Leveraging the Honor Code: Public Goods Contributions under Oath

Jérôme Hergueux, Nicolas Jacquemet, Stéphane Luchini, Jason Shogren

Leveraging the Honor Code: Public Goods Contributions under Oath*

Jérôme Hergueux[†] Nicolas Jacquemet[‡] Stéphane Luchini[§] Jason F. Shogren[¶]

March 2022

Abstract

Public good games are at the core of many environmental challenges. In such social dilemmas, a large share of people endorse the norm of reciprocity. A growing literature complements this finding with the observation that many players exhibit a self-serving bias in reciprocation: "weak reciprocators" increase their contributions as a function of the effort level of the other players, but less than proportionally. In this paper, we build upon a growing literature on truth-telling to argue that weak reciprocity might be best conceived not as a preference, but rather as a symptom of an internal trade-off at the player level between *(i)* the truthful revelation of their private reciprocal preference, and *(ii)* the economic incentives they face (which foster free-riding). In truth-telling experiments, many players misrepresent private information when this is to their material benefit, but to a significantly lesser extent than what would be expected based on the profit-maximizing strategy. We apply this behavioral insight to strategic situations, and test whether the preference revelation properties of the classic voluntary contribution game can be improved by offering players the possibility to sign a classic truth-telling oath. Our results suggest that the honesty oath helps increase cooperation (by 33% in our experiment). Subjects under oath contribute in a way which is more consistent with *(i)* the contribution they expect from the other players and *(ii)* their normative views about the right contribution level. As a result, the distribution of social types elicited under oath differs from the one observed in the baseline: some free-riders, and many weak reciprocators, now behave as pure reciprocators.

Keywords: Truth-telling oath; Public goods; Social preferences; Reciprocity; Cooperation.

JEL Classification: C72; D83.

*Published in Environmental & Resource Economics Vol. 81 (3). Revised version of PSE WP n°2016-22. We are grateful to Anne l'Hôte for outstanding research assistance and Maxim Frolov and Ivan Ouss for their help in running the laboratory experimental sessions. Financial support from the chair "Economie Publique et Développement Durable" (Aix-Marseille University), from French National Research Agency (through the program Investissements d'Avenir grant ANR-10–LABX-93-0 and the EUR grant ANR-17-EURE-0001), and from the Attractivity grant (University of Strasbourg) are gratefully acknowledged. Shogren thanks the Rasmuson Chair at the University of Alaska-Anchorage for the support while working on this project.

[†]French National Center for Scientific Research (CNRS, BETA lab), Strasbourg, France and ETH Zurich, Center for Law and Economics, Zurich, Switzerland. jerome.hergueux@gess.ethz.ch

[‡]Paris School of Economics and Université Paris 1 Panthéon-Sorbonne. Centre d'Economie de la Sorbonne (CES), Maison des Sciences Economiques, 106-112 boulevard de l'Hôpital 75013 Paris. Nicolas.Jacquemet@univ-paris1.fr

[§]Aix-Marseille Univ. (Aix-Marseille School of Economics) and CNRS, 5-9 Boulevard Maurice Bourdet, 13001 Marseille, France. stephane.luchini@univ-amu.fr

[¶]Department of Economics, University of Wyoming, Laramie, WY 82071-3985, United States. JRamses@uwyo.edu

1 Introduction

From fighting climate change to preserving biodiversity, public good games are at the core of many environmental challenges (Ostrom, 1990; Dietz, Ostrom, and Stern, 2003). Unlike standard models of voluntary public good provision, accumulated evidence from behavioral economics has established that selfish preferences – and thus free-riding – are a minority in the population (Henrich, Boyd, Bowles, Camerer, Fehr, Gintis, and McElreath, 2001). Similarly, altruistic preferences – and thus unconditional cooperation – are even more rarely found in the literature. Instead, most individuals endorse the norm of reciprocity in social dilemma situations (Gouldner, 1960; Dufwenberg and Kirchsteiger, 2004; Sobel, 2005; Falk and Fischbacher, 2006). These reciprocal agents are defined as "conditional cooperators": they are willing to contribute to the public good, but only to the extent that other players do so as well (see Chaudhuri, 2011, for a survey).

In contrast to those polar ideal types, previous experimental research on social preferences has established that the behavior of a substantial fraction of players in public good games can be described as neither pure free-riding, nor reciprocal. Instead, those players behave as "conditional cooperators with a self-serving bias" (Fischbacher, Gächter, and Fehr, 2001), or "weak conditional cooperators" (Fallucchi, Luccasen, and Turocy, 2018): their contribution positively increases as a function of that of the other players, but less than proportionally. Above and beyond selfish preferences, this literature has identified weak reciprocation as an important determinant of the fragility of cooperation in social dilemma situations (Fischbacher and Gächter, 2010). In the presence of weak reciprocators whom they think are not doing their "fair share", reciprocal players progressively withdraw their willingness to contribute, which can lead to the breakdown of cooperation.

The empirical facts are clear, but their interpretation is not. Should we think of weak reciprocity as a distinct type of social preference that should enter theoretical models alongside selfishness, reciprocity and altruism? In this paper, we investigate whether weak reciprocity may be best conceived not as a preference, but rather as the observable consequence of the internal trade-off between two conflicting objectives any player who endorses the norm of reciprocity will face: agents balance the truthful revelation of their reciprocity preference with the goal of maximizing their own private payoff. The reason is straightforward: social preferences are private information, and while some reciprocal players might incur a psychological cost for deviating from their preference, some may be willing to downplay or misrepresent it for their material benefit. Given sufficiently high economic incentives provides enough temptation, a reciprocal player might therefore behave as a weak reciprocator, or even a free-rider.

This interpretation is grounded in a growing literature on preferences for truth-telling ((see Abeler, Nosenzo, and Raymond, 2019, for a survey). This literature has clearly established that people are typically willing to misrepresent private information when this is to their material benefit, but most people incur a psychological cost from lying which mitigates (and sometimes eliminates) misreporting (Abeler, Becker, and Falk, 2014). As a result, many players lie if this

yields material gains, but to a significantly lesser extent than what would be expected based on the profit-maximizing strategy (Mazar, Amir, and Ariely, 2008; Fischbacher and Föllmi-Heusi, 2013; Kajackaite and Gneezy, 2017; Gneezy, Kajackaite, and Sobel, 2018).

The existing literature on truth-telling studies people's willingness to misreport private information when faced with material incentives to do so (e.g., misreporting the result of the roll of a die which determines individual earnings). By contrast, the core novelty of this paper is to apply the main behavioral insight from this literature on honesty and truth-telling to a strategic environment like public good provision. In non strategic experiments, many subjects decide to report private information in a way that lies in between their true private signal and the payoff maximizing strategy. We hypothesize that the same applies to the decision of whether to privately provide a public good: most players endorse reciprocity as a social norm, but many exhibit a self-serving bias in actual behavior.

In this paper, we therefore provide the first empirical test of whether the preference revelation properties of the classic voluntary contribution game can be improved by offering players the possibility to sign a solemn truth-telling oath. Our classic oath mechanism aims at strengthening players' intrinsic commitment to honesty in the absence of punishment, reputation effects, and social pressure. In the experiment, taking the oath is a voluntary choice made in private, which merely asks players to commit to "*tell the truth and always provide honest answers*". We use the oath as a commitment device to test whether a commitment to truth-telling can create the intrinsic motivation necessary for players to behave according to their underlying social preferences, despite economic incentives to the contrary. Our working assumption builds on commitment theory in social psychology (see, e.g., Kiesler, 1971; Joule, Girandola, and Bernard, 2007, for a detailed discussion of the foundations of commitment theory device), which posits that subjects under oath become relatively more committed to behaving according to their private preference. As a result, they become less likely to consider trading off the truthful revelation of this preference against their material benefits. Two important consequences follow: *(i)* a decrease in the cognitive resources necessary to solve the contribution decision task, and *(ii)* an increase in average contributions to the public good.

Our lab experiment combines a traditional voluntary contribution mechanism with an elicitation of first-order beliefs, normative views about how people should behave in the game, and subject-level measures of preference for conditional cooperation. When subjects are under oath, unconditional contributions to the public good increase by 33% on average. Importantly, the unconditional contribution choices of subjects who are under oath correlate significantly more with *(i)* the contribution they expect from the other players, and *(ii)* their normative views on the appropriate contribution level. Accordingly, our truth-telling treatment significantly affects the elicited distribution of social types in the game: the share of (perfect) reciprocators increases by 57% under oath, while the share of weak reciprocators and free-riders decrease by 14% and 38%, respectively. Finally, we provide evidence from decision times that players invest significantly less

cognitive resources in the game when under oath, which provides direct evidence that the treatment improves cooperation by inducing players to ponder less on the material consequences of their contribution choices. We conclude that players under oath are more cooperative on average because the honesty oath mechanism makes them less likely to misrepresent their "true" social type, even when this comes at a financial cost to themselves given the strategic environment.

2 Design of the experiment

Among potential alternative designs, we follow a large strand of the literature in behavioral environmental economics and model the protection of nature as a context-neutral public goods game (see, e.g. Barrett and Danneberg, 2012; Oliver, Pike, Huang, and Shogren, 2014; Feige, Ehrart, and Krämer, 2018). We therefore testbed the effect of a truth-telling oath on the preference revelation properties of the standard public goods game. Obviously, this begs the question of whether behavior observed in the lab will generalize to field applications. While we do not answer this question directly, the existing evidence tends to support the external (or "ecological") validity of the standard public goods game in the environmental context (see, e.g. Rustagi, Engel, and Kosfeld, 2010; Fehr and Leibbrandt, 2011). Similarly, the external validity of the other-regarding behaviors elicited in the lab has been assessed by an extensive empirical literature (see, e.g., Cialdini, 2001; Gachter, 2007, for examples in different domains). Finally, the literature on commitment in social psychology – which mostly relies on contextualized or field experiments – suggests that the associated treatment effects are not only strong (e.g. Geller, Kalsher, Rudd, and Lehman, 1989, in the context of safety-belt use), but may also be long-lasting (see Peer and Feldman, 2020, who establish that a honesty pledge has long-lasting effects in a controlled experiment where players make repeated ethical decisions).

