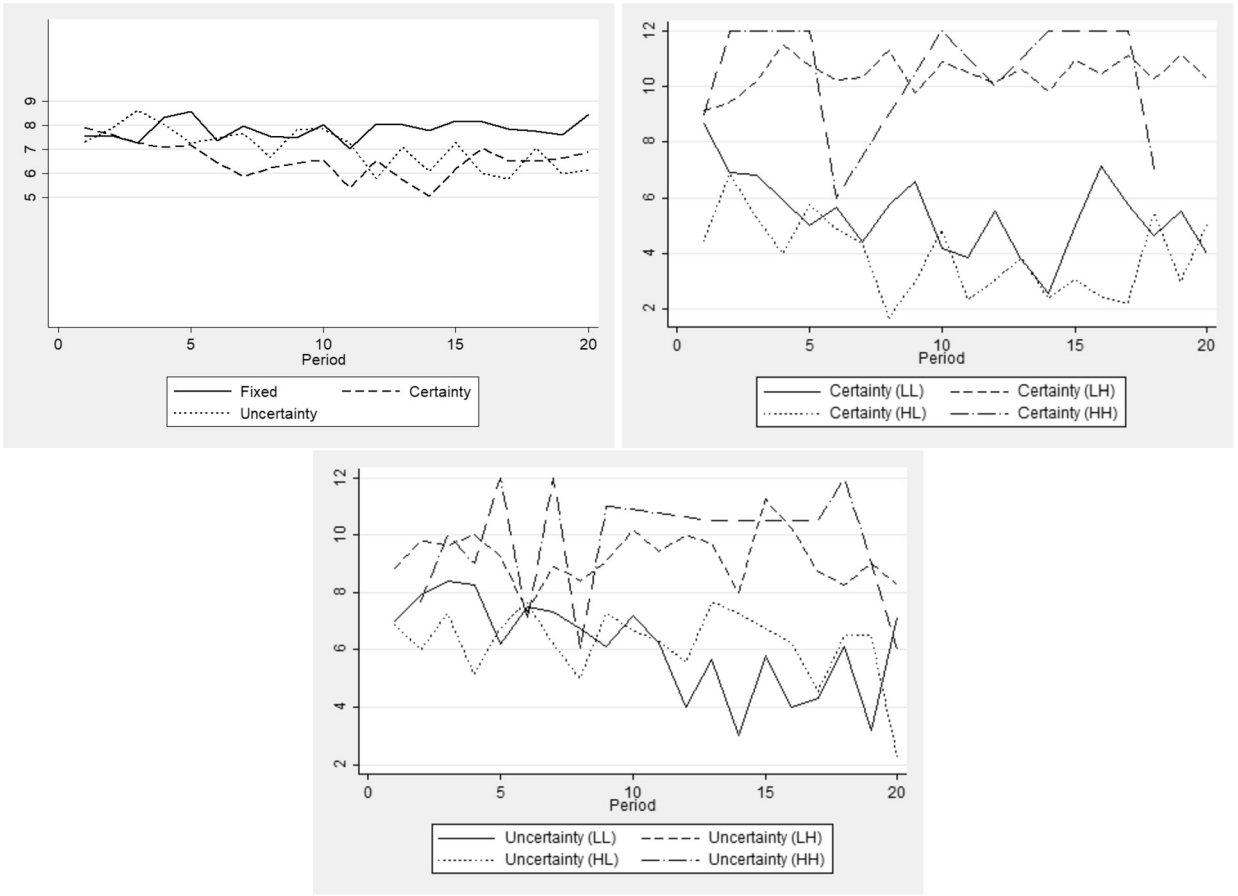

Fig. 12. Average contributions in successful agreements.

Fig. 14 shows that efficiency gains of the pledge-and-review treatments over PG obtain over all 20 periods. Indeed, because contributions in the PG treatment fall drastically over time, these relative gains increase over time. The negative time trends in PG ($p < 0.01$) and U ($p = 0.095$) are significantly more negative than the time trend in F ($p < 0.01$, $p < 0.01$); but only the trend in PG is significantly more negative than that in C ($p = 0.021$). When we separate payoffs in the C and U treatments by disagreement payoff pairs, the only significant time trend we observe is in U when both subjects have low disagreement payoffs (negative, $p < 0.01$).

