HEIDELBERG UNIVERSITY
DEPARTMENT OF ECONOMICS

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

(Un)Trustworthy Pledges and Cooperation in Social Dilemmas

Timo Goeschl

Alice Soldà

AWI DISCUSSION PAPER SERIES NO. 728
May 2023

(Un)Trustworthy Pledges and Cooperation in Social Dilemmas*

Timo Goeschl† Alice Soldà‡

May 9, 2023

Abstract

Pledges feature in international climate cooperation since the 2015 Paris Agreement. We explore how differences in pledgers' trustworthiness affect outcomes in a social dilemma that parallels climate change. In an online experiment, two participants interact with a randomly matched third player in a repeat maintenance game with a pledge stage. Treatments vary whether participants are matched with a player that is more or less trustworthy as revealed by behavior in a promise-keeping game; and whether they observe that trustworthiness. We find that participants knowingly matched with more trustworthy players cooperate more than participants matched with less trustworthy players (knowingly or unknowingly), but also more than participants unknowingly matched with more trustworthy players. In contrast, participants knowingly matched with less trustworthy players do not cooperate less than participants who are unknowingly so. Our findings suggest that the use of pledges, as per the Paris Agreement, can leverage the power of trustworthiness to enhance cooperation. (154 words)

Keywords: Social dilemmas; cooperation; pre-play communication; credibility; pledges; group formation.

* This research has been funded by the Federal Ministry of Education and Science (grant no. 01LA1806A).

†ZEW - Leibniz Center for European Economic Research; Chair of Environmental Economics, Alfred-Weber-Institut for Economics, Heidelberg University, Bergheimerstraße 20, 69115 Heidelberg, Germany. E-mail: goeschl@uni-heidelberg.de

‡(corresponding author), Department of Economics, Ghent University, Sint Pietersplein 6, 9000, Gent, Belgium. E-mail: alice.solda@ugent.be

1 Introduction

The 2015 Paris Agreement introduced a number of institutional innovations in international efforts to limit global greenhouse gas (GHG) emissions, the underlying cause of anthropogenic climate change. One key innovation was the introduction of a new construct termed *Intended Nationally Determined Contributions* (INDCs) (Falkner, 2016). INDCs are unilateral declarations that individual countries are expected to submit and that specify the country's intended future pathway of national GHG emissions. Each INDC should represent reductions in the country's emissions relative to business-as-usual and thus make an individual contribution to a collective decarbonization process that is supposed to limit global warming to no more than 2 degrees relative to pre-industrial levels.[1]

To the student of social dilemmas, INDCs share many feature of pledges: public statements by parties in which they announce how they will behave in the social dilemma in the future. Pledges constitute a form of structured pre-play communication. As such, they are both public and non-binding, but they also typically contain promises of future cooperative behavior (Charness and Dufwenberg, 2006; Vanberg, 2008). Similarly, INDCs are publicly announced declarations as to emissions reductions, but also do not constitute commitments enforceable by other parties to the dilemma.

In light of the parallels between INDCs and pledge, it is not surprising that there are similar expectations and skepticism associated with pledges in social dilemmas and INDCs in international climate change. In social dilemmas, the exchange of non-binding promises is thought to be conducive to subsequent cooperation (Charness and Dufwenberg, 2006; Vanberg, 2008), even though the experimental evidence on the specific ability of pledges to enhance cooperation is mixed: Pledges enhance cooperation in social dilemmas in some studies (Chen and Komorita, 1994; Pogrebna et al., 2011; Koessler et al., 2021), but not in others (McEvoy et al., 2022; Barrett and Dannenberg, 2016). Similarly, cautious optimism about countries' likely adherence to their publicly stated INDCs (Pauw and Klein, 2021) is tempered by the recognition that INDCs represent a form of 'cheap talk' and that other countries' trust in a country's INDC depends largely on their belief in the trustworthiness of the pledge made (Averchenkova and Bassi, 2016).

The question of how to make pledges trustworthy lies at the center of a literature that has studied the prerequisites for pledges in social dilemmas to foster cooperation in a group. Pledges can be trustworthy for a number of reasons. With commitment devices, for instance, pledgers themselves change their incentive structure such that fulfilling the pledge aligns with their self-interest (Reischmann and Oechssler, 2018). This extends to mechanism such as 'pledge-and-review' (Barrett and Dannenberg, 2016) that can be

[1] Assessments of the first round of INDCs submitted concluded that collectively, the INDCs would instead lead to between 2.7°C and 3.6°C warming (Höhne et al., 2017; Rogelj et al., 2016).

designed to implement pledges only after unanimous agreement (Harstad, 2023). The incentive structure also changes when the threat of peer punishment is present, offering another pathway to supporting other parties' confidence in the pledge (Lippert and Tremewan, 2021).

In many settings, however, structures that support trustworthiness of pledges can be difficult to set up. Credible commitments (Williamson, 1983), procedural provisions, and organizing peer punishment (Diekmann, 1985) are typically costly, both in terms of resources and time. In the case of INDCs, it also involves overcoming legal and political obstacles: Sovereign states cannot be easily held to their promises for reasons of international law Bauer et al. (2020) and getting states to punish each other effectively raises coordination and cooperation problems of its own (Cherry et al., 2021).[2] In such settings, the effectiveness of pledges is likely to depend to a large degree on access to information about pledgers' trustworthiness and on the record of trustworthiness contained in this information (e.g. Bolton et al., 2005; Pogrebna et al., 2011; Goeschl and Jarke, 2017). Such information can come, for instance, from observing the pledger's previous behavior in similar settings. This has been done in the context of climate cooperation where countries' likelihood of honoring their INDCs has been rated by examining past behavior in honoring other international agreements (Averchenkova and Bassi, 2016). On the other hand, having access to information about past behavior in a different setting need not affect outcomes if participants deem such information as uninformative about behavior in the dilemma at hand.

In this paper, we explore experimentally how a pledger's trustworthiness and other parties' awareness of the pledger's trustworthiness affect outcomes in a social dilemma with a pledge stage. The design features subjects playing five rounds of a variant of the public goods game, the maintenance game (MG) (Gächter et al., 2017), augmented by a pledge stage in fixed groups of three members. For one of the three members, the experimenter has a measure of their trustworthiness based on their previous behavior as a trustee in a cognate, but unrelated promise-keeping game (PKG) styled after Charness and Dufwenberg (2006): In the paper (but not in the experiment), trustees who make and keep the promise in the game are referred to as 'more trustworthy' and those who make and break the promise as 'less trustworthy' (Ismayilov and Potters, 2016). We have four treatment conditions, based on a two-by-two design. One dimension varies group composition: Whether the other two group members in the augmented maintenance game are matched with a more or less trustworthy trustee. The other dimension varies information: whether the other two group members observe the trustee's behavior in the promise-keeping game or not. This design allows us to measure – in a social dilemma augmented by a pledge stage – the causal effects of trustworthiness of a group member

[2]In fact, countries are always free to withdraw from the Agreement, as the USA did in 2017.

on efficiency, as measured by aggregate contributions by the other two group members.

