Cooperation under oath: A case for context-dependent preferences*

João Vaz[a] and Jason Shogren[b]

March 2023

Abstract

Previous work has found that the contextual manipulation of a truth-telling oath is effective at promoting pro-social behavior in situations of intrapersonal conflict (e.g., lying games) and conjecture that the mechanism behind changes in behavior is that the oath makes individuals properly represent their underlying preferences. There are nevertheless solemn oaths that consider other-regarding statements meant to regulate behavior in situations marked by strategic conflict whose impact remains untested. Here, we examine whether an other-regarding oath impacts behavior in the simultaneous and sequential versions of the prisoners' dilemma game and explore whether that impact, if observed, could be attributed to a change of preferences. We observe (1) that the oath significantly impacts cooperation by oath-takers across variants of the prisoners' dilemma and (2) an overwhelming transfer of reported strategies by oath-takers moving second from selfish to conditionally cooperative. Since strategies of second movers are elicited independently of beliefs, our results lend strong support to the hypothesis that the other-regarding oath works through a change of preferences. We provide direct proof of an instance where preferences for cooperation in strategic settings are context dependent.

Keywords: Social dilemma; Cooperation; Oath; Context-dependent preferences.
JEL Classification: C72, D83.

*Financial support provided by the University of Wyoming College of Business Academy of Research Excellence, the University of Wyoming Center for Global Studies, and the Stroock Chair of Natural Resource and Environmental Economics.

[a]Department of Economics, University of Gothenburg, Vasagatan 1, 411-24, Gothenburg, Sweden. joao.vaz@economics.gu.se

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

1 Introduction

Economists have long been trying to find solutions to align private interests to social goals. Among the most researched non-price interventions are contextual manipulations of the decision-making environment. Several studies have found evidence that pro-social contexts significantly increase cooperation in social dilemmas.[1] This evidence has cast doubt on utility theory in general and on social preference theories in particular (e.g., Weber et al.; 2004; Levitt and List; 2007). However, as noted by Fehr and Schmidt (2006) and Rabin (1998), context-dependent behavior does not imply that preferences are labile. Although some authors interpret context effects as a manifestation of variable preferences (e.g., McCusker and Carnevale; 1995; van Dijk and Wilke; 2000; Iturbe-Ormaetxe et al.; 2011; Chang et al.; 2019; Gächter et al.; 2022), others have demonstrated that changes in behavior are compatible with stable social preferences through changes in expectations (e.g., Sonnemans et al.; 1998; Dufwenberg et al.; 2011; Ellingsen et al.; 2012; Dreber et al.; 2013; Fosgaard et al.; 2014). Despite the importance of the variable-preference result for economic theory and for the question of why people cooperate in social dilemmas, direct empirical support for context-dependent preferences in strategic settings remains elusive.

More recently, researchers have examined the role of a solemn oath to honesty as a device against self-interest. In experimental studies, the oath procedure involves a pre-experiment request to sign a truth-telling oath, so that subsequent decisions can be evaluated within the oath-taking context. The honesty oath has been applied to situations of honest reveal of stated preferences (Carlsson et al.; 2013; Jacquemet et al.; 2013, 2017), compliance in tax-evasion games (Jacquemet et al.; 2020), and truthful communication in

[1] We broadly define changes in context as manipulations of otherwise "neutral" experimental environments, in which the decision situation is logically equivalent. Examples of pro-social contexts in social dilemmas include community framing of game (Kay and Ross; 2003; Liberman et al.; 2004; Rege and Telle; 2004; Dufwenberg et al.; 2011; Ellingsen et al.; 2012) and action labels (Andreoni; 1995; McCusker and Carnevale; 1995; Sonnemans et al.; 1998; Park; 2000; van Dijk and Wilke; 2000; Cubitt et al.; 2011; Fosgaard et al.; 2014; Khadjavi and Lange; 2015; Gächter et al.; 2022), moral suasion (Dal Bó and Dal Bó; 2014; Fellner et al.; 2013; Ito et al.; 2018; Konow; 2019), and action recommendations (Marks et al.; 1999; Croson and Marks; 2001; Galbiati and Vertova; 2008; Bicchieri et al.; 2021). Here, we are only concerned with the effect of context on cooperation in strategic settings. See Tversky and Kahneman (1981) for a seminal paper on how context affects individual choice. For a review, see Levin et al. (1998).

coordination games (Jacquemet et al.; 2018). These studies find that the oath can improve behavior toward social objectives.

The working hypothesis regarding the mechanism behind changes in behavior is that the honesty oath creates the incentive for individuals to behave according to their underlying preference.[2] Based on the empirical insight that a large share of individuals is neither purely selfish nor altruistic, but instead exhibits an intermediate degree of prosocial tendencies, the assumption is that the oath induces individuals to properly represent their social preference by making them more resistant to profit-maximizing behavior. This hypothesis is supported by Hergueux et al. (2022), who, in a public goods game, find that weak reciprocators (who contribute less than the contribution expected of others) are more likely to match the average contribution of other players when under oath, and by Jacquemet et al. (2020), who, in a tax-evasion game, find that compliance increases at the intensive margin among individuals with weak preferences for lying. In both studies, researchers find no appeal to a change of preference among the two polar types of pure selfishness (free riders and "chronic" liars) and pure altruism (unconditional cooperators and never liars).

The existing literature considers an oath statement that is independent of relational concerns. The honesty oath offers a commitment meant to regulate intrapersonal conflict, such as the incentive to misrepresent private information in favor of material gain, though not necessarily (or explicitly) at the expense of others. Examples of self-monitoring statements can be found in professional oaths, such as the MBA oath or the Dutch Bankers oath, which exhort the honest disclosure of financial activities or compliance with professional standards. There are nevertheless oath statements that entail an explicit commitment to a counterparty with whom we may have a conflict of interest. Examples include swearing fidelity in marriage vows, putting the interests of clients first in fiduciary oaths, or to do no harm to patients in medical oaths. These solemn pledges consider other-regarding statements that bind us to behavior that directly impacts the well-being of others in strategic interactions.

Although the honesty oath has been consistently shown to be effective in situations marked by intrapersonal conflict (e.g., lying games), it remains an open question whether an

[2] According to the socio-psychology theory of commitment, the oath is a commitment device that makes salient to the oath-taker the type of behavior that would be consistent with their private preference (see, e.g., Kiesler; 1971; Joule et al.; 2007).

other-regarding oath can help overcome self-interest in situations of interdependent choice, both from those taking the oath and from those whom the oath is intended to. Moreover, despite the support for the hypothesis that the honesty oath aligns behavior to underlying preferences, that an other-regarding oath operates through a similar mechanism remains undetermined. In contrast to a truth-telling statement, an other-regarding statement is directly connected not just to own outcomes, but to those of others. It is unclear whether the emphasis on the well-being of others only affects behavior that is already tied to self (i.e., pre-existing preferences), or whether oath-keeping behavior requires a change of other-regarding, or social, preferences.

The objective of this paper is twofold. First, evaluate whether an other-regarding oath has an impact on cooperation in strategic situations. And second, evaluate whether that impact, if observed, could be attributed to a change of preferences. To do so, we consider the effect of the oath in a prisoners' dilemma (PD) game, which represents the paradigm of human cooperation in the face of strategic conflict. Unlike previous work, where the oath is private and administered to all subjects before performing a task, in a two-player interaction, we ask that only one of the players sign a public oath. This allows us to explore differentiated impacts of the oath environment on both oath-takers and oath-recipients.[3] To understand not only whether the oath impacts behavior, but also whether there is support for a change of preferences, we consider both the simultaneous and sequential versions of the PD. In the sequential PD (SPD), one player moves before the other and the second mover observes the decision of the first mover before making her own decision. Since second movers can condition their action on that of first movers, we are guaranteed to observe a best reply, therefore allowing for the revelation of preferences over outcomes in the PD.

The oath statement that we ask subjects to swear is "to help others at all times"—a statement that can be found, for instance, in the Scout Oath, which is identical across countries, and whose pledge is in line with other pronouncements that prescribe doing service

[3]Previous work treats audience effects of the oath as confounds which might obscure private motives. Since the objective of our study is to examine the impact of the oath environment on both oath-takers and oath-recipients, we do not regard the concern for the reaction of others as a confound. For models of image-related motivations, see Bénabou and Tirole (2006) and Andreoni and Bernheim (2009). For examples of experimental studies examining audience effects, see Hoffman et al. (1996), Andreoni and Bernheim (2009), Filiz-Ozbay and Ozbay (2014), Savikhin Samek and Sheremeta (2014), and Kessler et al. (2021).

unto others, such as the Hippocratic Oath to "do no harm," that physicians take before practicing medicine. The oath is administered prior to learning the instructions of the game and, although players are left to decide what it means to help others and oneself during the course of the experiment, "to help others" has a direct connection to the PD, insofar as defection is harmful to one's opponent. This "expressive" cue can be contrasted with previous experiments that consider the honesty oath "to tell the truth and always provide honest answers" in strategic environments such as coordination (Jacquemet et al.; 2018) and public good (Hergueux et al.; 2022) games. To compare our findings to those of other researchers and to examine whether the oath-taking environment alone, irrespective of its informational content, is enough to prevent defection, we also test the effectiveness of the honesty statement in the simultaneous version of the PD.

The game is one shot in order to eliminate the incentive to maintain credibility in future interactions. Since there is no economic penalty for not taking the oath nor for failure to meet its provisions, the oath is costless and nonbinding. Because in a PD individuals have a strictly dominant strategy to defect, standard economic theory predicts that the oath be ineffective. On the other hand, experimental studies show that individuals often make cooperative choices and that changes in decision-making context impact behavior. There are two main mechanisms that could explain potential changes in behavior following the oath.

One explanation is provided by social preference theories, such as inequity aversion (Fehr and Schmidt; 1999; Bolton and Ockenfels; 2000) or intentions-based reciprocity (Rabin; 1993; Falk and Fischbacher; 2006). These theories concur with the hypothesis that experimental subjects are conditional cooperators, who interpret games where payoffs have a PD structure as a coordination game in terms of utilities. Because a coordination game has multiple equilibria, context acts as an equilibrium-selection device that changes behavior through changes in expectations about what others are likely to do. As a result, the oath might work to increase cooperation (at the intensive margin) by individuals with pre-existing conditional social preferences. We call this the coordination hypothesis.

An alternative explanation is that the oath works through a change of preferences. Theories of context-dependent preferences applied to social interactions include team reasoning (e.g., Sugden; 1993; Bacharach; 1999) and appropriateness of behavior for complying with social norms (e.g., March; 1994; Montgomery; 1998; Weber et al.; 2004). Although dif-

ferent theories provide different explanations of how context affects intrinsic motivations, they all predict that changes in behavior result from a change in the utility value placed on cooperation in a PD. According to these theories, the oath might activate a stronger desire to cooperate independently of a change in expectations. We call this the variable-preference hypothesis.

To help us evaluate the implications of an oath in simultaneous and sequential play under the coordination and variable preference hypotheses, we consider a one-parameter extension of the PD model to include an extra value for cooperation. This parameter reflects different types of players that differ in terms of their best response to defection or cooperation by their opponent. Assuming that players do not know whether their opponent is a cooperative type (high extra value) or a selfish type (low extra value), players are engaged in a game of incomplete information. Depending on types, the game might transition from a PD to a coordination game in terms of utilities. Because types encode different explanations as to why individuals perceive a PD game differently than its standard representation, the baseline model can be considered as a reduced-form formalization of different theories that reconcile experimental regularities.

In the experiment, we contrast behavior of oath and no-oath groups across the simultaneous and sequential variants of the PD. We employ a between-subject design, in which we compare responses of players moving simultaneously in the PD and moving first or second in the SPD, given the different roles assumed by players in oath groups (i.e., oath-takers and oath-recipients). We elicit players preferences, or types, in the SPD by employing the strategy method (Selten; 1967), whereby second movers are asked to report a contingent strategy—whether to cooperate or to defect against cooperation and defection by the first mover.[4]

Our analysis provides three main findings. First, the other-regarding oath significantly impacts cooperation by oath-takers across variants of the PD. Our results indicate that oath-takers increase cooperation by 23% in the simultaneous version of the PD and that, in sequential treatments, oath-takers are 37% more likely to initiate cooperation when moving

[4]Brandts and Charness (2000) examine whether the strategy method, as opposed to the direct-response method, induces differences in rates of cooperation and find no statistical difference. Nevertheless, statistical differences have been documented in other contexts (e.g., Iriberri and Rey-Biel; 2011). For a review of papers implementing the strategy method, see Brandts and Charness (2011).

first and 53% more likely to reciprocate cooperation when moving second. Despite the significant increase of cooperation by oath-takers, we find that behavior of oath-recipients was unaffected by the oath in both simultaneous and sequential interactions. Overall, we find support for the hypothesis that the oath can improve cooperative behavior in strategic situations, though only upon those directly bound by the behavioral prescription of the oath.