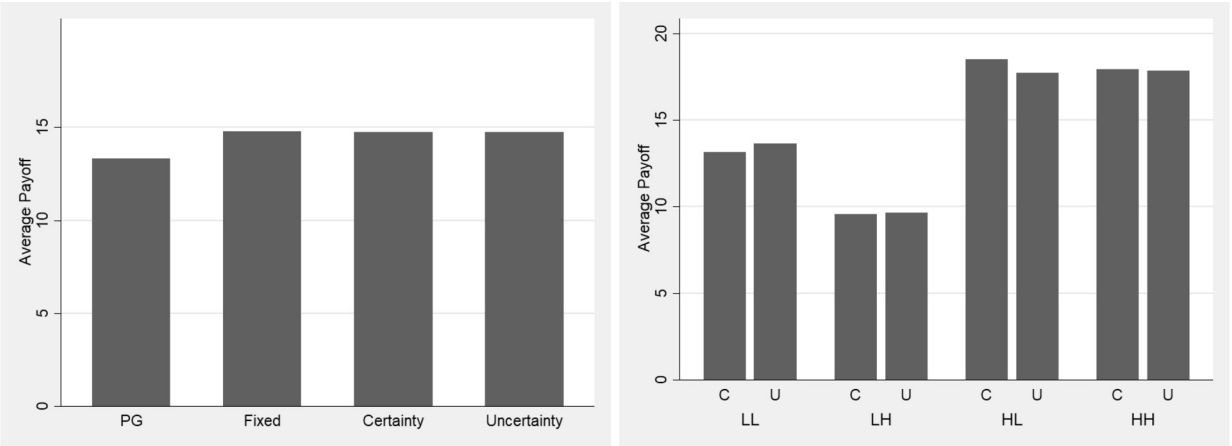

Fig. 13. Average Payoffs.