We study the effect of a truth-telling oath in a between-subjects design. The oath is designed according to the accumulated knowledge in the social psychology of commitment (see Section 2.2 below). The oath-taking procedure is implemented before the start of the experiment while subjects are not yet informed about its content, and is the only difference between the Oath and the Baseline treatments.

2.1 Decision tasks

The experiment followed a 3-step design. The public goods game was implemented in two distinct steps.[1] At the end of the experiment, subjects answered a standard socio-demographic questionnaire.

Step 1: Unconditional public goods game. Subjects participated in a classic public goods game in groups of 4 players, each with a 10 Euro endowment. Each Euro invested in the common

[1] We provide an English translation of the original experimental instructions in French in the Appendix, Section A.

project yielded a return of 0.4 Euro to each group member. Each subject made an unconditional contribution (either 0, 1, 2, ..., 10 Euros) to the common project, and individual earnings were determined according to the following payoff function:

$$\pi_i = 10 - contrib_i + 0.4 \sum_{j=1}^{4} contrib_j$$

Immediately after the decision screen, subjects were asked to report *(i)* their normative opinion on how much people should contribute to the public good, *(ii.a)* whether they had an idea about how much the other group members actually contributed, and if so *(ii.b)* their belief about how much the other group members contributed on average.

Step 2: Conditional public goods game. We next elicited *conditional* contributions to the common project to reveal subjects' underlying social type: free-rider, reciprocator, or altruist (Fischbacher, Gächter, and Fehr, 2001). For the conditional contributions decision, each subject stated their intended contribution for each possible value (0, 1, 2, ..., 10 Euros) of the average contribution of the three other group members. At the end of the experiment, the computer randomly drew two subjects from the group to be bound by their unconditional contribution decision from Step 1. For the other two group members, the conditional contribution from Step 2 was implemented for the calculation of individual payoffs. This procedure ensures that both decision steps were incentive compatible and equally important for the calculation of subjects' private earnings.

Step 3: Social values survey. At the end of the experiment, subjects answered a questionnaire asking (i) standard demographic questions, and (ii) social preferences questions taken from the *World Value Survey* (WVS), the *General Social Survey* (GSS) and the *German Socio-Economic Panel* (GSEP) — the three sources commonly used in the empirical literature. All questions were mandatory and none was remunerated:

(i) To what extent do you consider it justifiable to free-ride on public social allowances (cooperation variable; 10 points scale, WVS question);

(ii) Do you think people are mostly looking out for themselves as opposed to trying to help each other (altruism variable; 10 points scale, WVS question);

(iii) Do you think people would try to take advantage of you if they got a chance as opposed to trying to be fair (fairness variable; 10 points scale, WVS question);

(iv) Do you think most people can be trusted or that one needs to be very careful when dealing with people (trust variable; binary answer, WVS and GSS question);

(v) How much do you trust people in general (general trust variable; 4 points scale, GSEP question);

(vi) How much do you trust people you just met (trust in strangers variable; 4 points scale, GSEP question);

(vii) Do you generally see yourself as fully prepared to take risks as opposed to generally trying to avoid taking risks (risk aversion variable; 10 points scale question taken from Dohmen, Falk, Huffman, Sunde, Schupp, and Wagner, 2011).

2.2 Treatment variable: Design of the truth-telling oath procedure

The design of our truth-telling oath procedure follows Jacquemet, Joule, Luchini, and Shogren (2013), who provide a review of the commitment literature in social psychology (see also Jacquemet, James, Luchini, and Shogren, 2011). This literature shows how past decisions increase the likelihood that a person will behave in a specific way in the future if these past-and-future decisions are motivationally aligned. If the decisions are aligned, then voluntary compliance with the first decision can induce commitment towards the second decision, in which commitment is defined as a free will "binding of the individual to behavioral acts" (Kiesler and Sakumura, 1966). The mechanism is best illustrated in the context of foot-in-the door techniques, in which a small request (called a 'preparatory action') is made prior to a more substantial one (the target behavior). In line with the prediction from commitment theory, accumulated evidence in social psychology shows that compliance with the first request typically leads to significantly higher compliance with the second one. For instance, people behave more altruistically if they had to help someone in the past, and recycle more if they committed to based on a signed pledge (see Burger, 1999, for a meta-analyis). One interpretation of the strong behavioral effect of commitment is self-attribution (Bem, 1972): because they complied with the first request, decision-makers infer that they are generally willing to comply with this kind of requests and thus behave accordingly more frequently in the future. An alternative interpretation is cognitive dissonance (Festinger, 1957), which is avoided by complying with the second request conditional on compliance with the first one — whatever the reason behind compliance with the first request.

Common to both interpretations is the key feature that commitment is only produced by decisions that are freely made by decision-makers — as decisions made for external reasons do not trigger the psychological process leading to commitment (Joule, Girandola, and Bernard, 2007). A large strand of literature also shows that commitment is stronger if it is publicly expressed and signed (see, e.g., Pallack, Cook, and Sullivan, 1980; Katzev and Wang, 1994). In line with these insights from the social psychology of commitment, we expect a truth-telling oath that is freely

signed by subjects to commit them to truth-telling in future decisions.[2] We only depart from the commitment literature by relying on a private commitment decision. The main reason for this choice is to make sure that compliance with the oath is not due to external social pressure from the monitor or other subjects. Similarly, subjects never learn the signing decision of others at any stage of the experiment.

The oath is implemented as a between-subjects treatment variable at the session level, and follows a strict procedure: subjects were invited one by one to enter a separate room before entering the laboratory (where they were informed about the instructions of the experiment). In this room, a monitor privately offered subjects a form entitled "solemn oath".[3] The word "oath" is written on the form but never said aloud. To ensure the credibility of the form, we placed the Paris School of Economics logo on top, together the topic designation and the research number. Subjects were instructed to read the "form" carefully and decide whether they would like to sign it or not. They were explicitly told that signing the oath was not mandatory, and that their participation and earnings would not depend on their decision.

Regardless of whether subjects signed the oath, the monitor thanked them and invited them to enter the lab. At this stage, subjects randomly draw the name of the computer in front of which they would seat during the session, and which will be associated with their decisions. Subjects' decision were therefore recorded together with their decision times (this timer was not visible to subjects), but this data could not be linked back to individuals. We scripted what the monitor said when offering the oath to standardize the phrasing of the procedure. To avoid subject communication prior to the experiment, one monitor stayed with the other participants until all subjects had been presented with the oath. Subjects who were waiting for their turn could neither see nor hear what was happening at the oath-desk. Most subjects agreed to sign the oath, despite the lack of pressure to do so. The proportion of subjects who agreed to do so was also highly stable across sessions, with 0 to 2 subjects refusing to sign per session, resulting in a 95% agreement rate overall. We therefore pool all subjects in the Oath treatment together in our analysis and present intention-to-treat estimates of the impact of the oath on cooperative behavior.

[2]This exact same procedure has been shown to foster truth-telling in economic lying games like the sender-receiver game (Jacquemet, Luchini, Rosaz, and Shogren, 2018), or a coin flipping task both in the lab (Beck, Bühren, Frank, and Khachatryan, 2020) and in the field (Jacquemet, James, Luchini, Murphy, and Shogren, 2021). This increased likelihood of truthfully reporting information has implications in many different areas in economics. Several studies confirm that this change in behavior extends to preference revelation for non-market goods both in the lab (Jacquemet, James, Luchini, and Shogren, 2017) and in the field (Carlsson, Kataria, Krupnick, Lampi, Lofgren, Qin, Sterner, and Chung, 2013), to self-reported income in a tax evasion experiment in the lab (Jacquemet, Luchini, Malézieux, and Shogren, 2020) and in the field (Koessler, Torgler, Feld, and Frey, 2019), to self-reported happiness (Carlsson and Kataria, 2018) and to communication in coordination games (Jacquemet, Luchini, Shogren, and Zylbersztejn, 2017).

[3]An English translation of the original form in French is provided in the Appendix, Section B. Section C reports the detailed implementation of the oath procedure.

2.3 Experimental procedures

Given our goal to elicit social preferences in isolation from learning effects and strategic concerns, each step was only played once.[4] Subjects were only informed about the outcomes associated with their decisions and that of the other players at the very end of the experiment. The experiment was implemented under strict anonymity: at the beginning of the experiment, subjects randomly picked a computer name which determined their assignment to computer terminals. We recorded subjects' decisions together with their decision times on each screen (this timer was not visible to subjects), but our design prevented us from being able to link this data back to individuals. Overall, we conducted 3 and 6 BASELINE sessions with 180 subjects in November 2010 and November 2011, and 6 OATH sessions with 120 subjects in June 2013.

Once all subjects got randomly assigned to a computer, the monitor read the experimental instructions aloud, and subjects were then left to use all devices at their disposal to check their own understanding (access to the text, examples and earnings calculator). On the decision interface, the first screen provided subjects with general information about the experiment. Next the screen described the game, followed by examples and the associated payoffs for each player. The following screen provided an earnings calculator, which is an interactive page that allowed subjects to explore hypothetical scenarios of interest before making their contribution decisions in the public goods game. On all screens, including decision-making ones, a "review description button" provided subjects with a direct access to the instructions displayed at the beginning of the game. Subjects were only informed of their earnings at the very end of the experiment, which were paid privately in cash together with a 5 Euro show-up fee.

