Climate Leadership by Conditional Commitments

By Leif Helland[a], Jon Hovi[b], and Håkon Sælen[c]

a Department of Economics, and Center for Experimental Studies and Research (CESAR), BI Norwegian Business School; leif.helland@bi.no

b Department of Political Science, University of Oslo, and Käte Hamburger Kolleg/Centre of Global Cooperation Research, University of Duisburg-Essen.

c CICERO Center for International Climate Research, and Department of Political Science, University of Oslo.

Under the 2015 Paris climate agreement, each Party sets its own mitigation target by submitting a Nationally Determined Contribution (NDC) every five years. An important question is whether including conditional components in NDCs might enhance the agreement's effectiveness. We report the results of a closely controlled laboratory experiment – based on a mixed sequential-simultaneous public good game with one leader and three followers – that helps answer this question. The experiment investigates how two factors influence the effectiveness of leadership based on intrinsically conditional commitments. Measuring effectiveness in terms of followers' and total contributions, we find that it may help if the conditional promise is credible and if its implementation influences followers' welfare substantially. Importantly, however, for both factors we find a significant effect only if the leader does not reap disproportionate gains from the group's efforts. These findings have important implications concerning the future success of the Paris agreement.

JEL Classification: C72; C92; F55; F64; H41

1 Introduction

In the 2015 Paris Agreement under the United Nations Framework Convention on Climate Change (UNFCCC), each Party sets its own mitigation target by submitting a so-called Nationally Determined Contribution (NDC) every five years. Many Parties have stated conditions for the full implementation of their first-round NDC (UNFCCC 2015). In particular, some Parties have made their commitments conditional on the level of mitigation undertaken by other Parties. We refer to such cases as intrinsic conditions. In contrast, extrinsic conditions make mitigation commitments conditional on other countries' non-mitigation efforts, such as financial and technological support. An important question is whether intrinsic conditions, extrinsic conditions, or both, might enhance the effectiveness of the Paris agreement.

While extrinsic conditions have figured prominently in the UNFCCC in relation to action by developing countries,[1] intrinsic conditions – the focus of this paper – have less of a record in climate negotiations. However, as part of their 2020 pledges under the Cancun agreement, the European Union and Norway promised to cut emissions an additional 10% conditional on strong mitigation commitments by other Parties (UNFCCC 2011b). These intrinsically conditional commitments had little (if any) effect on other countries; hence, they were not implemented.

[1] For example, the Cancun agreement states that developing countries' mitigation shall be "supported and enabled by technology, financing, and capacity-building" (UNFCCC 2011a).

Intrinsically conditional commitments also constitute a central element in Victor's (2011) club approach to climate cooperation outside the UNFCCC. Victor (2011) suggests that cooperation should begin with agreements between small groups of enthusiastic countries. The "backbone" of his approach is a series of conditional offers, whereby enthusiastic countries (leaders) would outline what they are willing and able to do, conditional on what other countries (followers) offer and implement. Moreover, reluctant countries would be enticed to join the club via "exclusive and contingent" measures, such as preferential market access for club members.[2]

Hence, an important question concerning climate cooperation both inside and outside the UNFCCC process is whether and, if so, under which conditions an intrinsically conditional commitment by one or a few countries (or actors) might increase others' willingness to make deep emission cuts.

We present the results from a closely controlled laboratory experiment specifically designed to inform the conditionality debate. Thus, we respond to the recent call by Finus *et al.* (2017) for more research on mechanisms that might trigger more urgent collective action on climate change. This paper contributes in four main ways:

First, our experiment is – to the best of our knowledge – the first one to study leadership by conditional commitments. The basic preference configuration underlying the NDC process resembles the one found in one of the most widely studied games in experimental economics, the voluntary contribution mechanism game, also known as the public goods game. We study a novel variant of this game, with one leader and three followers, to assess the effectiveness of leadership by intrinsically conditional commitments. We measure effectiveness in terms of influence on followers' and total contributions to the public good.

Second, our results suggest that effectiveness is enhanced if the leader's conditional promise is credible, that is, if followers have reason to believe that fulfilling the leader's stated condition will actually cause the leader's promise to be implemented. Effectiveness is also enhanced if the leader can influence the followers' welfare substantially by implementing its conditional promise. It is well known that behaviour in the lab often deviates substantially from predictions derived from standard game-theoretic assumptions (see the next section). It is therefore interesting that, concerning the two above-mentioned factors, our results are largely in keeping with predictions derived from standard assumptions.

Third, and most importantly, our results do deviate from standard game-theoretic predictions in one significant respect: For both of the aforementioned factors we find a significant effect only if the leader does not reap disproportionate gains from the group's collective efforts. This finding suggests that efficient contribution norms do not easily evolve in groups where leaders benefit significantly more than followers from collective efforts.

Finally, even though the factors we study have a substantial effect on contributions, the outcome remains severely suboptimal even under favorable conditions. Thus, our results indicate that leadership by conditional commitments can only bring about efficient mitigation levels if supplemented by other measures. This finding provides some caution to the most optimistic supporters of the Paris agreement.

The environment we study is highly stylized: The game's structure is public knowledge; the sequence of moves, the time horizon, time periods, payoffs, and contributions are all unambiguously defined; subjects can observe behaviour without delay or noise; and all decisions are anonymous. This stylized environment only slightly resembles real-world settings where conditional commitments are or can be used. Thus, the external validity of our results should be checked through field studies if and when relevant field data become available. This being said,

[2] Victor (2011) also suggests that agreements should (a) be nonbinding, (b) entail high flexibility concerning choice of policy strategies, and (c) focus on policies that governments actually control, rather than on emission levels (which in large part depend on factors beyond governmental control).

we provide some suggestive evidence in online appendix A indicating that our subjects' behaviour does not deviate significantly from that of elite decision makers.[3]

An experimental design built on a stylized environment has its advantages. Empirical field data from international negotiations are not only scant; they also suffer from well-known problems such as endogeneity, selection on unobservables, omitted third variables, and reverse causality. Experiments permit randomization over treatments and truly exogenous variation in the explanatory variables; hence, the conditions (if any) under which conditional commitments might be effective can be investigated by systematically manipulating the structure of interactions.

In section 2, we review relevant literature. In section 3, we outline our model and treatments. In section 4, we provide details about the experiment's design and implementation. In section 5, we present our results. Finally, in section 6 we conclude and discuss some important implications of our results for the process established by the Paris agreement.

2 Related Literature

Much game-theoretic research on international environmental agreements is based on (variants of) a two-stage game, where countries first decide whether to participate in the agreement and then determine their mitigation efforts. Classic contributions include Barrett (1994, 2003) and Carraro and Siniscalco (1993, 1998). Most of these contributions predict very low levels of participation. Thus, many economists and game-theorists were surprised by the very broad participation in the Paris agreement. However, attempts have also been made to explain how higher participation levels might be achieved (for recent reviews, see Carattini *et al.* 2017; Hovi *et al.* 2015). We contribute to these attempts by exploring how (a particular form of) leadership might enhance cooperation.

Our research draws on and contributes to five more specific strands of literature. The first consists of theoretic (mostly game-theoretic) work on leadership in the form of unilateral emissions reductions. This strand offers very little support for the conjecture that unilateral action will induce other countries to follow suit. Using a two-country model, Hoel (1991) demonstrates that if one country (the leader) undertakes unilateral emissions reductions, the other country (the follower, which is assumed to be motivated by self-interest) may well increase its own emissions. The reason is that the leader's unilateral action diminishes the follower's marginal benefit of emissions reductions. Hoel also finds that unilateral emissions reductions may cause international climate change negotiations to result in an agreement with higher total emissions than if both countries act selfishly (in which case no unilateral action will occur).

Several more recent studies support Hoel's results. For example, Buchholz *et al.* (1998) find that other countries' free riding will likely offset unilateral efforts by one or a few countries. Thus, in their model (which closely resembles Hoel's) a coalition acting unilaterally can generate net benefits to its members only if it includes all major emitters. Similarly, using a coalition model, Holtsmark (2013) shows that if one country were to announce ambitious and unconditional emissions reductions before international negotiations take place, this may reduce the ambition of the international agreement.[4] Lastly, using an incomplete-information model, Konrad and Thum (2014) find that a unilateral and unconditional commitment to reducing emissions diminishes the gains from global cooperation and hence makes it more difficult to reach an effective international agreement. In contrast to these pessimistic findings, Buchholz and Sandler (2017) demonstrate

[3] The online appendix can be found on the home pages of this journal.

[4] Common to these game-theoretic studies is that they ignore the possibility of "no-regret" options for reducing emissions of greenhouse gases. Ott and Oberthür (1999) suggest that a leader might cause global emissions reductions by demonstrating such options' attractiveness to other countries. However, this alleged effect would seem to presuppose that the leader has superior knowledge concerning no-regret options —a rather strong assumption.

that incorporating ideas from behavioural economics – in particular a desire for reciprocity and a "warm-glow-of-giving" – entails that leadership might influence followers' contributions positively.

A second strand consists of game-theoretic work focusing on the prospects for transforming the climate change mitigation game from a social-dilemma game to a coordination game. The underlying assumption is that countries are much better at solving coordination games than they are at solving social-dilemma games. For example, Barrett (2003) shows how trade restrictions and technology standards might serve this function. Moreover, Barrett and Dannenberg (2012) find that a looming climate disaster with a known emissions threshold could transform the climate cooperation dilemma into a coordination game.