Our approach relates to other studies of pledges as a form of structured, non-binding pre-play communication in social dilemmas in which group members announce future play to other group members. Chen and Komorita (1994) compare group efficiency in the linear public goods game and find that pledges do not raise efficiency, while binding commitments do. The assessment in the literature has since become more optimistic: For example, Koessler et al. (2021) allow subjects to make a non-binding or binding pledge of contributing 75% of their endowment and find that the option of the non-binding pledge is both popular and as effective as the binding pledge. We differ from these papers in that we take the pledge stage as a given, in line with the empirical reality of the 2015 Paris Agreement, but vary group composition and information. In terms of the former, our approach mirrors studies that have manipulated group composition with respect to pro-social behavior through exogenous sorting (see Guido et al. (2019) for a survey) based on observed participant behavior. Like Burlando and Guala (2005), Gächter and Thöni (2005), and De Oliveira et al. (2015), we avoid strategic behavior by not making explicit that the sorting mechanism is based on performance in a previous task, shutting down possible signaling (Heinz and Schumacher, 2017). Like De Oliveira et al. (2015), one treatment dimensions varies whether group members are informed about the group composition. We differ from these papers by choosing to manipulate only the type of one out of three players, allowing a clean comparison of the manipulation across groups, and by manipulating the information about a single group member. We also differ in the specific task-game combination (promise-keeping game followed by a maintenance game with a pledge stage). Finally, we build on a literature that has been examining – and broadly confirming – the stability of pro-social preferences across different game forms Blanco et al. (2011); Yamagishi et al. (2013); Dariel and Nikiforakis (2014); Dreber et al. (2014) [3] We extend this literature by showing that pro-social behavior is stable between two games, the PKG and the MG, that are comparatively little studied in themselves, and not before in this combination and sequence.

We disentangle three effects on efficiency: The Composition Effect of being matched with a more or less trustworthy party; the Information Effect of learning about the party's trustworthiness; and the Pledge Effect of receiving a pledge from a more or less trustworthy party. Our behaviorally informed hypotheses predict that the Composition Effect is positive: Efficiency in the augmented maintenance game will be higher across all rounds when the trustee is more trustworthy, even when trustworthiness is not disclosed. The existence of the effect rests on three hypotheses. One is that it is common knowledge that trustworthiness and cooperative behavior are positively correlated. The second hypothe-

[3]Dreber et al. (2014) also find a that giving in a dictator game correlates with cooperation in the repeated prisoners' dilemma game, but only when no cooperative equilibria exist. They conclude that the underlying mechanisms of cooperative behavior differ between the two game forms.

sis is that it is common knowledge that trustworthiness is associated with a smaller gap between pledged and actual behavior. The third is that conditional cooperation is the dominant behavior among the population of players. Since the experimental manipulation changes the trustworthiness of only one of the three group members, the magnitude of the composition effect could be small. The Information Effect is predicted to amplify the Composition Effect: Compared to groups with no information about the trustee's trustworthiness, providing this information increases efficiency when trustworthiness is high and decreases efficiency when it is low. The Information Effect is predicted to be present right from the first round of the social dilemma because the two other group members can condition their pledges and their actions on the trustworthiness information received. Finally, the Pledge Effect is predicted to be negative: For the same pledge by a trustee, the other group members will withdraw fewer tokens if the trustee is more, rather than less, trustworthy.

Drawing on choice data of 795 participants in an online implementation of the design, we arrive at four main findings. First, efficiency tends to be higher in groups with a more trustworthy trustee even when the trustee's trustworthiness is not observable. This is in line with the predicted Composition Effect, but the difference does not rise to statistical significance. Second, revealing a trustee's trustworthiness boosts efficiency when that trustworthiness is high. This effect is substantial, highly significant, and present from round 1 of the interaction. This is in line with the predicted Information Effect. Third, the Information Effect is asymmetric: Revealing at trustee's trustworthiness does not reduce efficiency compared to not revealing that presence when that trustworthiness is low. Fourth, the same pledge leads to a more cooperative response by the other two group members when the trustee is known to be more, rather than less, trustworthy. This is in line with the predicted Pledge Effect. Overall, we conclude that for social dilemmas with a pledge stage, efficiency is served by having access to information that documents the trustworthiness of other players, thus helping pledges to achieve their intended outcome.

Our results are significant for four reasons. First, our results lend support to the view that cooperative pledges can positively affect efficiency in social dilemmas despite constituting non-binding pre-play communication. This contrasts with more pessimistic assessments in the literature (Chen and Komorita, 1994; Barrett and Dannenberg, 2016, e.g.) and sides with more optimistic assessments (Koessler et al., 2021). Second we show that information about the trustworthiness of another group member matters for the efficiency in social dilemmas. This finding is important in light of the fact that earlier papers derived in a related settings such as De Oliveira et al. (2015) do not find an Information Effect. Third, we show that it is specifically information about the presence of a more trustworthy group member that is critical for determining efficiency. Fourth, the clean identification through a successful exogenous manipulation of trustworthiness in the group and through an efficiency measure that excludes the third group member's

behavior provides high validity.

The results also have implications for judging one of the key innovations in the 2015 Paris Agreement. They suggest that in a world of sovereign states that differ in their track record of honoring their promises, a positive track record is conducive towards enhanced cooperation and should be publicized. Concerns that such publication will trigger a parallel decrease in cooperation when that track record is negative are given less support by our results. One important mechanism behind this enhanced cooperation is that countries take another country's INDC more seriously when that country has a positive track record and respond more cooperatively towards that pledge. Previous conduct in international agreements can therefore enhance outcomes when trustworthy states participate.

2 Experimental Design

In order to examine the effect of past promise-keeping behavior on cooperation, we designed a 2x2 experiment in which we exogenously manipulate (i) the group composition in terms of promise-keeping behavior and (ii) whether past promise-keeping behavior can be observed or not. To implement the manipulation, the experiment is composed of two steps. In the first step, we use a modified version of the one-shot promise-keeping game (PKG) from Charness and Dufwenberg (2006) to elicit promise-keeping behavior. In the second step, we use the behavior in the PKG to put together groups that play a repeated maintenance game (MG) with a pledge stage in fixed formation and measure participants' cooperation in the MG.

