In: Climatic Change. 06 June 2016 (online) DOI: 10.1007/s10584-016-1711-4

An Experimental Investigation into ‘Pledge and Review’ in Climate Negotiations

Scott Barrett,[1*] Astrid Dannenberg[2,3]

[1]Columbia University School of International and Public Affairs & Earth Institute

*Correspondence to: sb3116@columbia.edu; cell +1-646-300-1437.

[2]Department of Economics, University of Kassel

[3]Department of Economics, University of Gothenburg

June 2016

Abstract. A novelty of the new Paris Agreement is the inclusion of a process for assessment and review of countries’ nationally determined pledges and contributions. The intent is to reveal whether similar countries are making comparable pledges, whether the totality of such pledges will achieve the global goal, and whether, over the coming years, the contributions actually made by countries will equal or exceed their pledges. The intent is also to provide an opportunity for countries to express their approval, or disapproval, of the pledges and contributions made by individual countries. Here we report the results of a lab experiment on the effects of such a process in a game in which players choose a group target, declare their individual pledges, and then make voluntary contributions to supply a public good. Our results show that a review process is more likely to affect targets and pledges than actual contributions. Even when a review process increases average contributions, the effect is relatively small. As the window for achieving the 2 °C goal will close soon, our results suggest that, rather than merely implement the Paris Agreement, negotiators should begin now to develop complementary approaches to limiting emissions, including the adoption of agreements that are designed differently than the one adopted in Paris.

Keywords. Climate change, negotiations, experiment, pledge and review, game theory

Acknowledgements: We thank Alessandro Tavoni and two anonymous reviewers for comments and James Rising for developing the software for our spinning wheel. We are also grateful to the Magdeburg Experimental Laboratory for Economic Research team at Magdeburg University for support in conducting the experiment. This work was supported by the Princeton Institute for International and Regional Studies research community on Communicating Uncertainty: Science, Institutions, and Ethics in the Politics of Global Climate Change and the European Union (EU) Horizon 2020 program, action ERC-2014-STG, Project HUCO, grant number 636746.

1. Introduction

For 25 years, countries have been trying to negotiate agreements to limit global emissions of greenhouse gases, and yet all this time emissions have continued to increase. The 2009 Copenhagen conference invited countries to submit quantified, nationally determined emission reduction targets aimed at limiting mean global temperature change to 2 ˚C. However, the submissions made subsequently fell far short of the levels needed to meet this goal (Rogelj et al 2010), and so countries decided to negotiate a new agreement. In subsequent conferences, countries were urged to submit pledges for emission reductions known as "intended nationally determined contributions," to include a reference point for the emission target and a time frame for meeting it. As the negotiations advanced, it became clear that the new treaty's main novel feature would be a procedure for pledge and peer review, and the agreement ultimately adopted in Paris retains this feature. Negotiators have long appreciated the need for monitoring and verification (Thompson 2006), but as Aldy (2014: 283) has noted, the review process adopted by the UNFCCC before Paris did "not include a formal peer review mechanism." Paris moved the review process a step closer in this direction. "In order to build mutual trust and confidence," Article 13 of the Agreement establishes a "transparency framework" for the "tracking" of a country's "progress towards achieving [its] individual nationally determined contributions," with the information supplied being subject to a "technical expert review," the purpose of which is to determine whether a party has achieved its "nationally determined contributions" and to identify "areas of improvement for the Party…." Moreover, the agreement requires that each party "participate in a facilitative, multilateral consideration of progress with respect to… implementation and achievement of its nationally determined contribution." Article 14 goes on to say that parties shall also "periodically take stock of the implementation of [the] Agreement to assess [their] collective progress towards achieving" the 2 °C goal.

Will the new agreement work any better than the approaches tried previously? It will take years to know. The end-dates for the intended nationally determined contributions declared in Paris occur in 2025 and 2030, so it will take a decade or more to know whether the pledges made there are actually fulfilled. Even then, estimation of the *effect* of the new agreement will be difficult, since we will never be able to observe the "counterfactual"—the emissions that would have come about had Paris never been negotiated. In other contexts, humans have been shown to be sensitive to social feedback, even when the feedback does not involve a direct cost (Masclet et al 2003; López-Pérez and Vorsatz 2010; see the electronic supplementary material for a review). However, the climate problem differs in important ways from the settings studied previously.

Here we report the results of a new experiment designed to capture key features of the climate problem and the design of the new Paris Agreement. First, the players in our experiment can choose between "cheap" cooperation and "expensive" cooperation—a crucial distinction, as only expensive cooperation can stabilize concentrations and so reach the Paris Agreement's collective goal of "[h]olding the increase in the global average temperature to well below 2 °C above pre-industrial levels." Second, our game is played not only between individuals within a group but between the group and Nature. To avert a "catastrophic" outcome, players must undertake "expensive" cooperation, and the more they contribute collectively, the more they reduce the probability of triggering a "dangerous" outcome. Third, the players in our game choose more than their contributions. They also choose their collective goal (a value that is akin to the "carbon budget" associated with the 2 °C goal), which relates to the game they are playing against Nature (avoiding "dangerous" climate change), and their individual "pledges," which are represented in our experiment by non-binding declarations about a player's intentions to contribute in the future. A "review" in our experiment represents a judgment made by other players about an individual player's behavior. The process of "pledge and review" in our game allows players to be "judged" for both their "ambition" (their individual pledges relative to the collective goal) and for their contributions relative to their pledges. Transparency about contributions is thus critical to the process of peer review. Finally, our experiment explores the implications of varying the timing at which a review takes place. Timing was clearly considered to be important to some negotiators, as an earlier draft of the treaty distinguished between an "ex ante" review, conducted after pledges had been submitted but before contributions were made, and a "strategic" review, undertaken after contributions had been made.[1] The final agreement uses different language but still emphasizes the need for "tracking of progress."