The third strand contains political science and economics work that is more empirically oriented than the contributions in the first and second strands. While political scientists studying climate leadership and minilateralism have been more concerned with fairness than with effectiveness,[5] Skodvin and Andresen (2006) present a case study of the EU's attempt to exert leadership by saving the Kyoto Protocol after the US repudiation in 2001. They conclude that, although EU leadership was instrumental to Kyoto's entry into force, the resulting agreement was a mini-regime with "miniscule impact on climate change abatement". Their conclusion is supported by a recent econometric study by Almer and Winkler (2017), who find "very little evidence" for the hypothesis that Kyoto 1 influenced the emissions in the major Annex B emitters with binding targets. In another quantitative study, McLean and Stone (2012) find that Kyoto 1 is best understood as a case of the "Europeanization of international politics", whereby the EU was able to emerge as a key agenda setter, while its member countries subordinated their domestic climate politics to international cooperation.

Combining simulations with case studies, Underdal *et al.* (2012) focus specifically on leadership by conditional commitments. They argue that such leadership can work–but only under rather strict conditions. In particular, they find that successful leadership requires that two conditions be fulfilled: First, the leader must promise to undertake substantial additional emissions reductions if other countries fulfill the stated requirements. Second, the leader's promise must be credible, so that followers expect the leader to implement its promise of additional emissions reductions if (and only if) other countries fulfill the leader's stated conditions. It may be noted that these conditions, which resemble the corresponding conditions necessary for a threat to be effective (e.g., see Schelling 1960), also motivate our experiment.[6]

The fourth strand comprises experimental studies on public goods games with a provision threshold. Such games typically contain efficient equilibria, which might facilitate cooperation, especially with a sequential protocol (Erev and Rapoport 1990). However, threshold uncertainty can make coordination difficult and might constrain contributions even in the presence of a contribution threshold (Barrett and Dannenberg 2012, 2014; Dannenberg *et al.* 2015). On the other hand, voting concerning subjects' aggregate and/or individual contributions seem to practically guarantee successful coordination in threshold public goods games (Feige *et al.* 2014). These results are important for our experiment, where a conditional contribution by a leader can transform the game amongst the followers into a coordination game. Tavoni *et al.* (2011) investigates a public goods game in which total contributions below a given threshold makes everyone loose their remaining money with 50% probability. They find that in this environment heterogenous endowments makes success less likely, while communication has the opposite effect. Using a similar set-up Milinski *et al.* (2008) demonstrate that increasing the contingent failure probability increases the probability of succeeding in reaching the threshold. A recent experiment

[5] See e.g., Eckersley (2012) and Maltais (2014).

[6] Weischer et al. (2012) elaborate on the conditions for a promise to be effective in the context of climate change.

finds that cooperation is larger and more stable if it affects the probability rather than the size of damages (Köke *et al.* 2015).

Finally, the fifth strand consists of a small but growing body of experimental research on leadership in public goods games. It is well known that subjects' behaviour in public goods experiments tends to deviate systematically from standard game-theoretic predictions, which are based on the assumptions of purely self-interested motivation and common knowledge of rationality. In particular, subjects in public goods experiments contribute and (when given the opportunity) punish substantially more than suggested by the stark zero-contribution, zero-punishment predictions of standard game theory. The reasons for these deviations have been extensively explored in the literature (see e.g., Chauduri 2011; Fehr and Gächter 2000; Kosfeld *et al.* 2009; McEvoy 2010; McEvoy *et al.* 2011; Ostrom 2000; and Ostrom *et al.* 1992).

A number of other findings from experimental economics also provide relevant background for our experiment. For small groups (4 to 10 subjects) and sizable marginal per capita return on contributions (MPCR between 0.30 and 0.75), group size does not significantly affect contribution behaviour. In contrast, the MPCR, controlled for group size, significantly influences contributions in small-group, high-MPCR settings (Isaach and Walker 1988). Weimann *et al.* (2012) find that this relationship holds even for sizable groups (40 to 60 subjects) and for very low MPCRs (0.02 and 0.04). This finding indicates that small-group behaviour in the lab is also relevant for large-scale problems where the marginal benefits of individual contributions to a public good are negligible, as is typically the case for global emissions reductions.

Cherry *et al.* (2005) find that heterogeneously endowed subjects contribute significantly less than homogeneously endowed subjects do. In contrast, Reuben and Riedl (2013) find that both heterogeneous endowments and heterogeneous returns produce approximately a doubling of contributions relative to the contributions in homogeneous groups when no punishment is available. However, when punishments are introduced, the increase in contributions is substantially weaker with heterogeneous MPCRs than with heterogeneous endowments. The authors conclude that subjects converge on efficient contribution norms even when endowments differ, but not when subjects benefit unevenly from public goods provision. According to Reuben and Riedl (2013), uneven benefits give rise to conflicting contribution norms.[7] Such conflicting norms hamper cooperation. There are important differences in the design of these studies that may account for the differences in results. In Cherry *et al.* (2005), the experiment is one shot in groups of 4 while endowments are either earned or randomly allocated. In Reuben and Reidl (2013), endowments are random and the design is a 10 period partner-matching with groups of 3. However, Ruben and Reidl (2013) only analyze the final 5 periods. Furthermore, Cherry *et al.* (2005) use 4 earning levels, while Reuben and Reidl (2013) use only 2.[8]

Most public goods experiments implement simultaneous moves. In contrast, only a handful lets one group member (a leader) make its contribution decision before the other group members (the followers). Güth *et al.* (2007) find that experiments with (unconditional) leadership trigger higher average contributions than standard public goods experiments with simultaneous moves do. This difference in contributions is statistically significant, yet substantially moderate. Thus, while unconditional leadership enhances cooperation, it comes nowhere near fully solving the underlying collective action problem.[9] This result is supported by Levati *et al.* (2007), who find

[7] Fisher et al (1995) find that a player's MPCR has a strong positive effect on that player's contribution, but find no effect of MPCR heterogeneity on group contributions.

[8] An early study that investigates heterogenous MPCRs is Fischer et al. (1995). They use a combined between- and within subjects design and find that high MPCR types tend to contribute more than low MPCR types.

[9] Compared to the baseline in which all subjects choose simultaneously, average contributions (over all periods and all groups) increase by 13.5 percentage points (from 40 percent in the baseline). In an additional treatment the leader is granted the right to exclude one member of the group from consuming the public benefits in the next period. This treatment increases average contributions by 39 percentage points compared to the baseline.

that the effect of unconditional leadership is even weaker (but still significant) when subjects' endowments differ and this difference is public knowledge.

Gächter *et al.* (2010) find that reciprocator types contribute significantly more than self-interested types acting in the role of leader do.[10] A substantial part of this effect, however, is due to so-called false consensus. Reciprocator types initially tend to overestimate the number of other reciprocators in the population and hence contribute substantially in the first round. However, they are disappointed when other followers' contributions prove lower than expected. Disappointment due to false consensus may, at least partly, explain why average contributions are falling over time in the experiments such as the ones conducted by Güth *et al.* (2007) and Levati *et al.* (2007)[11]

In contrast, Rivas and Sutter (2011) find a substantial effect of leadership on contributions when leaders are permitted to self-select (rather than being allocated) into the leader role.[12] Moreover, with voluntary leadership, average contributions do not appear to be falling over time. These findings lend some support to the conjecture that enthusiastic leaders may make a difference. However, in the set-up of Rivas and Sutter leaders are not permitted to condition their contributions on follower behaviour—which is the focus of our experiment.

Of the contributions reviewed here, only Underdal *et al.* (2012) consider leadership by conditional commitments (as we do). Using Güth *et al.*'s unconditional leadership treatment as baseline, we introduce several novel treatments that aim at pinpointing the conditions under which leadership by intrinsically conditional commitments can or will be effective. Our treatments introduce changes step by step, so that only a single experimental design element differs from one treatment to the next.

3 Model and Treatments

Consider a three-stage one-shot game where one player is randomly selected as leader (L), while the other $n-1$ players are followers (F). Each player is endowed with z_k units of a numéraire good (with $k = \{L, F\}$).

In stage one, the leader decides how much of its endowment to contribute to a public account for the group. In our eight main treatments (T3 through T10), the leader can also promise to top up its contribution in stage three, provided that the followers' average contribution exceeds a minimum specified by the leader.[13]

In stage two, followers—having observed the leader's contribution and conditional promise (if any)—decide simultaneously how much of their endowment they will contribute to the public account; thus, player i's contribution c_i must satisfy $c_i \in [0, z_k]$. Once made, the followers' aggregate contributions are observed by the n players.[14]

Given our motivation, however, this sanctioning mechanism is not very interesting. A viable global climate is a pure public good, and excluding states from its benefits is not feasible.

[10] In this study the distribution of types is extracted using the strategy method (proposed by Selten 1967) prior to actual decision-making in the experiment.

[11] As suggested by an anonymous reviewer, false consensus might also explain why a small country such as Switzerland chose to present a highly ambitious INDC very early in the process leading up to the Paris agreement. Carattini *et al.* (2017) suggest that the Scandinavian countries' introduction of carbon taxes in the early 1990s (Finland 1990, Sweden and Norway 1991, Denmark 1992) may have contributed to starting a reciprocating process that eventually facilitated the Kyoto Protocol and even the Paris Agreement. However, they explicitly state that they are unable to decide whether false consensus played a role in this case.