Step 1: Promise-Keeping Game. Charness and Dufwenberg (2006) construct a PKG that consists of a trust game with a preceding promise opportunity. We implement the PKG by randomly allocating participants to one of two roles: trustor or trustee. A trustor chooses between a safe option *Out* and a risky option *In*. Choosing *Out* returns 20 cents to herself and 20 cents to the trustee she is matched with.[4] When choosing *In*, the game moves to the trustee. A trustee chooses between a selfish option *Don't Roll* and a cooperative option *Roll*. Choosing *Don't Roll* delivers 10 cents to the trustor he is matched with and 100 cents to himself. Choosing *Roll*, the trustee earns 50 cents for sure while the trustor earns 100 cents with probability $5/6$ or 10 cents with probability $1/6$. The defining feature of the PKG is that prior to the trustor making a decision, the trustee has a promise-making opportunity. This promise-making opportunity takes the form of structured communication: The trustee can either stay silent (*NoPromise*) or announce to the trustor that he will choose *Roll* (*Promise*). The subsequent choices by the trustor

[4]For the sake of clarification, we use 'she/her' when referring to the trustor and 'he/his' when referring to the trustee.

between *In* and *Out* and by the trustee between *Roll* and *Don't Roll* take place without knowledge of the other party's decision. The decision tree of the PKG is depicted in Figure 1. The outcomes of the PKG are not disclosed to the participants until after the second stage of the experiment.

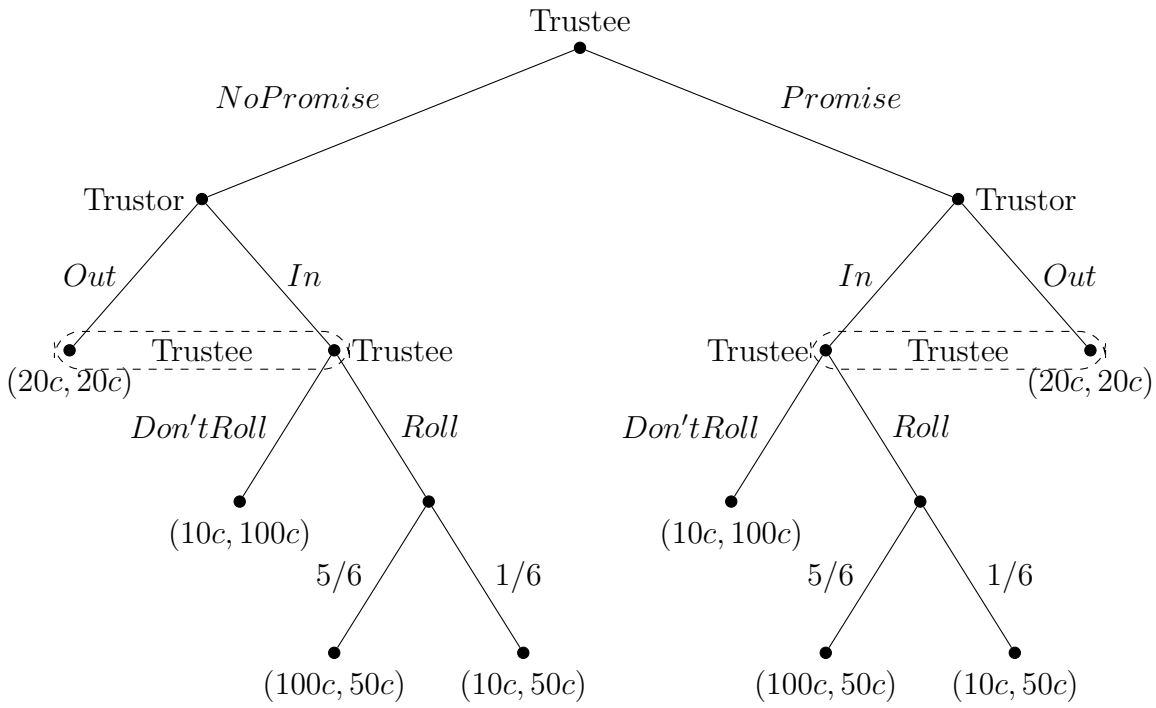


Figure 1: Modified Promise-Keeping Game.

Step 2: Maintenance Game. The Maintenance Game in step 2 of the experiment is a version of the linear public goods game in which the action space of participants consists of *withdrawing* from a public good that already exists rather than the standard case of contributing to a public good that does not exist yet. It is seen as a more fitting parable for social dilemmas such as climate change and has been shown to pose greater challenges to human cooperation (Gächter et al., 2017, 2022). This makes it a suitably challenging environment for detecting the treatement effects. Our modified version changes group size and MPCR and adds a pledge stage. Participants interact in groups of three and can withdraw tokens from a pre-existing public good. The public good contains 60 tokens. Each participant can withdraw up to 20 tokens from the public good. Each unit withdrawn yields 10 cents for the participant who withdrew it and nothing for the others. Each unit left in the public good yield 5 cents to every participant in the group. Hence, the socially optimal solution is to leave all 60 tokens in the public good, yielding 300 cents to each group member, while the Nash Equilibrium solution of the MG is for each payoff-maximizing participant to withdraw 20 tokens from the public good, yielding 200 cents to each group member.

The group composition in the MG is based on participants' role in the PKG: A group is always composed of two trustors and one trustee. Participants are aware that the trustor encountered in their MG is never the same as the trustor encountered in their PKG.[5] In our experiment, each period of the MG consists of four stages: a 'pledge stage', an 'information stage', a 'taking stage' and a 'feedback stage'. In the pledge stage, group members simultaneously indicate the number of units between 0 and 20 they intend to withdraw from the pre-existing public good. In the information stage, pledges are made public. In the taking stage, participants are reminded about their group members' pledges and make their actual withdrawal decisions. In the feedback stage, participants receive feedback on the pledges, withdrawal decisions and earnings of everyone in the group for the current period. Groups play the maintenance game for five periods in a fixed configuration.

Matching. The objective of the procedures is to create groups of three in the MG step of the experiment that did not interact in the PKG step. These procedures need to succeed in an online environment of the experiment, putting a bonus on minimal waiting times for participants. The resulting approach assembles participants into sets of six subjects at the beginning of the experiment, based on their arrival time. In each set, two participants are randomly assigned the role of trustees and the remaining four are assigned the role of trustors. For the PKG, both trustee are randomly matched with two trustors each and then complete that experimental step. Before the MG, both trustees are then rematched with the remaining two trustors in their set, as illustrated in Figure 2. As a result of this matching procedure, there is no prior payoff-dependent interactions between group members in the first round of the maintenance game and waiting times are kept short.

[5]The exact matching procedure used in the experiment is detailed in the next subsection.

a) Matching Promise-Keeping Game

b) Matching Maintenance Game

Figure 2: Matching Procedure.