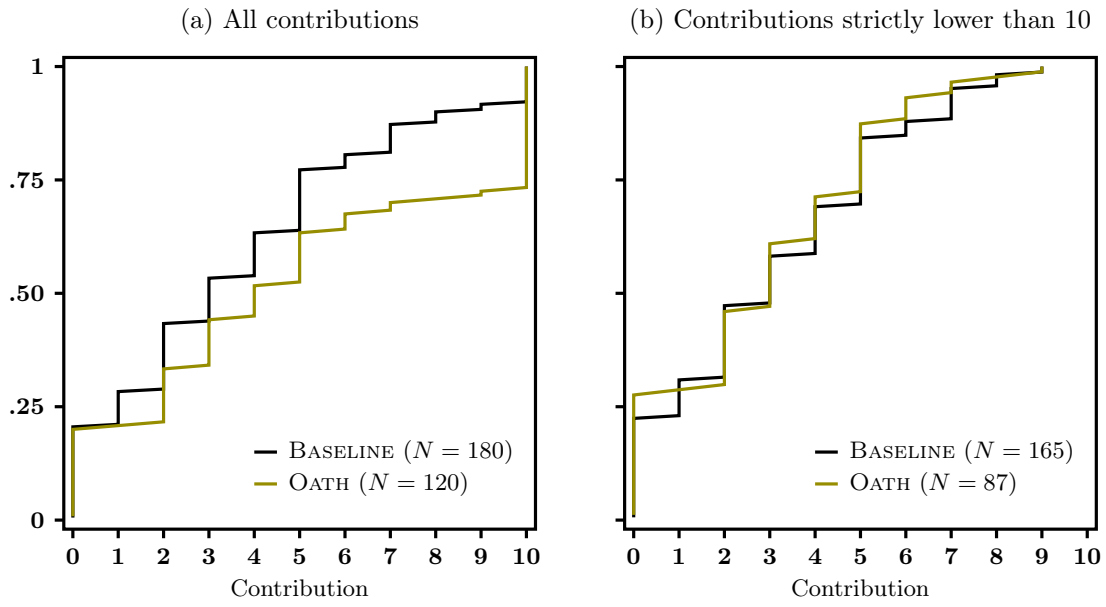
3 Main result: Public goods contributions under oath

Comparing the level of contributions to the public good between the two treatments, we find a strong effect of signing the oath: on average, our treatment induces a 33.1% increase in unconditional contributions to the public good. The mean unconditional contribution is 4.85 in OATH, up from 3.65 in BASELINE ($p = .011$, Wilcoxon-Mann-Whitney (WMW) test).

Figure 1.a presents the empirical distribution function (EDF) of contributions by treatment. Although subjects use the entire range of possible contributions in both treatments, the EDF of contributions under oath first order dominates the EDF of contributions in the BASELINE

[4]The experiment is computerized based on an own-developed decision-making interface described in Hergueux and Jacquemet (2015), from whom we borrow the public goods lab decisions for the BASELINE treatment. The time elapsed between the experimental sessions for both treatments does not result in significant differences in the balance of individual characteristics (see Appendix E). Comparing behavior within the BASELINE between the Nov. 2010 and Nov. 2011 sessions, we also fail to find any significant difference in unconditional contributions (the results are available from the authors upon request). All sessions took place at the experimental economics laboratory of the University Paris 1 Pantheon-Sorbonne, and subjects were recruited via an on-line registration system based on ORSEE (Greiner, 2015).

Figure 1: Contributions to the public good by treatment (empirical distribution functions)



($p = .001$).[5] The main effect of the oath is to induce significantly more subjects to choose the top contribution level: 27.5% contribute 10 in Oath relative to 8.3% in the Baseline; a 230.0% increase, which is significant at the 1% level ($p < .001$, proportion test). The remaining of the distribution is about the same as in the Baseline treatment. In particular, we observe no change in zero contributions: 20.0% in Oath, 20.5% in Baseline. Figure 1.b provides a closer look at this part of the distribution, based on the EDF of contributions by treatment for the sub-sample of contributions strictly lower than 10. The two treatments are similar — no first-order dominance of contributions in Oath ($p = .734$, bootstrap KS test). The mean contribution among those that are lower than 10 is also similar between treatments: 3.1 in Oath and 2.9 in Baseline ($p = .678$, WMW test). As a result, the oath seems to work largely by turning unconditional contributions that are lower than 10 into maximal contributions.

4 Why do unconditional contributions increase under oath?

The increase in the level of contributions under oath mainly comes from an increase in the proportion of subjects who contribute the totality of their endowment to the public good. In this section, we turn to the primitives that drive this behavioral change. We start by analyzing the effect of the oath on subjects' beliefs about the contributions of others, their normative opinion on

[5]This result comes from a bootstrap version of the univariate Kolmogorov-Smirnov test (bootstrap KS test hereafter). This modified test provides correct coverage even when the distributions being compared are not entirely continuous and, unlike the traditional Kolmogorov-Smirnov test, allows for ties (see Abadie, 2002; Sekhon, 2011).

the right contribution level, and the relationship between their contribution decisions and those beliefs and normative views. Next, we leverage the data from the conditional public goods game to study whether and how the oath impacted the distribution of the elicited social types in the game. Finally, we pinpoint the behavioral mechanism we posit behind our results – a simplification of the decision making problem for subjects under oath – through a detailed analysis of response times, a proxy for subjects' investment of cognitive resources in the decision task.

4.1 Relationship of contributions to first-order beliefs and normative opinions

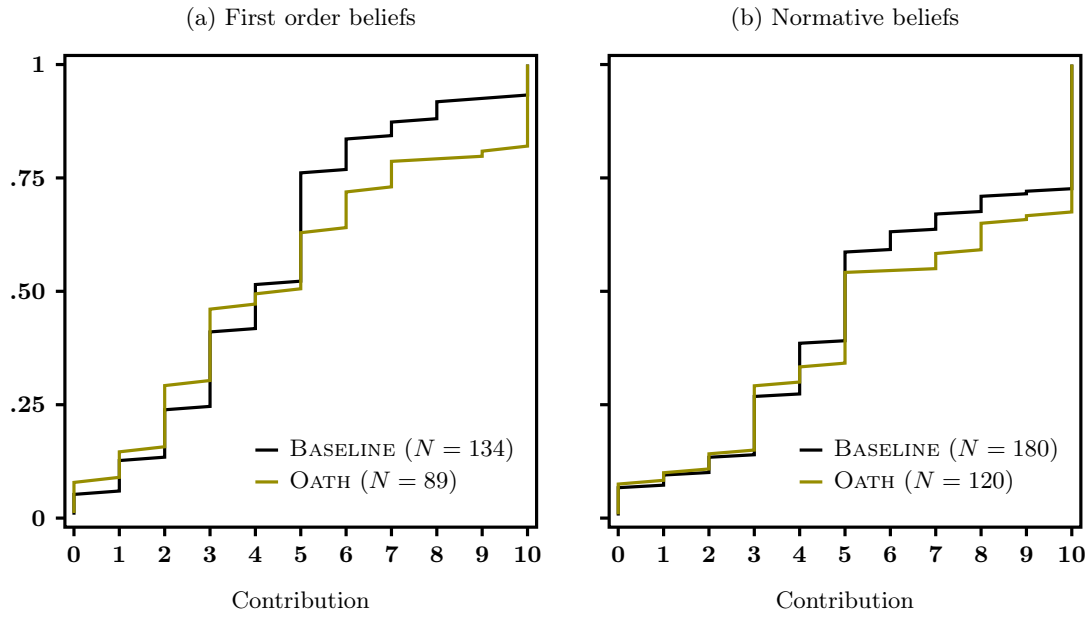
Because most individuals endorse the norm of reciprocity in situations of social dilemma, contributions in the unconditional public goods game typically depend on their beliefs about the behavior of the other players (first order beliefs). Similarly, in such situations, individuals typically try to behave in a way which they think is socially appropriate, even when this comes at a cost to themselves. Herein, we analyse how subjects' first order beliefs and normative opinions about the right contribution level drive the observed change in contributions under oath. First-order beliefs regarding the mean contribution of others and normative views with respect to the right contribution level are elicited right after the unconditional contribution decision, before subjects are asked for their conditional contribution decisions (see Step 1 in Section 2.1).[6]

Figure 2.a reports the cumulative distribution function of first order beliefs in each treatment, while Figure 2.b reports the cumulative distribution function of normative views. Overall, beliefs are not significantly different between the two treatments. The estimated average contribution of the other players is 4.34 in Baseline and 4.80 in Oath. This increase is small in magnitude and not statistically significant ($p = 0.522$, WMW test). That said, this average result conceals some interesting changes in the distribution of beliefs. Those are very similar at the bottom-end, but we observe a clear shift at the upper-end, whereby a larger share of subjects expect the other players to contribute their full endowment under oath (rather than choosing mid-contribution levels). This shift is statistically significant ($p = .012$, Fisher test). As a result, the distribution of beliefs in Oath first order dominates that of Baseline ($p = 0.063$, bootstrap KS test). This suggests that even though the oath was offered privately and anonymously to subjects, our oath mechanism had a small effect on the distribution of first order beliefs, which may have influenced their unconditional contribution decisions.