Though a laboratory experiment obviously cannot tell us what will actually happen in the wake of Paris, it can provide a comparison between a situation without a review process and situations with a review process, and so can show whether a process of review causes proposals, pledges, and, most important of all, actual contributions to increase.

2. Experimental design

Our analysis is based on a laboratory experiment of a game played by groups of five players. In this game, every player is endowed with 5 black poker chips worth €.10 each and 15 red poker chips worth €1.00 each. Hence, the group has 100 chips overall (25 black chips and 75 red chips), and both types of chip can be invested to "mitigate climate change." We can think of the black chips as a low-cost technology for "ordinary abatement" and the red chips as a high-cost technology for removing carbon dioxide from the atmosphere (Keith 2009). Contributions of either type of chip by any player gives every player in the group a return equal to €.05. This is the marginal benefit of avoiding "gradual" climate change. The game also involves "catastrophic" climate change, the avoidance of which is feasible but requires using both the low- and the high-cost technologies. If 50 or fewer chips are contributed overall, a threshold will be crossed, causing each player to lose €20. If the group

[1] We refer specifically to the 12 February 2015 text of the Ad Hoc Working Group on the Durban Platform for Enhanced Action, available at https://unfccc.int/files/bodies/awg/application/pdf/negotiating_text_12022015@2200.pdf.

contributes more than 50 chips, the probability of crossing the threshold declines linearly as more and more chips are contributed, reaching zero if and when all of the group members contribute all of their chips. To have any chance of avoiding "catastrophe," the players must thus contribute expensive chips and not only their cheap chips. This game design makes contributing chips a prisoners' dilemma: from the group's perspective, it is best for everyone to *contribute* all of their chips but from any individual's perspective it is best to *keep* all of his or her chips.[2]

The experiment was presented in a neutral frame as regards context and language to avoid any potential bias; there was no mention of "climate change," "cooperation," or "catastrophe" (instructions can be found in the SI). The game was played in stages. First, individuals made "proposals" for a group target knowing that the median value would be selected as the group's "target." Second, each player pledged an amount he or she intended to contribute subsequently. Third, the players made their actual contributions over two stages. It was common knowledge that the targets and pledges were non-binding and that all values would be revealed to every member of the group after each stage. Because players were allowed to contribute over two stages, players could see how much their co-players had contributed in the first stage before deciding how much to contribute in the second stage.

The game just described represents the *No-Review* treatment in which the players lack an explicit mechanism for expressing their judgment about other players' behavior. There were also three treatments that incorporated an explicit review process (see Fig. 1). In each of these treatments, every player "graded" all of the other players plus him or herself. Grades were on a scale from 1 to 6, with a grade of 1 being "very good" and 6 "insufficient." (Our experiment was conducted in Germany, and German students are familiar with this grading scale from their high school days.) After the grades had been submitted, the average grade given to every player was revealed to the group. The grades that the players gave to themselves were not revealed publicly. This grading scheme did not affect payoffs directly, but it did provide a vehicle for "peer review" by allowing the players to signal their approval or disapproval of the choices made by the members of their group. In the *Ex-Ante-Review* treatment, the review was done after the pledges but before the contributions were made. In the *Mid-Point-Review*, the review was done between the first and second contribution stage. Finally, in the *Ex-Post-Review* treatment, the review came after the second contribution stage.

In this game, the incentive to contribute red chips depends very much on players' expectations or "beliefs" for how many red chips their co-players will contribute. This is because contributions of red chips are very costly to individuals and even inefficient for the group so long as fewer than 50 chips are contributed in total, and at least some players must contribute red chips in addition to black chips in

[2] The game-theoretic model underlying our experiment is a specific representation of a more general theoretical model (Barrett 2013), and has been used in previous experimental investigations (Barrett and Dannenberg 2012, 2014). For details, see our electronic supplementary materials. For different but related experiments on "dangerous" climate change, see Milinski et al (2008), Tavoni et al (2011), and Dannenberg et al (2015).

order for the group contribution to top 50. To obtain an estimate of each player's expectations, just before contributions were chosen, players were asked to guess how many chips their co-players would contribute on average. To ensure that estimates reflected players' actual expectations, they were given a reward of €1 for correct guesses (meaning guesses that were within a range defined by the actual mean plus or minus 1).

In addition to the 20 poker chips, each player was given an "endowment fund" of €19 to ensure that he or she could not be left out of pocket.[3] The endowment fund could not be used to purchase chips, and so you can think of it as representing a country's "capital stock," a resource that cannot be used to mitigate climate change but that would be at risk should dangerous climate change occur. Given this fund, a player's worst possible payoff in the experiment was €0, and her best possible payoff was €38.50. The full cooperative payoff to each player was €24 and the Nash equilibrium payoff was €14.50. After the game was played, the participants were asked to complete a follow-up questionnaire. After that, the threshold was determined by the randomized spin of a computer wheel, with the "ends" set at 50 and 100. The wheel, representing Nature, determined whether, for those groups contributing between 50 and 100 chips, the players would lose the €20. Depending on the outcome of the spin, the players were then given their final payout in cash.

The experimental sessions were held in a computer lab at the University of Magdeburg, using undergraduate students recruited from the general student population. In total, 195 students participated in the experiment, each student taking part in one treatment only. In each session, 20 or 25 subjects were seated at linked computers (game software Ztree; see Fischbacher 2007) and randomly assigned to five-person groups.[4] Throughout the game, each player was identified by a different letter, from A to E. The experimental instructions handed out to the students included several numerical examples and control questions. The control questions tested subjects' understanding of the game to ensure that they were aware of the available strategies and the implications of making different choices. After reading the instructions and answering the control questions correctly, every subject first played the game in three practice rounds. It was common knowledge that the composition of every group would be changed between these rounds. It was also common knowledge that group composition would be changed again before the game was played for real.