Second, we find an overwhelming transfer of reported strategies by oath-takers moving second from unconditional defection to conditional cooperation. From a roughly equal proportion of selfish and conditionally cooperative types, we observe a 60% reduction of unconditional defectors, which is offset by an equivalent increase in the number of conditional cooperators. This result lends strong support to the variable-preference hypothesis, wherein the other-regarding oath works to increase cooperation at the extensive margin, by increasing the number of individuals who favor cooperative outcomes.

And third, we find that there is no impact of the honesty oath on cooperation by either oath-takers or oath-recipients. This result contrasts to previous work that finds that the honesty oath is effective at promoting prosocial behavior. Our data suggest that behavioral changes are sensitive to the expressiveness of the oath statement and not to the oath environment by itself.

Altogether, our findings are relevant for understanding the determinants of cooperation. The observation that cooperation increases following the other-regarding oath is of practical significance, considering the widespread use of solemn pledges to resolve conflict in strategic interactions (e.g., nonbinding international agreements and employment codes of conduct). Moreover, the result that the oath transforms selfish tendencies into cooperative dispositions is especially important for the debate of whether preferences are dependent on the oath-taking context, in particular, and on context in strategic settings, in general. Although several studies present data consistent with variable-preference theories, few provide unambiguous empirical support for the preference mechanism. Here, we present direct proof of an instance where preferences for cooperation in strategic settings are context dependent.

The remainder of this paper is organized as follows. Section 2 reviews related literature; Section 3 introduces the theoretical framework motivating our hypotheses; Section 4 describes our experiment; Section 5 presents our results; Section 6 offers a discussion; and Section 7 concludes.

2 Related Literature

Our study is related to the literature examining the impact of interventions aimed at promoting cooperation in social dilemmas (see Footnote 1). Because the oath procedure consists of a public commitment to prescribed behavior, our work relates to research examining the impact of non-binding verbal commitments, such as promises. The positive impact of promises on behavior has been extensively documented in experimental literature (e.g., Ellingsen and Johannesson; 2004; Charness and Dufwenberg; 2006; Vanberg; 2008; Ismayilov and Potters; 2016; Ederer and Stremitzer; 2017; Di Bartolomeo et al.; 2019; Mischkowski et al.; 2019; Ederer and Schneider; 2022). Nevertheless, beyond its informational content and the solemnity of the act, the oath that we implement is distinct from verbal messages in promise experiments for two main reasons.[5]

First, taking the oath precedes knowledge of the circumstances that individuals will be facing in a decision task. In promise experiments, subjects are asked to endogenously commit to an action by deciding whether to make a promise or not. Promises are elicited or solicited within the individual's intention and belief in her ability to perform a specific action. The behavioral prescription of the oath to a general class of "good" behavior is such that virtually every individual agrees to take it (in our study, the oath uptake is of 92.3% across all oath treatments), which allows for causal inference of the impact of the oath.[6] And second, the oath can be viewed as the first of a two-part strategy to achieve commitment. In social psychology, commitment is achieved through a preparatory act de-

[5] In general, an oath is a very serious promise of a heavy moral weight to act in accordance with certain principles, made verbally and publicly along with symbolic gestures, whereby the oath-taker lays her integrity on the line by offering a warranty (religious or secular, such as one's honor and conscience) should she break her word (Metz; 2013). Oath statements usually provide declarations of intent with best-endeavor clauses (to promote, support, encourage, or make efforts toward a desirable activity or goal) and are often bound to a general recipient, such as professional peers, the nation, society as a whole, or the common good. In contrast, promises are often unceremonious and typically provide action-oriented statements to a specific individual without the assurance of a moral penalty for reneging on the promise. For further details, see Sheinman (2011), Schlesinger (2011), and Metz (2013).

[6] Ismayilov and Potters (2016), for instance, show that the correlation between free-form promises and cooperation is due to self-selection and suggest that the effect of promises is indistinguishable from cheap-talk communication. Charness and Dufwenberg (2010) elicit promises (exogenously) by a third party and find limited support for the effect of promises on cooperation.

signed to induce a change in subsequent decisions (Kiesler; 1971). The oath works similar to foot-in-the-door techniques which involve a low-cost request that increases the probability of acceptance of subsequent high-cost requests.[7] As a result, unlike promises, the oath implements commitment before game play and is better interpreted as an institutional, or contextual, device that changes the target-decision environment (i.e., being under oath or not), rather than a within-game commitment technology.

Other related work includes experiments that examine cooperation in sequential social dilemmas.[8] Our paper joins the literature documenting that cooperation by followers is influenced by the decisions of leaders (e.g., Clark and Sefton; 2001; Moxnes and Van der Heijden; 2003; Gächter and Renner; 2018), as well as the literature that elicits preferences for conditional cooperation from the behavior of second movers (e.g., Fischbacher et al.; 2001; Kocher et al.; 2008). With regard to preference elicitation in the SPD, several studies evaluate whether preferences for conditional cooperation can be attributed to intentions- or outcome-based concerns (e.g., Blanco et al.; 2014; Miettinen et al.; 2020). In contrast to these experiments, our objective is not to assess the nature of pre-existing conditional preferences (e.g., reciprocity or inequity aversion), but rather to examine whether there is an appeal to a change of preferences following the oath.

Our findings also relate to studies that contrast behavior in the PD and in the SPD. Consistent with the view that a large number of individuals are conditional cooperators, behavior in the PD should respond to a belief about the behavior of other players. Based on models in which players are uncertain about their opponents' type (e.g., Bolle and Ockenfels; 1990; Falk and Fischbacher; 2006), the absence of strategic uncertainty in sequential interactions dictates that cooperation in the SPD should be weakly greater than in the PD. Our study relates to empirical work examining this prediction, which finds that the positive effect of sequentiality is sensitive to game parameters and the subject pool (e.g., Ahn

[7] Two classic examples in the experimental literature include Freedman and Fraser (1966), who found that women who were asked to answer questions about soap products were more likely to accept a subsequent, larger request that a group of men enter their home to take an inventory of the products they owned, and Harris (1972), who found that people asked for the time or for directions were more likely to a give a dime than those presented with an outright request.

[8] Earlier experiments of sequential social dilemmas include the SPD (Bolle and Ockenfels; 1990; Clark and Sefton; 2001), the gift-exchange game (Fehr et al.; 1993), the trust game (Berg et al.; 1995), or the public good games with a first mover (Potters et al.; 2007).

et al.; 2003, 2007; Khadjavi and Lange; 2015).[9] Different from these studies, we compare behavior in the PD and in the SPD by manipulating context, as well as the role assumed by different players within the oath-taking environment (i.e., oath-takers or oath-recipients as either first or second movers).

Finally, our work relates to the literature examining the mechanism by which context changes behavior in social dilemmas. In contrast to earlier studies, our experiment disentangles preferences from beliefs in determining behavior rather than focusing on cooperation decisions alone (see Gerlach and Jaeger (2016) and Alekseev et al. (2017) for two reviews of context effects). In relation to preferences, our paper joins the small set of empirical studies that elicit belief-independent preferences following changes in context using the strategy method in public goods settings (e.g., Frackenpohl et al.; 2016; Fosgaard et al.; 2017; Gächter et al.; 2017; Fosgaard et al.; 2019; Martinsson et al.; 2019; Hergueux et al.; 2022; Isler et al.; 2021; Gächter et al.; 2022). Consistent with our result, Frackenpohl et al. (2016), Fosgaard et al. (2017), and Gächter et al. (2022) find that preferences for conditional cooperation are sensitive to (give/take) institutional frames, although Fosgaard et al. (2014) in a similar setting find that behavioral changes are instead explained by changes in expectations. More directly connected to our analysis are studies that examine context effects in a PD and use the SPD to isolate the preference channel. Yamagishi and Kiyonari (2000) elicit group membership in a PD and use the SPD to articulate that in-group favoritism is driven by expectations, not preferences, and Ellingsen et al. (2012) examine the role of community framing in a PD and find that the framing effect disappears in sequential play, which is a result consistent with the coordination hypothesis. In contrast, we consider the contextual manipulation of the oath and find strong support for preferences to be oath, or context, dependent.

3 Theoretical Background

Consider two players, player 1 (she) and player 2 (he), who play the PD game in Table 1. Labels C and D indicate cooperation and defection. Material payoffs satisfy $x > y > z > 0$ and $2y > x$. The rational prediction is that players choose their dominant strategy D,

[9] For studies that compare cooperation in the PD and SPD with repeated interactions, see Oskamp (1974), Kartal and Müller (2021), and Ghidoni and Suetens (2022).

which leads to the Pareto-dominated outcome (D, D), as opposed to the socially optimum outcome where both players choose C. If players were to play the game sequentially, equilibrium play would entail defection by both first and second movers. Any context that does not change monetary incentives should be ineffective, since players have a dominant strategy to defect.

Table 1: The PD game

$1\backslash 2$	C	D
C	$y\backslash y$	$0\backslash x$
D	$x\backslash 0$	$z\backslash z$

3.1 Baseline with Social Preferences

Because players may be influenced by motives other than material gain, they need not consider D as a dominant strategy. To allow for a meaningful discussion about the mechanism behind oath effects, we consider an extension of the PD game to include social preferences.

There are several models of social preferences that reconcile the evidence that individuals favor cooperation. These models often focus on whether behavior originates from payoff-distribution concerns or whether behavior is motivated by the intentions of others. Our objective is not to examine the nature of preferences for cooperation, but rather to explore whether an oath changes behavior due to a change of preference. For simplicity, we account for social preferences by introducing a θ-parameter utility function, which assigns an extra value to the cooperative outcome. The magnitude of the extra value θ allows for the emergence of two types of players, who differ in terms of their best response to the actions of the other player: unconditional defectors, who favor D regardless, and conditional cooperators, who favor C if reciprocated.[10] Since players involved in a one-shot PD do

[10]Note that unconditional cooperators (who always favor cooperation) are precluded by design. This is motivated by the low frequency of unconditional cooperation in PD experiments (e.g., Bolle and Ockenfels; 1990; Clark and Sefton; 2001; Blanco et al.; 2011), which is also observed in our data. To account for the presence of unconditional cooperation, we could consider an alternative model with an extra value for the cooperative *choice*. A similar set of properties as those derived in the θ-model would follow. The main difference being that the presence of conditional cooperators would be sensitive to the relative magnitude of

not know which type of player they are playing against, they are involved in a game with incomplete information. The recast of the PD in terms of a game with incomplete information is borrowed from Bolle and Ockenfels (1990), but can be considered as a special case of the general model of social preferences examined in Charness and Rabin (2002).

3.1.1 θ-Model

Suppose that players favor the cooperative outcome by means of a θ utility parameter as described in Table 2. To account for preference heterogeneity, we let each player i for $i \in \{1, 2\}$ have their own θ_i, which is assumed to be perfectly known by oneself only. Since θ_i affects the ranking of outcomes, we may separate players into two meta types that differ in terms of their best response to the actions of the other player. The type-partition is dictated by the threshold $x - y$, such that, if $\theta_i < x - y$, then defection is a strictly dominant strategy (unconditional defector), and if $\theta_i \geq x - y$, then player i perceives the game as a stag-hunt, in which it is optimal that players coordinate their actions (conditional cooperator).

Table 2: The θ-model

$1\backslash 2$	C	D
C	$y + \theta_1 \backslash y + \theta_2$	$0 \backslash x$
D	$x \backslash 0$	$z \backslash z$

The model is a two-player incomplete information game that transitions from a PD to a coordination game depending on the type of players. The choice of D by all types of both players is always an equilibrium point. But since for a sufficiently large θ_i, players would favor the cooperative outcome, we may look for an equilibrium in which players choose C whenever θ_i exceeds a critical value and defect otherwise. For a sharper prediction of play, we make the following assumption.

Assumption 1: *$\theta_i \geq 0$ is a continuous variable drawn from a continuous and strictly increasing function $F(\theta)$ on $[0, \infty)$. This distribution is common knowledge.*

monetary payoffs.

Let π_j denote the probability that player $j = 3 - i$ cooperates. Player i chooses C if her expected utility U_i from cooperation, $U_i(C) = \pi_j(y + \theta_i)$, exceeds her expected utility from defection, $U_i(D) = \pi_j x + (1 - \pi_j)z$, or whenever θ_i satisfies:

$$\theta_i \geq x - y + z\,\frac{(1 - \pi_j)}{\pi_j}. \tag{1}$$

Imposing symmetry, we may find the equilibrium cutoff type, θ^*, above which cooperation is optimal from the solution to the following equation:

$$1 - F(\theta^*) = \frac{z}{z + y - x + \theta^*} \tag{2}$$

where $1 - F(\theta^*)$ is the probability that $\theta_i \geq \theta^*$.