[12] Compared to the simultaneous-choice baseline (with average contributions of 40 percent) voluntary leadership increases average contributions by almost 23 percentage points.

[13] In T1 and T2, the leader cannot make such a conditional promise.

[14] If followers move in a pre-determined sequence and promises are non-binding, a unique equilibrium exists in which no player contributes (by backwards induction). If promises are binding, a cooperative equilibrium exist in

In stage three, the leader's contribution can be increased (unless the leader contributed its entire endowment in stage one). In some of our treatments, the leader is free to choose whether it will top up its contribution and, if so, by how much. In other treatments, a computer program automatically implements the leader's conditional promise whenever followers fulfill the leader's condition. All treatments except T1 include this third stage (T1 consists of stages one and two only, and replicates Güth *et al.* 2007).

Contributions are multiplied by a factor (greater than unity and less than the number of players in the group) before being divided on all group members—either evenly or relative to the players' endowments (see Table 1). Unless a contribution can be pivotal for increasing one or more other players' contributions, it is a strictly dominant strategy to contribute zero units to the public account (assuming rationality, self-interested motivation, and complete information).

Our 10 treatments were designed to study under what conditions leadership by conditional commitment will effectively enhance followers' contributions (Table 1). We are particularly interested in the effects of (1) giving the leader the possibility to explicitly state its conditions for topping up; (2) making the leader's conditional promise binding (i.e., fully credible); (3) expanding the leader's endowment; and (4) increasing the leader's MPCR.[15]

Player i's payoff π_i equals:

$$\pi_i = z_k - c_i + \alpha_k \sum_{i=1}^{n} c_i$$

Here the first right-side term (z_k) represents player i's endowment, the second (c_i) represents player i's contribution,[16] and the third represents player i's benefit from its own and others' contributions, with α_k representing the MPCR. In all treatments, $n = 4$ and $z_F = 100$. The values of z_L, α_L, and α_F vary across treatments (see Table 1). Our design keeps the social return on contributions to the public good constant as we vary α_L and α_F over treatments.[17,18]

addition to the non-cooperative one. With a pre-determined sequence of moves and binding promises, however, a unique subgame-perfect equilibrium exists. This is in contrast to the case where followers move simultaneously, and where there may be multiple ways to play the cooperative equilibrium (see below).

[15] Interpreting (3) and (4) in a climate context: A leader can be "big" in two ways; by having a large endowment, which can be interpreted as having a large capacity to emit; and by having a large marginal benefit of abatement, which can be interpreted as having a large population benefiting from it.

[16] For the leader, c_i represents the sum of its contribution in stage 1 and its contribution in stage 3.

[17] Specifically, $\alpha_L + 3\alpha_F = 1.6$ both when $\alpha_L = \alpha_F = 0.4$ and when $\alpha_L = 0.64$ and $\alpha_F = 0.32$.

[18] More generally, our design enables us to vary each of our three main parameters (credibility, leverage and even/uneven distribution of the gains from cooperation), while keeping the other two constant. To do this, we need a total of eight treatments (T3 through T10).

Treatment	Short description	Detailed description	z_L	α_L	α_F
T1	Baseline/control	Standard sequential-simultaneous public goods game. Leader moves first and followers move simultaneously after having observed the leader's contribution.	100	0.4	0.4
T2	Implicit conditionality	As T1, except that leader can top up (i.e., make a second contribution decision) in stage three, after having observed the followers contributions.	100	0.4	0.4
T3	Explicit conditionality, nonbinding promise	As T2, except that leader can make a nonbinding, conditional promise to top up. The condition is that the followers' average contribution must exceed a minimum chosen by the leader.	100	0.4	0.4
T4	Explicit conditionality, binding promise	As T3, except that the leader's promise is binding, in the sense that if the leader's stated condition is fulfilled, then the promise is automatically implemented by the computer.	100	0.4	0.4
T5	Explicit conditionality, binding promise, public gains shared unevenly	As T4, except that the leader's share of the gains from the public account equals twice that of a follower.	100	0.64	0.32
T6	Explicit conditionality, binding promise, big leader, public gains shared evenly	As T4 except that the leader's endowment equals twice that of a follower.	200	0.4	0.4
T7	Explicit conditionality, binding promise, big leader, public gains shared unevenly	As T4, except that the leader's endowment equals twice that of a follower and that the leader's share of the gains equals twice that of a follower.	200	0.64	0.32
T8	Explicit conditionality, nonbinding promise, public gains shared unevenly	As T5, except that the leader's promise is nonbinding.	100	0.64	0.32
T9	Explicit conditionality, nonbinding promise, big leader, public gains shared evenly	As T6, except that the leader's promise is nonbinding.	200	0.4	0.4
T10	Explicit conditionality, nonbinding promise, big leader, public gains shared unevenly	As T7, except that the leader's promise is nonbinding.	200	0.64	0.32

Table 1: *Treatments.* z_L*: the leaders endowment;* α_k*: MPCR for player* k $(k = L, F)$

We begin by considering a situation in which the assumptions of what we might call the "standard model" apply: In this situation, it is common knowledge that all n players are rational and purely self-regarding. Based on these assumptions, what will be the game's subgame-perfect equilibrium? The answer depends on whether the leader's promise is binding.

First, consider the case where the leader's promise is nonbinding, so that in stage three, the leader is free to choose whether it will keep or violate its promise (if any) from stage one. Using backward induction we find that in stage three, the leader will contribute zero. The reason is that the marginal cost of contributing one unit is 1, whereas the marginal private benefit of contributing one unit is only α_L (< 1). Moreover, contributing a positive amount in stage three cannot influence followers' contributions, simply because followers have no decisions to make after stage two (in the one-shot game).

In stage two, no follower will make a contribution, because $\alpha_F < 1$ and because followers anticipate that, regardless of their decisions, the leader will contribute zero in stage three.

Finally, in stage one the leader will contribute nothing, because $\alpha_L < 1$ and because the leader anticipates that, regardless of the leader's stage-one contribution (and promise, if any), no follower will make a contribution in stage two.

It follows that for T1 and T2 (in which the leader can make no promise at all) as well as for T3, T8, T9, and T10 (in which the leader can make only a nonbinding promise), the unique subgame-perfect equilibrium of the standard model is that all n players contribute nothing.[19] Thus, in these treatments each player's equilibrium payoff equals its endowment z_k. Because $\alpha_k > 1/n$ in our design, this subgame-perfect equilibrium is Pareto dominated by the non-equilibrium outcomes wherein all players contribute their entire endowment. Note that many such outcomes exist, because the leader can divide its contribution of z_L units between stage one and stage three in many different ways.[20]

Backward induction shows that in a finitely repeated game, the stage-game equilibrium will be played in every period. Thus, in T1, T2, T3, T8, T9, and T10 the subgame-perfect equilibrium in the finitely repeated game is that all players contribute zero units in every period.

Next, turn to the case where the leader's promise is binding (T4 through T7). For this case, our experimental design makes it public knowledge that if followers fulfill the leader's condition, then the leader's promise to make an additional contribution in stage three will be automatically implemented by the computer. In all treatments where the leader's promise is binding, cooperative equilibria exist. In particular, the leader can—by choosing its stage-one contribution and promise appropriately—create a coordination game for the followers. Denote the leader's conditional contribution b and its minimum requirement for the followers' average contribution c^*. Follower i's return from contributing to the public account will then equal:

$$c_i(\alpha_F - 1) \; if \; c_i < (n-1)c^* - \sum_{j \neq i} c_j$$

and

$$c_i(\alpha_F - 1) + b\alpha_F \; if \; c_i \geq (n-1)c^* - \sum_{j \neq i} c_j$$

where c_j is follower j's contribution. Notice that follower i's return function shifts vertically at the point where i's contribution is pivotal for triggering implementation of the leader's conditional

[19] What would the leader promise in equilibrium? Because the promise is nonbinding, it is costless to make any promise as well as to violate it (cheap talk). Any promise is thus consistent with equilibrium behavior. Experimental evidence suggests that most people do not use cheap talk to mislead others (Ostrom 2000: 141).

[20] When the leader must choose an integer between 0 and 100 (as in our experiment), exactly 101 such Pareto-optimal outcomes exist.

contribution. Denote this point c_i^*. The return function is non-negative if $c_i^*(\alpha_F - 1) + b\alpha_F \geq 0$. Solving for b gives:

$$b \geq \frac{(1-\alpha_F)}{\alpha_F} c_i^* \tag{1}$$

When condition (1) holds, a follower has no incentive to deviate unilaterally from c_i^*; hence, c_i^* constitutes a best reply given the other players' contributions (and the leader's promise). Dropping the subscript on c^*, condition (1) gives the combinations of b and c^* with which the leader will create a coordination game for the followers. All combinations of follower contributions that exactly meet the leader's stated minimum constitute equilibria of the stage 2 game; however, zero contributions by all players is also an equilibrium. The positive-contribution equilibria Pareto dominate the zero-contribution equilibrium whenever condition (1) is a strict inequality. However, zero contribution is the maximin strategy for followers.