Treatment variations. The purpose of the design is to measure the effect of past promise-keeping behavior on others' cooperation. This requires manipulating group composition and the information given to participants about their group composition. Exogenous matching in the MG ensures meeting the former objective: Some couples of former trustors will be matched with a third group member who kept his promise as a trustee in the PKG (treatment *Promise kept*). Others will be matched with a third group member who did not (treatment *Promise broken*). Exogenously varying the information provided in the MG to former trustors about the matched trustee's decisions in the PKG ensures meeting the latter objective.[6] Some couples of formers trustors will learn whether the third group members made a promise to choose *Roll* as a trustee in the PKG or not and if so, whether he kept it (treatment *Observable*). Others will not learn anything about the third group member's behavior as a trustee in the PKG (treatment *Hidden*). To limit demand effects, participants are provided with this information in isolation at the beginning of the first round. It then remain at the top of participants' screen until the end of the fifth period.

Payouts Payoffs are determined at the end of the experiment and are the sum of two parts, each corresponding to one of the steps. For the PKG, participants chosen as trustors are paid according to the decisions of the trustee they are matched with. Trustee are paid according to the decisions of one of the two trustors he is matched with equal probability. For the MG, one of the five rounds is randomly selected for payment.

[6]Note that participants are informed at the beginning of the experiment that their decisions in the first part of the experiment can affect their earnings in the second part.

3 Hypotheses

Like in other social dilemmas, behavior in the MG reflects selfish payoff considerations as well as conditional and unconditional cooperation, even though cooperative drivers are less forceful in the MG compared to the standard public goods game (Gächter et al., 2017). These considerations inform our two main hypotheses, which were pre-registered[7].

The primary target of our paper, and the focus of the hypotheses, is the efficiency of the groups in the MG. Recall that the MG always matches two participants that assumed the role of a trustor in the PKG with one participant who assumed the role of a trustee in the PKG. Our hypotheses rely on two underlying assumptions regarding the behavior of the trustees. The first is a positive correlation between trustworthiness and cooperativeness. In line with previous experiments (De Oliveira et al., 2015), we find that trustworthy trustees withdraw, on average, fewer tokens than less trustworthy ones. The second is a positive correlation between trustworthiness and *credible* pledges, that is, how close pledges are to actual withdrawal decisions. In line with previous experiments (Cagala et al., 2019), we find that more trustworthy trustees make pledges that are more credible on average than those of less trustworthy trustees. These behavioral differences between more and less trustworthy trustees can confound the treatment effect of manipulating trustworthiness in the group with the treatment and provide a rationale for excluding the behavior of trustees exogenously allocated to the group. The two trustors' withdrawal decisions, or equivalently their payoffs, therefore constitute our outcome variable for identifying the impact of varying the trustworthiness of a co-player on cooperation, and therefore efficiency, in the MG with a pledge stage.

The first hypothesis concerns the treatment effect of being matched with a trustee player who is more trustworthy, i.e. that kept his promise in the PKG, even when that trustworthiness is not disclosed to other group members. Trustworthiness structurally matters if the propensity to keep promises and the propensity to cooperate are positively correlated. In this case, matching trustors with a more trustworthy trustee also means matching trustors, on average, with a more cooperative trustee. The required experimental variation comes from the *Promise kept* and *Promise broken* treatments under the *Hidden* condition. If conditional cooperation is the dominant behavioral type in among participants, this means that groups allocated a more trustworthy trustees exhibit, on average, high cooperation levels even if the trustors are unaware of the trustee's behavior in the PKG. Averaged over all five rounds of the MG, we therefore predict fewer tokens withdrawn and higher efficiency when the trustee is more trustworthy in the *Hidden* condition. We term this is the "Composition Effect".

[7] All our preregistered hypotheses can be found at https://aspredicted.org/ZMG_SHP. We re-arranged them in the paper for improved readibility.

Hypothesis 1 (Composition Effect) Everything else equal, average efficiency in the MG with a pledge stage is predicted to be higher (withdrawals lower) when trustors are matched with a more rather than a less trustworthy trustee when trustworthiness is not observable.[8]

The second of our hypotheses concerns the treatment effect on trustors of being informed about the trustworthiness of the trustee in the group, from the start of the MG to the end. Previous experimental evidence from a three-player linear public goods game did not find evidence that information about whether the group contains one, two, or three cooperators affects efficiency (De Oliveira et al., 2015). This is surprising because knowledge about the co-player's past behavior in social dilemmas has been shown elsewhere to matter significantly for efficiency (Fischbacher et al., 2001; Fischbacher and Gächter, 2010). The reason is that trustors' beliefs about the trustee's propensity to cooperate are shaped by evidence on the trustee's behavior in the PKG. In concrete terms, promise-keeping trustees will, on balance, be perceived as likely cooperators while promise-breaking trustee will be perceived as likely defectors in the social dilemma. Since beliefs about co-players' cooperation are strong and positive predictors of own cooperation, we predict less withdrawals and therefore higher efficiency when the presence of a more trustworthy trustee is disclosed compared to when it is not. Likewise, when the presence of a less trustworthy trustee is disclosed, we predict higher withdrawals and therefore lower efficiency. We term this the "Information Effect". While the Composition Effect of Hypothesis 1 will arise over time and therefore be detectable only across all rounds, we predict the Information Effect to be detectable as soon as information about trustworthiness is made available. In the present MG, trustors are informed before the first round that their group contains a more (less) trustworthy trustee. As a result, the causal effect of trustworthiness can establish itself from the start rather than over time, as in Hypothesis 1.

Hypothesis 2 (Information Effect) Everything else equal in the MG with a pledge stage, efficiency is higher (lower) in groups matched with a more (less) trustworthy trustee when trustworthiness is observable rather than unobservable. This difference in efficiency is detectable both in round 1 and in the average across all rounds.[9]

In addition to our two pre-registered hypotheses, we explore the mechanism behind how the information about trustee's trustworthiness affects the relationship between the trustee's pledges and trustors' withdrawals. While pledges constitute a form of 'cheap talk', their *intended* function is to signal future cooperative behavior to other group members and thus to foster more cooperation. When a more and a less trustworthy trustee

[8]The Composition Effect corresponds to hypothesis 3 in the pre-registration.
[9]The Information Effect corresponds to hypotheses 4 and 5 in the pre-registration.

pledges the same amount of withdrawal, we expect trustors, as conditional cooperators, to withdraw less when matched with a trustee revealed as more trustworthy. The reason is that trustors will expect the more trustworthy trustee to be withdrawing less than the less trustworthy trustee, to which they respond with withdrawing less.

Hypothesis 3 (Pledge Effect) Everything else equal, including the trustee's pledge, trustors withdraw less when matched with a trustee revealed as more trustworthy than when matched with a trustee revealed as less trustworthy.

These three hypotheses are the object of tests in section 5.

4 Implementation

Like the hypotheses, the experimental design and procedures were pre-registered.[10]

Participants. The experiment was programmed using oTree (Chen et al. 2016) and conducted online on Amazon MTurk. A total of 1,839 U.S. residents took part in the experiment. Participation was restricted to individuals over 18 years of age, who completed at least 300 HITs with an approval rate of at least 99%. Overall, 32.19% of the participants were female and the average age was 39.47 years (SD = 11.29).