Proposition 1. *If θ^* in (2) exists, we get an equilibrium in which a type θ_i player chooses C when $\theta_i \geq \theta^*$ and chooses D otherwise. The choice of D by all types of both players is always an equilibrium point.*

Let us now consider equilibrium behavior in the SPD game. Suppose that player 2 moves second. If player 2 is a conditional cooperator ($\theta_2 \geq x - y$), he would reciprocate cooperation by the first mover with cooperation, and defection with defection. Player 2 therefore adopts the following strategy:

$$s_2(\theta_2) = \begin{cases} C & \text{if player 1 chose } C \text{ and if } \theta_2 \geq x - y \\ D & \text{otherwise.} \end{cases} \tag{3}$$

Let $\overline{\pi}$ denote the probability that a second mover cooperates against cooperation: $\overline{\pi} = 1 - F(x - y)$, which corresponds to the mass of conditional cooperators. Player 1, who moves first, would choose C if her expected utility from cooperation, $U_1(C) = \overline{\pi}(y + \theta_1)$, exceeds her expected utility from defection, $U_1(D) = z$. We may find the critical cooperation threshold for a player who moves first from the following equation:[11]

$$\hat{\theta} = \max\{0, -y + z/\overline{\pi}\} \tag{4}$$

[11] Provided that $F(x-y) < 1$. If $F(x-y) = 1$, then there are no conditional cooperators, and a first-mover would always choose D, or $\hat{\theta} = \infty$.

Proposition 2. *In the SPD game, player 2, who moves second, behaves according to* (3) *and a type* θ_1 *player who moves first chooses* C *whenever* $\theta_1 \geq \hat{\theta}$ *for* $\hat{\theta}$ *in* (4) *and* D *otherwise*.

It follows from (2) and (4) that $\hat{\theta} < \theta^*$. That is, the proportion of first movers who cooperate in a sequential game is larger than the proportion of players who cooperate in a simultaneous game.

3.2 Oath

Suppose that, prior to learning about the game, one of the players is asked to swear under oath to being cooperative and that their decision is observed by the other player. This situation parallels the case of adding a communication stage before gameplay. Because the model assumes incomplete information about types, a pre-play interaction may create signaling opportunities. However, given that the oath is administered before learning about the game, we might risk invoking an incorrect equilibrium criterion about the oath-taking decision. Since types reflect preferences over outcomes, which are unknown at the oath-taking stage, it is extreme to appeal to the notion of sequential rationality in informing the decision of whether to take the oath or not. We therefore make the following assumption:

Assumption 2: *Taking an oath is independent of types*.

Assumption 2 is in principle testable, particularly if we find that there is scope for separation of types in the data. We return to this discussion after presenting the experimental results.

The following two subsections explore the behavioral implications of the oath given our two hypotheses: (1) the coordination hypothesis and (2) the variable preference hypothesis.

3.2.1 Coordination Hypothesis

Suppose that the oath does not change preferences. Because not all individuals have a dominant strategy to defect, some players perceive the PD as a game with multiple equilibria. That is the case in the θ-model, in which, according to Proposition 1, there are two equi-

libria for conditional cooperators whose type exceeds θ^*. In that case, the oath need not be ineffective, but instead act as a selection device of the cooperative equilibrium. Moreover, because the oath does not change preferences, the selection device argument should apply with equal force to both oath-takers and oath-recipients of type $\theta \geq \theta^*$.

The way in which the oath environment affects equilibrium selection depends on the specific variety of social preferences that transform the PD into a game with multiple equilibria. For instance, if individuals have an intrinsic desire for equity, fairness, or reciprocity, the oath may help identify what outcomes are perceived as equitable, fair, and kind. In that case, the oath acts as a cluster of self-fulfilling expectations that promotes cooperation due a change of "empirical" expectations (i.e., first-order beliefs that a certain behavior will be followed).[12] Alternatively, if individuals have a propensity toward guilt aversion, the oath may induce individuals to believe that their partner anticipates cooperation. In that case, the oath promotes cooperation due to changes of "normative" expectations (i.e., second-order beliefs that a certain behavior should be followed). In both instances, changes in behavior operate through changes in expectations of players who exhibit conditional social preferences.

The coordination hypothesis thus states that the oath facilitates the attainment of the cooperative equilibrium in simultaneous play, without affecting types, by acting as a coordination device for players of type $\theta \geq \theta^*$.

The coordination hypothesis nevertheless predicts no impact of the oath in sequential interactions. In the SPD, second movers choose their strategy independently of beliefs. If an oath does not change preferences (e.g., the desire for fairness, or the feeling of guilt from not matching expected cooperation), second movers should respond to observed cooperation and defection in the same way, with or without the oath. As for first movers, their behavior depends on the belief about the proportion of conditional cooperators. However, given that (1) preferences of second movers have not changed and (2) first movers anticipate that the choice of second movers is belief-independent (i.e., there is no strategic uncertainty), then the choice of first movers should also be unaffected by the oath. Therefore, the coordination hypothesis states that behavior of first and second movers in the SPD following the oath should remain that described in Proposition 2.

[12] That is, the oath might indicate the focal equilibrium pf the game Schelling (1980).

3.2.2 Variable Preference Hypothesis

Suppose that taking the oath affects the degree to which a player favors cooperation. Consistent with how we introduced social preferences in the baseline model, an increase in preference for cooperation can be formalized by an extra value to the cooperative outcome, which is equivalent to a change in the value of θ.[13,14]

The simplest way to capture a change of types without imposing homogeneous effects is to assume that θ is drawn from a different distribution, G. To rationalize that θ is nondecreasing following the oath, it is enough to assume that G first-order stochastically dominates the baseline distribution of types, F. Since the oath environment features an other-regarding commitment from the oath-taker to the oath-recipient, we assume that only the type of the oath-taker is drawn from a different distribution. We could extend the analysis to the case where the oath-recipient's type also changes. This would not affect the overall direction of the results, although it remains a testable assumption. To simplify exposition, we let player 1 be the oath-taker and player 2 be the oath-recipient.

Assumption 3: *Oath-takers of type θ_1 are distributed according to a continuous and strictly increasing distribution function G on $[0, \infty)$, which first-order stochastically dominates distribution F. That is, $G(\theta_1) \leq F(\theta_1)$ for all $\theta_1 \in [0, \infty)$. Oath-recipients of type θ_2 are distributed according to F. Both distributions are common knowledge.*

Let us consider the PD game. As in the baseline model, the choice of D by all types of both players remains an equilibrium point. As for the cooperative equilibrium, since types of players are drawn from different distributions, equilibrium play entails two threshold types, θ_1^o and θ_2^o for the oath-taker and the oath-recipient. The probability that players

[13] As in the baseline model, we could have assumed an extra value to the cooperative choice, in which the oath would result in the emergence of unconditional cooperators. Again, we rule out this case by design.

[14] Extra values can be thought of as a gratification for "doing the right thing." There are several models that capture a change of preference following a positive change in context, such as models of social identity (e.g., Akerlof and Kranton; 2000, 2005) and norm compliance (e.g., Andreoni and Bernheim; 2009; Krupka and Weber; 2013; López-Pérez; 2008), in which utility depends on how actions conform with one's social identity or with the norm of a specific context, or models of a belief-independent nonpecuniary or psychological cost from lying (Ellingsen and Johannesson; 2004). We return to theories of context-depedent preferences in the discussion section.

cooperate under the candidate for equilibrium is $\pi_1^o = 1 - G(\theta_1^o)$ for oath-takers and $\pi_2^o = 1 - F(\theta_2^o)$ for oath-recipients. The two equilibrium thresholds are found from the solution to the following system of equations, which marks the indifference of expected utilities from choosing C and D for the oath-taker (5) and for the oath-recipient (6):

$$1 - F(\theta_2^o) = \frac{z}{z + y - x + \theta_1^o}, \tag{5}$$

$$1 - G(\theta_1^o) = \frac{z}{z + y - x + \theta_2^o}. \tag{6}$$

Proposition 3.

(a) If θ_i^o for $i = \{1, 2\}$ in (7) and (8) exist, we get an equilibrium in which a type θ_i player chooses C when $\theta_i \geq \theta_i^o$ and chooses D otherwise. The choice of D by all types of both players is always an equilibrium point.

(b) If θ_i^o for $i = \{1, 2\}$ and θ^* from Proposition 1 exist, then $\theta_i^o \leq \theta^*$.

The proof is in Appendix A.

From Proposition 3(b), it follows that both oath-takers and oath-recipients are more likely to choose C in simultaneous play with an oath than in the baseline case.[15]

Let us now consider the SPD game. A type θ_i player who moves second follows the same strategy as in (3) (without the oath). Only now the proportion of players choosing C in response to C is different across oath-takers ($\overline{\pi}_1$) and oath-recipients ($\overline{\pi}_2$). The proportion of oath-recipients who cooperate against cooperation is the same as in the baseline model, $\overline{\pi} = \overline{\pi}_2 = 1 - F(x - y)$, but that of oath-takers is $\overline{\pi}_1 = 1 - G(x - y)$, which is greater than $\overline{\pi}_2$ since $G \leq F$.

As for the behavior of first movers, the choice of cooperation is expected to differ between oath-takers and oath-receivers. Oath-takers who move first cooperate if $\theta_1 \geq \hat{\theta}_1^o$, where

$$\hat{\theta}_1^o = \max\{0, -y + z/\overline{\pi}_2\}, \tag{7}$$

which is equal to $\hat{\theta}$ in (4), and oath-recipients who move first cooperate if $\theta_2 \geq \hat{\theta}_2^o$, where

$$\hat{\theta}_2^o = \max\{0, -y + z/\overline{\pi}_1\}. \tag{8}$$

[15]The probability that both players cooperate in the baseline case is $[1 - F(\theta^*)]^2$. The probability that players cooperate following an oath is $[1 - G(\theta_1^o)\,][1 - F(\theta_2^o)\,] \geq [1 - F(\theta^*]^2$.

Proposition 4. *In the SPD, player i, for $i \in \{1, 2\}$, who moves second, behaves according to* (3)*. A type θ_i player who moves first chooses C whenever $\theta_i \geq \hat{\theta}_i^o$, as defined in* (7) *and* (8)*, and chooses D otherwise.*

Altogether, the variable-preference hypothesis makes three key predictions regarding behavior in the PD and in the SPD. First, both oath-takers and oath-recipients cooperate more when moving simultaneously (Proposition 3). Second, the behavior of second movers is the same for oath-recipients but different for oath-takers, since they are more likely to be conditional cooperators (Proposition 4). And third, first-move cooperation is expected to be larger by both oath-takers and oath-recipients (Proposition 4), although for different reasons: oath-recipients cooperate more because they expect a larger proportion of conditional cooperators moving second; and oath-takers cooperate more because they are more likely to be of a type that exceeds the first-move cooperation threshold.

4 Experiment: Design and Procedures

The basic decision situation is a PD game where players choose between cooperation and defection. The experiment is cast along two main dimensions. In the first dimension, we consider no-oath and oath groups. In the second dimension, we consider simultaneous-move and sequential-move games. In oath groups, one player is the potential oath-taker and her match is the oath-recipient. In oath treatments with sequential play, we consider two sub-treatments in which the potential oath-taker moves first or second. For an overview of all control and treatment groups, see Table 4.

The experiment was administered on Amazon Mechanical Turk (MTurk), which is an online platform that connects employers with workers who are asked to perform simple tasks privately and anonymously at any location.[16] To ensure high-quality data collection,

[16]MTurk has become increasingly popular for running experiments in behavioral research. See Dimant et al. (2020) and Jacquemet et al. (2021) for two recent applications. Many studies point to the robustness, generalizability, and reproducibility of laboratory findings in online environments (Horton et al.; 2011; Arechar et al.; 2018; **?**; Snowberg and Yariv; 2018). Horton et al. (2011), in particular, compare behavior in a PD played online and in the laboratory, and find that MTurk reproduces the levels of cooperation found in the physical laboratory.

we applied the following restrictions to the participant pool: approval rate by employers greater than 97%, more than 10,000 tasks completed, and could participate only once.[17] We used Qualtrics surveys to document decision-making that participants completed individually, and emulated interactions through post-hoc matching.[18]

In all treatments, participants were shown the instructions of the PD game that they were to play against the player they had been matched with. Each participant had to choose between options A and B. If both choose A, each gets \$5.5, and if both choose B, each gets \$3. If one player chooses A and the other chooses B, the one choosing A gets \$0.5 and the one choosing B gets \$8. Table 3 depicts the game in the experiment. Option A is the cooperative choice and option B is defection. In all treatments, we used letters A and B so that participants do not associate labels with particular meanings. In the text, we revert to using letters C and D.

Table 3: PD in experiment

$1\backslash 2$	A	B
A	$5.5\backslash 5.5$	$0.5\backslash 8$
B	$8\backslash 0.5$	$3\backslash 3$

The instructions block was followed by a decision task where players selected their preferred choice. In sequential treatments we adopted the strategy method, whereby second movers chose one of four contingent strategies: CC, CD, DC, and DD. The first entry corresponds to their preferred choice against cooperation by the first mover and the second entry corresponds to their preferred choice against defection by the first mover. Based on their preferred strategy, we classify individuals as unconditional cooperators (CC), conditional defectors (CD), and unconditional defectors (DD). DC is less intuitive as it prescribes defection in response to cooperation (which makes sense from a payoff-maximizing perspective), but cooperation against defection (which would make sense if the individual were altruistic). For symmetry, we call them conditional defectors.