Turning our attention to normative views, we again observe no average treatment effect on normative opinions about the right contribution level, which is equal to 6.08 in Oath as opposed

[6]With respect to first order beliefs, we ask subjects to report (i) whether they thought about the behavior of the other group members while making their own contribution decision and, if so, (ii) how much they believed the others group members actually contributed on average. The data is thus only available for those subjects who answered yes to the first question. We observe no change in the proportion of subjects who declared thinking about the contributions of the other players while making their own decision between treatments, *i.e.*, 74.4% in the Baseline as opposed to 74.2% in Oath (Fischer exact test: $p = 0.840$). Appendix F, provides additional statistics on these sub-groups. With respect to normative views, we ask subjects how much they think people *should* contribute to the common project.

Figure 2: Distribution of first-order beliefs and normative views by treatment



to 5.73 in Baseline ($p = .434$, WMW test). In contrast to first-order beliefs, we further observe no significant difference in the share of subjects who think that the right contribution level is the maximal one, *i.e.*, 27.7% in Baseline as opposed to 33.3% in Oath ($p = 0.304$, WMW test). This slight difference mainly comes from a shift to the right in the share of subjects who declare that contributions should be strictly between 5 and 10.

Overall, we conclude from Figure 2 that the oath *(i)* only had a marginal impact on the distribution of subjects' first order beliefs, and *(ii)* left their normative views about the appropriate contribution level unchanged. We now turn to whether our oath mechanism impacted the strength of the relationship between unconditional contribution decisions, first-order beliefs and normative opinions. To that end, we build two measures. The first measure, which we call the *reciprocity gap*, quantifies the strength of reciprocal behavior at the subject level by taking the difference between unconditional contribution decisions and first order beliefs about the average contribution of the other players. This gap is null for a subject who behaves as a perfect reciprocator (because the unconditional contribution perfectly matches the average contribution expected from the other players), and negative for weak reciprocators and free-riders. Following the same logic, we also construct a *normative gap* by computing the difference between the unconditional contribution decision of each subject and their normative view on the right contribution level.

The marginal distribution of these gaps is summarized in Table 1, where we distinguish subjects from whom the gap is negative (*i.e.*, actual unconditional contributions are lower than the predicted value), zero (the actual and predicted contributions perfectly match), or positive. For both gaps,

Table 1: Distribution of the reciprocity and normative gaps, by treatment

	Reciprocity gap				Normative gap			
	N	Negative	Zero	Positive	N	Negative	Zero	Positive
Baseline	134	44.0	29.9	26.1	180	57.0	35.2	7.8
Oath	89	25.8	47.2	27.0	120	41.7	50.0	8.3
p-value (Fisher test)		.024	.011	1		.010	.012	1

Note. For each treatment, the table reports the marginal distribution of the reciprocity gap (left-hand side) and the normative gap (right-hand side). The last row provides the results from a Fisher exact test of the difference in distribution between treatments (pooling all other groups together).

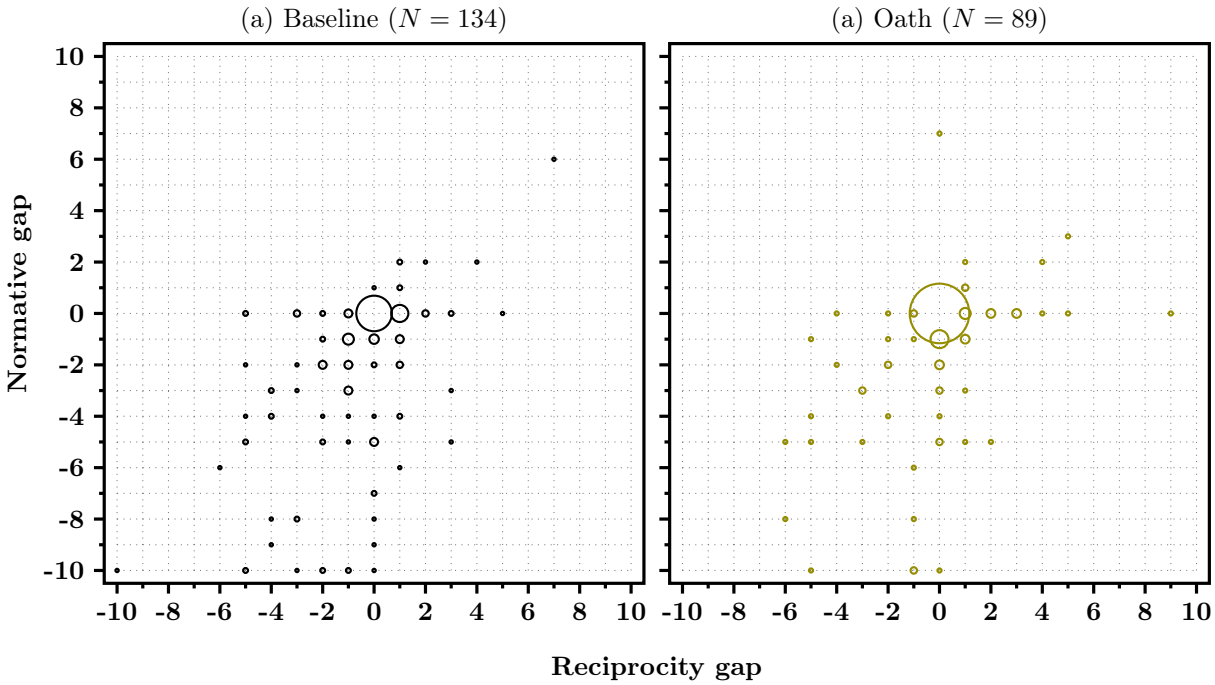
we observe a large share of subjects (about a third) whose unconditional contribution perfectly matches the reciprocal or normative contribution in the Baseline. The oath induces a sharp increase in the share of subjects who behave as perfect reciprocators, as well as the share of subjects who follow their normative views when choosing their unconditional contribution (both changes are statistically significant according to the Fisher exact test presented in the last row of the table). This increase in the share of subjects who exhibit zero reciprocity and normative gaps comes from a decrease in the share of subjects with a negative gap, while the share of subjects who exceed the reciprocal or normative contribution remains fairly stable between treatments. These changes lead to an overall increase in the correlation between subjects' unconditional contributions and both the estimated contribution of the other players (from .70 in Baseline to .81 in Oath, $p = .076$, Fisher's z-transformation test) and their normative view on the right contribution level (from .47 in Baseline to .74 in Oath, $p < 0.001$, Fisher's z-transformation test).

The joint distributions of the gaps at the subject level are presented in Figure 3 (the size of the dots are proportional to the proportion of subjects within treatments, to account for the differences in sample sizes). This figure clearly shows that subjects whose gap is zero on either dimensions are very often the same. The proportion of subjects whose unconditional contribution coincides with both their reciprocal preferences and their normative views drastically increases under oath: from 16.4% of subjects in Baseline to 28.1% under oath ($p = .054$, proportion test). Subjects under oath thus appear less likely to trade-off their reciprocal or normative preference against their private payoff when deciding on their contribution level.

4.2 Distribution of social types under oath

The evidence presented in Section 4.1 is consistent with the idea that the oath reduces the share of players who are willing to misrepresent their private reciprocal preference for material gains. However, the unconditional nature of subjects' contribution decision in Step 1 of the experiment does not allow us to precisely identify the social type of each subject. To do so, one needs to elicit players' contribution decision in all possible states of the world (that is, for each possible average contribution level of the other players). This is what we do in Step 2 of the experiment, where we

Figure 3: Joint distribution of the reciprocity and normative gaps in each treatment



Note. In each treatment, the figure reports the normative gap and the reciprocity gap computed for each subject. The size of the dots depends on the proportion of subjects for a given combination.

ask subjects to report *conditional* contribution decisions to the public good (Fischbacher, Gächter, and Fehr, 2001). This alternative decision making procedure has the advantage of removing strategic uncertainty[7], and providing the full conditional contribution schedule of each subject (which is necessary to identify their underlying social type).

Table 2 provides an overview of the average conditional contribution schedule of our subjects in each treatment. The first line in each panel reports the average conditional contribution for each level of the contribution of the other players. For any positive contribution of the other players, subjects always contribute more under oath. The last three rows in each panel disaggregate these average conditional contributions according to several behavioral patterns: (i) the share of subjects who contribute their whole endowment for each level of the contribution of others (i.e., contribute 10), (ii) the share of subjects who behave as perfect reciprocators (i.e., perfectly match the contribution of others), and (iii) the share of subjects who do not contribute at all (i.e., free-ride). The second row shows that the share of subjects who contribute 10 to the public good is typically small (lower than 10%) and does not differ either between treatments or according to the average contribution of the other group members, except when this contribution is itself 10. In this case, 35% of subjects contribute 10 in the Baseline, as opposed to 55% in Oath. Similarly, the share of subjects who contribute 0 is pretty stable across both dimensions, except when the

[7]It does not, however, remove strategic behavior: there is still room for players to increase their private material gains at the expense of the other players.

Table 2: Conditional contributions by treatment

Others' average contribution		0	1	2	3	4	5	6	7	8	9	10
Baseline	Average contribution	0.94	1.46	1.91	2.54	3.00	3.61	3.99	4.29	4.66	5.09	5.36
	Contribute 10 (%)	5.0	1.7	1.1	0.6	1.1	0.6	2.2	1.7	3.9	10.0	35.0
	Reciprocate (%)	77.8	37.2	34.4	30.0	28.9	37.8	25.0	20.6	20.6	20.0	35.0
	Contribute 0 (%)	77.8	36.7	27.8	19.4	18.9	15.0	15.0	17.8	19.4	19.4	27.8
Oath	Average contribution	0.86	1.56	2.13	2.62	3.14	3.78	4.39	4.66	5.34	5.69	6.57
	Contribute 10 (%)	4.2	3.3	3.3	3.3	3.3	4.2	3.3	5.8	8.3	9.2	55.0
	Reciprocate (%)	83.3	42.5	39.2	39.2	35.0	45.0	33.3	33.3	34.2	34.2	55.0
	Contribute 0 (%)	83.3	35.0	25.8	20.8	20.8	16.7	13.3	14.2	15.0	15.0	20.8

Note. For each treatment, the first row provides the average contribution observed against the average level of contribution of other group members (in column). The second row provides the share of subjects who contribute their full endowment, the third reports the share of subjects whose conditional contribution perfectly matches the contribution of others, and the last row reports the share of subjects who contribute 0.