3. Results

Figure 2 presents mean values for the targets, pledges, and contributions. For every treatment, the mean target exceeds the mean pledge, which in turn exceeds the

[3] The first five groups played the game with an endowment fund of €15. However, payoffs turned out to be lower than we expected, and so we increased the endowment fund to €19 for the remaining groups. Statistical tests show that this change did not affect the participants' behavior in the game (Whitney-Wilcoxon rank-sum test, P > .10 for the chosen targets, pledges, and contributions).

[4] We aimed to have 10 groups per treatment, but due to no-shows in one session, only nine groups played the Mid-Point-Review treatment.

mean contribution. In short, individual pledges fell below the group target and contributions fell below the pledges.[5] Figure 2 also shows that the mean values for targets, pledges, and contributions are higher for the three review treatments compared to the control without review, but the differences are generally small. Statistical analyses of these data (see our electronic supplementary materials) show that the differences in targets between *No-Review* and *Ex-Ante-Review* and between *No-Review* and *Ex-Post-Review* are significant (Mann-Whitney-Wilcoxon rank-sum test (MWW), $P < .05$ each). Moreover, the differences in pledges between *No-Review* and the three review treatments are at least weakly significant (MWW test, $P < .10$ each). The differences in contributions between *No-Review* and *Ex-Ante-Review* as well as between *No-Review* and *Mid-Point-Review* are not statistically significant (MWW test, $P > .30$ each).

The largest aggregate contributions are found in the *Ex-Post-Review*. For this treatment, the average contribution is 19 percent higher than in *No-Review* and this difference is on the borderline of statistical significance (MWW test, P = .112).[6] On average, this means that the probability of catastrophe decreases from 84 percent in *No-Review* to 62 percent in *Ex-Post-Review*. Figure 3 shows both the distribution of group contributions and the median value. In *Ex-Post-Review*, the median is above 70 while in the remaining three treatments it is around 60.

Regression analysis reveals the critical chain of causality that underpins the effects of the review process (Table 1). The review process increases individual proposals for the group targets (and, hence, group targets) directly, with the effect being statistically significant for the *Ex-Ante-Review* and *Ex-Post-Review* treatments. The review process does not have any other direct effects, but it does have indirect effects. First, the review process increases pledges indirectly, because pledges increase with the group target. Second, the review process increases players' expectations for how much their co-players will contribute, as these expectations depend on the pledges made by other players (which in turn depend on the group targets). Finally, the review process increases contributions indirectly. It does this, first, by increasing pledges (which depend on targets), as people who pledge more tend to contribute more; and, second, by increasing players' expectations for how much their co-players will contribute, as contributions increase in these expectations. The review process does not affect contributions directly.

Figure 4 arranges all groups according to their contribution level, from lowest to highest. It also shows the corresponding group values for pledges and expectations. All groups with high contributions have high pledges and expectations. Groups with low pledges and expectations tend to have low contributions. However, not all groups

[5] Contributions relative to the monetary endowment (in our experiment, the value of all chips given to a player at the start of the game), ranges from 46% (*No-Review*) to 60% (*Ex-Post-Review*), which is close to what has been observed previously in one-shot public goods games (Ledyard 1995).

[6] As one of our reviewers pointed out, the difference in contributions between *No-Review* and *Ex-Post-Review* is weakly significant according to a one-sided t-test (P = .086). The results for all other comparisons remain qualitatively unchanged using either a two-sided or one-sided t-test instead of the non-parametric MWW test.

with high pledges and expectations have high contributions. High pledges and high expectations thus appear to be necessary but not sufficient for high contributions.

The behavior of individuals mirrors these observations about groups. Figure 5 shows that, with one exception (in the *Ex-Ante-Review* treatment), individuals who pledged to give a low contribution gave a low contribution, but that the players who pledged high sometimes gave a high contribution and sometimes gave a low contribution. Similarly, Figure 6 shows that players with low expectations tended to contribute low, but that the players with high expectations sometimes contributed low and sometimes contributed high. As observed for group behavior, high pledges and expectations are a necessary condition for high contributions by individuals, but they are not sufficient.

Table 2 shows the grades that players on average gave to their co-players and to themselves. Average grades were better when the reviews were given earlier rather than later in the process (falling from 2.1 in *Ex-Ante-Review* to 3.3 in *Mid-Point-Review* to 3.6 in *Ex-Post-Review*; remember that higher values imply a worse grade), arguably because things generally looked better earlier in the game. The grades that subjects gave to themselves were better than the grades they received by their co-players, implying that the players applied different standards to themselves than to their co-players. A plausible explanation for this is "self-serving bias," a tendency for people to perceive themselves in a more positive light than others do (Baumeister 1998). However, we do not find evidence that the difference between peer and self-assessment has any effect on behavior. Subjects who gave themselves a grade that was much better than the grade given to them by their peers did not behave differently in subsequent stages than the subjects who gave themselves a grade that was closer to the one given to them by their peers.

Regression analysis (Table 3) shows that higher pledges cause players to be given a better grade only in the *Ex-Ante-Review* treatment, where players have no other information to go on. In the *Mid-Point-Review*, a player's grade is affected only by her first period contribution, not her pledge, indicating that players care about actions, not words. Finally, in the *Ex-Post Review*, a player's grade is significantly affected by his first- *and* second-stage contributions *as well as by his pledge*. In this case, however, the coefficient on pledges has the *opposite* sign compared with the *Ex-Ante-Review* treatment. This is because, in the *Ex-Post-Review* treatment, a player's peers can see whether his contributions correspond to his pledge. The data show that people who pledged to make a low contribution tended to contribute very little, but that people who pledged to make a high contribution sometimes contributed very few chips. People who gave high pledges were thus graded down because their contributions often fell short of their pledges.