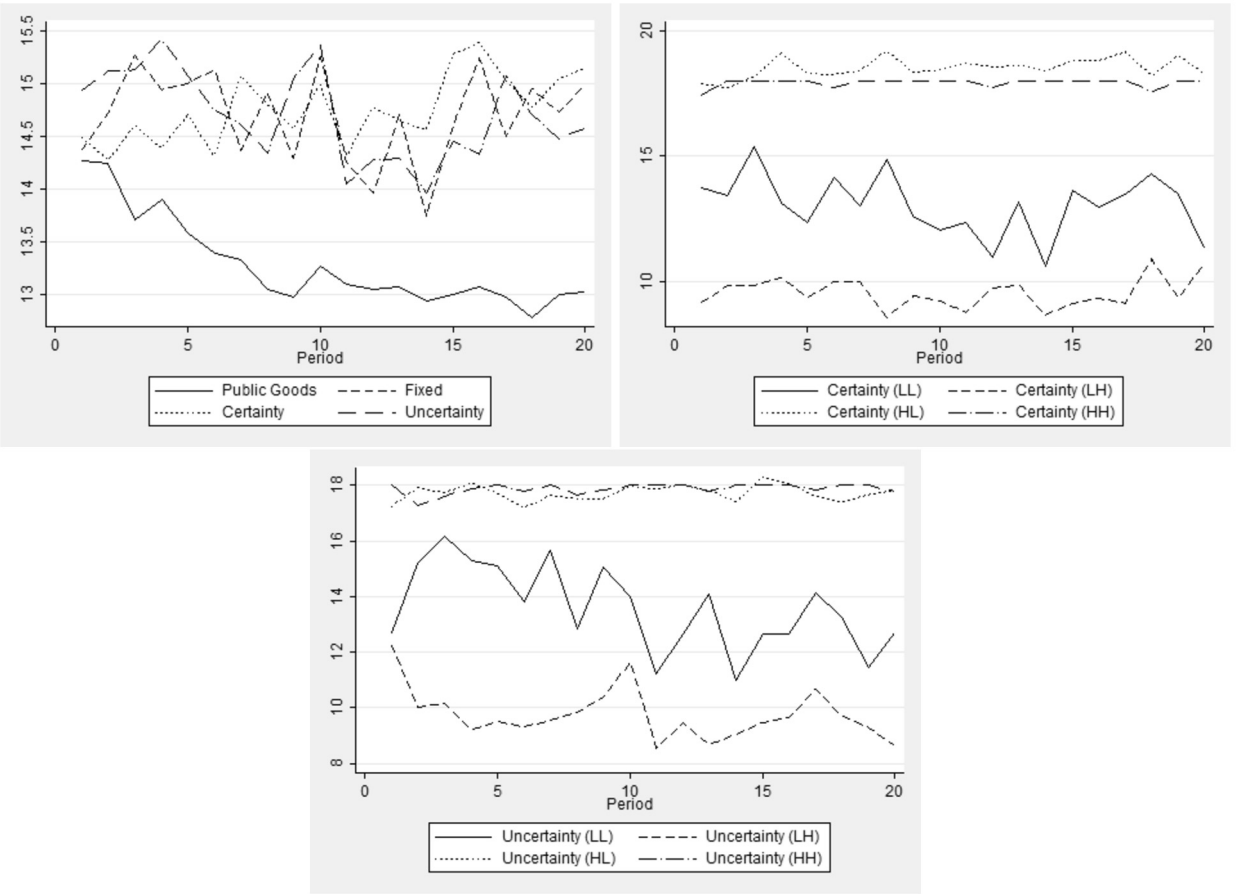

Fig. 14. Average payoffs by treatment.

Result 4. *Payoffs:* Payoffs are higher in F, C, U, than PG. There is no evidence of a difference in payoffs between the P&R treatments.

5.4.1. Optimal strategies

Computing expected payoffs from different pledges (or contributions in PG) based on observed pledge and voting frequencies shows us what strategies are payoff maximizing given the behaviour of other subjects, and can help explain the time trends we observe. These are shown in Fig. 15. Expected payoffs in PG decline mechanically in contributions, with

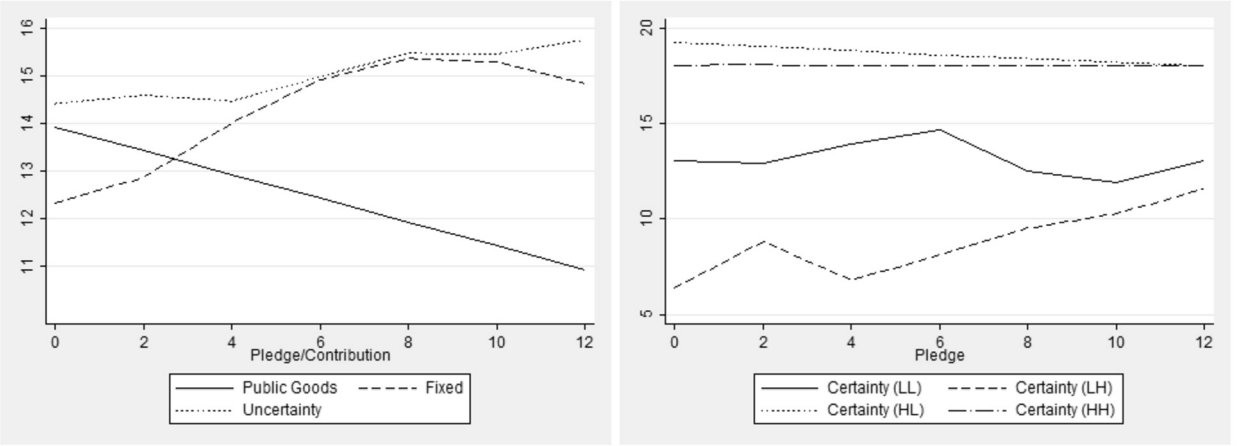

Fig. 15. Expected payoffs.

the slope determined by the MPRC. In F, expected payoffs increase steeply until half the endowment, then flatten off, with the optimal pledge of 8 only narrowly beating 6, 10, and 12. In U, expected payoffs increase almost monotonically, and it is optimal to pledge the full endowment. In both treatments contributing nothing is the worst choice, but incentives to increase pledges are much stronger in F where expected payoffs are 25% higher for the optimal choice, compared to only 10% in U.