Detailed procedure. After accepting the HIT, participants were redirected to the oTree interface and grouped into sets of six based on arrival time. The matching procedures were implemented as described above and both steps of the experiment conducted in close succession. Participants had to answer a comprehension questionnaire correctly after the presentation of the instructions in order to proceed. During the experiment, participants could re-read the instructions at any time by clicking on a reminder button on their screen.[11]

Earnings. The experiment took an average of 20 minutes. Participants were paid the sum of their earnings in Part 1 and for one randomly selected round in Part 2, in addition to a $1 participation reward. The average payoff was $3.61 (SD = 0.59). Participant earnings were denominated directly in cents. All participants were paid less than 48 hours after the completion of the experiment.

Exclusion rule. We pre-registered that we would exclude from the analyses observations from participants from groups in which at least one group member dropout of the

[10] https://aspredicted.org/ZMG_SHP

[11] The screens used in the experiment are provided in Appendix A-2.

experiment for the periods after the drop occurred. If the drop occurred during the first period, we excluded all observations for this group. Hence, the following analyses were conducted on the sub-sample of 816 participants for which no group members dropped during the first period.

5 Results

This section is organised as follow. First, we provide some summary statistics to ensure that our randomization worked. Second, we present evidence that our manipulation of group composition worked: More (less) trustworthy trustees were more (less) cooperative group members and pledge more (less) credibly. Finally, we present our main results on withdrawal decisions and pledges.

5.1 Descriptive statistics

Summary statistics. Table 1 provides summary statistics on trustees' behavior prior to the treatment manipulation. In both treatments, most participants made a promise to choose the cooperative option (83.04% and 88.24%). Out of these participants, most of them chose to keep that promise (62.37% and 80.74%). Using two-sided Fisher exact tests, we find no significant difference between treatments in the percentage of participants who made a promise ($p = 0.282$). One random difference is that despite identical instructions, participants in the Hidden treatment were less likely to keep their promise than participants in the Observable treatment, ($p = 0.002$).

Table 1: Summary statistics (B players)

	Hidden	Observable
Promise made	83.04%	88.24%
	(N=93)	(N=135)
Promise kept	62.37%	80.74%
	(N=58)	(N=109)
Promise broken	37.63%	19.26%
	(N=35)	(N=26)

Note: Table 1 displays the percentage of participants who made a promise to choose the cooperative action and the percentage of these participants who chose to keep their promise and the percentage of those who chose not to keep their promise.

5.2 Manipulation checks

Hypotheses 1 through 3 rely on two specific premises of trustee behavior in the PKG and in the MG being correlated. These premises, both based on prior evidence, need to hold in order for the hypotheses to be able to fail. If the premises do not hold, our design simply failed to manipulate the composition of the groups in the intended way.

To ascertain whether the manipulation succeeded, we first test whether we replicate the observation by Cagala et al. (2019) that on average, promise-keeping trustees withdraw less in the MG than promise-breaking trustees. Figure 3 shows the average tokens taken by trustees for each round, depending on the treatment condition. A simple visual inspection already suggests that the manipulation succeeded: Groups that were assigned a less trustworthy trustee in the MG were also assigned a group member that was, on average, less cooperative in the MG than a more trustworthy trustee.

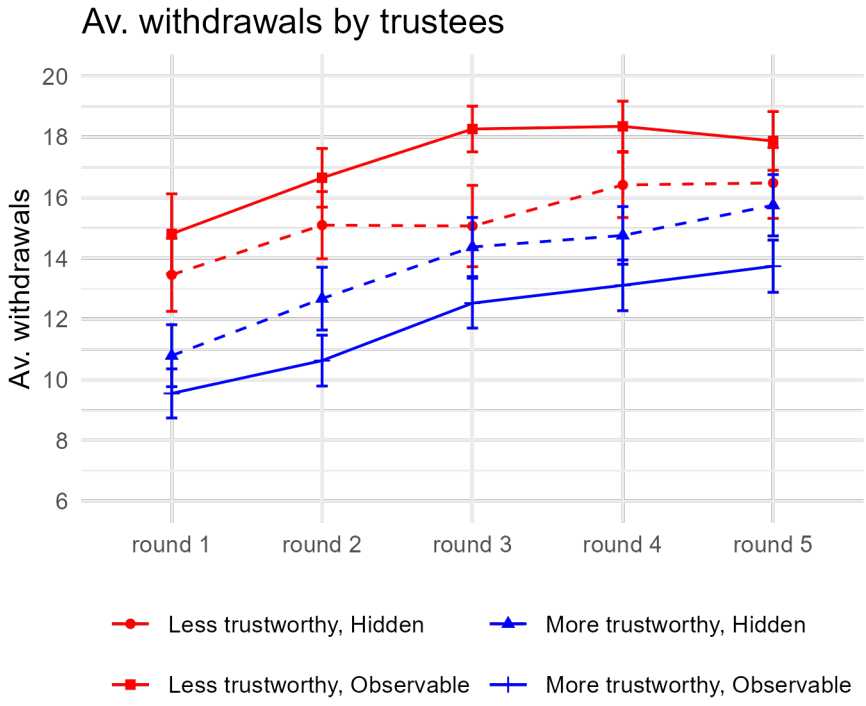


Figure 3: Withdrawals by (former) trustees in the Maintenance Game by round, for each treatment condition (see legend). Vertical bars indicate standard errors.

We formally interrogate the experimental data by comparing takings data of more and less trustworthy trustees in a pooled fashion. Table A-2 shows that on average, more trustworthy trustees withdrew 4.05 fewer tokens than less trustworthy trustees in the first round and 3.35 fewer tokens across all rounds, and the differences are highly significant (MW tests: $p < 0.001$ in both cases). Our exogenous manipulation of groups therefore succeeded with respect to takings behavior.

Secondly, we test whether we can replicate the premise that trustees identified as

more trustworthy in the PKG also make more credible pledges in the MG.[12] Credibility is measured by the difference between the amount of tokens taken w_i and the stated pledges $\bar{w}_i$ ($cred = \bar{w}_i - w_i$) in each round. A negative difference indicates by how much the trustee withdrew more than pledged. As before, we compare the credibility of more and less trustworthy trustees in a pooled fashion. Figure 4 displays the average credibility for trustees who broke their promise (red line) and trustees who kept their promise (blue line) in the promise-keeping game pooled across treatments in each round of the maintenance game.