We observe a remarkably high variation in contributions in all treatments, ranging from 35-78 in *No-Review* to 54-85 in *Ex-Ante-Review*, 30-92 in *Mid-Point-Review*, and 25-95 in *Ex-Post-Review* (see Figure 3). Some groups contributed so little as to make "catastrophe" inevitable, whereas other groups contributed so much that the risk of "catastrophe" was remote.

To explore this variation in group-behavior more systematically, we divided the groups into three categories (Table 4). “Successful” groups (11 in total) contributed at least 75 chips in total; these groups had a better than even chance of avoiding catastrophe. “Intermediate” groups (22) contributed between 50 and 75 chips; these groups were more likely than not to trigger “catastrophe.” Finally, “unsuccessful” groups (6) contributed 50 or fewer chips; these groups were sure to trigger “catastrophe.” Focusing on the contrast between the successful and unsuccessful groups, the successful groups chose a higher target (MWW test, P = .0432), pledged to contribute more (MWW test, P = .0180), had higher expectations about other players’ contributions (MWW test, P = .0025), and made higher first-stage contributions (MWW test, P = .0009).

To further explore the effect of group composition, we defined “free riders” as players who contributed five or fewer chips in the first stage. In the successful groups, free riding was rare. Only one group had a free rider; the average across all these groups was just 0.09. In the unsuccessful groups, the average number of free riders was much higher (1.66), with some groups having as many as three free riders. The difference in free riding between the successful and unsuccessful groups was highly significant (MWW test, P = .0004). It thus seems that the presence of one or two “free riders” virtually guarantees a bad overall outcome. This is not only because the free riders fail to contribute. It is also because the behavior of the free riders causes the conditional cooperators to reduce *their* contributions in the second stage.

There is some controversy as to whether the 2 °C goal first endorsed by the parties to the Framework Convention in Copenhagen and Cancun is “scientifically meaningful,” let alone achievable (Victor and Kennel 2014). In our experiment, targets expressed in terms of a group’s total contribution of chips are indisputably meaningful, and the ones actually chosen are all achievable by design. The goal agreed in Paris—to limit mean global temperature change “to well below 2 °C above pre-industrial levels”—is even more ambitious than the earlier one, but does it herald stronger future collective action? In our experiment, groups that chose higher targets tended to contribute more. Groups that chose the maximum target of 100 chips contributed on average 70 chips, whereas groups that chose a lower target contributed just 59 chips on average. However, comparison of the “successful” and “unsuccessful” groups shows that the group target is only one of a multiple of preconditions for successful action. Successful groups—the ones that have a better than even chance of avoiding “catastrophe”—not only chose an ambitious target; their members also made equally ambitious pledges, had positive expectations for how much their co-players will contribute, and undertook substantial early action. Whether the ambitious goal agreed in Paris turns out to be a harbinger of substantial future emission reductions may thus depend on whether it raises expectations for national action and whether countries fulfill these expectations by taking visible strong action over the next few years.

4. Conclusions

As in our experiment, analyses of the contributions pledged in the run up to the Paris conference predict that they will fall short of achieving the 2 °C goal chosen by the same group of countries (International Energy Agency 2015, UNFCCC Secretariat 2015). Actual contributions may even come in below pledges as happened in our experiment.

Of course, our experiment focused on only one particular aspect of the pledge and review mechanism, namely its potential to change behavior under perfect information about players' contributions. Whether the pledge and review mechanism adopted in Paris will turn out to be effective may also depend on factors we did not consider in our experiment, such as transparency, public attention, and comparability. However, it is not obvious that a consideration of other factors will favor cooperation. Unlike in our experiment, the pledges submitted by countries in the run-up to Paris were expressed in different terms (total emissions, emissions intensity, emissions relative to business as usual, emissions with or without international offsets, and so on), making it difficult to know whether similar countries are pledging to make similar sacrifices (Aldy and Pizer 2014). Also, countries may be more interested in a country's effort, which is imperfectly correlated with its emissions, whereas in our experiment effort and contributions are equivalent. Cooperation by more than five countries will be needed to stabilize atmospheric concentrations of greenhouse gases, and free rider incentives generally increase with group size. Finally, efforts by a subset of countries to limit emissions may be further undermined by "globalization." For example, should only a sub-group of countries limits its emissions, market prices, including energy prices, will change, causing production of greenhouse gas intensive goods to shift towards the countries that do not limit their emissions. Similarly, the drop in fossil fuel prices brought about by a sub-group's efforts to limit emissions may increase the amount of fossil fuels consumed by other countries. Both of these responses lead to "leakage" (Felder and Rutherford 1993). Future research may show how a pledge and review mechanism fares under these alternative conditions.

We find that the pledge and review process may lead to small increases in contributions, and we find no evidence that it is harmful to cooperation. Other kinds of non-binding institution have been found to undermine cooperation by adding another source of frustration to the game (Dannenberg 2016). The implication of our research is thus not that the pledge and review mechanism should be replaced, but that it should be combined with other measures.

Our results for "successful" and "unsuccessful" groups might seem to suggest that conditional cooperators would do better by shunning the "free riders" and forming a "club" of their own—a group of likeminded countries that can deny non-members the benefits of the club members' actions (Keohane and Victor 2011). However, emission reductions are a global public good, and no country can be excluded from benefiting from the emission reductions achieved by club members. To limit climate change, clubs must therefore focus on something like cooperation in the development of a new technology or special trade arrangements—and then leverage the supply of this "good" for the purpose of getting all countries to limit emissions (Nordhaus 2015). A related but somewhat different approach emphasizes the need for agreements to focus on choices involving individual gases and sectors that facilitate coordination, with conditional cooperators offering a combination of sticks and carrots to broaden participation and increase contributions (Barrett 2003). As our experiment shows that it would be imprudent to solely rely on a review process to change the behavior of free riders, these approaches deserve more serious consideration. The priority, we believe, should be to develop coordination agreements, including the effort already underway to negotiate an amendment to limit HFCs in the Montreal Protocol, as these would be complementary to the Paris Agreement.