Incentives are clearest in C for subjects with a low disagreement payoff matched with a subject with a high disagreement payoff. In this case, the optimal choice of pledging everything increases expected payoffs by 81% over contributing nothing. In these games, expected payoffs for subjects with the high disagreement payoff decline gradually from around 19 to 18 ECU as they increase their pledges. When both subjects have low disagreement payoffs, expected payoffs peak in the middle at 14.7 ECU with a pledge of 6, compared to 13 ECU when pledging either all or nothing. When both disagreement payoffs are high, expected payoffs are 18 ECU or very close for all pledge levels.

6. Discussion and conclusion

We test the performance of three stylised versions of the pledge-and-review procedure in the laboratory. We begin by implementing a standard public goods game (the PG treatment) and, step-by-step, add voting over pledged contributions with fixed homogeneous disagreement payoffs should unanimity not be reached (F), then heterogeneity in disagreement payoffs (C), and finally uncertainty over the disagreement payoffs (U).

We find that the pledge-and-review institution increases efficiency over the public goods treatment, and neither heterogeneity nor uncertainty over disagreement payoffs negate this result. In contrast to our game theoretic predictions, the highest contributions are made in F, that is, *absent* uncertainty. In addition, F saw no decrease in contributions, whereas they declined over time in both C, and U. In the following we discuss possible explanations for these trends, and the underperfomance of pledge-and-review under uncertainty.

A widely accepted explanation for the decline in contributions over repetitions of experimental public goods game is the presence of conditional cooperators who prefer to contribute when others do, but prefer not to otherwise: initially this type of player contributes a positive amount, which reduces over time as they match with less cooperative types (Fischbacher and Gächter, 2010).[14] Introducing the review stage allows conditional cooperators to test, risk free, the cooperativeness of each new partner, meaning there is no reason to reduce pledges even if one's belief in the average level of cooperativeness in the population declines over time. Rather than relying on binding individual contributions, as in the standard PG treatment, the mechanism introduces binding pairs of contributions. Even though pledges are unconditional, actual contributions are conditional on the combination of pledges being satisfactory. This (limited) conditionality allows players to pledge high contributions without risking low payoffs due to non-contributing partners.[15] In other words, the P&R institution means that conditional cooperators can always play their ideal strategy, whereby they cooperate with cooperative types and contribute nothing when matched with non-contributors. Note that such a strategy can be a Nash equilibrium, but not subgame perfect. This reduces inefficient outcomes where conditional cooperators are matched, but do not contribute because they have no way of identifying each others' types.

So why does this mechanism become less effective over time in the heterogenous treatments? In C, this trend is driven by games where both players have a low disagreement payoff, and by declining contributions of high disagreement payoff

[14] Conditional cooperators have been found to make up a plurality of experimental subjects both in the lab (Fischbacher and Gächter, 2010) and the field (Rustagi et al., 2010).

[15] We thank an anonymous reviewer for pointing out this characterization of the mechanism.

players matched with a player with a low disagreement payoff. The latter trend can be explained simply by subjects learning to exploit their stronger bargaining position: because it is always in the financial self interest of the low payoff player to vote yes, the high payoff player can reduce their contribution until any fairness norm is violated, triggering costly punishment. This is similar to dynamics that have been observed in Ultimatum games (Roth et al., 1991). In C_{LL}, because it is always in the financial interests of both players to vote yes, and both players know this (91% of votes in these games were in favour), the game boils down to the standard PG game, with the usual decline in contributions.

Similar to C_{LL}, the strategy of testing the cooperativeness of one's partner also fails in U because of the 50% probability of learning one has a low disagreement payoff and being compelled to vote yes in order to avoid a very low final payoff. Even taking into account this complication, the low level of contributions in U relative to the theoretical prediction of full contributions is surprising: subjects in laboratory experiments typically cooperate *more* than equilibrium based on self-interested preferences would suggest rather than *less*.