Decision tasks were followed by a series of demographic questions. At the end, and in

[17] See Robinson et al. (2019) for best practices using MTurk for experimental research.

[18] Distinct surveys were posted into separate batches with a specified number of assignments for unique pairings.

each treatment, participants were matched in order of arrival of completed surveys, their decisions were paired, and they were informed of the decision of their match and their payment.

In oath treatments, potential oath-takers were asked to agree to take a solemn oath prior to reading the instructions of the game. We examined the role of the oath statement: "I swear upon my honor to help others at all times," which prescribes behavior that has a connection with the task and is representative of the type of oath we explored in the theoretical model. To compare our results to those of other experiments, we also examined the role of the truth-telling statement, which read: "I swear upon my honor to tell the truth and always provide honest answers." When presenting our results, we name the primary oath statement as the "cooperation oath" and that of the truth-telling statement as the "honesty oath."

After deciding whether to take the oath or not, potential oath-takers were told that (1) they had been randomly selected to take the oath, (2) their match had not been given the option of taking the oath, (3) their match would be shown the oath statement that they were asked to sign, and (4) their match would know whether they signed the oath or not. Oath-recipients were shown the oath statement that their match was asked to sign and they were told that (1) their match had been randomly selected to take the oath, (2) their match signed the oath or not, and (3) their match knows that they had not been given the option of taking the oath.[19] Participants then proceeded to the instructions block and all subsequent blocks, which were the same as those in no-oath treatments. The experimental instructions can be found in the online appendix.

5 Results

In this section, we present our results. First, we provide an overview of our sample, including adjustments to the pool of subjects and summary statistics (Section 5.1). Then, we present the results of simultaneous-move treatments, which include the effect of the cooperation and the honesty statements on oath-takers and oath-recipients (Section 5.2).

[19] In oath treatments, we recruited potential oath-takers first and, based on the number of subjects who accepted and refused the oath, we created two separate batches for oath-recipients with the associated number of assignments for unique pairings, where each batch contained a distinct survey indicating whether one's match had accepted or refused the oath.

Finally, we examine the impact of the cooperation oath in sequential treatments on first-move decisions (Section 5.3) and on second-move decisions (Section 5.4).

5.1 Overview of Sample

We recruited a total of 858 MTurk workers in June and July 2019. The average age was 35 and the majority of participants were male (62%), white (62%), attended higher education (88%), were never married (52%), and were US citizens (78%).[20] These characteristics were similar across groups. The experiment lasted about 5 minutes and participants earned on average $5.8 (including a $1.5 show-up fee).[21]

Although signing the oath was not mandatory, we found an acceptance rate of 92.3% across all oath treatments.[22] Because the decision to take the oath is communicated to the other participant, two distinct surveys were administered to oath-recipients. Given the large uptake of the oath, we do not have enough data on oath-recipients whose match refused the oath to make a statistical assessment of their behavior. For that reason, we dropped 23 observations from the analysis corresponding to oath-recipients who had been matched with non-takers. We nonetheless consider all potential oath-takers since that allows us to make the distinction between the effect of treatment assignment (being given the option to take the oath) and the effect of treatment status (accepting the oath). Overall, our statistical analysis considers a total of 835 participants. See Table 4 for the distribution of subjects across groups.[23,24]

[20] For details on descriptive statistics, see the online appendix.

[21] Average hourly wages on MTurk (ignoring idle time) is $6.19/h and the average requester pays $11/h (Hara et al.; 2018).

[22] This is line with previous experiments involving an oath procedure. See Jacquemet et al. (2013, 2017, 2018).

[23] There were 19 unmatched participants. In some cases, MTurk considered batches of assignments complete but some workers failed to successfully submit their task. We dropped the associated observations, excluded those workers from accepting any other assignment, and paid unmatched participants the maximum amount they would have gotten given their choice.

[24] The simultaneous-move cooperation-oath treatment contains relatively more observations because in the first batch of oath-recipients we uploaded a version of the survey that did not elicit data on three control variables: (1) time spent at instructions and decision-task blocks, (2) number of children, and (3) whether opponent having taken the oath affected their decision (self-reported at the end of the survey). We therefore posted a second batch for oath-recipients and oath-takers. We maintained responses from the first batch since

Table 4: Treatments and players

Description	**Timing**	**# Observations**			
		Base	POT	OT	OR
No oath	Simultaneous	79	—	—	—
Cooperation oath	Simultaneous	—	100	90	107
Honesty oath	Simultaneous	—	70	70	71
No oath	Sequential	2×70	—	—	—
Cooperation oath	Sequential / OT first	—	70	64	64
Cooperation oath	Sequential / OR first	—	71	63	63

Note: POT refers to potential oath-takers, OT to oath-takers, and OR to oath-recipients whose match accepted the oath. The no-oath sequential group considers 70 first movers and 70 second movers.

5.2 Simultaneous

Table 5 shows the results of simultaneous treatments. Mean cooperation of subjects in the control (no-oath) group is 50.63%. We compare behavior in the control group against behavior in four distinct treatment groups: oath-takers and oath-recipients in cooperation- and honesty-oath treatments.

Let us start with the cooperation oath. The cooperation rate of all potential oath-takers is 59%. The difference with respect to control is not statistically significant (one-sided proportions test, $z = -1.1180$, $p = 0.1318$). Among potential oath-takers, 90% took the cooperation oath. If we consider cooperation among those who took the oath (effect on treated), mean cooperation is 62.22% and the difference with respect to control is statistically significant at the 10% level ($z = -1.5175$, $p = 0.06$). As for oath-recipients, mean cooperation is 50.47%. The difference with respect to control is not statistically significant ($z = 0.0223$, $p = 0.5089$).

In honesty-oath treatments, all potential oath-takers took the oath. Mean cooperation among oath-takers is 58.57% and not statistically different from mean cooperation in the control group ($z = -0.9710$, $p = 0.1658$). As for oath-recipients, 57.75% cooperated, but the difference with respect to control is again not significant ($z = 0.8728$, $p = 0.1914$).

The results indicate that the cooperation oath produces a small and positive effect upon oath-takers (roughly 12pp increase in cooperation). Oath-recipients, on the other hand,

the only information missing is on the three control variables and not on their actual decisions.

Table 5: Treatment effects: Simultaneous treatments

		Cooperation					
		Mean	SD	n	Diff.	SE	p-value
No Oath		0.5063	[0.5032]	79	—	—	—
Coop. Oath	POT	0.5900	[0.4943]	100	-0.0837	(0.0747)	0.1318
	OT	0.6222	[0.4875]	90	-0.1159*	(0.0760)	0.0646
	OR	0.5047	[0.5023]	107	0.0017	(0.0742)	0.5089
Honesty Oath	OT	0.5857	[0.4962]	70	-0.0794	(0.0814)	0.1658
	OR	0.5775	[0.4975]	70	-0.0711	(0.0812)	0.1914

Note: POT refers to potential oath-takers, OT to oath-takers, and OR to oath-recipients whose match accepted the oath. Standard deviations of mean cooperation in square parentheses and standard errors of the difference in means in parentheses. p-values for one-sided proportions test (outcome in treatment larger than in control). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

were unaffected by the fact that their opponent took the oath. We find a direct effect of the oath upon those who take it, but no indirect effect upon those who receive it. The results also indicate that there is no impact of the honesty oath on cooperation by oath-takers nor by oath-recipients.

In the following sections, we return to focusing on the cooperation oath alone.

5.3 Sequential First Mover

Table 6 reports the results of first movers in no-oath and oath groups. Mean cooperation of first movers without the oath is 45.71%.[25] The cooperation rate of potential oath-takers is 61.43% — a significant increase of 15.71pp relative to control (one-sided proportions test, $z = -1.8641$, $p = 0.0312$). If we consider the subsample of those who took the oath, cooperation increases significantly by 16.79pp ($z = -1.9466$, $p = 0.0258$). As for oath-recipients who move first and whose opponent took the oath, their cooperation rate increases to 50.40% — a 5.08pp increase relative to control. This difference is not statistically significant ($z = -0.5854$, $p = 0.2791$).

[25] A smaller cooperation rate than observed in the simultaneous-move control group (50.63%). This difference is not significant; two-sided proportions test, $z = 0.5996$, $p = 0.5487$.

Table 6: Treatment effects: Sequential treatments (first movers)

		Cooperation					
		Mean	SD	n	Diff.	SE	p-value
No Oath		0.4571	[0.5018]	70	—	—	—
	POT	0.6143	[0.4903]	70	-0.1571**	(0.0832)	0.0312
Coop. Oath	OT	0.6250	[0.4880]	64	-0.1679**	(0.0849)	0.0258
	OR	0.5079	[0.5040]	63	-0.0508	(0.0867)	0.2791

Note: POT refers to potential oath-takers, OT to oath-takers, and OR to oath-recipients whose match accepted the oath. Standard deviations of mean cooperation in square parentheses and standard errors of the difference in means in parentheses. p-values for one-sided proportions test (outcome in treatment larger than in control). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

The first-move results reinforce the pattern observed in simultaneous treatments: following the oath, oath-takers cooperate more (diffect effect), but oath-recipients do not (no indirect effect).

5.4 Sequential Second Mover

We now consider the results of second movers in oath and no-oath groups. When players move second, they are asked to report one of four contingent strategies: CC, CD, DC, and DD. Table 7 shows the proportion of subjects choosing any of the four strategies.

In the no-oath group, 41.43% subjects chose CD (conditional cooperators), 38.57% chose DD (unconditional defectors), 11.43% chose CC (unconditional cooperators), and 8.6% chose DC (conditional defectors). This is the distribution of strategies against which we compare the strategies adopted in oath groups.

Among potential oath-takers, the proportion of CDs significantly increases by 24.77pp (two-sided proportions test, $z = -2.95$ $p = 0.0032$) and the proportion of DDs significantly decreases by 23.08pp ($z = 3.0881$, $p = 0.0020$). Although the proportion of CCs increases (by 4.06pp) and that of DCs decreases (by 5.75pp), these changes are not statistically significant. The direction and significance of the preceding results is maintained if we consider the subsample of those who took the oath (effect on treated). Figure 1a illustrates the distribution of strategies adopted by oath-takers against that of the no-oath control group. The null hypothesis that the two distributions are equal is rejected (two-

sample Kolmogorov-Smirnov (KS) test; $p = 0.005$).

Among oath-recipients who move second and whose opponent took the oath, the distribution of strategies of oath-recipients follows a similar pattern to that of oath-takers: the number of CCs and CDs increases, and the number of DCs and DDs decreases. None of these changes is significant, save for the 7.01pp decrease in the number of DCs.[26] Figure 1b illustrates the distribution of oath-recipients as second movers against the no-oath control group. The null that the distribution of strategies in control and oath-recipient groups is equal cannot be rejected (KS test; $p = 0.151$).