References

Aldy, JE (2014) The crucial role of policy surveillance in international climate policy. *Climatic Change* 126: 279-292.

Aldy JE, Pizer WA (2014) Alternative metrics for comparing domestic climate change mitigation efforts and the emerging international climate policy architecture. *Rev Environ Econ & Policy* 6: 86-109.

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

Barrett S (2013) Climate treaties and approaching catastrophes. *J Envtl Econ & Mgmt* 66: 235-250.

Barrett S, Dannenberg A (2012) Climate negotiations under scientific uncertainty. *Proc Natl Acad Sci USA* 109: 17372-17376.

Barrett S, Dannenberg A (2014) Sensitivity of collective action to uncertainty about climate tipping points. *Nature Clim Chang* 4: 36-39.

Baumeister RF (1998) The Self. In: Gilbert D, Fiske S, Lindzey G (eds) The Handbook of Social Psychology. McGraw-Hill, Boston.

Dannenberg A (2016) Nonbinding Agreements in Public Goods Experiments. *Oxford Economic Papers*, forthcoming.

Dannenberg A, Löschel A, Paolacci G, Reif C, Tavoni A (2015) On the provision of public goods with probabilistic and ambiguous thresholds. *Envtl & Resource Econ* 61**:** 365-383.

Felder S, Rutherford TF (1993), Unilateral CO_2 reductions and carbon leakage: the consequences of international trade in oil and basic materials. *J Envtl Econ & Mgmt* 25: 162-176.

Fischbacher U (2007) Z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Econ* 10: 171-178.

International Energy Agency (2015) World energy outlook special report: executive summary*,* Paris: International Energy Agency.

Keith DW (2009) Why capture CO_2 from the atmosphere? *Science* 325(5948):1654–1655.

Keohane RO, Victor DG (2011) The regime complex for climate change. *Perspectives on Politics* 9: 7-23.

Ledyard JO (1995) Public goods: a survey of experimental research. In Kagel JH, Roth AE (eds) The handbook of experimental economics. Princeton University Press, Princeton, pp 111-194.

López-Pérez R, Vorsatz M (2010) On approval and disapproval: theory and experiments. *J Econ Psychology* 31: 527-541.

Masclet, D., C. Noussair, S. Tucker, M. C. Villeval (2003) Monetary and nonmonetary punishment in the voluntary contributions mechanism. *Am Econ Rev* 93: 366-380.

Milinski M, Sommerfeld RD, Krambeck HJ, Reed FA, Marotzke J (2008) The collective-risk social dilemma and the prevention of simulated dangerous climate change. *Proc Natl Acad Sci USA* 105: 2291-2294.

Nordhaus W (2015) Climate clubs: overcoming free-riding in international climate policy. *Am Econ Rev* 105: 1339-1370.

Rogelj J, Nabel J, Chen C, Hare W, Markmann K, Meinshausen M, Schaeffer M, Macey K, Höhne N, Copenhagen Accord pledges are paltry (2010) *Nature* 464**:** 1126-1128.

Tavoni A, Dannenberg A, Kallis G, Löschel A (2011) Inequality, communication and the avoidance of disastrous climate change in a public goods game. *Proc Natl Acad Sci USA 108*: 11825-11829.

Thompson A (2006) Management under anarchy: the international politics of climate change. *Climatic Change* 78: 7-29.

UNFCCC Secretariat (2015) Synthesis report on the aggregate effect of the intended nationally determined contributions. 30 October 2015 at http://unfccc.int/resource/docs/2015/cop21/eng/07.pdf.

Table 1. Linear regressions of individual proposals, pledges, beliefs, and contributions

Variables	Proposal	Pledge	Belief	Contribution
Treatment dummies (Baseline: *No-Review*)				
Ex-Ante-Review	12.94**	0.72	0.73	-0.94
	(3.11)	(0.79)	(0.89)	(1.25)
Mid-Point-Review	5.92	0.89	-0.08	0.10
	(4.79)	(0.73)	(0.92)	(1.31)
Ex-Post-Review	13.14**	0.70	1.09	-0.38
	(4.04)	(1.24)	(0.83)	(1.40)
Target		0.21**	0.05	-0.10
		(0.05)	(0.06)	(0.06)
Others average pledge			0.51**	0.04
			(0.18)	(0.25)
Own pledge				0.31**
				(0.11)
Belief				0.77**
				(0.11)
Constant	79.10**	-2.63	1.27	4.68
	(2.76)	(4.58)	(2.89)	(3.92)
Observations	195	195	195	195
R-squared	0.08**	0.25**	0.21**	0.36**

Numbers show coefficients from Ordinary-Least-Squares regression models. Numbers in parentheses are robust standard errors clustered at the group level. Levels of significance: ** P < .01, * P < .05. Definitions of variables: Proposal = individuals' proposals for collective contribution target, Target = groups' collective contribution target, Pledge = individuals' announced contributions, Belief = individuals' expectations of others' contributions.