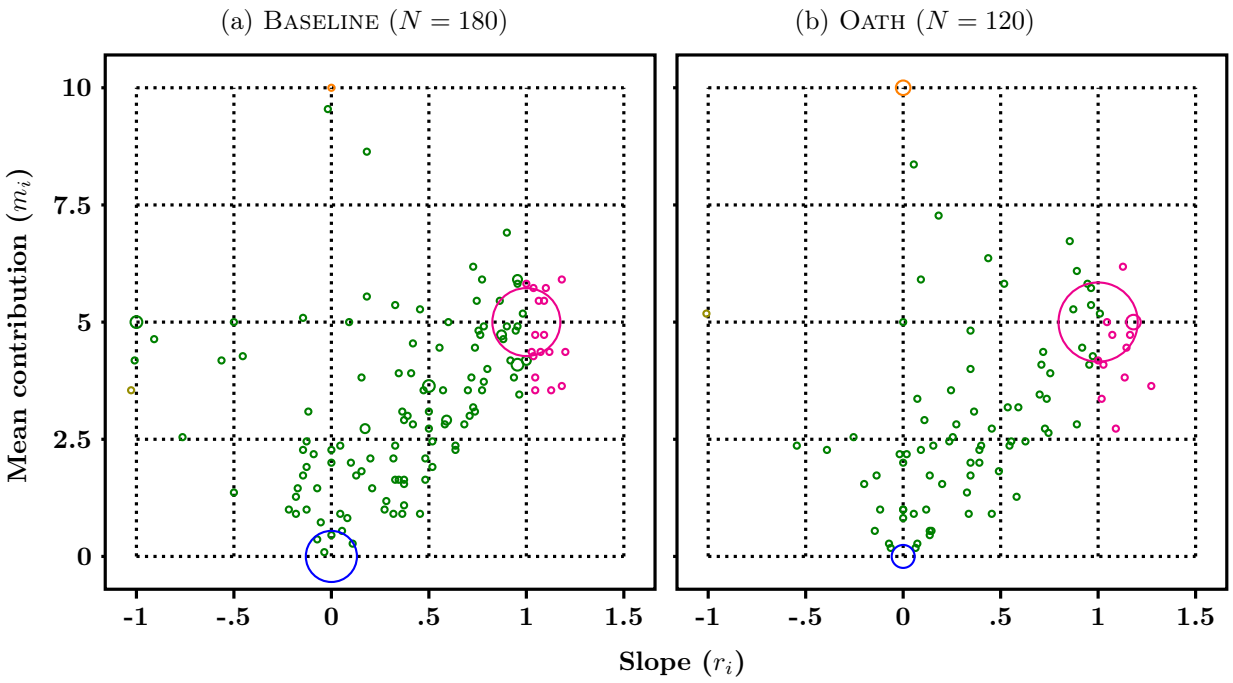
contribution of the other players is itself 0 (with a shift from 77.8% in the Baseline to 83.3% under oath). As shown in the third row of the table, the main driving force of these two changes is a sharp increase in the fraction of subjects who behave as perfect conditional cooperators under oath.

To study conditional contribution schedules at the individual level, we reduce subjects' reaction functions to two parameters, following Fischbacher and Gächter (2010): *(i)* the slope of the reaction functions to the average contribution of the other group members ("reciprocity", denoted r_i), and *(ii)* the average proportion of the endowment that is conditionally contributed across all 11 conditional contributions decisions ("mean contribution", denoted m_i).[8]

The joint distributions of these two parameters over individual subjects in each treatment are reported in Figure 4 — in which the size of the dots is proportional to the number of subjects. There is a large heterogeneity in conditional cooperation in both treatments, but the distribution in Oath is shifted to the right compared to the one in Baseline. This suggests that reciprocation plays a stronger role in contribution behavior under oath. To better characterize this change in reciprocity, we classify subjects into four exclusive social types. We only depart from Fischbacher, Gächter, and Fehr (2001) in that we fully automate the classification rule. By contrast, they identify "weak reciprocators" (called 'hump shaped' contributors in their setting) from a visual examination of their conditional contribution patterns. Our classification is consistent with that of Fallucchi, Luccasen, and Turocy (2018), who use hierarchical cluster analysis to separate subjects into exclusive behavioral types based on a meta-analysis of 6 seminal public goods experiments on conditional cooperation:

[8]The parameter r_i is estimated at the subject level from a linear regression in which the dependent variable is the reported level of conditional contribution and the independent variable is the value $\{0, 1, 2, \dots 10\}$ of the possible average contribution of the three other members. The estimated reciprocity parameter r_i is the slope of the regression line for each subject. The parameter m_i is calculated directly as the mean of the individual conditional contributions.

Figure 4: Conditional cooperation pattern, by treatment



Note. In each treatment, the figure reports the conditional cooperation parameters computed for each subject. The size of the dots is proportional to the number of subjects. To facilitate the reading, observations are classified by different colors that correspond to the types put forward in Table 3: Free-riders are indicated in blue, altruists in orange, reciprocators in pink and weak reciprocators in green.

1. Free-riders do not contribute to the public good, irrespective of the average contribution level of the other players: $m_i = 0$;
2. Altruists contribute their entire endowment, irrespective of the average contribution level of the other players: $m_i = 10$;
3. Reciprocators match the contribution level of the other players: $\{r_i \geq 1\}$;
4. Weak reciprocators under-match the contribution level of the other players: $\{r_i < 1\}$.[9]

Table 3 reports the resulting distribution of social types by treatment, along with the results from Fisher tests of the difference in the proportions of each social type between treatments. The analysis of our subjects' conditional contribution schedules largely confirms the evidence presented in Section 4.1: the honesty oath procedure induces a 57% increase in the proportion of subjects who behave as perfect reciprocators in the sample (from 21.7% to 34.1%, $p = .023$). As expected,

[9]This classification rule does not allow for any decision error, as it requires free-riders to never contribute and altruists to always contribute their entire endowment. We replicate the analysis with different classification rules in Appendix D, and show that the results are similar. Further, out of 300 subjects in this experiment, 39 actually have a negative estimated r_i. Excluding them from the analysis does not change the nature of our conclusions.

Table 3: Distribution of social types in the conditional public goods game

Social Type				Baseline ($N = 180$)	Oath ($N = 120$)	Fisher test
Free rider	$\{r_i = 0\}$	&	$\{m_i = 0\}$	9.4	5.8	$p = .286$
Altruist	$\{r_i = 0\}$	&	$\{m_i = 10\}$	0.6	3.3	$p = .085$
Reciprocator	$\{r_i \geq 1\}$	&	$\{0 < m_i < 10\}$	21.7	34.1	$p = .023$
Weak reciprocator	$\{r_i < 1\}$	&	$\{0 < m_i < 10\}$	65.0	55.8	$p = .117$

Note. The table reports the definition of social types based on the average conditional contribution, m, and the slope of the contribution schedule, r; along with the distribution of social types observed in the conditional public goods game. The last column reports Fisher exact tests of the difference in the frequency of each sub-type by treatment (with all other types pooled).

this change in proportions is counterbalanced by a decrease in the share of weak reciprocators (from 65% to 56%) and free-riders (from 9.4% to 5.8%). Taken together, the decrease in the share of both types under oath is statistically significant ($p = .022$, Fisher test). We also observe a marginally significant increase in the share of altruists (from 0.6% to 3.3%), but our sample size prevents us from drawing too strong a conclusion in this respect. Overall, the shift in the distribution of social types between treatments is statistically significant at $p = .015$ (Fisher test).

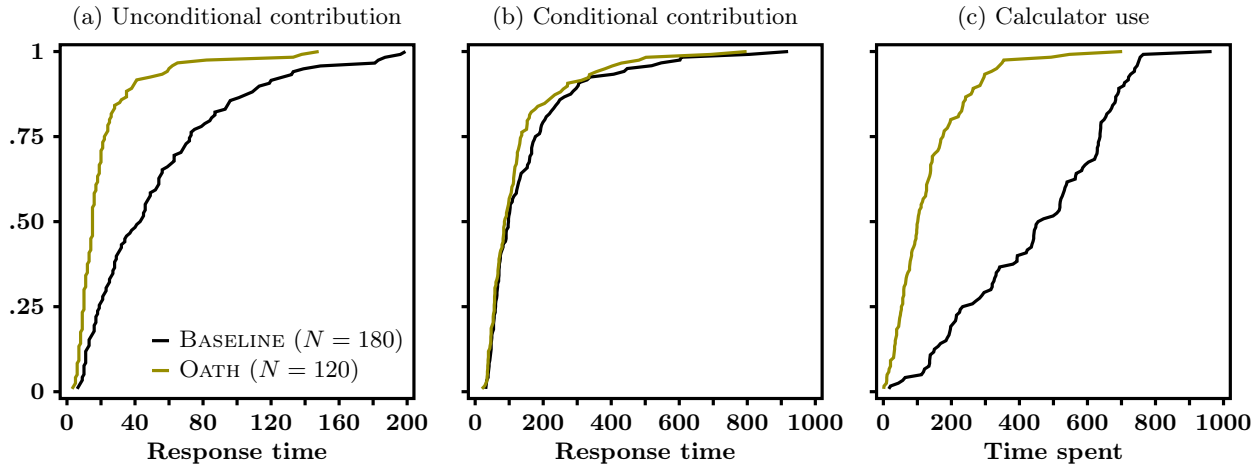
4.3 Behavioral mechanism: Evidence from response times

The honesty oath increases public goods contributions by increasing the share of players who decide to match the effort level they expect from others and/or their normative view on the right contribution level. Accordingly, the oath treatment increases the share of players who behave as perfect reciprocators, at the expense of weak reciprocators and, to a smaller extent, free riders. This evidence is consistent with our working assumption: while most players endorse the norm of reciprocity, many are willing to trade-off the truthful revelation of this private preference against their material gains, which gives rise to "*conditional cooperat[ion] with a self-serving bias*" (Fischbacher, Gächter, and Fehr, 2001, p.401). According to this interpretation, the honesty oath works by simplifying subjects' decision making problem: by putting more weight on the truthful revelation of their private social preference, the oath mechanism makes players less likely to ponder on the material consequences of their contribution choice. Players under oath should thus invest less cognitive resources in their contribution decisions.