In our view, the most likely explanation is altruistic voting: the theoretical prediction in this treatment relies on the fact that players must contribute fully in order to remove the incentive of a partner who turns out to have a high disagreement payoff to vote no if one's own disagreement payoff turns out to be low. However, in our experiments, many high payoff subjects voted yes against their own interests, undermining this mechanism. This notion is supported by the fact that in U, while contributing fully is the best-response to the empirical distribution of subjects' choices for a payoff-maximizing individual, the payoff gain over contributing nothing is much smaller than in F, and may not sufficiently compensate those who engage in altruism or costly punishment in the voting stage. These weaker incentives to maintain high pledges in U can explain their decline relative to F, as conditional cooperators become frustrated with less cooperative partners.

The "problem" of altruistic voting is also present in C, but may be exacerbated in U by intention-based fairness issues. Pledging some fixed amount with a low disagreement pay-off is in one's self-interest, so it is less attributable to being kind than pledging the same amount with a high disagreement pay-off. So, if intentions matter, then falling short by some amount with a high disagreement pay-off should be punished less than falling short that same amount with a lower disagreement pay-off. In the Uncertainty treatment, the expected disagreement pay-off is 12, whereas in the Certainty treatment it is either 6 or 18. Hence, in the Uncertainty treatment, we should see less costly punishment and more altruistic voting of subjects with a high disagreement pay-off vis-à-vis partners with a low disagreement pay-off than in the Certainty treatment. Indeed, the data are consistent with this reasoning. If this effect is anticipated then, as in Exley (2015), subjects may latch onto uncertainty as an excuse not to contribute much to the public good.

One might think that risk-aversion may in some way play a role in explaining our results in U. However, our theoretical and empirical results are not affected by allowing for non-risk-neutrality. As shown in Online Appendix B2, the result that all SPNE involve full contributions holds for all non-risk-seeking agents. In fact, it requires an unreasonable degree of risk-seeking to overturn this finding. For example, with the utility function $u(x) = x^{1-r}/(1-r)$, the result holds provided $r > -1.49$, which Holt and Laury (2002) find to be true for at least 99% of their subjects. To check empirically for an effect of risk aversion, we run a series of regressions of subject choices using our survey measure of risk preferences (see Table 19 in Online Appendix D). We first regress the pledge/contribution choice on treatment dummies and the risk variable we elicited in the questionnaire, then include interaction terms. Coefficients on the risk variable and interaction terms are close to zero, and not statistically significant. The same is true for voting decisions (where we additionally control for own and partner contributions).

Another possible explanation would be that subjects are unable to follow the logic of backward induction. However, the consistently high pledges of subjects with low disagreement payoffs that were paired with subjects with high disagreement payoffs in the C treatment speak against this explanation. Furthermore, the declining pledges for the high disagreement payoff players in this case also suggest they consider, or learn to consider, the voting stage when determining their pledges.

The disagreement payoff in our P&R treatments emulates the expected payoff from the off-equilibrium path action 'reject' in Harstad (2021a). In the heterogeneous C and U treatments, the disagreement payoffs that lead to a rejection of a pledge vector are high. Hence, the higher pledges (and contributions) in the fixed treatment are counterbalanced by the fact that subjects agree to implement the public good only when they count in C and U and, while payoffs in all three pledge-and-review treatments are higher than in the public goods treatment, we do not find significant payoff differences between the three pledge-and-review treatments.

With or without uncertainty, the experiment reveals that the Paris Agreement's pledge-and-review institution may be useful: It allows for behaviourally reasonable strategies with higher contributions than are typically sustained in public goods games without review.

Appendix. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.geb.2021.08.003.

References

Ambrus, Attila, Greiner, Ben, 2019. Individual, dictator, and democratic punishment in public good games with perfect and imperfect observability. J. Public Econ. 178, 104053.

Anderson, Lisa R., Mellor, Jennifer M., Milyo, Jeffrey, 2008. Inequality and public good provision: an experimental analysis. J. Socio-Econ. 37, 1010–1028.