Table 2. Grades

	Average received grade	Average own grade
Ex-Ante-Review	2.1	1.4
Mid-Point-Review	3.3	2.2
Ex-Post-Review	3.6	2.4

Table 3. Linear regressions of average received grades

Variables	*Ex-Ante-Review*	*Mid-Point-Review*	*Ex-Post-Review*
Proposal	-0.00	0.00	-0.01
	(0.01)	(0.01)	(0.01)
Pledge	-0.33**	-0.03	0.16**
	(0.04)	(0.08)	(0.04)
First-stage contribution		-0.20**	-0.31**
		(0.04)	(0.05)
Second-stage contribution			-0.34**
			(0.06)
Constant	8.90**	5.89**	5.55**
	(1.34)	(0.85)	(0.55)
Observations	50	45	50
R-squared	0.76**	0.63**	0.72**

Numbers show coefficients from Ordinary-Least-Squares regression models. Numbers in parentheses are robust standard errors clustered at the group level. Levels of significance: ** P < .01, * P < .05. Definitions of variables: Proposal = individuals' proposals for collective contribution target, Pledge = individuals' announced contributions, First-stage contribution = individuals' contributions in the first stage of the game, Second-stage contribution = individuals' contributions in the second stage of the game.

Table 4. Comparison between groups with different performance

Group performance	Definition	Number of groups (%)	Target	Sum of pledges	Average belief	Average first-step contribution	Average number of 1^{st}-stage free-riders (max number)
Successful	Q>=75	11 (28%)	93.6	91.4	16.8	12.6	.09 (1)
Intermediate	50<Q<75	22 (56%)	91.4	84.6	14.6	9.2	.41 (2)
Unsuccessful	Q<=50	6 (15%)	85.3	74.5	11.9	5.8	1.66 (3)

Fig. 1. Timeline for the experiment

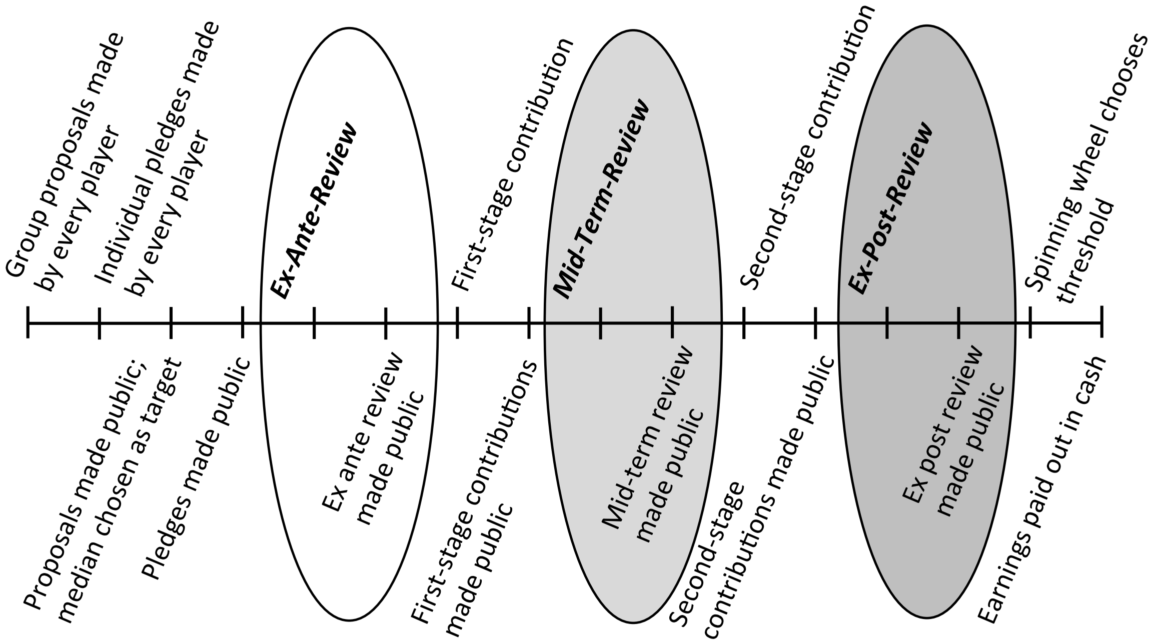

Group proposals made by every player
Individual pledges made by every player
Ex-Ante-Review
First-stage contribution
Mid-Term-Review
Second-stage contribution
Ex-Post-Review
Spinning wheel chooses threshold
Proposals made public; median chosen as target
Pledges made public
Ex ante review made public
First-stage contributions made public
Mid-term review made public
Second-stage contributions made public
Ex post review made public
Earnings paid out in cash