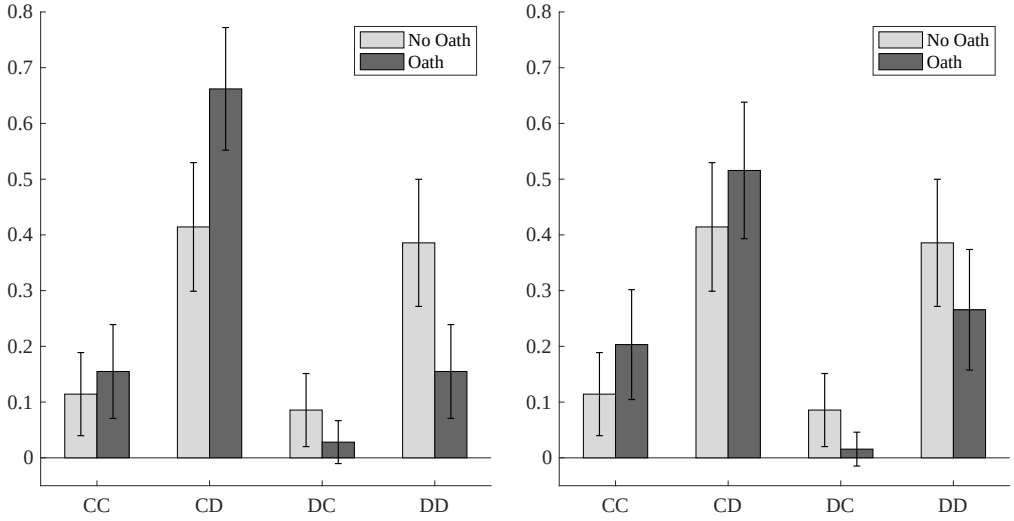
Table 7: Treatment effects: Sequential treatments (second movers)

			Strategy					
			Mean	SD	n	Diff.	SE	p-value
No Oath		CC	0.1143	[0.3205]	8	—	—	—
		CD	0.4143	[0.4962]	29	—	—	—
		DC	0.0857	[0.2820]	6	—	—	—
		DD	0.3857	[0.4903]	27	—	—	—
Coop. Oath	POT	CC	0.1549	[0.3644]	11	-0.0406	(0.0574)	0.4798
		CD	0.6620	[0.4764]	47	-0.2477***	(0.0814)	0.0032
		DC	0.0282	[0.1666]	2	0.0575	(0.0388)	0.1397
		DD	0.1549	[0.3644]	11	0.2308***	(0.0723)	0.0020
	OT	CC	0.1587	[0.3684]	10	-0.0444	(0.0597)	0.4544
		CD	0.6508	[0.4805]	41	-0.2365***	(0.0841)	0.0064
		DC	0.0317	[0.1767]	2	0.0540	(0.0401)	0.1912
		DD	0.1587	[0.3684]	10	0.2270***	(0.0742)	0.0035
	OR	CC	0.2000	[0.4029]	13	-0.0888	(0.0630)	0.1577
		CD	0.5429	[0.5018]	33	-0.1013	(0.0858)	0.2399
		DC	0.0143	[0.1195]	1	0.0701*	(0.0369)	0.0686
		DD	0.2429	[0.4319]	17	0.1201	(0.0802)	0.1392

Note: POT refers to potential oath-takers, OT to oath-takers, and OR to oath-recipients whose match accepted the oath. Standard deviations of mean cooperation in square parentheses and standard errors of the difference in means in parentheses. p-values for two-sided proportions test. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

[26] Along with the decrease in DCs among oath-takers, this might suggest that individuals make more conscientious decisions in oath treatments.

Figure 1: Distribution of types in experiment



(a) Oath-takers

(b) Oath-recipients

Note: Distribution of types of second movers in no-oath (control) group (light grey in panels A and B) against the distribution of oath-takers (dark-grey in A) and against the distribution of oath-recipients (dark-grey in B). Each staple indicates the fraction of second movers that chose the corresponding strategy. Error bars indicate 95% confidence intervals.

6 Discussion

The discussion of our results is organized as follows. First, we contrast our findings to other experimental results in the literature and discuss the empirical support for the coordination and variable-preference hypotheses (Section 6.1). Second, we calibrate the theoretical model introduced in Section 3 with our experimental results (Section 6.2) and discuss the implications of model assumptions for deriving the empirical distribution of types (Section 6.3). Finally, we discuss how alternative behavioral theories and models of social preferences relate to our results (Section 6.4).

6.1 Main Findings

Our results indicate that the cooperation oath has a direct effect upon those who take it, but not upon those who receive it. We find that oath-takers are more likely to cooperate in simultaneous (12pp increase) and sequential interactions, both when moving first (17pp) and second (24pp). We find no effect of the honesty oath in simultaneous play.

There are three main experimental regularities observed in social dilemmas. First, although free riding is a dominant strategy, a substantial share of individuals cooperates in one-shot social dilemmas, but free riding is also frequently observed (e.g., Dawes; 1980; Dawes and Thaler; 1988). Second, a large proportion of individuals are conditional cooperators (e.g., Fischbacher et al.; 2001; Kocher et al.; 2008; Chaudhuri; 2011). And third, prosocial manipulations of the game cause substantial increases in cooperation (e.g., Dawes; 1980; Sally; 1995; Alekseev et al.; 2017).

Our results are in line with previous studies. In the baseline case, we find that roughly 50% of subjects cooperate in the PD and in the SPD when moving first, and that 40% of individuals report a preference for conditional cooperation in the SPD when moving second. Following the cooperation oath, our results indicate a significant increase in cooperation by oath-takers in different roles across variants of the PD.

Our experiment was designed to evaluate (1) whether the oath has an impact on cooperation in a PD (which we find support for), and (2) whether that impact, if observed, could be attributed to a change of preferences.

The significant change in the distribution of strategies adopted by oath-takers relative to the baseline case gives support to the variable-preference hypothesis. If preferences had remained the same, second movers would have maintained their ordering of all four outcomes in the SPD. Instead, we find a significant decrease in unconditional defection (23pp) and a significant increase in conditional cooperation (24pp).[27]

Note that statistical support for the variable-preference hypothesis does not rule out the competing coordination hypothesis. The coordination hypothesis establishes that the oath acts as a selection device in the presence of multiple equilibria. Given the presence

[27]The significant increase in cooperation by oath-takers moving first is also consistent with a change of preference, since the competing coordination hypothesis predicts that no change in behavior should be observed in sequential play.

of conditional cooperators in the baseline group ($> 40\%$), this hypothesis predicts that cooperation increases in the PD at the intensive margin among conditional cooperators. It is conceivable to have a simultaneous increase in the *number* of conditional cooperators (i.e., extensive margin) and in the *choice* of cooperation among pre-existing conditional cooperators (i.e., intensive margin).

However, our results indicate that the coordination hypothesis has limited scope for explaining the behavior of pre-existing conditional cooperators. In the baseline case, by pooling the share of unconditional (11.4%) and conditional cooperators (41.4%), we find that the proportion of cooperative types amounts to 52.9% of our sample, which is roughly equal to the cooperation rate observed in the PD (50.6%). These figures suggest that there is no room for cooperation to increase at the intensive margin, since all conditional cooperators cooperate without the oath.[28]

As a robustness check, we tested the impact of the honesty oath to understand whether the oath environment alone, irrespective of the content of the oath script, was enough to promote cooperation in a PD. We find that the honesty oath had no impact on behavior of neither oath-takers nor oath-recipients. This result can be contrasted to that of Hergueux et al. (2022), who find that the honesty oath had a positive impact on cooperation in a public goods game. Despite the difference in results, our findings are not strictly incompatible. The authors argue that the honesty oath leads individuals to behave according to their underlying preference for conditional cooperation. Our results, on the other hand, suggest that there is no room for conditional cooperators to adhere to cooperation beyond baseline compliance.

[28]This argument assumes that types are independent of the role assigned to players in the PD and in the SPD. Findings from previous experiments indicate that behavior across roles of the same game is consistent with stable social preferences. Blanco et al. (2011) and Blanco et al. (2014) show a strong correlation between first and second-move choices in the SPD using a within-subject design. Altmann et al. (2008) and Gächter et al. (2012) have a similar result for the trust game and for the sequential voluntary contribution game, respectively. Blanco et al. (2014), in particular, show how the correlation of first- and second-move decisions can be based on a non-belief, preference-based motive. Krupka and Weber (2013) find that behavior is constant across different versions of the same dictator game and conclude that stable social preferences are consistent with observed choices.

6.2 Evaluation of the Results

Here, we evaluate the experimental results considering the theoretical framework introduced in Section 3.

In the baseline case, the choices of second movers reveal that 41% of individuals are conditional cooperators (CD), 39% are unconditional defectors (DD), 11% are unconditional cooperators (CC), 9% are conditional defectors (DC). In the model, unconditional cooperation and conditional defection were precluded due to their lack of frequency in SPD experiments. Consistent with our results, we observe that 80% of observations refer to conditional cooperation and unconditional defection. The data are well accounted by the θ-model, which assigns an additional value to the cooperative outcome. For the subsequent evaluation, we drop observations CC and DC and evaluate the remaining data under the assumption that the θ-model holds true. The model parameters were calibrated using the monetary payoffs in the experiment after normalizing the "sucker" payoff to zero (by subtracting 0.5 from all payoff entries), so that $x = 7.5$, $y = 5$, and $z = 2.5$.

Table 8 contains the data used in the following analysis. In the baseline case, the proportion of CD gives an estimate for $\overline{\pi} = 1 - F(x - y)$. The proportion of C choices of first movers gives an estimate for $1 - F(\hat{\theta})$. The proportion of C choices in the PD gives an estimate for $1 - F(\theta^*)$. From equation (4) find $\hat{\theta} = -0.17$ (95% CI: $[-1.15, 1.46]$), which implies $\hat{\theta} = 0$, and from equation (2) find $\theta^* = 4.94$ (95% CI: $[4.05, 6.31]$).

The above calibration has two main implications. First, since the first-move cooperation threshold $\hat{\theta}$ is zero and the proportion of first-move defection (i.e., $F(\hat{\theta}) = 0.54$) is weakly greater than the proportion of unconditional defectors (i.e., 0.48), it follows that the probability mass of unconditional defectors lies at a θ-value of zero. And second, since the proportion of conditional cooperators is roughly equal to the cooperation rate in the PD ($\simeq 0.51$), the probability mass of conditional cooperators lies above the simultaneous-move cooperation threshold θ^*. Therefore, the distribution of θ is such that $1/2$ of θs are zero and the remaining $1/2$ is larger than 5. The density function should be virtually zero between these values (see Figure 2). In the absence of an oath, there are two types of individuals, one with $\theta = 0$ and the other with large θ values.[29]

[29] In a similar setting, Bolle and Ockenfels (1990) also find that only a distribution which allows a large mass near zero and a large mass on an extreme value of θ passes a statistical test.

Table 8: Distribution of choices

		Base	OT	OR
PD	C	0.5063	0.6222	0.5047
		(0.056)	(0.051)	(0.048)
1 in SPD	C	0.4571	0.6250	0.5079
		(0.060)	(0.061)	(0.063)
2 in SPD	CD	0.5179	0.8039	0.6600
		(0.067)	(0.056)	(0.067)

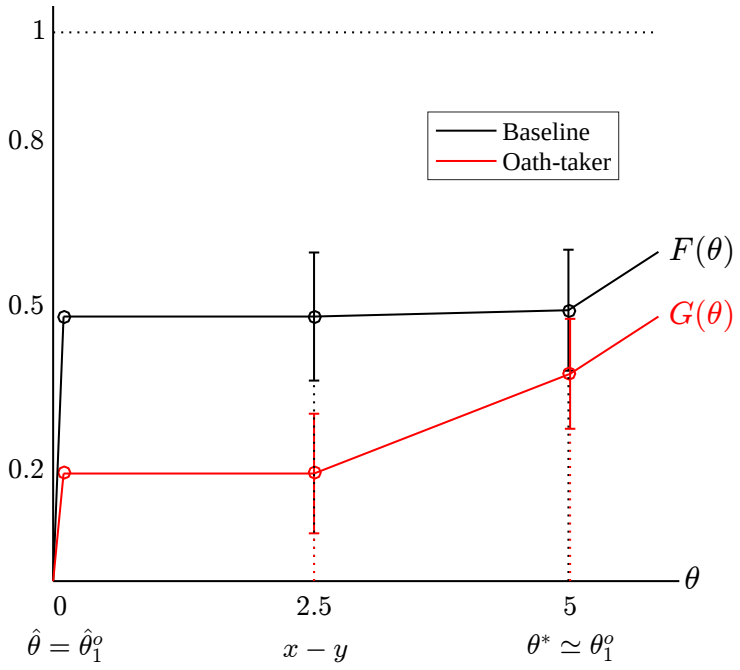
Note: Standard errors in parentheses.

We now turn to the case of the oath. Since the choices of oath-recipients moving second are not significantly different from those observed in the baseline case, types of oath-recipients are drawn from the same distribution, F. In contrast, the choices of oath-takers moving second are significantly different, which implies that types of oath-takers are drawn from a different distribution, G. Below, we focus on deriving G.

Again, Table 8 summarizes the data used in the following analysis. The proportion of CD gives an estimate for $\overline{\pi}_1 = 1 - G(x - y)$ and $\overline{\pi}_2 = 1 - F(x - y)$ for oath-takers and oath-recipients; the proportion of C choices by first movers gives an estimate for $1 - G(\hat{\theta}_1^o)$ and $1 - F(\hat{\theta}_2^o)$; and the proportion of C choices in the PD gives an estimate for $1 - G(\theta_1^o)$ and $1 - F(\theta_2^o)$. From equation (7) find $\hat{\theta}_1^o = -1.21$ (95% CI: $[-1.81, -0.27]$), which implies $\hat{\theta}_1^o = 0$, and from equation (5) find $\theta_1^o = 4.95$ (95% CI: $[4.17, 6.1]$).

Similar to the baseline case, the fact that the first-move cooperation threshold $\hat{\theta}_1^o$ is zero and the proportion of first-move defection (i.e., $G(\hat{\theta}_1^o) = 0.38$) is greater than the proportion of unconditional defectors (i.e., 0.20) implies that the mass of unconditional defectors lies at zero. However, since the 0.62 cooperation rate by oath-takers in the PD is smaller than the 0.8 proportion of conditional cooperators, it follows that roughly 18% of conditional cooperators defected when under oath. As a result, θ values of oath-takers follow a three-point distribution, such that $1/5$ of θs are zero; $1/5$ lie between 2.5 and 5; and $3/5$ are larger than 5. See Figure 2.