Given that the leader sets b and c^* according to equation (1) and followers coordinate to meet c^*, the leader's marginal benefit with regard to c^* is $\alpha_L(n-1)$ while the marginal cost is $\frac{(1-\alpha_L)(1-\alpha_F)}{\alpha_F}$. Rearranging gives the following conditions for profits to be increasing in c^*:

$$\frac{\alpha_L}{(1-\alpha_L)} > \frac{(1-\alpha_F)}{\alpha_F(n-1)} \tag{2}$$

This condition is always satisfied in our experiment. Furthermore, when followers fail to coordinate, the leader's marginal benefit and marginal cost with regard to c^* are both zero. The leader will maximize its payoff by maximizing c^* subject to its own budget constraint, the followers' budget constraint, and the condition in equation (1). Solving yields

$$c^* = \min\left[z_L \times \frac{\alpha_F}{(1-\alpha_F)}, z_F\right] \tag{3}$$

$$b = \frac{(1-\alpha_F)}{\alpha_F} c^* \tag{4}$$

The numeric equilibrium solutions in terms of the leader's contribution and the required average follower contributions are listed in Table 2 for the treatments with binding promises.

Treatment	Equilibria $(c_L, \bar{c}_F)$	Percent of potential
T4	$(0,0)$ and $(100,66)$	74.5
T5	$(0,0)$ and $(100,47)$	60.3
T6	$(0,0)$ and $(150,100)$	90.0
T7	$(0,0)$ and $(200,94)$	96.4

Table 2: *Equilibrium leader contribution and equilibrium average follower contributions in treatments with binding promise*

Note that more than two equilibria typically exist: unless the leader's condition requires all followers to contribute their entire endowment, the followers can take on the costs involved in satisfying the leader's stated condition in (many) different ways. Given this coordination problem, it is by no means obvious that the followers will manage to settle on any particular positive-contribution equilibrium.[21] Treatment 6 provides an exception to the multiplicity of positive contribution equilibria. In this treatment followers are required to contribute their entire endowment in equilibrium, reducing the number of positive contribution equilibria to one. To ensure that the positive-contributions equilibrium is strictly Pareto dominant, the leader must set the ratio of b to c^* slightly higher than implied by equations (3) and (4).

The last column in Table 2 provides the sum of contributions in the positive contributions equilibrium, as a percentage of the sum of endowments. This percentage can be taken as a measure of efficiency, since the payoffs are increasing linearly in the sum of contributions. As can be seen, achieving a high efficiency in public goods provision requires a high leader endowment and access to a commitment technology.

In our experiment, the costs of public goods provision are linear. In online appendix B, we explore a model in which costs are increasing in own contributions (e.g., see Barrett 1994; 2002). We demonstrate that modified versions of conditions (1) and (2) can be satisfied for increasing marginal costs.

4 Implementation

As explained in the previous section, we ran 10 treatments (including the control treatment). Each treatment consisted of 16 periods. To avoid "envy effects," we let each subject act as leader for four (subsequent) periods. Which subject acted as leader in which four periods was determined randomly.[22]

We recruited a total of 408 subjects for the experiment, 176 subjects from the general student population at BI Norwegian Business School and 232 subjects from the general student population at Appalachian State University, Boone. While the empirical record remains thin, the existing evidence seems to indicate that elites do not differ radically from students in terms of self regard or strategic reasoning (Hafner-Burton *et al.* 2014; LeVeck *et al.* 2014; for more details, see online appendix A). The number of groups included in a session varied from 3 to 7. No subject participated in more than one session. We ran a total of 21 sessions for the experiment, striving to balance the US and Norwegian sessions over the 10 treatments. The sessions were conducted between May 2013 and May 2014.

We implemented a partner design in which the four-subject groups were formed randomly at the beginning of each treatment and remained constant for that treatment's 16 periods. Subjects only received feedback about behaviour in their own group. In each period, all subjects received information about the leader contribution prior to entering stage 2. After followers had made their (simultaneous) contribution decisions in stage 2, all subjects were informed about these follower contributions. When relevant (T2-T10), all subjects were informed about the leaders' stage three contribution after it had been made. Finally at the end of each period all subjects

[21] An anonymous reviewer commented that the leader could ensure that the followers do NOT encounter such a coordination problem, by formulating its conditional commitment on the form: "I contribute an extra amount of x, conditional on EACH of the other countries contributing an amount of y". While the reviewer is obviously right, we have never seen a conditional commitment resembling this formulation in the real world.

[22] Literature on the possible effects of role reversal is scarce; however, the scant evidence that does exist seems to point in the direction of no significant effect. For example, Hall (2013) finds no significant difference in behavior between role-reversal and single-role protocols for a trust game. Ball et al. (1991) report a weak role-switching effect in a bilateral-bargaining experiment. Their interpretation is that role-switching facilitates improved decision-making (in a game-theoretic sense) because it helps people focus on their adversary's decisions.

received feedback in the form of a statistic covering decisions and payoffs in the current and all previous periods. Subjects' anonymity was preserved throughout.

All sessions were computerized, and the experiment was programmed in z-Tree (Fischbacher 2007). In each session the administrator, having seated the subjects at randomly drawn cubicles in the lab, distributed the instructions and read them aloud. Sample instructions and screen shots are included in online appendix C. The session began after subjects had answered a set of control questions designed to ensure they understood the payoff structure. Each session lasted about one hour. In the experiment an Experimental Currency Unit (ECU) was used. The instructions made the exchange rate from ECU to USD or NOK public knowledge.

Subjects received their earnings in cash and privately, at the end of the session, which lasted on average around one hour. Subject earnings averaged around USD 40 / NOK 250 in the Oslo sessions, and USD 24 / NOK 150 in the Boone sessions.[23]

5 Results

We are particularly interested in how the average follower contributions varies by treatment. In addition, we study how often the leader creates a coordination game for the followers and whether followers are able to coordinate by meeting (or exceeding) the leader's stated condition.

5.1 Average follower contributions

Figure 1 shows the average follower contributions and the average leader contribution for all of our 10 treatments. Seven main features stand out.

First, the average follower contributions varies considerably across treatments. In particular, it is more than three times higher in treatment 6 (the maximum) than in treatment 3 (the minimum). Thus, the variables defining our treatments seem to influence the followers' behaviour.

Second, the average leader contribution also varies considerably across treatments. Thus, the variables defining our treatments seem to influence the leader's behaviour as well.

Third, the average follower contributions are positively correlated with the average leader contribution. This finding suggests that the leader's behaviour influences the followers' behaviour (and possibly vice versa).

Fourth, the average follower contributions are not higher in treatments 2 and 3 than in treatment 1. Thus, giving the leader the opportunity to top up or to top up and make a conditional promise does not by itself enhance public goods provision.

Fifth, and consistent with the equilibrium of our model, the average follower contributions are higher in treatments where (1) the leader has a large endowment and (2) implementation of the leader's promise is automatic, than in treatments with neither of these two features. For example, compare treatment 6 (uneven endowment and automatic implementation) with treatment 3 (even endowment and voluntary implementation). The impact of coordination is analyzed further below.

Sixth, only one treatment (treatment 6) displays average follower contributions higher than 50% of the endowment. Thus, public goods provision remains moderate even under favorable conditions.

Finally, the general pattern is that leaders contribute more than half of their total contributions in stage one.[24] The exception is when leaders have a higher endowment than their

[23] Differences in earnings are due to the particular rules of the two labs, reflecting the alternative hourly wage for a student in the two locations.

[24] One might wonder why a rational leader would contribute anything at all in stage one. As suggested by an anonymous reviewer, the motive might be reputation building or a desire to signal a will to cooperate, thereby

followers and pledges are binding. Under such conditions, at least half of what leaders contribute is contributed in stage three. This is what one would expect to see if conditional commitments primarily work for leaders with high endowment and access to a commitment technology.

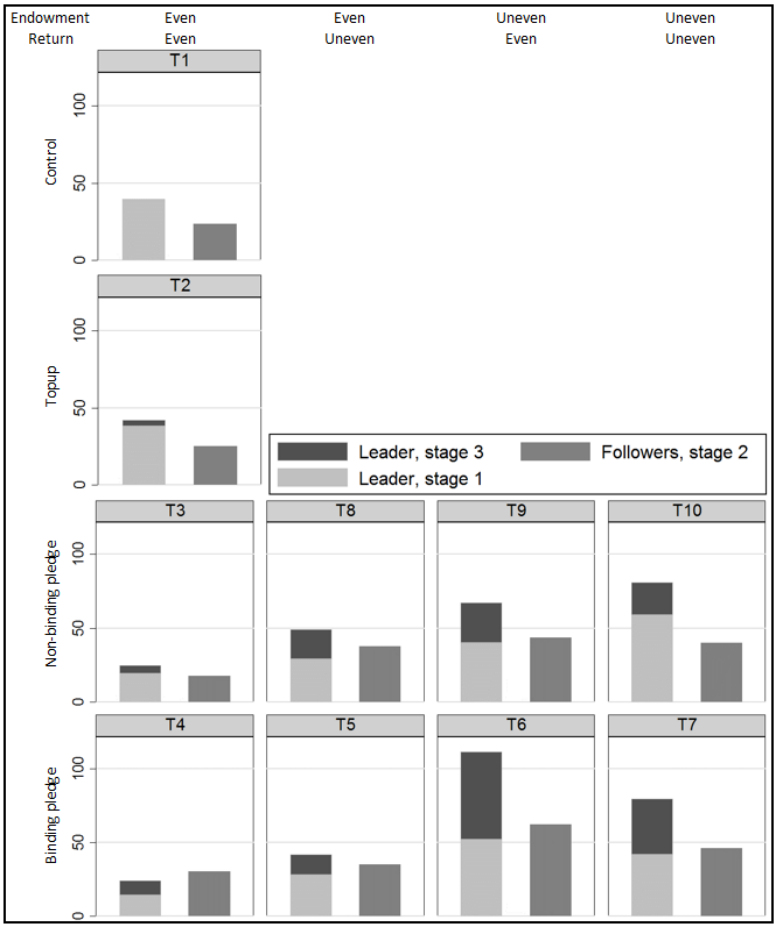

Figure 1: *Average leader and follower contributions by treatment*

5.2 Treatment regressions

In this section we analyze our data using a series of regressions, paying particular attention to interaction effects. In the regressions individual decisions are used as the unit of analysis. We run the regressions with individual random effects, and cluster standard errors at the group level to control for within-group interactions.[25] The results from the regression analysis are consistent

making the conditional promise more credible. We leave for future research a more in-depth analysis of such motives.