Fig. 2. Group averages for targets, pledges, and contributions by treatment

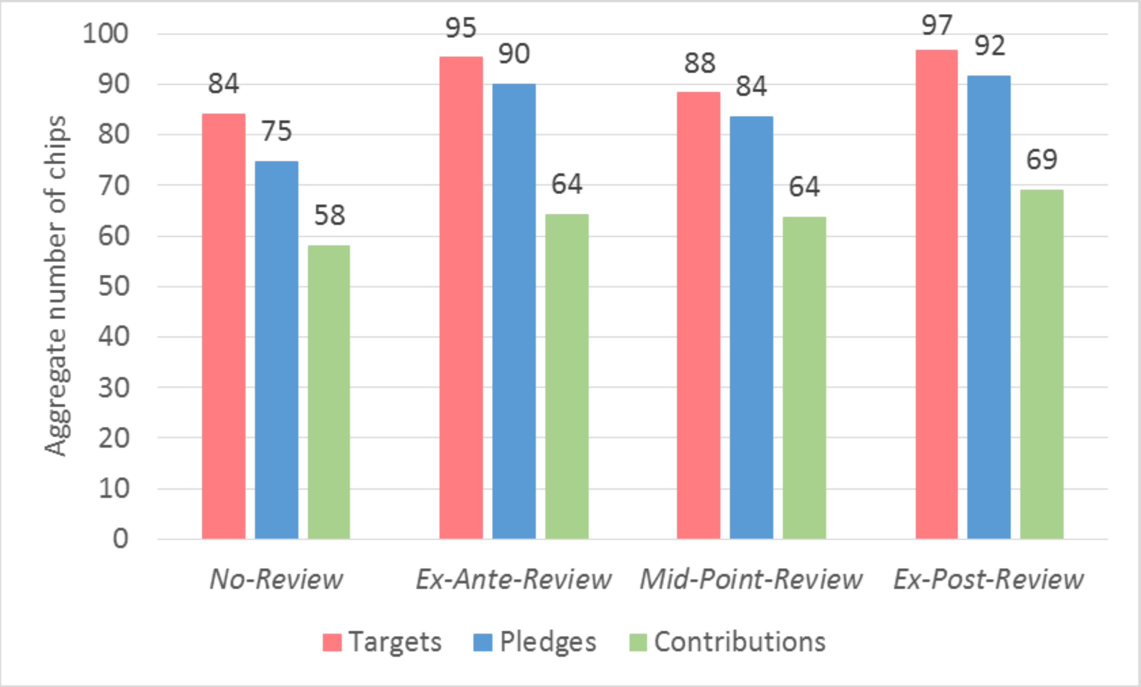
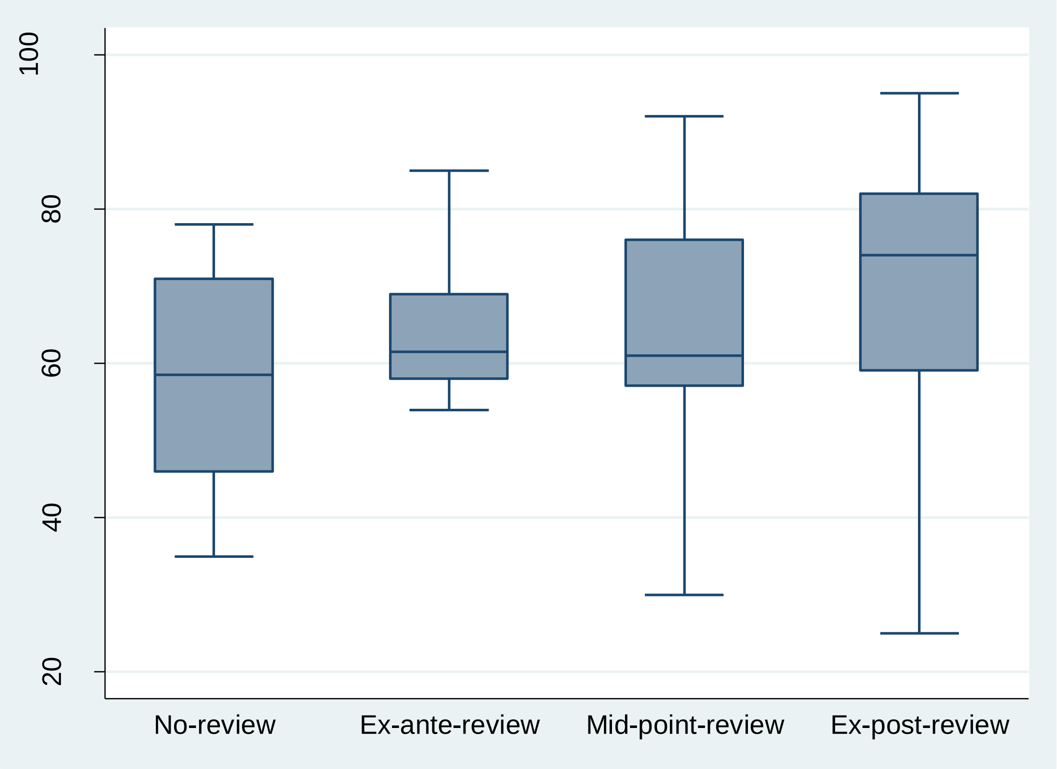


Fig. 3. Distribution of group contributions by treatment



The vertical line shows the range of group contributions from the minimum to the maximum value. The horizontal line in the box represents the median value. Fifty percent of the observations, above and below the median, lie in the box.