We provide suggestive evidence supporting this interpretation by using response times — *i.e.*, the time elapsed between the appearance of the decision screen and subjects' actual choice — as a proxy for the investment of cognitive resources in the contribution decision. The existing literature has shown that response times are typically longer the harder is the decision, notably due to conflicting motivations (Krajbich, Bartling, Hare, and Fehr, 2015).[10] The distribution

[10]Note that the way we use decision times for this test is distinct from that of a separate literature on public goods concerned about whether cooperative strategies are, in general, more "intuitive" than free-riding (Rubinstein, 2007; Rand, Greene, and Nowak, 2012; Lotito, Migheli, and Ortona, 2013). By contrast to this literature, our response

Figure 5: Empirical distribution functions of decision times in the public goods game (unconditional and conditional) and time spent using the earnings calculator



of decision times in the unconditional public goods game is provided in Figure 5a. Consistent with our hypothesis, decision times drop significantly in Oath relative to Baseline: the median response time is 44.5 seconds in Baseline, as opposed to 15 seconds in Oath (a 66.3% reduction in decision time). As a result, the EDF of decision times in Oath first order dominates the EDF observed in Baseline ($p < .001$, bootstrap KS test). Interestingly however, when we focus on the conditional contribution decision task (Figure 5.b), the first order dominance of the EDF of decision times in Oath becomes marginal and statistically insignificant ($p = .268$). This may be due to the fact that, in this decision, subjects' have to type in 11 contribution decisions as opposed to one, which effectively imposes a downward limit to their observed decision times.

Last, Figure 5.c reports the EDF of the time spent by subjects on the earnings calculator screen which allows them to explore possible scenarios of interest before making their decision in Step 1 of the experiment. Consistent with the hypothesis that subjects give less weight to the monetary consequences of their decision, subjects spend considerably less time with the earnings calculator under oath. The median time spent with the calculator in Baseline is 463 seconds, as opposed to only 101 seconds in Oath – a 80% decrease in usage time. As a result, The EDF in Oath first order dominates the EDF in Baseline with $p < .001$ (bootstrap KS test).

times data is self-paced – we do not impose any time pressure on subjects – and merely reflects subjects' investment of cognitive resources in the decision making task. In other words, we do not seek to test whether deliberative or intuitive decision making are causally related to cooperation levels through a manipulation of decision times (see Evans and Rand, 2019, for a review).

5 Conclusion

The extant literature on the private provision of public goods has shown that the majority of players endorse the norm of reciprocity in social dilemma situations, while a minority endorses either free-riding, or altruism (Chaudhuri, 2011). In practice however, many players behave as none of those ideal types. Rather, those individuals can be described as "weak reciprocators": they increase their effort level as a function of that of others, but less than proportionally. Beyond its empirical prevalence, weak reciprocation has been shown to have important negative consequences for the stability of cooperation at the group level (Fischbacher and Gächter, 2010).

In this paper, we argue that weak reciprocation might be best conceived not as a preference, but rather as the observable consequence of a trade-off between two conflicting objectives at the player level: staying true to one's private social preference on the one hand, and maximizing one's private material payoff on the other. Weak reciprocation in this context is consistent with the recent literature on truth-telling showing that many players usually lie, but significantly less than what would be expected based on the profit-maximizing strategy – i.e., most people have an intrinsic preference for being honest, which mitigates (and sometimes eliminates) misreporting (Abeler, Nosenzo, and Raymond, 2019). The novelty of our paper resides in the fact that we apply this behavioral insight to a strategic decision making game such as a the public goods game. In such a game, the trade-off between players' preference for staying honest and their private material gain can lead reciprocal subjects to behave as weak reciprocators, or even free-riders. As a result, as long as players face strong enough monetary incentives to free-ride, actual contributions to the public good may not fully reveal their underlying private social preference.

The open question we explore herein is whether one can improve the preference revelation properties of the public goods game by using a cheap, non-market commitment mechanism — the solemn truth-telling oath, which asks players to think about "honesty" in their contributions. We test whether a classical truth-telling oath can induce more cooperation in a standard public good provision experiment. Our hypothesis is that the oath will create an intrinsic commitment to behave according to one's private social preference and normative views: subjects become less likely to trade-off the truthful revelation of their preference against their material benefits. Taking the oath should therefore induce an increase in the average contributions made to the public good.

Consistent with our hypothesis, we observe a 33% rise in contribution levels in the truth-telling oath treatment. We then discuss the behavioral channel through which this effect occurs. First, we show that players under oath are significantly more likely to match the contribution they expect from others and/or their normative view about the right contribution level. As a result, the empirical distribution of the elicited social types is modified in the oath treatment, with a 57% increase in the proportion of subjects who behave as perfect reciprocators, largely counterbalanced by a decrease in the share of weak reciprocators and free-riders. Last, we analyze data from decision times to pinpoint the mechanism we posit behind our result: players in the honesty oath treatment invest significantly less cognitive resources in the decision task, suggesting

that they ponder less on the material consequences of their contribution choices. We conclude that the classical solemn oath of honesty may be a cost-effective mechanism to induce more cooperation from players in the public goods like situations which lie at the core of many environmental and social challenges. This is true not because the oath induces a change in players' private social preferences, but rather because it creates the intrinsic commitment necessary for many to behave according to their reciprocal norm, despite economic incentives to the contrary.

References

Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Model," Journal of the American Statistical Association, 97(457), 284–292.

Abeler, J., A. Becker, and A. Falk (2014): "Representative evidence on lying costs," Journal of Public Economics, 113, 96–104.

Abeler, J., D. Nosenzo, and C. Raymond (2019): "Preferences for truth-telling," Econometrica, 87(4), 1115–1153.

Barrett, S., and A. Dannenberg (2012): "Climate Negotiations Under Scientific Uncertainty," Proceedings of the National Academy of Sciences, 109(43), 17372–17376.

Beck, T., C. Bühren, B. Frank, and E. Khachatryan (2020): "Can Honesty Oaths, Peer Interaction, or Monitoring Mitigate Lying?," Journal of Business Ethics, 163(3), 467–484.

Bem, D. (1972): "Self-Perception Theory," in Advances in Experimental Social Psychology, ed. by L. Berkowitz, vol. 6. Academic Press, New York.

Burger, J. M. (1999): "The Foot-in-the-Door Compliance Procedure: A Multiple-Process Analysis and Review," Personality and Social Psychology Review, 3(4), 303–325.

Carlsson, F., and M. Kataria (2018): "Do People Exaggerate How Happy They Are? Using a Promise to Induce Truth-Telling," Oxford Economic Papers, 70(3), 784–798.

Carlsson, F., M. Kataria, A. Krupnick, E. Lampi, A. Lofgren, P. Qin, T. Sterner, and S. Chung (2013): "The Truth, the Whole Truth, and Nothing but the Truth - A Multiple Country Test of an Oath Script," Journal of Economic Behavior & Organization, 89, 105–121.

Chaudhuri, A. (2011): "Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature," Experimental economics, 14(1), 47–83.

Cialdini, R. (2001): Influence: Science and Practice. Allyn & Bacon, Boston (MA).

Dietz, T., E. Ostrom, and P. C. Stern (2003): "The struggle to govern the commons," science, 302(5652), 1907–1912.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011): "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences," Journal of the European Economic Association, 9(3), 522–550.

Dufwenberg, M., and G. Kirchsteiger (2004): "A theory of sequential reciprocity," Games and economic behavior, 47(2), 268–298.

Evans, A. M., and D. G. Rand (2019): "Cooperation and decision time," Current opinion in psychology, 26, 67–71.

Falk, A., and U. Fischbacher (2006): "A theory of reciprocity," Games and economic behavior, 54(2), 293–315.

Fallucchi, F., R. A. Luccasen, and T. L. Turocy (2018): "Identifying Discrete Behavioural Types: A Re-Analysis of Public Goods Game Contributions by Hierarchical Clustering," Journal of the Economic Science Association, pp. 238–254.

Fehr, E., and A. Leibbrandt (2011): "A Field Study on Cooperativeness and Impatience in the Tragedy of the Commons," Journal of Public Economics, 95(9-10), 1144–1155.

Feige, C., K.-M. Ehrart, and Krämer (2018): "Climate Negotiations in the Lab: A Threshold Public Goods Game with Heterogeneous Contributions Costs and Non-Binding Voting," Environmental & Resource Economics, 70(2), 343–362.