[25] We performed a series of robustness tests using alternative model specifications. Using group averages as the dependent variable (rather than individual decisions), with group random effects and standard errors clustered on groups, does not qualitatively alter results. Inclusion of fixed-period effects (alone or in addition to random effects) does not qualitatively alter results either. Again, this holds both for individual decisions and for group averages as dependents, and for full sample analysis as well as for split sample analysis.

with the results from non-parametric tests using group level data as units of analysis (see online appendix D).

	Model 1		Model 2		Model 3			
					$MPCRs$:			
					Even		Uneven	
Binding	7.66	**	8.07	**	13.35	***	4.10	
	(3.66)		(3.26)		(4.58)		(3.92)	
Endowment	16.91	***	10.76	***	19.48	***	5.24	
	(3.68)		(3.51)		(5.32)		(4.46)	
Returns	1.83		0.59					
	(3.76)		(3.39)					
Top up	4.26		3.82					
	(5.80)		(4.85)					
Promise	-0.71		1.77					
	(5.84)		(5.01)					
Lab	12.34	***	10.45	***	2.40		14.18	***
	(3.19)		(2.84)		(5.36)		(4.14)	
Unconditional c_L			0.12	***	0.08		0.07	
			(0.03)		(0.05)		(0.04)	
Lagged c_L			0.07	***	0.11	***	0.06	**
			(0.02)		(0.03)		(0.03)	
Constant	14.68	***	8.41	***	11.25	***	20.64	***
	(3.74)		(3.13)		(3.65)		(4.14)	
R^2	0.131		0.197		0.295		0.101	
Subjects	408		408		160		152	
Observations	4896		4590		1800		1710	

Table 3: *Follower contributions. Random (individual) effects GLS regressions. (Robust standard errors clustered on groups). *10%; **5%; ***1%.*

Table 2 reports the results of four GLS regressions. Model 1 includes our five institutional variables: Binding (scores 1 if the promise is implemented automatically, 0 otherwise), Endowment (scores 1 if endowments are uneven, 0 otherwise), Returns (scores 1 if MPCR are uneven, 0 otherwise), Top-up (scores 1 if leader can top up, 0 otherwise), and Promise (scores 1 if leader can make a promise, 0 otherwise). It also includes the control variable Lab (scores 1 for Boone sessions, 0 for Oslo sessions).

Both Endowment and Binding have a positive and significant effect on average follower contributions. Top-up, Promise, and Returns have no significant effect in Model 1. Lab has a significant positive effect.

A main finding in the experimental literature on public goods provision is that a significant fraction of subjects reciprocate the actions of others.[26] To account for this behavioural regularity, Model 2 adds the leader's unconditional contribution in the current period and the leader's total contribution in the previous period as control variables.[27] Because data for the leader's total

[26] Chadhuri (2011) provides a thorough review of experimental results. Theories of reciprocity are provided in e.g. Sugden (1984) for public goods games, and in Falk and Fischbacher (2006) for a more general setting.

[27] Both variables are observable at the contribution stage of followers, and are therefore potentially subject to follower reciprocation of leader behavior.

contribution in the previous period is undefined for period 1, the number of observations is lower for Model 2 than for Model 1. Both additional controls have a positive and significant effect. Concerning the other variables, the most important change from Model 1 to Model 2 is that the effect of Endowment declines in magnitude. However, both Binding and Endowment retains significantly positive effects. Thus, it seems that these two variables' effect on average follower contributions is partly mediated by the average leader contribution.

Non-parametric tests (see online appendix D) indicate the presence of statistical interaction; hence, we also analyze our data separately for treatments with even returns and for treatments with uneven returns. The results are presented in Model 3 and confirm that interaction effects are indeed present. With uneven returns ($\alpha_L = 0.64$), both Binding and Endowment have only weak positive effects that are not statistically significant at conventional levels. However, with even returns ($\alpha_L = 0.4$), each variable's effect increases by a factor of about three. They also become strongly significant. Concerning the controls, it is worth noting that Lab is no longer significant. Finally, R-squared is higher than in any of our other regressions.[28]

In summary, our regressions confirm that leading by conditional commitment can enhance followers' contributions to a public good. They also confirm that this effect depends on the institutional setting. In particular, leading by conditional commitment is most likely to induce followers to contribute (more) if the leader has both credibility and ability to influence followers' payoff substantially, while the benefits from cooperation are shared evenly.[29]

We summarize these findings in two results:

Result 1 *Leadership by conditional commitment enhances public goods provision under some conditions, yet falls substantially short of solving the collective-action problem faced by the subjects in our experiment.*

Result 2 *Endowment and binding promises interact with returns concerning their effect on average follower contributions.*

In all treatments, follower contributions are declining over time, as can be seen in Figure 2. The steepness of the decline varies by treatment. In every treatment the decline is significantly different from zero at conventional levels.[30] and we find no interaction between treatment variables and temporal decline (see appendix E)". Hence, it appears that treatments shift the entire contribution curve vertically while not affecting its slope, which remains negative throughout. Declining contributions have been documented in countless variants of public goods games, and is considered a "core fact" (Ostrom 2000).[31]

[28] Model 3 includes only the treatments where conditional promises can be made, which excludes T1 and T2. In the remaining treatments, the variables Topup and Promise are constants (equal to 1), and hence excluded from the regression.

[29] At face value the result that pledges only works if leaders have access to a commitment technology might seem trivial. However, a large experimental literature demonstrates that threats and promises that are non-credible under assumptions of pure-self regard and rationality, are nevertheless frequently enforced by behavioral mechanisms (see e.g. the surveys in Chaudhuri 2011, and Fehr and Fischbacher 2005). For such reasons, we regard the finding that access to a commitment technology is essential for effective leadership as non-trivial.

[30] Tested by regressing follower contributions on periods as a running variable.

[31] It might seem surprising that a significant decline is present even in treatments where leaders can make binding pledges. We speculate that the observed decay is due to followers reciprocating each other's declining contributions, in parallel to what is commonly observed in public goods experiments without a sequential structure of moves.

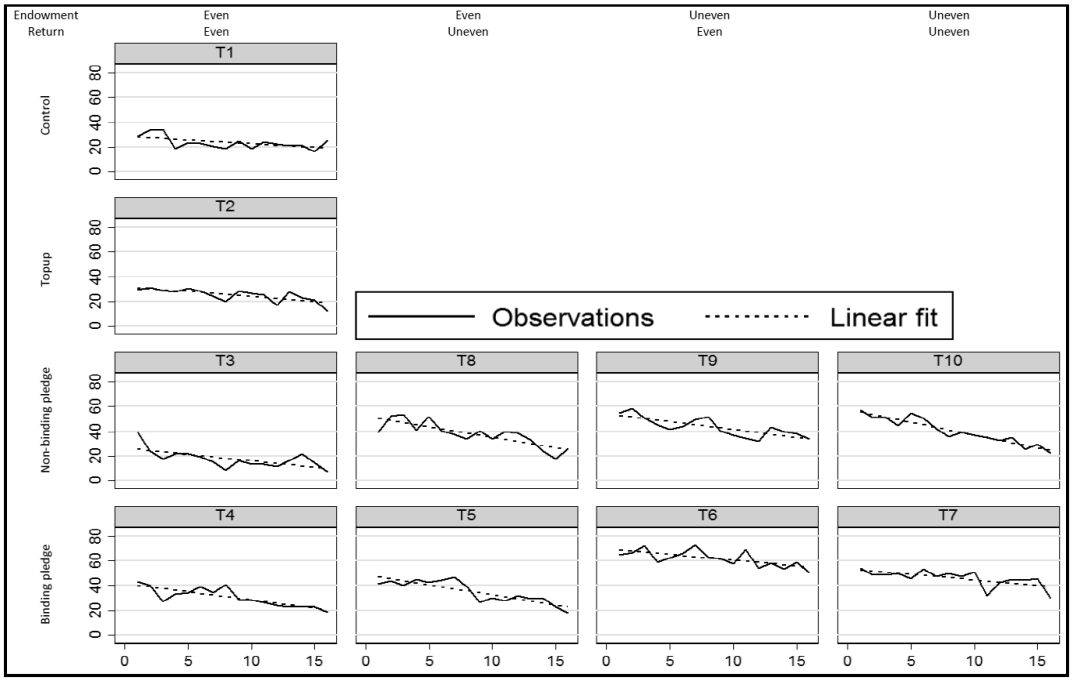

Figure 2: *Time-paths of average play by treatment*

5.3 Coordination

As shown in Section 4, the leader can—by fulfilling condition (1)—create a coordination game for the followers, assuming that the leader's promise is binding. For a target of 100 ECUs in T6 there is only one way to play the positive-contributions equilibrium. In treatments T4, T5, and T7, the positive-contributions equilibrium requires a lower target and can be played in multiple ways. Thus, we expect leaders to set higher targets more frequently in T6 than in the other three treatments with binding promises.