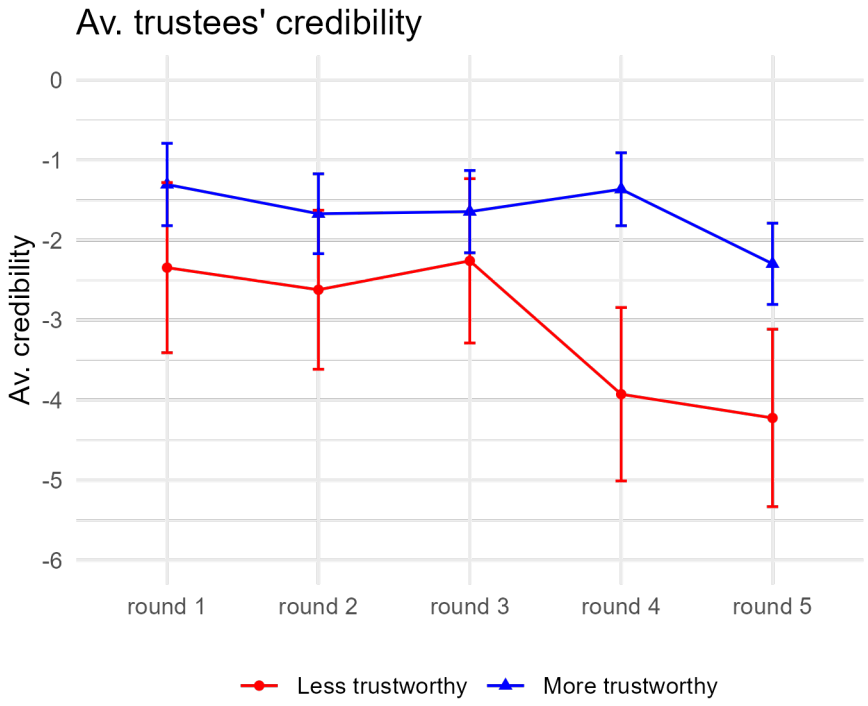


Figure 4: Average difference between the amount of tokens taken and the stated pledges for B players who broke their promise and B players who kept their promise in the Hidden treatment. Vertical bars indicate standard errors.

Figure 4 shows that on average, both more and less trustworthy trustees take more tokens than they pledged they would: The red (less) and the blue (more trustworthy) line are below 0 throughout the game. Figure 4 also suggests that less trustworthy trustee are less credible than more trustworthy ones, especially in later stages of the game: The red curve is always below the blue curve, with a clear divergence in rounds 4 and 5. A formal test of the premise Table A-2 shows that on average, more trustworthy trustees take 1.74 tokens more than they pledged while less trustworthy trustees take 3.06 tokens more. The difference is not significant when looking at all rounds together (MW test: $p = 0.285$), but clearly so in the last two rounds (MW test: -4.07 vs. -1.85; $p = 0.034$). This suggests that those who keep and those who break their promise issue pledges with

[12]This is pre-registered as Hypothesis 6.

comparable credibility in the early rounds of the MG before the promise-breakers start to diverge from the pledges over time.

A random effect GLS regression using credibility as the dependent variable supports this interpretation (see Table 2). The independent variables include a dummy variable that equals 1 for a more trustworthy trustee and a time trend (round). Of the three specifications tested, the best performing model (3) features a common intercept of just over 1 token for both more and less trustworthy trustees. The statistically significant decline in credibility over the rounds of around half a token that is largely offset by a significant interaction effect that is present when the trustee kept the promise.[13] This is consistent with less trustworthy trustees making early credible pledges for strategic reasons while more trustworthy trustee pledging credibly based on type.

Table 2: Effect of trustee's trustworthiness on own credibility.

Dep. var:	Trustees' credibility in round t		
	(1)	(2)	(3)
More trustworthy trustee	1.359	0.440	–
	(0.706)	(1.053)	
Round	–	-0.498*	-0.560**
		(0.232)	(0.179)
More trustworthy trustee*Round	–	0.319	0.403*
		(0.271)	(0.182)
Constant	-3.040	-1.605	-1.283**
	(0.604)	(0.900)	(0.466)
Obs.	1048	1048	1048
Clusters	228	228	228

Note: Table 2 displays the GLS coefficients of the random effects regression. Standard errors in parentheses. *p <0.05, **p <0.01, ***p <0.001.

Based on the tests on withdrawal and pledging behavior, we conclude that the exogenous manipulation of group composition succeeded. Groups assigned a more (less) trustworthy group member are also assigned a more (less) cooperative and credible group member.[14]

[13]The experimental evidence does in fact not allow to reject the hypothesis that more trustworthy trustees actually maintain the same level of credibility throughout. Testing whether the combined effect of the round and interaction effects is different from zero leads to a p-value of 0.227.

[14]This provides a test for hypothesis 1 in the pre-registration and the results show that this hypothesis is supported by the data.

5.3 Main Results

The results for our tests of Hypotheses 1 through 3 are visually prefigured by Figure 5. The top panel of Figure 5 displays the average final withdrawal decisions of trustees across rounds for each treatment separately. The bottom panel likewise displays their average pledges across rounds. For both panels, the red lines represent a group assigned a less trustworthy, the blue lines those assigned a more trustworthy trustee. Solid lines represent groups in which trustee's trustworthiness was observable, dotted lines those where it was hidden. We will refer to Figure 5 throughout this section.

Inspecting Figure 5, we first of all see patterns broadly familiar from other repeated social dilemmas. Participants started by taking out somewhere between 11 and 15 (out of 20 tokens) in round 1, depending on the treatment. This corresponds to contribution of between 5 and 9 tokens, or around 25% to 45% of endowment. On average, this is comparable, but somewhat less than in the standard VCM, corroborating claims that the MG presents a more challenging environment for cooperation (Gächter et al., 2017, 2022). As the game progressed, takings grew and, conversely, efficiency declined, irrespective of the treatment. However, coordination persisted up to and including the final round. Both observations align with the patterns observed in the VCM and other social dilemmas (Fischbacher and Gächter, 2010).