Battaglini, Marco, Harstad, Bård, 2016. Participation and duration of environmental agreements. J. Polit. Econ. 124, 160–204.

Brandts, Jordi, Charness, Gary, 2011. The strategy versus the direct-response method: a first survey of experimental comparisons. Exp. Econ. 14, 375–398.
Cardenas, Juan-Camilo, 2003. Real wealth and experimental cooperation: experiments in the field lab. J. Dev. Econ. 70, 263–289.
Casari, Marco, Luini, Luigi, 2009. Cooperation under alternative punishment institutions: an experiment. J. Econ. Behav. Organ. 71, 273–282.
Chan, Kenneth S., Mestelman, Stuart, Moir, Rob, Muller, R. Andrew, 1996. The voluntary provision of public goods under varying income distributions. Can. J. Econ., 54–69.
Chan, Kenneth S., Mestelman, Stuart, Moir, Robert, Muller, R. Andrew, 1999. Heterogeneity and the voluntary provision of public goods. Exp. Econ. 2, 5–30.
Cherry, Todd L., Kroll, Stephan, Shogren, Jason F., 2005. The impact of endowment heterogeneity and origin on public good contributions: evidence from the lab. J. Econ. Behav. Organ. 57, 357–365.
Cinyabuguma, Matthias, Page, Talbot, Putterman, Louis, 2005. Cooperation under the threat of expulsion in a public goods experiment. J. Public Econ. 89, 1421–1435.
Dannenberg, Astrid, Gallier, Carlo, 2019. The choice of institutions to solve cooperation problems: a survey of experimental research. Exp. Econ., 1–34.
Decker, Torsten, Stiehler, Andreas, Strobel, Martin, 2003. A comparison of punishment rules in repeated public good games: an experimental study. J. Confl. Resolut. 47, 751–772.
Dutta, Prajit K., Radner, Roy, 2004. Self-enforcing climate-change treaties. Proc. Natl. Acad. Sci. 101, 5174–5179.
Dutta, Prajit K., Radner, Roy, 2006. Population growth and technological change in a global warming model. Econ. Theory 29, 251–270.
Dutta, Prajit K., Radner, Roy, 2019. The Paris Accord can be effective if the Green Climate Fund is effective. Mimeo, Columbia University.
Exley, Christine L., 2015. Excusing selfishness in charitable giving: the role of risk. Rev. Econ. Stud. 83, 587–628.
Fagerland, Morten W., Sandvik, Leiv, 2009. The Wilcoxon–Mann–Whitney test under scrutiny. Stat. Med. 28, 1487–1497.
Falk, Armin, Becker, Anke, Dohmen, Thomas J., Huffman, David, Sunde, Uwe, 2016. The preference survey module: a validated instrument for measuring risk, time, and social preferences. IZA Discussion Paper No. 9674.
Fehr, Ernst, Schmidt, Klaus, 1999. A theory of fairness, competition, and cooperation. Q. J. Austrian Econ. 114, 817–868.
Fellner-Röhling, Gerlinde, Kröger, Sabine, Seki, Erika, 2020. Public good production in heterogeneous groups: an experimental analysis on the relation between external return and information. J. Behav. Exp. Econ. 84, 101481.
Fischbacher, Urs, 2007. z-Tree: Zurich toolbox for ready-made economic experiments. Exp. Econ. 10, 171–178.
Fischbacher, Urs, Gächter, Simon, 2010. Social preferences, beliefs, and the dynamics of free riding in public goods experiments. Am. Econ. Rev. 100, 541–556.
Fischbacher, Urs, Schudy, Simeon, Teyssier, Sabrina, 2014. Heterogeneous reactions to heterogeneity in returns from public goods. Soc. Choice Welf. 43, 195–217.
Frederick, Shane, 2005. Cognitive reflection and decision making. J. Econ. Perspect. 19, 25–42.
Greiner, Ben, 2015. Subject pool recruitment procedures: organizing experiments with ORSEE. J. Econ. Sci. Assoc. 1, 114–125.
Harstad, Bård, 2012. Climate contracts: a game of emissions, investments, negotiations, and renegotiations. Rev. Econ. Stud. 79, 1527–1557.