Figure 2 displays the cumulative frequency of targets for the four treatments with binding promises. As can be seen, higher targets are set on average in T6 than in T4, T5, and T7. Indeed, while almost 30% of the targets in T6 are 90 ECUs or higher, only 2–12% of the targets in T4, T5 and T7 are at or above this level.

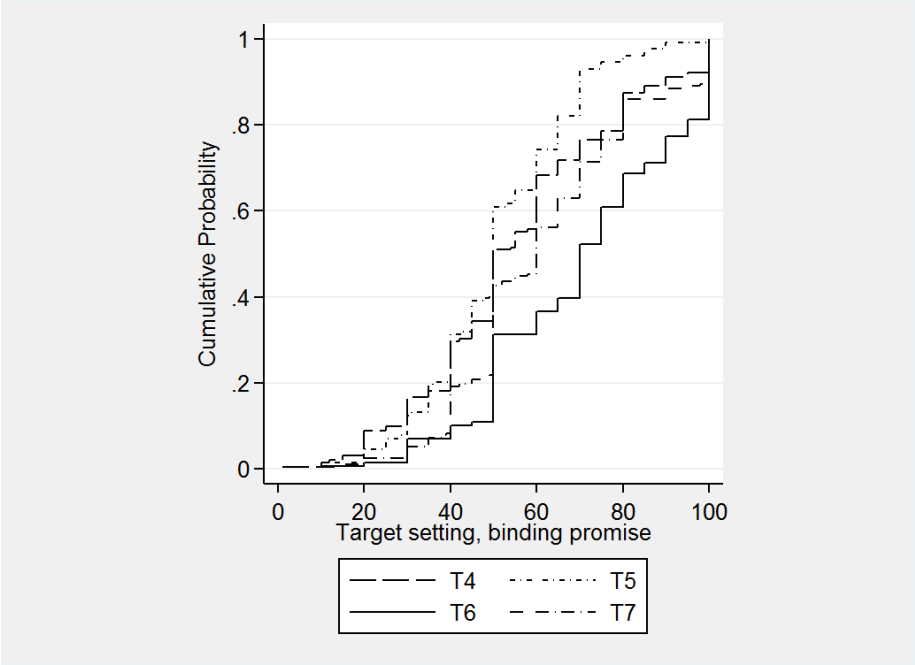

Figure 3: *Cumulative frequencies of target setting in treatments with binding promises*

Do promises work? Figure 3 shows (a) the proportion of rounds in which condition (1) is fulfilled and (b) the proportion of rounds in which condition (1) is fulfilled *and* followers meet or exceed the leader's stated target. Treatments with nonbinding promises are included for comparison.

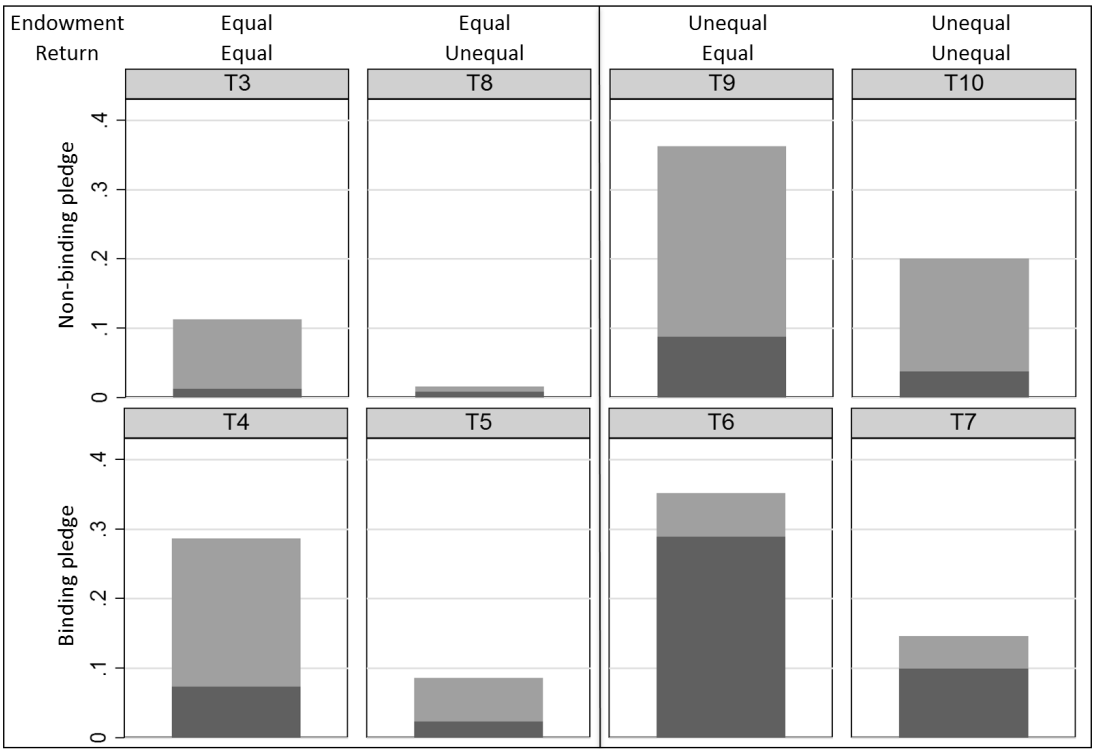

Figure 4: *Light grey bars: the proportion of rounds in which the leader fulfills condition (1). Dark grey bars: the proportion of rounds in which condition (1) is fulfilled* and *followers meet or exceed the leader's target.*

Figure 4 shows that condition 1 is met more often when returns are distributed evenly than when returns are distributed unevenly (compare T4 versus T5, T6 versus T7, T3 versus T8, and T9 versus T10). The reason is that condition 1 requires a larger promise-target ratio when returns are distributed unevenly. The figure also shows that condition 1 is met more often when the leader's endowment is large than when it is small (compare T4 versus T6, T5 versus T7, T3 versus T9, and T8 versus T10). In contrast, whether the promise is binding has no systematic effect on whether condition 1 is met. However, the fact that equilibria with positive contributions exist only when promises are binding is reflected in the followers' behaviour: Binding commitments have a huge effect on whether the followers meet (or exceed) the target (compare treatments vertically). In addition, the size of the leader's endowment also affects whether the followers meet (or exceed) the target. Hence, it seems that the ability to influence followers' welfare substantially is important both because it affects the size of the leader's promise and because it affects how followers respond. In contrast, credibility is important largely because it generates equilibria with positive contributions, thereby causing followers to contribute more.

Result 3 *Leaders with a large endowment tend to promise more than do leaders with a small endowment. Followers tend to meet the leader's condition more often when the leader has a large endowment and when the leader's promise is credible.*

5.4 Trade-off in target setting

In our model, the optimal target is a function of the conditional commitment promised, and the optimal relationship between the two variables varies between treatments, as shown in section 3. Table 2 lists the theoretical equilibrium values for the two variables across the four treatments with binding promises. The mean values observed in the same groups are reported in Table 4. Comparing the two tables reveals that all the observed differences between treatments have the theoretically predicted sign. WRS tests indicate that most of the differences are significant (at the 1% level). For conditional promises, theory predicts five pair wise differences. All except one of these differences are significant (the exception being T6-T7). No difference is predicted between T4 and T5, and indeed the data shows no significant difference. For the targets, three of the six predicted differences are significant (T6-T4, T6-T5, and T7-T5). In sum, the variations across treatments are consistent with theory. However, the results also show that leaders set the targets too high relative to the conditional promises.[32]

Treatment	Mean conditional promise	Mean target
T4	57.7	55.7
T5	61.0	50.9
T6	87.1	70.5
T7	95.1	61.7

Table 4: *Mean conditional promises and mean targets in treatments with binding promises*

6 Discussion and Conclusions

Our results suggest that two factors influence the prospects for leadership by conditional commitments to enhance cooperation. First, it helps if the leader's promise is credible, that is, if followers have reason to believe that fulfilling the leader's stated condition will cause the leader to actually implement its promise. Without such credibility, the followers' incentive to fulfill the leader's condition is diluted.

Second, it also helps if the leader has a large endowment, that is, if it has the ability to influence followers' welfare substantially. Indeed, unless the leader has such ability, followers cannot benefit by joining forces to fulfill the leader's condition—even if the leader's promise is credible. This result suggests that the country most likely to influence others through conditional commitments is China, whose emissions are roughly three times those of the EU.