Figure 2: Distribution function of types



Note: Error bars indicate 95% confidence intervals.

6.3 Model Assumptions

In the evaluation of our results, we make two key assumptions that we discuss below. First, the distribution of types is common knowledge. And second, oath-taking is independent of types.

The distribution of types is common knowledge. A strong assumption in our model is that individuals form rational expectations about F and G. Although we cannot claim that F is common knowledge, our results suggest that oath-recipients did not form an accurate prediction of the distribution of oath-takers, G. Relative to the baseline case, neither the distribution of types of oath-recipients nor their behavior in the PD and as first movers in the SPD, changed following the oath, which could indicate that oath-recipients did not expect that oath-takers were populated by relatively more cooperative types (i.e., the oath was not credible).[30]

[30]The small increase in cooperation by oath-takers in the PD (12pp increase) suggests that oath-takers

The assumption of common knowledge is especially important for determining the behavior of first movers. Provided that F is known, individuals should be more willing to cooperate when they move first than when they move simultaneously, since, whereas first movers can be assured that conditional cooperators will cooperate in response to cooperation, that a conditional cooperator chooses to cooperate in the PD is not guaranteed. Instead, we find that cooperation rates in the PD and of first movers in the SPD are roughly the same, which suggests that first movers believe that there are more unconditional defectors than there are in reality.

The estimation of first-move cooperation threshold $\hat{\theta}$ (in the baseline) and $\hat{\theta}_1^o$ (of oath-takers) also relies heavily on the common knowledge assumption. Based on the distribution of reported strategies by second movers and on the choice of cooperation by first movers, our estimates indicate that both thresholds should be negative, and thus zero by assumption. Therefore, all types of first movers should have found it optimal to cooperate, including unconditional defectors. To the contrary, we observe that 37.5% and 54.3% of first movers defect with and without the oath, respectively. The observed defection rates could nevertheless be justified in two ways. First, the estimation of $\hat{\theta}$ and $\hat{\theta}_1^o$ is highly sensitive to the calibration of the z and y parameters. If we assume that the distribution of θ is independent of small changes to monetary payoffs, a value of $z > 3.125$ or $y < 4$ that preserves the PD structure of the game would result in a positive threshold for both the no-oath and oath cases. And second, if we relax the assumption of a lower bound of zero imposed on the value of θ, then unconditional defectors can exhibit not only individualistic ($\theta = 0$), but spiteful preferences ($\theta < 0$), whereby spiteful types enjoy increased well-being when others are worse off (Levine; 1998). Both explanations result in a mass of unconditional defectors below an admissible type-threshold, who would strictly prefer defection when moving first.

understood that the oath was noncredible. At the extensive margin, we found that the number of conditional cooperators increased following the oath. However, the rate of increase in the number of conditional cooperators (55%) was accompanied by a lower rate of increase in cooperation in the PD (23%). This suggests that, at the intensive margin, the newly converted conditional cooperators were not willing to chance cooperation. In the model, this is justified by the joint hypothesis that (1) oath-takers do not believe that the probability that oath-recipients cooperate has increased (i.e., the oath is noncredible) and that (2) the magnitude of θ is low among oath-takers (i.e., the oath provides a low cooperative thrust among new conditional cooperators). If either of the two were rejected, we would have observed a higher increase in cooperation by oath-takers.

We emphasize that our experiment only allows us to reject the joint hypothesis of the θ-value structure and the decision structure of a game with incomplete information. Our results point to the acceptance of an extra value for the cooperative outcome but nevertheless suggest that individuals do not form an accurate prediction of the distribution of types. The assumption of common knowledge is critical for the estimation of the two-point distribution of θ in the absence of an oath and the three-point distribution with the oath. Definitive support for the empirical distributions of θ would require the weakening of the common-knowledge assumption in an experimental setting (e.g., by providing individuals with the actual frequency of types before gameplay).

Oath-taking is independent of types. In the model, we assumed that oath-taking is independent of types. From a standard game-theoretical perspective, this is a reasonable assumption since individuals are asked to swear under oath prior to learning about the game.[31] Given that types reflect preferences over outcomes (which are unknown at the oath-taking stage), it would be forceful to extend sequential rationality to the oath-taking decision and allow for the signaling of types. However, because 10% of potential oath-takers across treatments did not take the oath, there could be type-dependent separation of the oath-taking decision in the data.[32]

If we assumed that the oath stage is embedded in the PD game and that the oath could signal types, there would be two possible explanations for the partial pooling observed in the data.[33] First, in line with the coordination hypothesis, the oath is a costless signal and partial pooling is the product of a "babbling" equilibrium, in which oath-taking is random and the oath is uninformative (Farrell and Rabin; 1996).[34] And second, in line with the

[31] At the oath-taking stage, potential oath-takers are only given information provided on the consent form. Specifically, that they and another participant will make a decision for a bonus payment and that their payment depends on the decision they make, as well as on the decision of a person they are paired with. At the oath-taking stage, potential oath-takers are unaware (1) that their partner will learn of whether they took the oath or not (this information is only provided afterward) and (2) of the specific task that they will be asked to perform. See online appendix.

[32] We reject the hypothesis that everyone takes the oath; the proportion of uptake of cooperation oath is 0.9 (95% CI: $[0.855, 0.935]$).

[33] Since types are assumed continuous and the message space is discrete, there can only exist pooling or partial pooling equilibria. No fully separating equilibrium exists.

[34] Babbling is the only possible equilibrium. The existence of an informative partial-pooling equilibrium with costless signaling rests on a monotonicity condition, which requires that preferences of sender and

variable-preference hypothesis, defection under oath is costly and the magnitude of the (psychological) cost could dictate that low-θ types choose to separate by not taking the oath.[35,36]

Although, on the one hand, our results suggest that oath-recipients do not update their posterior beliefs about the distribution of types of oath-takers (see discussion of common knowledge assumption), which is consistent with a babbling equilibrium, on the other hand, we find that the oath is costly, in the sense that oath-takers are of a more cooperative type (which could be formalized by a cost from defection instead of an extra value for cooperation). Our experiment does not allow us to conclude that potential oath-takers face a sufficiently large cost to make uncooperative types separate at the oath-taking stage.[37] Despite the 90% uptake of the oath (which suggests that all types of players accept to take the oath), conclusive support for the type-independent oath-taking assumption would require further examination.

6.4 Related Theories

Our evidence suggests that (1) cooperation in a PD is driven by the presence of conditional cooperators and that (2) the number of conditional cooperators significantly increases when individuals are under oath.

That individuals are conditional cooperators, who interpret the PD game as a coordination game in terms of utilities, can be justified by models of social preferences that

recipient be (somewhat) congruent, in the sense that both are positively correlated with respect to the state of the world whose uncertainty the sender's message mitigates (Crawford and Sobel; 1982). In the game, that refers to the type of the potential oath-taker. However, because types of oath-takers and oath-recipients are independent, the monotonicity condition is not met, and no partial-pooling equilibrium exists. In practice, any candidate for a partial-pooling equilibrium that entails higher cooperation rates under oath would involve profitable deviations by oath-non-takers who are unconditional defectors.

[35] Assuming that the cost is nonincreasing with types.

[36] The analogous argument for separation due to preferences is that of selection, in the sense that cooperative types are more likely to take the oath when offered to do so.

[37] If we compare the cooperation rate in the PD of those who accepted the oath (56/90) and of those who rejected the oath (3/10), this yields a Fisher exact test statistic of 0.0863, which is not significant at the 5% level. Nevertheless, even if the cooperation rate were significantly different, that would not necessarily indicate separation at the oath-taking stage, since cooperative types who refuse the oath and afterward are made aware that their match will know of their decision, might choose to defect instead.

incorporate normative principles, such as equity (e.g., Fehr and Schmidt; 1999; Bolton and Ockenfels; 2000) or reciprocity (e.g., Rabin; 1993; Falk and Fischbacher; 2006), or by models that are based on the desire for avoiding disapproval, due to a prosocial image of self (e.g., Bénabou and Tirole; 2006) or the avoidance of guilt from disappointing others (e.g., Battigalli and Dufwenberg; 2007). Many studies have tested the predictive power of alternative models of social preferences applied to a variety of social settings.[38] More recently, Miettinen et al. (2020) conduct a "horse race" between different models and find that both reciprocity and inequity-aversion motives perform well in explaining cooperation in an SPD. Our baseline data are consistent with these explanations.

More significant to our study are theories that relate to the mechanism underlying the effect of changes in context. Coordination-device theories assume that social preferences are stable and account for context effects through changes in expectations. According to these theories, conditional cooperators prefer cooperation to defection, but only if they believe that their partner will cooperate. Since we observe a belief-independent change in choices made by second movers, our data does not reconcile with explanations based on invariant preferences.

There are three classes of theories that assume that preferences are flexible and dependent on context which are consistent with our results. The first is prospect theory (Kahneman and Tversky; 1979; Tversky and Kahneman; 1981) and subsequent models of reference-dependent preferences (e.g., Tversky and Kahneman; 1991; Tversky and Simonson; 1993; Kőszegi and Rabin; 2006), which predict that choice depends on the subjective value placed on gains and losses relative to a reference point. Although formulated to explain individual-decision problems of choice under uncertainty, several authors have used prospect theory to interpret context effects in strategic environments.[39] A noteworthy example is Andreoni (1995), who suggest that positive frames emphasize the positive externality (i.e., gains) from contributions to a public good, which works to change players reference point and therefore behavior. Similarly, the oath may draw attention to the pos-

[38] See Charness and Rabin (2002) for an empirical comparison of existing theories. For a review of models of social preferences, see Fehr and Schmidt (2006).

[39] Examples include Andreoni (1995); van Dijk and Wilke (2000); Goeree and Offerman (2003); Goeree et al. (2003); Armantier and Treich (2009); and Iturbe-Ormaetxe et al. (2011). Iturbe-Ormaetxe et al. (2011), in particular, integrate prospect theory into a model of public-goods provision.

itive externality of cooperation in a PD and interfere with the subjective utility placed on cooperative outcomes.[40]

The second class of theories relates to group identity and corresponding models of social identity (e.g., Akerlof and Kranton; 2000, 2005) and team reasoning (e.g., Bacharach; 1999; Gold and Sugden; 2007).[41] These theories assume that individuals categorize themselves and others into groups and that positive contexts generate group identification. The perception of group identification causes individuals to attach a higher value to group, as opposed to individual, outcomes. The oath, by emphasizing cooperation, could elicit group membership and prompt individuals to shift their preference toward maximizing joint payoffs.

The final class of theories is that of social norms. Theories of social norms assume that individuals make rule-based decisions, and that different contexts are associated with the presence or strength of a norm that dictates what constitutes appropriate behavior.[42] Models of norm compliance include those developed by Andreoni and Bernheim (2009), Krupka and Weber (2013), and López-Pérez (2008), who assume that utility depends on both monetary outcomes and the degree to which certain actions comply with a social norm. In these models, changes in context operate to change the utility (or appropriateness) of specific actions.[43] As a result, oath-takers might acknowledge that the appropriate behavior is to comply with the provisions of the oath statement, which leads them to favor the cooperative outcome. This interpretation conforms with the null effect of the honesty oath and of

[40]Related theories include those developed by Bordalo et al. (2012, 2013), who examine the saliency of choice. Models of saliency consider the role of both involuntary and exogenous stimuli (e.g., nudges, frames), as well as voluntary and endogenous stimuli (e.g., the oath) in drawing attention to particular choices. For a review, see Bordalo et al. (2022).

[41]See also Brewer and Kramer (1986); Wit and Wilke (1992); and Sugden (1993). Group membership has been shown to affect cooperation (e.g., Eckel and Grossman; 2005; Goette et al.; 2006; Charness et al.; 2007) and coordination (e.g., Weber; 2006; Chen and Chen; 2011; Chen et al.; 2014). Field and laboratory experiments have shown how inducing social identity can shift time, risk, and other-regarding preferences (e.g., Chen and Li; 2009; Benjamin et al.; 2010).

[42]See for example Bicchieri (2005); Biel and Thøgersen (2007); and Bicchieri et al. (2014).

[43]These models are similar to social-identity models, in which utility depends on whether actions conform with one's identity. Chang et al. (2019) articulate how norm-based behavior agrees with the theoretical framework of social identity, considering that different social groups share a different set of normative prescriptions for behavior.

the cooperation oath on oath-recipients, who, by not having "constrained" their behavior to cooperation, could have experienced a weaker normative thrust to cooperate in the oath environment.

7 Conclusion

In a wide range of experimental settings, changes in context have been shown to produce a significant effect on behavior in social dilemmas. In this paper, we explored the implications of a solemn oath to cooperation applied to the simultaneous and sequential versions of the PD. The oath provides a contextual manipulation of the PD that exhorts individuals to abide to cooperative behavior. We find that the oath enhances cooperation by oath-takers, though not by oath-recipients. Despite only a moderate increase in cooperation in simultaneous play, we find that oath-takers are more likely to follow through, as well as to initiate, cooperation in sequential interactions. Our results conform with previous findings that subjects cooperate more when the context promotes adherence to pro-social behavior.