Festinger, L. (1957): A Theory of Cognitive Dissonance. Stanford University Press, CA.

Fischbacher, U., and F. Föllmi-Heusi (2013): "Lies in disguise: an experimental study on cheating," Journal of the European Economic Association, 11(3), 525–547.

Fischbacher, U., and S. Gächter (2010): "Social preferences, beliefs, and the dynamics of free riding in public goods experiments," American economic review, 100(1), 541–56.

Fischbacher, U., S. Gächter, and E. Fehr (2001): "Are people conditionally cooperative? Evidence from a public goods experiment," Economics letters, 71(3), 397–404.

Gachter, S. (2007): "Conditional Cooperation. Behavioral Regularities from the Lab and the Field and Their Policy Implications," in Economics and Psychology. A Promising New Cross-Disciplinary Field, ed. by B. S. Frey, and A. Stutzer, CESifo Seminar Series. MIT Press, Cambridge (MA).

Geller, E. S., M. J. Kalsher, J. R. Rudd, and G. R. Lehman (1989): "Promoting Safety Belt Use on a University Campus: An Integration of Commitment and Incentive Strategies1," Journal of Applied Social Psychology, 19(1), 3–19.

Gneezy, U., A. Kajackaite, and J. Sobel (2018): "Lying aversion and the size of the lie," American Economic Review, 108(2), 419–53.

Gouldner, A. W. (1960): "The norm of reciprocity: A preliminary statement," American sociological review, pp. 161–178.

Greiner, B. (2015): "Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE," Journal of the Economic Science Association, 1(1), 114–125.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath (2001): "In search of homo economicus: behavioral experiments in 15 small-scale societies," American Economic Review, 91(2), 73–78.

Hergueux, J., and N. Jacquemet (2015): "Social Preferences in the Online Laboratory: A Randomized Experiment," Experimental Economics, 18(2), 252–283.

Jacquemet, N., A. James, S. Luchini, and J. Shogren (2011): "Social Psychology and Environmental Economics: A New Look at Ex Ante Corrections of Biased Preference Evaluation," Environmental & Resource Economics, 48(3), 411–433.

——— (2017): "Referenda under Oath," Environmental & Resource Economics, 67(3), 479–504.

Jacquemet, N., A. G. James, S. Luchini, J. J. Murphy, and J. F. Shogren (2021): "Do truth-telling oaths improve honesty in crowd-working?," PloS ONE, 16(1), e0244958.

Jacquemet, N., R.-V. Joule, S. Luchini, and J. F. Shogren (2013): "Preference elicitation under oath," Journal of Environmental Economics and Management, 65(1), 110–132.

Jacquemet, N., S. Luchini, A. Malézieux, and J. Shogren (2020): "Who'll Stop Lying under Oath? Experimental Evidence from Tax Evasion Games," European Economic Review, 20, 103369.

Jacquemet, N., S. Luchini, J. Rosaz, and J. F. Shogren (2018): "Truth-Telling under Oath," Management Science, 65(1), 426–438.

Jacquemet, N., S. Luchini, J. Shogren, and A. Zylbersztejn (2017): "Coordination with Communication under Oath," Experimental Economics, 21(3), 627–649.

Joule, R.-V., F. Girandola, and F. Bernard (2007): "How Can People Be Induced to Willingly Change Their Behavior? The Path from Persuasive Communication to Binding Communication," Social and Personality Psychology Compass, 1(1), 493–505.

Kajackaite, A., and U. Gneezy (2017): "Incentives and cheating," Games and Economic Behavior, 102, 433–444.

Katzev, R., and T. Wang (1994): "Can Commitment Change Behavior ? A Case Study of Environmental Actions," Journal of Social Behavior and Personality, 9, 13–26.

Kiesler, C. (1971): The Psychology of Commitment. Experiments Liking Behavior to Belief. Academic Press, New York.

Kiesler, C., and J. Sakumura (1966): "A Test of a Model for Commitment," Journal of Personality and Social Psychology, 3(3), 349–353.

Koessler, A.-K., B. Torgler, L. P. Feld, and B. S. Frey (2019): "Commitment to Pay Taxes: Results from Field and Laboratory Experiments," European Economic Review, 115, 78–98.

Krajbich, I., B. Bartling, T. Hare, and E. Fehr (2015): "Rethinking Fast and Slow Based on a Critique of Reaction-Time Reverse Inference," Nature Communications, 6.

Lotito, G., M. Migheli, and G. Ortona (2013): "Is Cooperation Instinctive? Evidence from the Response Times in a Public Goods Game," Journal of Bioeconomics, 15(2), 123–133.

Mazar, N., O. Amir, and D. Ariely (2008): "The dishonesty of honest people: A theory of self-concept maintenance," Journal of marketing research, 45(6), 633–644.

Oliver, M., J. Pike, S. Huang, and J. Shogren (2014): "Climate Policy Coordination through Institutional Design: An Experimental Examination," in Toward a New Climate Agreement: Conflict, Resolution, and Governance, ed. by T. L. Cherry, J. Hovi, and D. McEvoy, pp. 108–127. Routledge Books, London.

Ostrom, E. (1990): Governing the commons: The evolution of institutions for collective action. Cambridge university press.

Pallack, M., D. Cook, and J. Sullivan (1980): "Commitment and Energy Conservation," in Applied Social Psychology Annual, ed. by L. Bickman, pp. 235–253. Beverly Hills, CA: Sage.

Peer, E., and Y. Feldman (2020): "Honesty Pledges for the Behaviorally-Based Regulation of Dishonesty," SSRN Working Paper.

Rand, D. G., J. D. Greene, and M. A. Nowak (2012): "Spontaneous Giving and Calculated Greed," Nature, 489(7416), 427–430.

Rubinstein, A. (2007): "Instinctive and Cognitive Reasoning: A Study of Response Times," Economic Journal, 117(523), 1243–1259.

Rustagi, D., S. Engel, and M. Kosfeld (2010): "Conditional Cooperation and Costly Monitoring Explain Success in Forest Commons Management," science, 330(6006), 961–965.

Sekhon, J. (2011): "Multivariate and Propensity Score Matching Software with Automated Balance Optimization," Journal of Statistical Software, 42(7), 1–52.

Sobel, J. (2005): "Interdependent preferences and reciprocity," Journal of economic literature, 43(2), 392–436.

Appendix

A Public goods game instructions

We provide below a screen-by-screen translation in English of the original instructions in French.

A.1 Screen 1: Description

In this section, groups of 4 participants (yourself and 3 other participants) are randomly formed.

At the beginning of this section, each member of the group receives 10€.

Each member of the group must then decide how many euros to keep for himself or herself and how many to invest in a common project.

Each euro invested in the common project by a member of the group yields a return of 0.40€ to each of the 4 group members (including yourself). In other words, the total amount of the contributions to the common project is multiplied by 1.6 before being evenly distributed between the 4 group members.

Your earnings in euros at the end of this section are given by:
10 - (your contribution to the common project) + 0.4 × (total contribution to the common project)

⇒ The next screen gives examples...

A.2 Screen 2: Examples

- If no one contributes to the common project, your earnings will be: 10 - 0 + 0.4 × 0 = 10€ and the earnings of the other group members will be: 10 - 0 + 0.4 × 0 = 10€.
- If everyone contributes 10€ to the common project, your earnings will be: 10 - 10 + 0.4 × 40 = 16€ and the earnings of the other group members will be: 10 - 10 + 0.4 × 40 = 16€.
- If the other group members contribute a total of 15€ to the common project and your contribution is 0€, your earnings will be: 10 - 0 + 0.4 × 15 = 16€. If you contribute 5€ to the common project, your earnings will be: 10 - 5 + 0.4 × 20 = 13€.
- If you contribute 4€ to the common project and the other group members contribute a total of 1€, your earnings will be: 10 - 4 + 0.4 × 5 = 8€. If the other group members contribute a total of 11€, your earnings will be: 10 - 4 + 0.4 × 15 = 12€.

⇒ In the next screen you can experiment with an earnings calculator in order to test all the possibilities you are interested in before making your decision.

A.3 Screen 3: Earnings calculator

The experiments you try on this screen will not affect your earnings in this section.

[EARNINGS CALCULATOR]

- My contribution
- Contribution of the other participants
- Use random values for the other participants

After each trial, click on the "restart" button in order to reinitialize the calculator.

A.4 Screen 4: How to enter your decision?

In the next screen, you will decide how much of your 10€ you want to invest in the common project. You will have to make this decision in 2 different situations:

- In the first situation, you simply have to decide how much of your 10€ you want to invest in the common project.
- In the second situation, you have the opportunity to choose your contribution depending on the contribution of the other group members.

Once you have made those 2 decisions, the decision that will be used to determine your contribution and calculate your earnings in this section will be randomly selected. You will therefore have to think carefully about both types of decisions.

A.5 Screen 5: Enter your decision 1/2

This is a decision screen. Once you have made your decision and clicked the "Next" button, you will not be able to go back to this screen again.

You have 10€ in your possession. How much do you want to invest in the common project?

A.6 Screen 6: Additional Questions: decision 1/2

Your answers to these additional questions do not affect your earnings.

1. Generally speaking, how much do you think people should contribute to the common project?

2. When you chose the amount of your contribution, did you try to anticipate the amount that the other group members would actually contribute to the common project?
Please choose only one of the following:

- Yes
- No

3. IF YES TO PREVIOUS QUESTION ⇒ How much did you think the other group members would contribute on average?

A.7 Screen 7: Enter your decision 2/2

This is a decision screen. Once you have made your decision and clicked the "Next" button, you will not be able to go back to this screen again.

You are now provided with a contribution table that lists each possible average contribution that the other group members could make (all integers between 0 and 10).