Each factor's effect, however, is present only if the leader does not reap disproportionate gains from the group's collective efforts—a result that concurs with previous findings from experimental work on simultaneous-move public goods games with punishment opportunities. These previous findings show that efficient contribution norms do not easily evolve in groups where some members benefit significantly more from cooperation than others do.

The importance of fairness has long been recognized in the scholarly literature; indeed, substantial evidence indicates that people might be prepared to make substantial material sacrifices if they can thereby ensure a fair distribution (Fehr and Gächter 2000). Fairness considerations

[32] In online appendix C, we also demonstrate the absence of a hump shaped relationship between Target and followers contribution. In addition we show that leaders in general prefer to lead by unconditional rather than by conditional contributions. This preference, however, disappears when leaders are given access to a commitment technology. Secondly, we show midrange targets are not more effective than high or low targets in generating follower contributions.

have motivated scholars working on international environmental agreements to focus on compensatory arrangements, such as side deals (Gosnell and Tavoni 2016) and technological or financial transfers (Aldy and Stavins 2012; Harstad 2012; 2016). In contexts characterized by widespread heterogeneity concerning endowments, gains, or both, the introduction of such compensatory arrangements might solve many concerns related to fairness. While compensatory arrangements are not included in our experiment, our results support the notion that they may play an important role for boosting contributions to a public good (such as climate change mitigation).

Our results shed light on why the EU's conditional commitment under the Cancun agreement largely failed to induce other major emitting countries to reciprocate. As argued by Underdal *et al.* (2012: 485), the EU's conditional promise to raise emissions reductions from 20% to 30% below 1990 levels was probably credible. However, only about 10% of global emissions come from sources within the EU countries.[33] Thus, the difference between reducing EU emissions 30% and reducing them 20% corresponds to an additional reduction of global emissions of only about 1%. It is understandable that other major emitting countries showed little interest in undertaking substantial and costly additional emissions reductions to secure such a modest global effect. In addition, these other major emitting countries faced huge coordination problems in fulfilling the EU's stated condition. First, the costs of fulfilling this condition could likely be split in many different ways (however, it is hard to be sure, because the EU's stated condition was rather vague). And second, competing norms exist concerning what is a fair division of the required mitigation burden. Thus, even in the unlikely event that all other major emitting countries desired joint action to fulfill the EU's condition, competing contribution norms could easily undermine their ability to coordinate.

Estimates suggest that the current NDCs under the Paris agreement will – if implemented – entail substantial reductions in global emissions. As suggested by Carattini *et al.* (2017), the many relatively ambitious NDCs may have been facilitated partly by local social norms and partly by leadership in the form of ambitious pledges submitted early in 2015 by the EU as well as by some non-EU countries (e.g., Switzerland), perhaps in the expectation that other countries would reciprocate.[34]

However, current pledges will unlikely suffice to reach Paris' goal of no more than 1.5°-2°C warming, compared to preindustrial times (e.g., see Young 2016). Thus, pledges seem to help but are unlikely to solve the global climate dilemma. Acknowledging this problem, the Paris agreement created a system for ratcheting up the member countries' NDCs over time. Our experiment enables us to comment on the likely success of this system. Intuitively, one might expect a system of pledge-and-review to motivate countries to make deeper contributions over time. However, our results do not support this conjecture; rather, in our experiment contributions decrease over periods, just like they do in standard public goods games. Moreover, previous experiments and simulations suggest that enforcement is required to ensure high or even increasing contributions over periods (e.g., Ostrom *et al.* 1992; Aakre *et al.* 2016; Sælen 2014). However, NDCs are not subject to enforcement (beyond naming and shaming); indeed, they are not even legally binding. Despite that Paris requires countries to submit gradually more ambitious NDCs over time, it is therefore far from obvious that further rounds of pledges will ensure sufficient mitigation to avoid "dangerous anthropogenic interference with the climate system" (UNFCCC 1992). In particular, one cannot take it for granted that all countries' current and future NDCs will prove credible, at least not unless a series of technological breakthroughs occur (Narita and Wagner 2017).[35] Thus, our results suggest some caution against excessive optimism concerning Paris.

[33] European Commission (2014).

[34] Using a sequential public good game with exogenous ordering, Cartwright and Patel (2010) show that actors who are placed early (enough) in the sequence and expect actors who are placed later in the sequence to imitate their choice will contribute.

[35] A country's ability to commit credibly to ambitious NDCs, for example by way of a strong climate law, may

Finally, this conclusion is reinforced by yet another finding: Even under favorable circumstances, leading by conditional commitments has only a limited effect in our experiment. Indeed, it comes nowhere near fully solving the underlying collective action problem in any of our treatments. At best, it motivates followers to contribute around half their endowment (on average); thus, the outcome invariably remains severely suboptimal. However, although our results indicate that leadership through intrinsically conditional commitments cannot overcome the problem of climate change, they also suggest that such leadership might at least serve as one helpful element in a bigger package of measures. A bigger package is exactly what Victor (2011) advocates. Thus, the parties to the Paris agreement – especially the major emitters – should seriously consider incorporating intrinsic conditional commitments in their future pledges under the Paris agreement.

Supplementary material

Supplementary material – the online Appendix, the data and the replication files – is available online at the OUP website.

Funding

This work was supported by the Department of Political Science at the University of Oslo [Småforsk 2014]; and the Norwegian Research Council [212996/F10].

Acknowledgements

Previous versions of this paper were presented at the International Studies Association's Annual Convention in Toronto, 26 March 2014 (panel WC30), at the Oslo Academy of Global Governance workshop, 25 September 2014, and at the 9th NCBEE Conference in Aarhus, 26 September 2014. We are indebted to the participants in and attendants of these events, to our two anonymous referees and the journal editor, and to Todd Cherry and Xinyuan Dai for helpful comments.

References

Aakre, S., Helland, L., and Hovi, J. (2016) When does (informal) enforcement work? *Journal of Conflict Resolution*, 60, 1312–1340.

Aldy, J.E., and Stavins, R.N. (2012) The promise and problems of pricing carbon: Theory and experience, *Journal of Environment and Development*, 21, 152–80.

Almer, C., and Winkler, R. (2017) Analyzing the effectiveness of international environmental policies: The case of the Kyoto Protocol, *Journal of Environmental Economics and Management*, 82, 125–151.

Ball, S., Bazerman, M., and Carroll, J. (1991) An evaluation of learning in the bilateral winner's curse, *Organizational behavior and Human Decision Processes*, 48, 1–22.

Barrett, S. (1994) Self-enforcing international environmental agreements, *Oxford Economic Papers*, 46, 878—894.

Barrett, S. (2002) Consensus treaties, *Journal of Institutional and Theoretical Economics*, 158, 529–547.

also depend on domestic political economy factors, such as the strength of the government (Fankhauser *et al.* 2015) and the relative strength of supportive and opposing lobby groups (Marchiori *et al.* 2017). It may also depend on international factors, such as other countries' commitments and whether the country concerned has hosted a UNFCCC meeting (Fankhauser *et al.* 2016).

Barrett, S. (2003) *Environment and statecraft: The strategy of environmental treaty-making.* Oxford University Press, Oxford.

Barrett, S., and Danneberg, A. (2012) Climate negotiations under scientific uncertainty, *Proceedings of the American Academy of Science*, 109,17372–76.

Barrett, S., and Dannenberg, A. (2014) Sensitivity of collective action to uncertainty about climate tipping points, *Nature Climate Change* 4, 36–39.

Buchholz, W., Haslbeck, C., and Sandler, T. (1998) When does partial cooperation pay? *Finanzarchiv*, 55, 1–20.

Buchholz, W., and Sandler, T. (2017). Successful leadership in global public good provision: Incorporating behavioural approaches. *Environmental and Resource Economics*, 67(3), 591-607.
Carattini, S., Levin, S., and Tavoni, A. (2017) *Cooperation in the climate commons*, GRI Working Paper 259, Grantham Research Institute on Climate Change and the Environment, London.

Carraro, C., and Siniscalco, D. (1993) Strategies for the international protection of the environment, *Journal of Public Economics*, 52, 309–28.

Carraro, C., and Siniscalco, D. (1998) International institutions and environmental policy: International environmental agreements: Incentives and political economy, *European Economic Review*, 42, 561–72.

Cartwright, E., and Patel, A. (2010) Imitation and the incentive to contribute early in a sequential public good game, *Journal of Public Economic Theory*, 12, 691–708.

Chaudhuri, A. (2011) Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature, *Experimental Economics*, 14, 47–83.

Cherry, T.L., Kroll, S., and Shogren, J. F. (2005) The impact of endowment heterogeneity and origin on public good contributions: Evidence from the lab, *Journal of Economic behaviour & Organization*, 57, 357–65.

Dannenberg, A., Löschel, A., Paolacci, G., Reif, C., and Tavoni, A. (2015) On the provision of public goods with probabilistic and ambiguous thresholds, *Environmental and Resource Economics*, 61, 365–83.

Eckersley, R. (2012) Moving forward in the climate negotiations: multilateralism or minilateralism? *Global Environmental Politics*, 12, 24–42.

Erev, I., and Rapaport, A. (1990) Provision of step-level public goods. The sequential-contribution mechanism, *Journal of Conflict Resolution*, 34, 401–425.

European Commission (2014) EU Greenhouse Gas Emissions and Targets.
Available from `http://ec.europa.eu/clima/policies/g-gas/index_en.htm` (accessed 13 May, 2017).