The main contribution of our work relates to the mechanism by which the oath induces a change in behavior. The observation that pro-social manipulations of context promote cooperative behavior does not imply that preferences are malleable. Jacquemet et al. (2020) and Hergueux et al. (2022) find support that an oath creates the intrinsic motivation necessary for individuals to behave according to their underlying social preferences. In other context-manipulation studies, many authors offer evidence that context influences behavior through changes in expectations of play. Relatively few studies present results indicative of a preference change in strategic settings, though often not providing direct proof. The novelty of our work is that we provide direct evidence that social preferences are oath, and therefore context, dependent. Among oath-takers, we find that a significant 23pp decrease in the proportion of unconditional defectors was followed by an increase of equal magnitude in the number of conditional cooperators.

It is important to note some of the limitations of our work. Although previous studies suggest that social preferences are stable across variants of the same game (e.g., Altmann et al.; 2008; Blanco et al.; 2011; Gächter et al.; 2012), it remains untested whether the set of preferences induced by the oath is constant across roles in the PD and in the SPD. In our analysis, we assumed that the oath is equally relevant in the simultaneous game, as

it is among first and second movers in the sequential game. However, the preference for conditional cooperation might differ depending on whether oath-takers know that that their partner cooperated (i.e., among second movers) versus if they believe that their partner will cooperate with some probability (i.e., simultaneous move). Therefore, an approach like that developed by Krupka and Weber (2013) to identify whether the strength of the conditional cooperation "norm" induced by the oath is invariant across roles in the PD and in the SPD would constitute valuable future research.

Another limitation is that the positive statement we make regarding the variable-preference hypothesis does not, in itself, rule out the coordination hypothesis. Our experiment was designed to test whether there is an appeal to a change of preference by eliminating the role of expectations in dictating behavior of second movers. Considering that preferences change toward conditional cooperation, expectations play a key role in how individuals behave in the simultaneous PD. It is therefore possible that both the variable-preference and coordination hypotheses hold concurrently. Further work disentangling the role of preferences and expectations on cooperation of both oath-takers and oath-recipients in simultaneous interactions would provide an important complement to our results. Especially since the rate of increase in conditional cooperators is not accompanied by the rate of increase in cooperation in the PD, which suggests meaningful belief-based effects of the oath.

It would also be important to investigate the robustness of our results. Although the overall size of our study is large, because of the variety of roles we assigned different players in the PD and in the SPD, it is desirable to increase the sample size of individual subtreatments. For instance, in the evaluation of choices made by oath-recipients, we find that there is not a significant change of preferences. However, the visual inspection of reported strategies by second movers in Figure 1 suggests that preferences of oath-recipients move in the same direction as that of oath-takers, though to a lesser degree. A higher sample size could detect if differences of a smaller magnitude are statistically important. Moreover, regarding the impact of the oath and the support for the preference channel, it is important to ascertain whether these effects are systematic (i.e., across strategic environments) and durable (i.e., across multiple interactions).

If our results are confirmed, the fact that a solemn oath changes behavior due a change of preferences is important from both a conceptual and practical perspective. A great deal of theoretical work is dedicated to the development of models of social preferences that are

consistent and general enough to accommodate multiple confounds in experimental games. Our results are nevertheless difficult to reconcile with models of cross-situational social preferences, since a large proportion of individuals might view one situation as relevant for social preferences (e.g., when under oath), though irrelevant in other situations. From a practical perspective, if a solemn oath activates cooperative dispositions irrespective of changes in expectations, the oath might constitute a stable and cost-effective instrument to galvanize individuals to cooperate in interactions where conflict is predicted to occur.

A Proof of Proposition 3

That (a) is true follows directly from the discussion in the main text. As for (b), equation (5) defines $\theta_1(\theta_2)$ with $\partial\theta_1/\partial\theta_2 = zf(\theta_2)/[1-F(\theta_2)]^2 > 0$. Substitute $\theta_1(\theta_2)$ into equation (6) and define

$$H(\theta_2) \equiv x - y + \frac{zG\big(\theta_1(\theta_2)\big)}{1 - G(\big(\theta_1(\theta_2)\big)} - \theta_2,$$

where $H(\cdot)$ is a continuous and differentiable function on $[0, \infty)$ and $(0, \infty)$, respectively. Let η be the set of solutions to $H(\theta_2) = 0$. By assumption, η is nonempty. Let $\theta_2^o \equiv \min \eta$, which Pareto dominates all other solutions. Since

$$H(0) = x - y + \frac{zG(x-y)}{1 - G(x-y)} > 0$$

and

$$H(\theta^*) = x - y + \frac{zG(\theta^*)}{1 - G(\theta^*)} - \theta^* \leq 0$$

because $G(\theta^*) \leq F(\theta^*)$, then by the intermediate value theorem $\theta_2^o \leq \theta^*$. In addition, because $\partial\theta_1/\partial\theta_2 > 0$, then $\theta_1^o = \theta_1(\theta_2^o)$ is such that $\theta_1^o \leq \theta^*$ $\square$

References

Ahn, T.-K., Lee, M., Ruttan, L. and Walker, J. (2007). Asymmetric payoffs in simultaneous and sequential prisoner's dilemma games, *Public Choice* **132**: 353–366.

Ahn, T.-K., Ostrom, E. and Walker, J. M. (2003). Heterogeneous preferences and collective action, *Public Choice* pp. 295–314.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity, *The Quarterly Journal of Economics* **115**(3): 715–753.

Akerlof, G. A. and Kranton, R. E. (2005). Identity and the economics of organizations, *Journal of Economic Perspectives* **19**(1): 9–32.

Alekseev, A., Charness, G. and Gneezy, U. (2017). Experimental methods: When and why contextual instructions are important, *Journal of Economic Behavior & Organization* **134**: 48–59.

Altmann, S., Dohmen, T. and Wibral, M. (2008). Do the reciprocal trust less?, *Economics Letters* **99**(3): 454–457.

Andreoni, J. (1995). Warm-glow versus cold-prickle: The effects of positive and negative framing on cooperation in experiments, *The Quarterly Journal of Economics* **110**(1): 1–21.

Andreoni, J. and Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects, *Econometrica* **77**(5): 1607–1636.

Arechar, A. A., Gächter, S. and Molleman, L. (2018). Conducting interactive experiments online, *Experimental Economics* **21**(1): 99–131.

Armantier, O. and Treich, N. (2009). Subjective probabilities in games: An application to the overbidding puzzle, *International Economic Review* **50**(4): 1079–1102.

Bacharach, M. (1999). Interactive team reasoning: A contribution to the theory of cooperation, *Research in Economics* **53**(2): 117–147.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in games, *American Economic Review* **97**(2): 170–176.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior, *American Economic Review* **96**(5): 1652–1678.

Benjamin, D. J., Choi, J. J. and Strickland, A. J. (2010). Social identity and preferences, *American Economic Review* **100**(4): 1913–1928.

Berg, J., Dickhaut, J. and McCabe, K. (1995). Trust, reciprocity, and social history, *Games and Economic Behavior* **10**(1): 122–142.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*, Cambridge University Press.

Bicchieri, C., Dimant, E. and Xiao, E. (2021). Deviant or wrong? the effects of norm information on the efficacy of punishment, *Journal of Economic Behavior & Organization* **188**: 209–235.

Bicchieri, C., Muldoon, R., Sontuoso, A. et al. (2014). Social norms, *The Stanford Encyclopedia of Philosophy* .

Biel, A. and Thøgersen, J. (2007). Activation of social norms in social dilemmas: A review of the evidence and reflections on the implications for environmental behaviour, *Journal of Economic Psychology* **28**(1): 93–112.

Blanco, M., Engelmann, D., Koch, A. K. and Normann, H.-T. (2014). Preferences and beliefs in a sequential social dilemma: A within-subjects analysis, *Games and Economic Behavior* **87**: 122–135.

Blanco, M., Engelmann, D. and Normann, H. T. (2011). A within-subject analysis of other-regarding preferences, *Games and Economic Behavior* **72**(2): 321–338.

Bolle, F. and Ockenfels, P. (1990). Prisoners' dilemma as a game with incomplete information, *Journal of Economic Psychology* **11**(1): 69–84.

Bolton, G. E. and Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition, *American Economic Review* **90**(1): 166–193.

Bordalo, P., Gennaioli, N. and Shleifer, A. (2012). Salience theory of choice under risk, *The Quarterly Journal of Economics* **127**(3): 1243–1285.

Bordalo, P., Gennaioli, N. and Shleifer, A. (2013). Salience and consumer choice, *Journal of Political Economy* **121**(5): 803–843.

Bordalo, P., Gennaioli, N. and Shleifer, A. (2022). Salience, *Annual Review of Economics* **14**: 521–544.

Brandts, J. and Charness, G. (2000). Hot vs. cold: Sequential responses and preference stability in experimental games, *Experimental Economics* **2**: 227–238.

Brandts, J. and Charness, G. (2011). The strategy versus the direct-response method: A first survey of experimental comparisons, *Experimental Economics* **14**(3): 375–398.

Brewer, M. B. and Kramer, R. M. (1986). Choice behavior in social dilemmas: Effects of social identity, group size, and decision framing., *Journal of Personality and Social Psychology* **50**(3): 543.

Carlsson, F., Kataria, M., Krupnick, A., Lampi, E., Löfgren, Å., Qin, P. and Sterner, T. (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.

Chang, D., Chen, R. and Krupka, E. (2019). Rhetoric matters: A social norms explanation for the anomaly of framing, *Games and Economic Behavior* **116**: 158–178.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership, *Econometrica* **74**(6): 1579–1601.

Charness, G. and Dufwenberg, M. (2010). Bare promises: An experiment, *Economics Letters* **107**(2): 281–283.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests, *The Quarterly Journal of Economics* **117**(3): 817–869.

Charness, G., Rigotti, L. and Rustichini, A. (2007). Individual behavior and group membership, *American Economic Review* **97**(4): 1340–1352.

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

Chen, R. and Chen, Y. (2011). The potential of social identity for equilibrium selection, *American Economic Review* **101**(6): 2562–2589.

Chen, Y. and Li, S. X. (2009). Group identity and social preferences, *American Economic Review* **99**(1): 431–57.

Chen, Y., Li, S. X., Liu, T. X. and Shih, M. (2014). Which hat to wear? Impact of natural identities on coordination and cooperation, *Games and Economic Behavior* **84**: 58–86.

Clark, K. and Sefton, M. (2001). The sequential prisoner's dilemma: Evidence on reciprocation, *The Economic Journal* **111**(468): 51–68.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission, *Econometrica* pp. 1431–1451.

Croson, R. and Marks, M. (2001). The effect of recommended contributions in the voluntary provision of public goods, *Economic Inquiry* **39**(2): 238–249.

Cubitt, R. P., Drouvelis, M. and Gächter, S. (2011). Framing and free riding: emotional responses and punishment in social dilemma games, *Experimental Economics* **14**: 254–272.

Dal Bó, E. and Dal Bó, P. (2014). "do the right thing:" The effects of moral suasion on cooperation, *Journal of Public Economics* **117**: 28–38.

Dawes, R. M. (1980). Social dilemmas, *Annual Review of Psychology* **31**(1): 169–193.

Dawes, R. M. and Thaler, R. H. (1988). Anomalies: Cooperation, *The Journal of Economic Perspectives* **2**(3): 187–197.

Di Bartolomeo, G., Dufwenberg, M., Papa, S. and Passarelli, F. (2019). Promises, expectations & causation, *Games and Economic Behavior* **113**: 137–146.

Dimant, E., Van Kleef, G. A. and Shalvi, S. (2020). Requiem for a nudge: Framing effects in nudging honesty, *Journal of Economic Behavior & Organization* **172**: 247–266.

Dreber, A., Ellingsen, T., Johannesson, M. and Rand, D. G. (2013). Do people care about social context? framing effects in dictator games, *Experimental Economics* **16**: 349–371.

Dufwenberg, M., Gächter, S. and Hennig-Schmidt, H. (2011). The framing of games and the psychology of play, *Games and Economic Behavior* **73**(2): 459–478.

Eckel, C. C. and Grossman, P. J. (2005). Managing diversity by creating team identity, *Journal of Economic Behavior & Organization* **58**(3): 371–392.

Ederer, F. and Schneider, F. (2022). Trust and promises over time, *American Economic Journal: Microeconomics* **14**(3): 304–20.

Ederer, F. and Stremitzer, A. (2017). Promises and expectations, *Games and Economic Behavior* **106**: 161–178.

Ellingsen, T. and Johannesson, M. (2004). Promises, threats and fairness, *The Economic Journal* **114**(495): 397–420.