Fig. 4. Total of pledges, expectations, and contributions

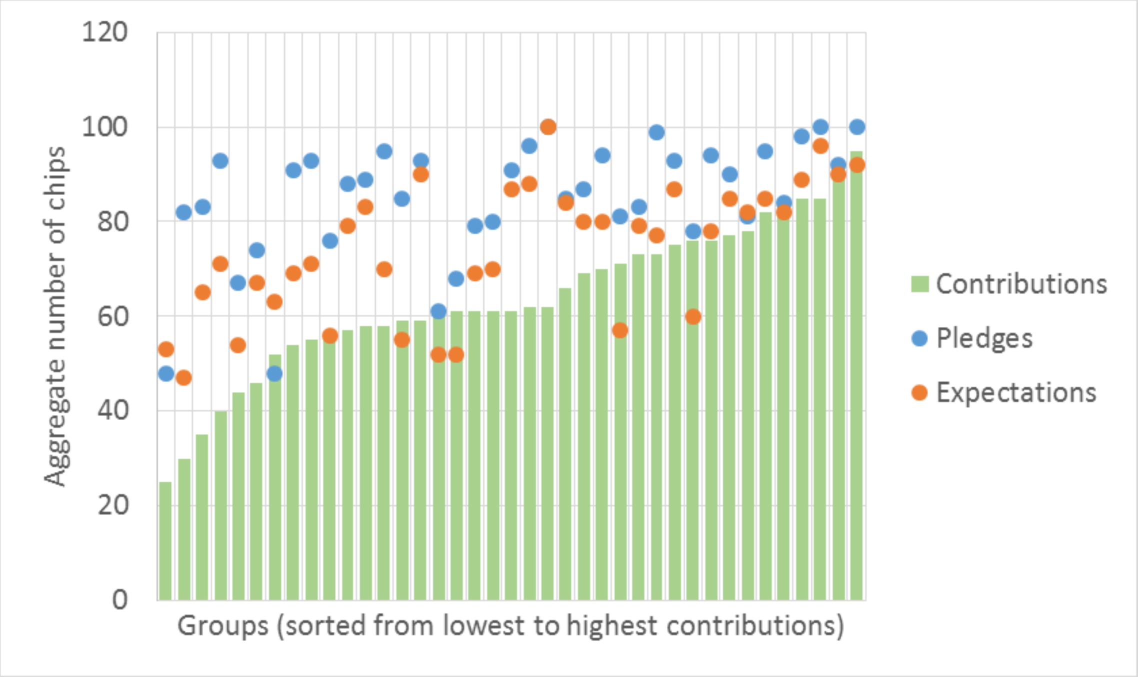
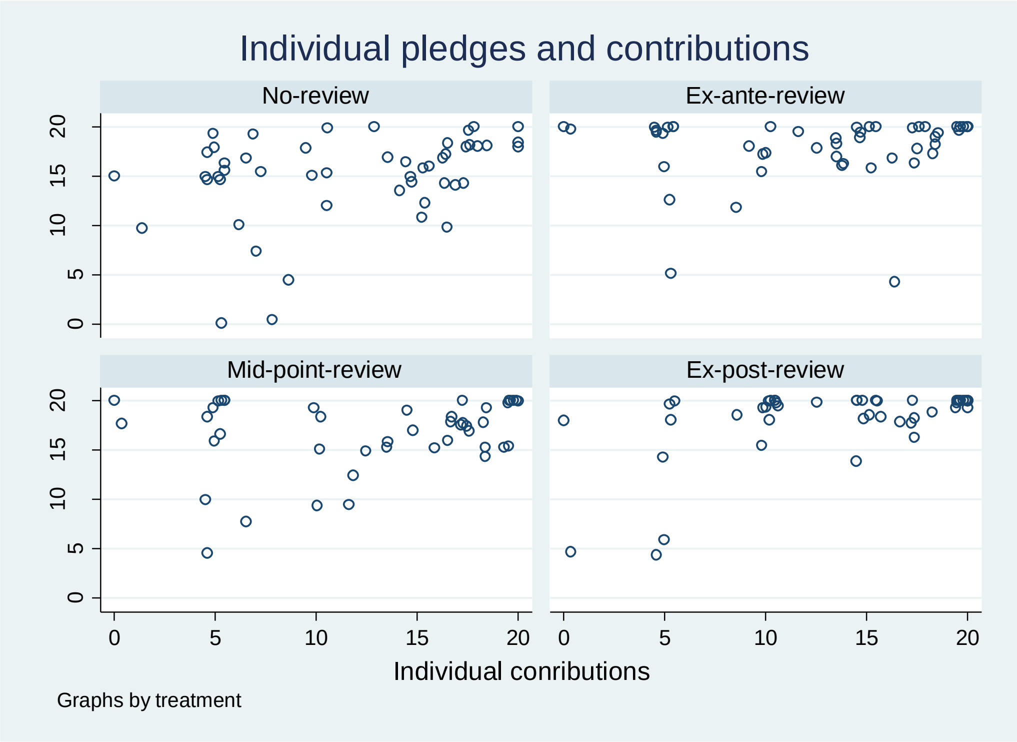
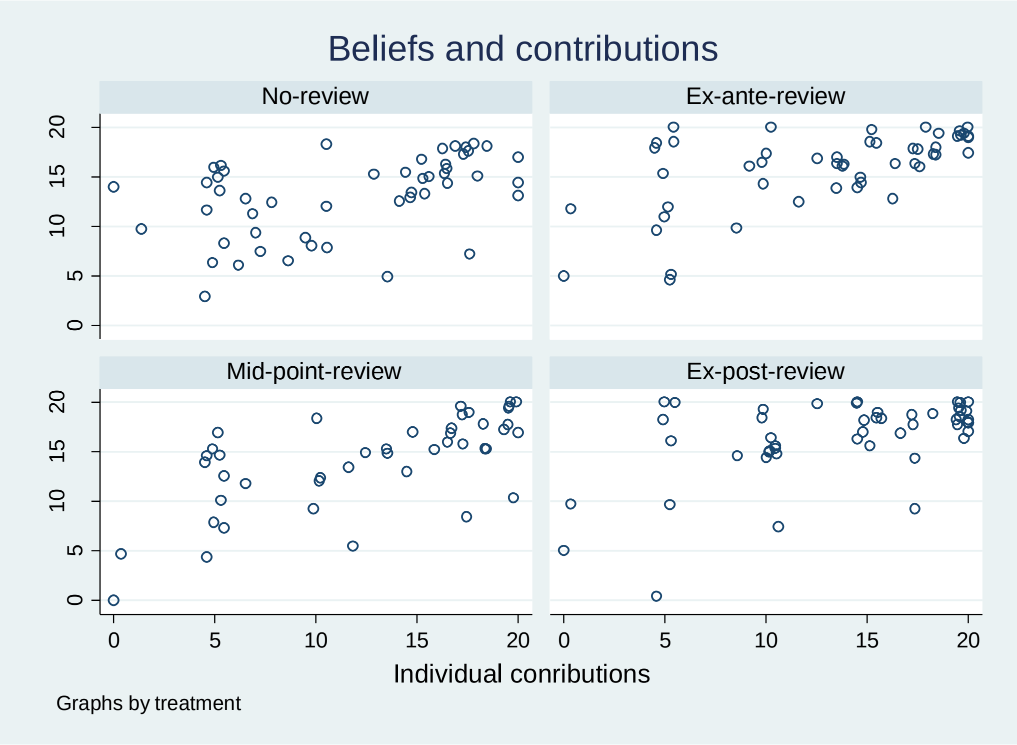


Fig. 5. Individual pledges and contributions



Note: Jitter (3%) has been added to make all data points visible.

Fig. 6. Beliefs and contributions



Note: Jitter (3%) has been added to make all data points visible.