For each possible average contribution of the other group members, how much do you want to invest in the common project?

If the other group members make an average contribution of: 0€ 1€ 2€ 3€ 4€ 5€ 6€ 7€ 8€ 9€ 10€
How much do you want to invest in the common project?

B Oath form used in the experiment

PARIS SCHOOL OF ECONOMICS
ÉCOLE D'ÉCONOMIE DE PARIS

SOLEMN OATH

I undersigned ..... swear upon my honor that, during the whole experiment, I will:

Tell the truth and always provide honest answers.

Paris, ... Signature.....

Paris School of Economics, 48 Boulevard Jourdan 75014 Paris – France.

C Oath procedure

Subjects wait in front of the laboratory room. The experimenter's assistant distributes consent forms and pens to participants.

The experimenter announces: *"You will enter the room one by one. Please wait until I call you to come in. Once in the room, I will take the signed consent form from you and you will draw the name of your computer. You will then enter the room and wait until everyone has completed this process."*

Inside the room, one participant enters and the experimenter says: *"Hi, please give me the consent form; thank you. Now please draw a paper with your computer name."*

Then, while giving the oath form to the subject: *"OK. We would also like you to sign this document, but please notice that signing it is not mandatory and will not affect either your participation in the experiment*

or your experimental earnings from the experiment. Please read this form carefully and decide whether you want to sign it or not."

Whatever the decision: get the oath form back and keep in mind the computer name drawn by the subject in case he/she decided not to sign the oath (in this case, record the computer name once the participant has left the room). *"Thank you, you can now enter the next room and settle in front of the computer you have just drawn. We will start in a few minutes."*

Hide the folder of signed/refused oaths so that they are not visible. Have the next subject enter the room and repeat the process.

Throughout this process, another experimenter waits in the laboratory to help participants find their way and tells them that they are not allowed to talk before the start of the experiment.

D Treatment effect on the distribution of social types using alternative classification rules

The classification rule used in the text (Section 4.2) is "very strict", in the sense that it requires that free riders exactly satisfy $m_i = 0$ (*i.e.*, the subject never makes a positive contribution, irrespective of the average contribution of the other members of the group) and that altruists exactly satisfy $m_i = 1$ (*i.e.*, the subject always contributes all of his endowment, irrespective of the average contribution of the other members of the group). To check for the consistency of our results, we also apply less stringent classifications rules, which notably allow subjects for some range of decision error. Specifically, the "strict" classification rule requires that free riders satisfy $m_i \leq 0.1$ (and, symmetrically, that altruists satisfy $m_i \geq 0.9$), while the "loose" classification rule requires that free riders satisfy $m_i \leq 0.2$ (and, symmetrically, that altruists satisfy $m_i \geq 0.8$).

Table A: Distribution of types in the conditional public goods: alternative classification rules

Strict (Fisher test: $p = .039$)						
Social Type				Baseline ($N = 180$)	Oath ($N = 120$)	Fisher test
Free rider	$\{r_i = 0\}$	&	$\{m_i \leq 1\}$	17.8	15.0	$p = .636$
Altruist	$\{r_i = 0\}$	&	$\{m_i \geq 9\}$	1.1	3.3	$p = .222$
Reciprocator	$\{r_i \geq 1\}$	&	$\{0 < m_i < 10\}$	21.7	34.2	$p = .023$
Weak reciprocator	$\{r_i < 1\}$	&	$\{0 < m_i < 10\}$	59.4	47.5	$p = .045$
Loose(Fisher test: $p = .041$)						
Social Type				Baseline ($N = 180$)	Oath ($N = 120$)	Fisher test
Free rider	$\{r_i = 0\}$	&	$\{m_i \leq 2\}$	28.3	22.5	$p = .284$
Altruist	$\{r_i = 0\}$	&	$\{m_i \geq 8\}$	1.7	4.2	$p = .273$
Reciprocator	$\{r_i \geq 1\}$	&	$\{0 < m_i < 10\}$	21.7	34.2	$p = .023$
Weak reciprocator	$\{r_i < 1\}$	&	$\{0 < m_i < 10\}$	48.3	39.2	$p = .125$

Table A replicates Table 3 in the text based on these two alternative classification rules. The same change in the overall distribution of social types between treatments can be observed under those modified classification rules. The main consequence of using a less conservative rule is to reclassify a fraction of

subjects considered as weak reciprocators according to the very strict classification rule as free-riders (to a lesser extent, some subjects are also reclassified as altruists, but the share of this social type always remains small, and the change is not statistically significant). The increase in the share of reciprocators is not affected by the change in the classification rule, and remains in line with a significant decrease in the share of weak reciprocators.

E Balancing tests of individual covariates between treatments

Table B reports the mean of our subjects' demographic characteristics by treatment (with the associated standard deviation in parenthesis). The two last columns report a chi-square test of the difference in the frequency of answers between treatments, and a WMW test of the difference in median answers, respectively. Age is measured in years. Education levels and monthly salary are reported in 7 incremental categories. Subjects reported how carefully they read the experimental instructions on a 5 points scale. All remaining variables are dichotomous.

Table B: Balance of individual covariates between treatments

	No Oath		Oath		Chi-square test	W-M-W test
	Obs	Mean	Obs	Mean		
Age	180	27.37	120	27.74	$p = .491$	$p = .974$
		(10.898)		(11.649)		
Female	180	0.53	120	0.58	$p = .421$	$p = .422$
		(0.500)		(0.500)		
Education level	179	4.21	120	3.79	$p = .005$	$p = .012$
		(1.476)		(1.478)		
Scientific major	100	0.16	68	0.16	$p = .976$	$p = .976$
		(0.368)		(0.371)		
Salary level	176	1.98	117	1.96	$p = .845$	$p = .929$
		(1.149)		(1.102)		
Education level father	116	3.47	110	3.29	$p = .991$	$p = .568$
		(2.240)		(2.198)		
Education level mother	118	3.15	114	3.02	$p = .810$	$p = .579$
		(2.037)		(2.000)		
Not born in France	180	0.24	120	0.18	$p = .186$	$p = .187$
		(0.428)		(0.382)		
Father not born in France	180	0.49	120	0.48	$p = .850$	$p = .628$
		(0.501)		(0.502)		
Mother not born in France	180	0.44	120	0.43	$p = .739$	$p = .740$
		(0.500)		(0.496)		
Student	180	0.64	120	0.59	$p = .409$	$p = .410$
		(0.482)		(0.494)		
Civic volunteer	180	0.21	120	0.19	$p = .768$	$p = .769$
		(0.405)		(0.395)		
Previous subject to similar experiment	180	0.41	120	0.38	$p = .564$	$p = .565$
		(0.494)		(0.491)		
Has read the instructions carefully	180	4.00	120	4.09	$p = .719$	$p = .217$
		(0.717)		(0.733)		

Table C additionally reports the average answers to the social value survey questions by treatment (see Step 3 in Section 2.1), along with WMW non-parametric tests of the difference between treatments (except for question *(iv)* which is dichotomous, and where we report a Fischer exact test of the difference

Table C: Self-reported social attitudes between treatments

		N	Scale	Baseline	Oath	Δ(Oath - Baseline)
(i)	Cooperation	287	0-10	3.71	4.01	$p = .167$
(ii)	Altruism	295	0-10	3.90	3.71	$p = .501$
(iii)	Fairness	290	0-10	5.69	5.39	$p = .291$
(iv)	Trust (WVS)	278	0-1	0.66	0.73	$p = .239$
(v)	General trust	291	0-4	2.43	2.34	$p = .256$
(vi)	Trust in strangers	295	0-4	2.04	1.93	$p = .214$
(vii)	Risk aversion	238	0-10	6.12	5.69	$p = .284$

in frequency by treatment). We can see no evidence that subjects' social values as measured by these questions differ between treatments.

F Behavior of subjects with no first-order belief

The first-order belief (FOB) variable is only available for those subjects who answered "yes" to the question about whether they thought about the behavior of the other players when deciding on their own contribution. As a result, the reciprocity gap (Section 4.3) can only be computed for this sub-sample of subjects. Within each treatment, we do not observe any significant difference in contributions between both sub-samples. In Baseline, the mean unconditional contribution is 3.6 for subjects who answered "no" and 3.7 for those who answered "yes", with $p = .923$. In Oath, the mean unconditional contribution is 5.3 for subjects who answered "no" and 4.7 for those who answered "yes", with $p = .355$. Similarly, within each treatment, Fisher exact tests indicate that the overall distribution of normative beliefs are similar in both sub-samples ($p = .911$ in Baseline and $p = .225$ in Oath).

Table D: Distribution of the normative gap by FOB and treatment

	Baseline				Oath			
	N	Negative	Zero	Positive	N	Negative	Zero	Positive
With FOB	134	57.5	36.5	6.0	89	46.1	47.2	6.7
Without FOB	46	54.4	32.6	13.0	31	29.0	58.1	12.9

Table D reports the distribution of the normative gaps separately for subjects who answered "yes" (first row) and "no" (second row) to the FOB question. Within each treatment, the differences between both sub-samples are not statistically significant (Fisher exact tests, $p = .332$ in Baseline and $p = .199$ in Oath). The normative gaps exhibit the same pattern between treatments: in both sub-samples, we observe an increase in the fraction of subjects who exhibit no normative gap in Oath as compared to Baseline. Interestingly, however, Fisher exact tests indicate that the overall change in the distribution of normative gaps between treatments is to a large extent driven by subjects who do not think about the behavior of the other players ($p = .235$ with FOB as opposed to $p = .058$ without FOB), as those subjects are twice as likely to behave according to their normative view under oath.