Ellingsen, T., Johannesson, M., Mollerstrom, J. and Munkhammar, S. (2012). Social framing effects: Preferences or beliefs?, *Games and Economic Behavior* **76**(1): 117–130.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity, *Games and Economic Behavior* **54**(2): 293–315.

Farrell, J. and Rabin, M. (1996). Cheap talk, *Journal of Economic Perspectives* **10**(3): 103–118.

Fehr, E., Kirchsteiger, G. and Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation, *The Quarterly Journal of Economics* **108**(2): 437–459.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation, *The Quarterly Journal of Economics* **114**(3): 817–868.

Fehr, E. and Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism–experimental evidence and new theories, *Handbook of the Economics of Giving, Altruism and Reciprocity*, Vol. 1, Elsevier, pp. 615–691.

Fellner, G., Sausgruber, R. and Traxler, C. (2013). Testing enforcement strategies in the field: Threat, moral appeal and social information, *Journal of the European Economic Association* **11**(3): 634–660.

Filiz-Ozbay, E. and Ozbay, E. Y. (2014). Effect of an audience in public goods provision, *Experimental Economics* **17**: 200–214.

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

Fosgaard, T. R., Hansen, L. G. and Wengström, E. (2014). Understanding the nature of cooperation variability, *Journal of Public Economics* **120**: 134–143.

Fosgaard, T. R., Hansen, L. G. and Wengström, E. (2017). Framing and misperception in public good experiments, *The Scandinavian Journal of Economics* **119**(2): 435–456.

Fosgaard, T. R., Hansen, L. G. and Wengström, E. (2019). Cooperation, framing, and political attitudes, *Journal of Economic Behavior & Organization* **158**: 416–427.

Frackenpohl, G., Hillenbrand, A. and Kube, S. (2016). Leadership effectiveness and institutional frames, *Experimental Economics* **19**: 842–863.

Freedman, J. L. and Fraser, S. C. (1966). Compliance without pressure: The foot-in-the-door technique., *Journal of Personality and Social Psychology* **4**(2): 195.

Gächter, S., Kölle, F. and Quercia, S. (2017). Reciprocity and the tragedies of maintaining and providing the commons, *Nature Human Behaviour* **1**(9): 650–656.

Gächter, S., Kölle, F. and Quercia, S. (2022). Preferences and perceptions in Provision and Maintenance public goods, *Games and Economic Behavior* **135**: 338–355.

Gächter, S., Nosenzo, D., Renner, E. and Sefton, M. (2012). Who makes a good leader? Cooperativeness, optimism, and leading-by-example, *Economic Inquiry* **50**(4): 953–967.

Gächter, S. and Renner, E. (2018). Leaders as role models and 'belief managers' in social dilemmas, *Journal of Economic Behavior & Organization* **154**: 321–334.

Galbiati, R. and Vertova, P. (2008). Obligations and cooperative behaviour in public good games, *Games and Economic Behavior* **64**(1): 146–170.

Gerlach, P. and Jaeger, B. (2016). Another frame, another game? Explaining framing effects in economic games, *Proceedings of Norms, Actions, Games (NAG 2016)*.

Ghidoni, R. and Suetens, S. (2022). The effect of sequentiality on cooperation in repeated games, *American Economic Journal: Microeconomics* **14**(4): 58–77.

Goeree, J. K., Holt, C. A. and Palfrey, T. R. (2003). Risk averse behavior in generalized matching pennies games, *Games and Economic Behavior* **45**(1): 97–113.

Goeree, J. K. and Offerman, T. (2003). Competitive bidding in auctions with private and common values, *The Economic Journal* **113**(489): 598–613.

Goette, L., Huffman, D. and Meier, S. (2006). The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups, *American Economic Review* **96**(2): 212–216.

Gold, N. and Sugden, R. (2007). *Theories of team agency*, Oxford University Press.

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C. and Bigham, J. P. (2018). A data-driven analysis of workers' earnings on Amazon Mechanical Turk, *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–14.

Harris, M. B. (1972). The effects of performing one altruistic act on the likelihood of performing another, *The Journal of Social Psychology* **88**(1): 65–73.

Hergueux, J., Jacquemet, N., Luchini, S. and Shogren, J. F. (2022). Leveraging the honor code: Public goods contributions under oath, *Environmental and Resource Economics* **81**(3): 591–616.

Hoffman, E., McCabe, K. and Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games, *The American Economic Review* **86**(3): 653–660.

Horton, J. J., Rand, D. G. and Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market, *Experimental Economics* **14**(3): 399–425.

Iriberri, N. and Rey-Biel, P. (2011). The role of role uncertainty in modified dictator games, *Experimental Economics* **14**: 160–180.

Isler, O., Gächter, S., Maule, A. J. and Starmer, C. (2021). Contextualised strong reciprocity explains selfless cooperation despite selfish intuitions and weak social heuristics, *Scientific Reports* **11**(1): 1–17.

Ismayilov, H. and Potters, J. (2016). Why do promises affect trustworthiness, or do they?, *Experimental Economics* **19**(2): 382–393.

Ito, K., Ida, T. and Tanaka, M. (2018). Moral suasion and economic incentives: Field experimental evidence from energy demand, *American Economic Journal: Economic Policy* **10**(1): 240–67.

Iturbe-Ormaetxe, I., Ponti, G., Tomás, J. and Ubeda, L. (2011). Framing effects in public goods: Prospect theory and experimental evidence, *Games and Economic Behavior* **72**(2): 439–447.

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

Jacquemet, N., James, A., Luchini, S. and Shogren, J. F. (2017). Referenda under oath, *Environmental and Resource Economics* **67**(3): 479–504.

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

Jacquemet, N., Luchini, S., Malezieux, A. and Shogren, J. F. (2020). Who'll stop lying under oath? Empirical evidence from tax evasion games, *European Economic Review* **124**: 103369.

Jacquemet, N., Luchini, S., Shogren, J. F. and Zylbersztejn, A. (2018). Coordination with communication under oath, *Experimental Economics* **21**(3): 627–649.

Joule, R.-V., Girandola, F. and Bernard, F. (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.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk, *Econometrica* **47**(2): 363–391.

Kartal, M. and Müller, W. (2021). A new approach to the analysis of cooperation under the shadow of the future: Theory and experimental evidence, *Available at SSRN 3222964* .

Kay, A. C. and Ross, L. (2003). The perceptual push: The interplay of implicit cues and explicit situational construals on behavioral intentions in the prisoner's dilemma, *Journal of Experimental Social Psychology* **39**(6): 634–643.

Kessler, J. B., Low, C. and Singhal, M. (2021). Social policy instruments and the compliance environment, *Journal of Economic Behavior & Organization* **192**: 248–267.

Khadjavi, M. and Lange, A. (2015). Doing good or doing harm: Experimental evidence on giving and taking in public good games, *Experimental Economics* **18**: 432–441.

Kiesler, C. A. (1971). *The Psychology of Commitment: Experiments Linking Behavior to Belief*, Academic Press.

Kocher, M. G., Cherry, T., Kroll, S., Netzer, R. J. and Sutter, M. (2008). Conditional cooperation on three continents, *Economics Letters* **101**(3): 175–178.

Konow, J. (2019). Can ethics instruction make economics students more pro-social?, *Journal of Economic Behavior & Organization* **166**: 724–734.

Kőszegi, B. and Rabin, M. (2006). A model of reference-dependent preferences, *The Quarterly Journal of Economics* **121**(4): 1133–1165.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary?, *Journal of the European Economic Association* **11**(3): 495–524.

Levin, I. P., Schneider, S. L. and Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects, *Organizational Behavior and Human Decision Processes* **76**(2): 149–188.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments, *Review of Economic Dynamics* **1**(3): 593–622.

Levitt, S. D. and List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world?, *Journal of Economic Perspectives* **21**(2): 153–174.

Liberman, V., Samuels, S. M. and Ross, L. (2004). The name of the game: predictive power of reputations versus situational labels in determining prisoner's dilemma game moves, *Personality and Social Psychology Bulletin* **30**(9): 1175–1185.

López-Pérez, R. (2008). Aversion to norm-breaking: A model, *Games and Economic Behavior* **64**(1): 237–267.

March, J. G. (1994). *Primer on decision making: How decisions happen*, New York: Free Press.

Marks, M. B., Schansberg, D. E. and Croson, R. T. (1999). Using suggested contributions in fundraising for public good, *Nonprofit Management and Leadership* **9**(4): 369–384.

Martinsson, P., Medhin, H. and Persson, E. (2019). Minimum levels and framing in public good provision, *Economic Inquiry* **57**(3): 1568–1581.

McCusker, C. and Carnevale, P. J. (1995). Framing in resource dilemmas: Loss aversion and the moderating effects of sanctions, *Organizational Behavior and Human Decision Processes* **61**(2): 190–201.

Metz, T. (2013). The ethics of swearing: The implications of moral theories for oath-breaking in economic contexts, *Review of Social Economy* **71**(2): 228–248.

Miettinen, T., Kosfeld, M., Fehr, E. and Weibull, J. (2020). Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions, *Journal of Economic Behavior & Organization* **173**: 1–25.

Mischkowski, D., Stone, R. and Stremitzer, A. (2019). Promises, expectations, and social cooperation, *The Journal of Law and Economics* **62**(4): 687–712.

Montgomery, J. D. (1998). Toward a role-theoretic conception of embeddedness, *American Journal of Sociology* **104**(1): AJSv104p92–125.

Moxnes, E. and Van der Heijden, E. (2003). The effect of leadership in a public bad experiment, *Journal of Conflict Resolution* **47**(6): 773–795.

Oskamp, S. (1974). Comparison of sequential and simultaneous responding, matrix, and strategy variables in a prisoner's dilemma game, *Journal of Conflict Resolution* **18**(1): 107–116.

Park, E.-S. (2000). Warm-glow versus cold-prickle: A further experimental study of framing effects on free-riding, *Journal of Economic Behavior & Organization* **43**(4): 405–421.

Potters, J., Sefton, M. and Vesterlund, L. (2007). Leading-by-example and signaling in voluntary contribution games: An experimental study, *Economic Theory* **33**: 169–182.

Rabin, M. (1993). Incorporating fairness into game: Theory and economics, *The American Economic Review* pp. 1281–1302.

Rabin, M. (1998). Psychology and economics, *Journal of Economic Literature* **36**(1): 11–46.

Rege, M. and Telle, K. (2004). The impact of social approval and framing on cooperation in public good situations, *Journal of Public Economics* **88**(7-8): 1625–1644.

Robinson, J., Rosenzweig, C., Moss, A. J. and Litman, L. (2019). Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool, *PloS One* **14**(12): e0226394.

Sally, D. (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992, *Rationality and Society* **7**(1): 58–92.

Savikhin Samek, A. and Sheremeta, R. M. (2014). Recognizing contributors: An experiment on public goods, *Experimental Economics* **17**: 673–690.

Schelling, T. C. (1980). *The Strategy of Conflict*, Harvard University Press.

Schlesinger, H. J. (2011). *Promises, Oaths, and Vows: On the Psychology of Promising*, Taylor & Francis.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments, *Tübingen: JCB Mohr* pp. 136–168.

Sheinman, H. (2011). *Promises and Agreements: Philosophical Essays*, Oxford University Press.

Snowberg, E. and Yariv, L. (2018). Testing the waters: Behavior across participant pools, *Technical report*, National Bureau of Economic Research.

Sonnemans, J., Schram, A. and Offerman, T. (1998). Public good provision and public bad prevention: The effect of framing, *Journal of Economic Behavior & Organization* **34**(1): 143–161.

Sugden, R. (1993). Thinking as a team: Towards an explanation of nonselfish behavior, *Social Philosophy and Policy* **10**(1): 69–89.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice, *Science* **211**(4481): 453–458.

Tversky, A. and Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model, *The Quarterly Journal of Economics* **106**(4): 1039–1061.

Tversky, A. and Simonson, I. (1993). Context-dependent preferences, *Management Science* **39**(10): 1179–1189.

van Dijk, E. and Wilke, H. (2000). Decision-induced focusing in social dilemmas: Give-some, keep-some, take-some, and leave-some dilemmas., *Journal of Personality and Social Psychology* **78**(1): 92.

Vanberg, C. (2008). Why do People Keep their Promises? An Experimental Test of Two Explanations, *Econometrica* **76**(6): 1467–1480.

Weber, J. M., Kopelman, S. and Messick, D. M. (2004). A conceptual review of decision making in social dilemmas: Applying a logic of appropriateness, *Personality and Social Psychology Review* **8**(3): 281–307.

Weber, R. A. (2006). Managing growth to achieve efficient coordination in large groups, *American Economic Review* **96**(1): 114–126.

Wit, A. P. and Wilke, H. A. (1992). The effect of social categorization on cooperation in three types of social dilemmas, *Journal of Economic Psychology* **13**(1): 135–151.

Yamagishi, T. and Kiyonari, T. (2000). The group as the container of generalized reciprocity, *Social Psychology Quarterly* pp. 116–132.