Harstad, Bård, 2016. The dynamics of climate agreements. J. Eur. Econ. Assoc. 14, 719–752.
Harstad, Bård, 2021a. A theory of pledge-and-review bargaining. https://www.sv.uio.no/econ/personer/vit/bardh/dokumenter/prb.pdf. (Accessed 18 August 2021).
Harstad, Bård, 2021b. Pledge-and-review bargaining: from Kyoto to Paris. https://www.sv.uio.no/econ/personer/vit/bardh/dokumenter/ndc.pdf. (Accessed 18 August 2021).
Heap, Shaun P. Hargreaves, Ramalingam, Abhijit, Stoddard, Brock V., 2016. Endowment inequality in public goods games: a re-examination. Econ. Lett. 146, 4–7.
Heike, Hennig-Schmidt, Irlenbusch, Bernd, Rilke, Rainer Michael, Walkowitz, Gari, 2018. Asymmetric outside options in ultimatum bargaining: a systematic analysis. Int. J. Game Theory 47, 301–329.
Hennig-Schmidt, Heike, Li, Zhu-Yu, Yang, Chaoliang, 2008. Why people reject advantageous offers—non-monotonic strategies in ultimatum bargaining: evaluating a video experiment run in pr China. J. Econ. Behav. Organ. 65, 373–384.
Herrmann, Benedikt, Thöni, Christian, Gächter, Simon, 2008. Antisocial punishment across societies. Science 319, 1362–1367.
Holt, Charles A., Laury, Susan K., 2002. Risk aversion and incentive effects. Am. Econ. Rev. 92, 1644–1655.
Kingsley, David C., 2016. Endowment heterogeneity and peer punishment in a public good experiment: cooperation and normative conflict. J. Behav. Exp. Econ. 60, 49–61.
Knez, Marc J., Camerer, Colin F., 1995. Outside options and social comparison in three-player ultimatum game experiments. Games Econ. Behav. 1, 65–94.
Kölle, Felix, 2015. Heterogeneity and cooperation: the role of capability and valuation on public goods provision. J. Econ. Behav. Organ. 109, 120–134.
le Sage, Sander, van der Heijden, Eline, 2015. The effect of voting on contributions in a public goods game. CentER Discussion Paper 2015-039.
Miller, Luis, Montero, Maria, Vanberg, Christoph, 2018. Legislative bargaining with heterogeneous disagreement values: theory and experiments. Games Econ. Behav. 107, 60–92.
Nikiforakis, Nikos, Noussair, Charles N., Wilkening, Tom, 2012. Normative conflict and feuds: the limits of self-enforcement. J. Public Econ. 96, 797–807.
Reischmann, Andreas, Oechssler, Joerg, 2018. The binary conditional contribution mechanism for public good provision in dynamic settings — theory and experimental evidence. J. Public Econ., 104–115.
Roth, Alvin E., Prasnikar, Vesna, Okuno-Fujiwara, Masahiro, Zamir, Shmuel, 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study. Am. Econ. Rev., 1068–1095.
Rustagi, Devesh, Engel, Stefanie, Kosfeld, Michael, 2010. Conditional cooperation and costly monitoring explain success in forest commons management. Science 330, 961–965.
Schlag, Karl H., 2008. A new method for constructing exact tests without making any assumptions. Department of Economics and Business Working Paper 1109. Universitat Pompeu Fabra.
Schlag, Karl H., 2015. Who gives direction to statistical testing? Best practice meets mathematically correct tests. https://homepage.univie.ac.at/karl.schlag/research/statistics/exact np 1.pdf. (Accessed 18 August 2021).
Van Miltenburg, Nynke, Buskens, Vincent, Barrera, Davide, Raub, Werner, 2014. Implementing punishment and reward in the public goods game: the effect of individual and collective decision rules. Int. J. Commons 8, 47–78.
Wilcox, Rand R., 2001. Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy. Springer.
Zelmer, Jennifer, 2003. Linear public goods experiments: a meta-analysis. Exp. Econ. 6, 299–310.