Falk, A., and Fischbacher, U. (2006) A theory of reciprocity, *Games and Economic Behavior*, 54, 293–315.

Fankhauser, S., Gennaioli, C., and Collins, M. (2015) The political economy of passing climate change legislation: Evidence from a survey, *Global Environmental Change*, 35, 52–61.

Fankhauser, S., Gennaioli, C., and Collins, M. (2016). Do international factors influence the passage of climate change legislation? *Climate Policy*, 16, 318–31.

Fehr, E., and Gächter, S. (2000) Cooperation and punishment in public goods experiments, *American Economic Review*, 90, 980–994.

Fehr, E., and Fischbacher, U. (2005) The economics of strong reciprocity, in H. Gintis, S. Bowles, R. Boyd and E. Fehr (eds) *The Foundations of Cooperation in Economic life. Moral Sentiments and Material Interests.* MIT Press, Cambridge, MA.

Feige, C., Ehrhart,K.-M., and Krämer, J. (2014): *Voting on contributions to a threshold public goods game: An experimental investigation.* Working Paper Series in Economics, 60, Karlsruher Institut für Technologie, Karlsruhe.

Finus, M., Cooper, P., and Almer, C. (2017) The use of international agreements in transnational environmental protection, *Oxford Economic Papers*, 69, 333–344.

Fischbacher, U. (2007) z-Tree: Zurich toolbox for ready-made economic experiments, *Experimental Economics*, 10, 171–178.

Fisher, J., Isaac, R.M., Schatzberg, J.W., and Walker, J.M. (1995) Heterogeneous demand for public goods: Behavior in the voluntary contributions mechanism, *Public Choice*, 85, 249–266.

Gächter, S., Nosenzo, D., Renner, E., and Sefton, M. (2010) Who makes a good leader? Cooperativeness, optimism, and leading-by-example, *Economic Inquiry*, 50, 953–967.

Gosnell, G., and Tavoni, A. (2016). *A bargaining experiment on heterogeneity and side deals in climate negotiations*, GRI Working Paper 249, Grantham Research Institute on Climate Change and the Environment, London.

Güth, W., Levatia, M.V., Sutter, M., and van der Heijden, E. (2007) Leading by example with and without exclusion power in voluntary contribution experiments, *Journal of Public Economics*, 91, 1023–1042.

Hafner-Burton, E.M., LeVeck, B.L., Victor, D.G., and Fowler, J.H. (2014) Decision maker preferences for international legal cooperation, *International Organization*, 68, 845–876.

Hall, D.T. (2013) Using role-reversal in laboratory experiments.
Available from `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2292100` (accessed 24 July, 2014).

Harstad, B. (2012) Climate contracts: A game of emissions, investments, negotiations, and renegotiations, *Review of Economic Studies*, 79, 1527–57.

Harstad, B. (2016) The dynamics of climate agreements, *Journal of the European Economic Association*, 14, 719–52.

Hoel, M. (1991) Global environmental problems: The effects of unilateral actions taken by one country, *Journal of Environmental Economics and Management*, 20, 55–70.

Holtsmark, B. (2013) International cooperation on climate change: Why is there so little progress? In Roger Fouquet (ed.) *Handbook on Energy and Climate Change.* Edward Elgar, Cheltenham.

Hovi, J., Ward, H., and Grundig, F. (2015) Hope or despair? Formal models of climate cooperation, *Environmental and Resource Economics*, 62, 665–88.

Isaac, R.M., and Walker, J. (1988) Communication and free-riding behavior: The voluntary contribution mechanism, *Economic Enquiry*, 26, 585–608.

Konrad, K. A., and Thum, M. (2014) Climate policy negotiations with incomplete information, *Economica*, 81, 244–256.

Kosfeld, M., Okada, A., and Riedl, A. (2009) Institution formation in public goods games, *American Economic Review*, 99, 1335–1355.

Köke, S., Lange, A., and Nicklisch, A. (2015) *Adversity is a school of wisdom: Experimental evidence on cooperative protection against stochastic losses*, WiSo-HH Working Paper Series, No. 22, University of Hamburg.

Levati, M.V., Sutter, M., and van der Heijden, E. (2007) Leading by example in a public goods experiment with heterogeneity and incomplete information, *Journal of Conflict Resolution*, 51, 793–818.

LeVeck, B.L., Hughes, D.A., Fowler, J.H., Hafner-Burton, E.M., and Victor, D.G. (2014) The role of self-interest in elite bargaining, *Proceedings of the National Academy of Sciences*, 111, 18356-541.

Marchiori, C., Dietz, S., and Tavoni, A. (2017) Domestic politics and the formation of international environmental agreements, *Journal of Environmental Economics and Management*, 81, 115-31.

Maltais, A. (2014) Failing international climate politics and the fairness of going first, *Political Studies*, 62, 618–633.

McEvoy, D.M. (2010) Not it: Opting out of voluntary coalitions that provide a public good, *Public Choice*, 142, 9–23.

McEvoy, D.M., Cherry, T.L., and Stranlund, J.K. (2011) *The endogenous formation of coalitions to provide public goods: Theory and experimental evidence*, Working paper, Department of Economics, Appalachian State University, Boone, NC.

McLean, E., and Stone, R. (2012) The Kyoto protocol: Two-level bargaining and European integration, *International Studies Quarterly*, 56, 99–113.

Milinski, M., Sommerfeld, R. D., Krambeck, H.-J., Reed, F. A., and Marotzke, J. (2008) The collective-risk social dilemma and the prevention of simulated dangerous climate change, *Proceedings of the National Academy of Sciences*, 105, 2291–94.

Narita, D., and Wagner, U.J (2017) Strategic uncertainty, indeterminacy, and the formation of international environmental agreements, *Oxford Economic Papers*, 69, 432–452.

Ostrom, E. (2000) Collective action and the evolution of social norms, *Journal of Economic Perspectives*, 14, 137–158.

Ostrom, E., Walker, J., and Gardner, R. (1992) Covenants with and without a sword: Self-governance is possible, *American Political Science Review*, 86, 404–417.

Ott, H. E., and Oberthür, S. (1999) *Breaking the impasse. Forging an EU leadership initiative on climate change*, Policy paper, Heinrich Böll Stiftung, Berlin.

Reuben, E., and Riedl, R. (2013) Enforcement of contribution norms in public good games with heterogeneous populations, *Games and Economic Behavior*, 77, 122–137.

Rivas, F., and Sutter, M. (2011) The benefits of voluntary leadership in experimental public goods games, *Economics Letters*, 112, 176–8.

Schelling, T. C. (1960) *The Strategy of Conflict*, Harvard University Press, Cambridge, MA.

Selten, R. (1967) Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments, in H. Sauermann (ed) *Beiträge zur experimentellen Wirtschaftsforschung*, Mohr, Tübingen.

Skodvin, T., and Andresen, S. (2006) Leadership revisited, *Global Environmental Politics*, 6, 13–27.

Sugden, R. (1984) Reciprocity: the supply of public goods through voluntary contributions, *The Economic Journal*, 94, 772–787.

Sælen, H. (2014) The effect of enforcement in the presence of strong reciprocity: an application of agent-based modelling, in T.L. Cherry, J. Hovi and D.M. McEvoy (eds) *Toward a New Climate Agreement. Conflict, Resolution and Governance*, Routledge, London.

Tavoni, A., Dannenberg, A., Kallis, G., and Löschel, A. (2011) Inequality, communication, and the avoidance of disastrous climate change in a public goods game, *Proceedings of the National Academy of Sciences*, 108, 11825–29.

UNFCCC (1992) The United Nations Framework Convention on Climate Change. Available from: http://unfccc.int/files/essential_background/background_publications_htmlpdf/application/pdf/conveng.pdf (accessed 30 April 2017)

UNFCCC (2011a) Decision 1/CP.16 The Cancun Agreements. Available from http://unfccc.int/resource/docs/2010/cop16/eng/07a01.pdf (accessed 2 June 2015)

UNFCCC (2011b) FCCC/SB/2011/INF.1/Rev.1. Compilation of economy-wide emission reduction targets to be implemented by Parties included in Annex I to the Convention. Available from http://unfccc.int/resource/docs/2011/sb/eng/inf01r01.pdf (accessed 2 June 2015)

UNFCCC (2015) INDCs as communicated by Parties. Available from http://www4.unfccc.int/submissions/indc/Submission%20Pages/submissions.aspx (accessed 13 May 2017)

Underdal, A., Hovi, J., Kallbekken, S., and Skodvin, T. (2012): Can conditional commitments break the climate change negotiations deadlock? *International Political Science Review*, 33, 475–493.

Victor, D. (2011) *Global Warming Gridlock. Creating More Effective Strategies for Protecting the Planet*, Cambridge University Press, Cambridge, MA.

Weischer, L., Morgan, J., and Patel, M. (2012) Climate clubs: Can small groups of countries make a big difference in addressing climate change? *Review of European Community and International Environmental Law* (RECIEL), 21, 177–192.

Weimann, J., Brosig-Koch, J., Hennig-Schmidt, H., Keser C., and Stahr, C. (2012) *Public-good experiments with large groups*, working paper, Faculty of Economics, Georg-August-Universität, Göttingen.

Young, O.R. (2016) The Paris agreement: Destined to succeed or doomed to fail? *Politics and Governance*, 4, 124–13.