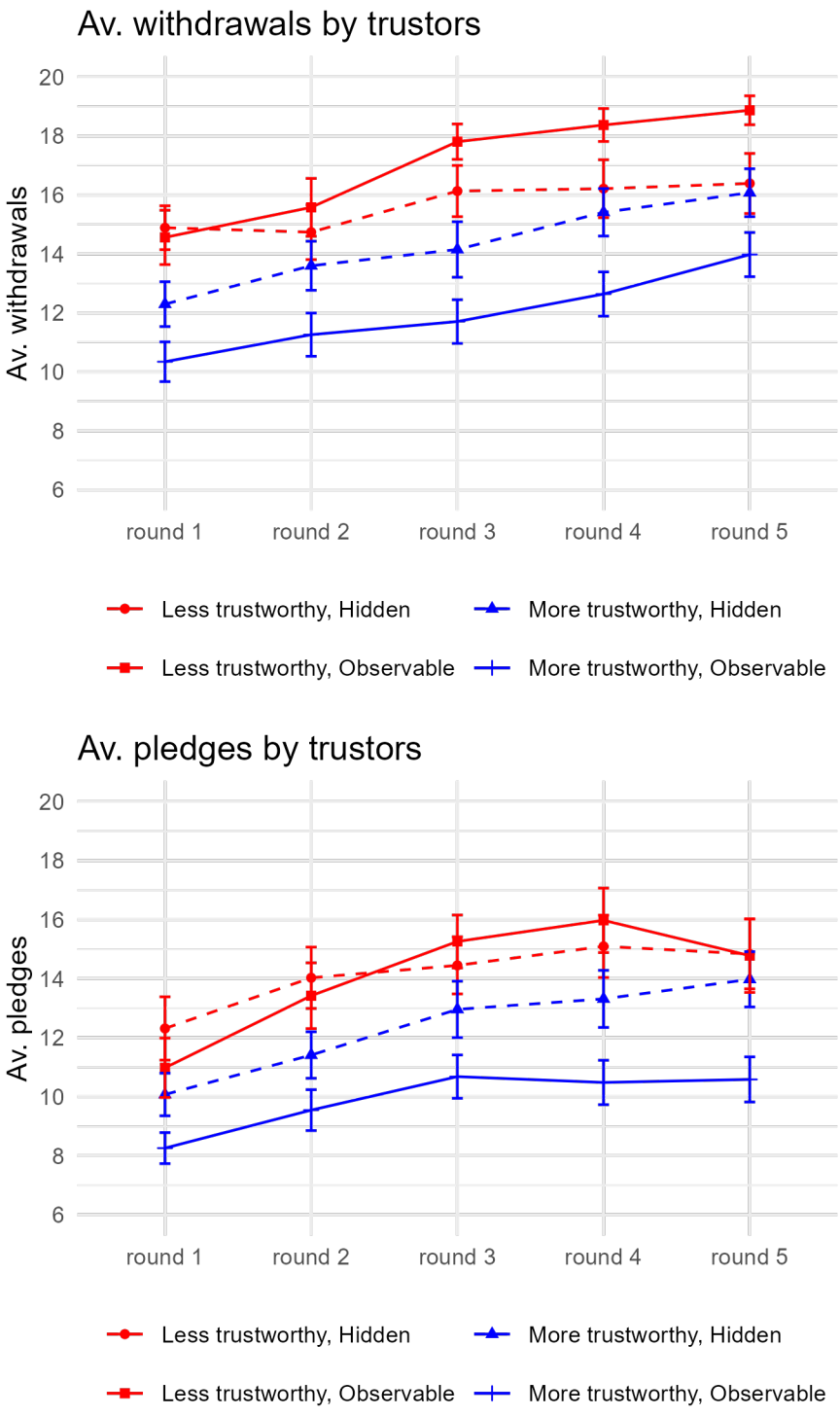


Figure 5: Average withdrawal decisions A players by treatments.

Hypothesis 1 predicts that in the Hidden treatment, when it is common knowledge that trustworthiness is not observable, trustors' average withdrawals across all rounds would be lower when matched with a more, rather than a less, trustworthy trustee. Conversely, predicted efficiency is higher when matched with a more trustworthy trustee. The reason was the Composition Effect of conditionally cooperating trustors being more likely to be

encountering cooperative trustees.

The experimental evidence does not support Hypothesis 1. Comparing the blue (more) and red (less trustworthy) dotted lines in Figure 5, we see that in the first round, trustors withdrew fewer tokens when matched with a more trustworthy trustee (Mann-Whitney test:[15] 14.89 (0.744) vs. 12.29 (0.760): $p = 0.026$). This pattern persists across the five rounds but on average, the difference is too small to pass tests of statistical significance (MW tests: 15.69 (0.766) vs. 14.24 (0.690): $p = 0.174$). The fact that differences in withdrawals between groups facing more and less trustworthy trustees do not grow over time also challenges the logic underlying Hypothesis 1: Stronger beliefs about being matched with a more cooperative group member, which are associated with more trustworthy trustees, would instead be expected to favor a divergence.

Result 1 (Composition Effect) Hypothesis 1 is not supported: When trustees' trustworthiness is not observable, efficiency in the Maintenance Game is, on average, not significantly higher when a group is matched with a more, rather than a less, trustworthy trustee.

Result 1 shows that when trustors interact with trustee without knowing whether they kept or broke their promise in the PKG, the impact of trustees' trustworthiness on efficiency is statistically weak and gets weaker across rounds. In other words, a possible Composition Effect is not strong enough to establish itself in a five-round MG with a pledge stage.

Hypothesis 2 predicts that in the Observable treatment, fewer tokens will be withdrawn in groups that observe that they are matched with a more trustworthy trustee compared to those that do not observe that their trustee is more trustworthy. Likewise, more tokens will be withdrawn when group members know that they are matched with a less trustworthy trustee. Hypothesis 2 also predicts that this Information Effect will be detectable not only on average, but already from round 1. While resting on strong conceptual foundations, this prediction runs counter with an important previous finding that the Information Effect is absent or at least weak (De Oliveira et al., 2015).

The experimental evidence supports Hypothesis 2, but only partially. Comparing the solid and dashed blue lines in Figure 5, the Information Effect for trustors matched with more trustworthy trustees stands out: Trustors took 2.1 fewer tokens in the first round (MW test: 10.16 (0.674) vs. 12.29 (0.760): $p = 0.058$) and an average of 2.4 fewer tokens per round across all rounds (MW test: 11.84 (0.649) vs. 14.24 (0.690): $p = 0.047$) when they observed the trustee's trustworthiness. This Information Effect means that making higher trustworthiness observable has social returns in terms of significantly higher effi-

[15] MW, hereafter. All p-values are two-sided.

ciency. For less trustworthy trustees, the Information Effect is absent: Comparing the solid and dashed red lines, we can see that trustors took about the same amount of tokens in round 1 (MW test: 14.89 (0.744) vs. 14.56 (0.922): $p = 0.738$) and, on average, across all rounds (MW test: 16.49 (0.609) vs. 15.69 (0.766): $p = 0.720$).

Result 2 (Information Effect) Hypothesis 2 is partially supported: The difference in efficiency between groups that are knowingly matched with a more trustworthy trustee in the Maintenance Game and those that are so matched unknowingly is positive and significant across all rounds. There is no difference in efficiency between groups that are knowingly matched with a less trustworthy trustee and those that are so matched unknowingly. The Information Effect therefore asymmetrically favors the observability of higher trustworthiness.

Result 2 is in line with the theoretical considerations informing Hypothesis 2: Conditional cooperators respond to another group member being revealed as more trustworthy in the PKG by cooperating more in the MG. However, they do not respond to another group member being revealed as less trustworthy in the PKG by cooperating less in the MG. Decreasing cooperation is conditional on that group member actually behaving less cooperatively in the MG. Our finding contrasts with that in De Oliveira et al. (2015), who do not find an information effect.

The Composition Effect and Information Effect jointly explain the difference between the blue (more) and red (less trustworthy) solid lines in Figure 5. When trustworthiness is observable, there are persistent differences between groups that know to be interacting with more and less trustworthy trustees in the MG. In the first round, the former groups withdrew significantly fewer tokens when matched with a more, rather than less, trustworthy trustee (MW tests: 10.16 (0.674) vs. 14.89 (0.744): $p < 0.001$). This pattern persisted throughout the game (MW test: 11.84 (0.649) vs. 15.69 (0.766): $p = 0.003$). When combined, the Composition Effect and Information Effect are therefore clearly conducive towards efficiency in the MG with a pledge stage.

Results 1 and 2 can be subjected to further statistical checks by combining the data across the four treatments and then testing for the Composition Effect and Information Effect while accounting for multiple testing along the way. The results are displayed in Table 3. Models (1) and (2) examine trustors' first round behavior on the basis of an OLS regression. The dependent variable is the average number of tokens withdrawn by trustors in the same group, with the unit of observation at the group level. The independent variables include dummy variables for each treatment. Compared to Model (1), Model (2) add the trustees' pledge behavior in the round to the analysis. Models (3) and (4) examine trustors' behavior throughout the entire game. We report the results of a random effects GLS regression, with the average number of tokens withdrawn by

trustors in the same group in round t as the dependent variable. The unit of observation is again at the group level, and there are independent variables dummy variables for each treatment. To account for the collapse of cooperation over time typically observe in VCM games (Chaudhuri, 2011; Fischbacher and Gächter, 2010), we allow for a linear time trend. As in the analysis of the first round behavior in Models (1) and (2), Model (4) is a version of Model (3) augmented by the trustees' pledge behavior. Throughout Table 3, the baseline (reference level) is trustors' takings in the treatment in which they are unknowingly matched with a less trustworthy trustee.

Table 3: Effect of trustee's trustworthiness on trustors' withdrawal decisions.

Dep. var:	av. token withdrawn by trustors in round t			
	first round		all rounds	
	(1)	(2)	(3)	(4)
Hidden*Less trustworthy trustee	Ref.	Ref.	Ref.	Ref.
Hidden*More trustworthy trustee	-2.593*	-2.241	-1.384	-1.323
	(1.316)	(1.267)	(1.240)	(1.063)
Obs*Less trustworthy trustee	-0.328	-0.692	0.863	0.581
	(1.592)	(1.532)	(1.494)	(1.281)
Obs*More trustworthy trustee	-4.730***	-4.157***	-3.879***	-3.615***
	(1.195)	(1.155)	(1.123)	(0.964)
Trustee's pledge	–	0.216***	–	0.158***
		(0.049)		(0.020)
Time trend	–	–	0.881***	0.744***
			(0.071)	(0.074)
Constant	14.89***	12.51***	13.18***	11.64***
	(1.039)	(1.134)	(0.999)	(0.884)
Obs.	228	228	1048	1048

Note: Table 3 displays the OLS coefficients of the linear regression (1) and (2), and the GLS coefficients of the random effects regression (3) and (4). Standard errors in parentheses. The unit of observation is at the group level. Stars indicates significant differences from the Hidden*Promise broken. *p <0.05, **p <0.01, ***p <0.001.

The econometric results are consistent, and add nuance, to our previous results. Starting with Result 1, the evidence remains non-supportive on the Composition Effect. This can be seen in the row that reports the coefficients for the treatment in which the trustee's higher trustworthiness is unobservable (Hidden*More trustworthy trustee). Since the reference treatment and this treatment are informationally equivalent in the first round, we

focus on the coefficients for Models (3) and (4) and recover negative coefficients equivalent to around 1.3 fewer tokens taken per round on average. These coefficients are not statistically significant, supporting Result 1. Similarly for Result 2, the evidence remains supportive of an asymmetric Information Effect: Comparing for the same level of trustworthiness the coefficients in the Observable and Hidden treatment, we find for high trustworthiness that the coefficients in round 1 (Models 1 and 2) and throughout the game (Models 3 and 4) are significantly lower when more trustworthiness is observable (Wald test: $p < 0.05$ in all models). When groups are matched with a less trustworthy trustee, the coefficients turn from negative to positive as play progresses from the first round (Models 1 and 2) to the entire game (Models 3 and 4), but do not reach statistical significance.

Table 3 also reaffirms the joint impact of the Composition and the Information Effect on efficiency. Compared to the baseline of being unknowingly matched with a less trustworthy trustee, trustors take about 4.7 fewer tokens (24% of endowment) on average when they know that they are matched with a more trustworthy trustee ($p < 0.001$).[16]

Moving on to the final test, Hypothesis 3 predicts a Pledge Effect: For the same pledge by a trustee, trustors withdraw less when matched with a trustee revealed as more trustworthy than when matched with a trustee revealed as less trustworthy. If confirmed, the Pledge Effect would demonstrate that information on trustworthiness is able to leverage the pledge stage of a social dilemma in order to facilitate efficiency.

Table 3 already alludes to the role that trustee's pledges have for trustors' withdrawal decisions: Averaged across all treatment conditions, trustors respond to less cooperative pledges by trustees by cooperating less in return. For every token pledged to be taken by the trustee, trustors withdraw roughly an additional 0.2 tokens, both in round 1 (Model 2) and across all rounds (Model 4).

Since the hypothesized Pledge Effect relies on trustors' knowledge of the trustee's trustworthiness, our analysis progresses beyond Table 3. Specifically, we test whether trustors' withdraw different amount of tokens in the Observable treatment when a more trustworthy trustee is in the group. For this treatment, Table 4 shows the results of a regression of trustors' withdrawals on trustworthiness and trustee's pledges. Model 1 reports on an OLS estimation for round 1 behavior, and Model 2 on a GLS estimation for average behavior across all rounds. We reconfirm the finding of Table 3 that trustors' taking behavior responds to the trustee's pledge: In round 1, trustors take an average of around 0.4 tokens more for every token that the trustee pledges to take and around 0.25 tokens more across all rounds. Table 4 also shows that there is a significant interaction effect between the trustee's pledge and their trustworthiness: when the trustee is more

[16]This provides a test for hypothesis 2 in the pre-registration and the results show that this hypothesis is supported by the data.

trustworthy, trustors raise their withdrawals by only half as much, around 0.2 tokens more in round 1 and around 0.1 tokens more across all around for an extra token pledged to be taken by the trustee. This interaction effect accords with the Pledge Effect predicted in Hypothesis 3: Higher observable trustworthiness of the trustee significantly reduces trustors' withdrawal for the same trustee pledge.

Table 4: Effect of trustee's pledge and trustworthiness on trustors' withdrawal decisions.

Dep var:	av. withdrawals by trustors in round t	
	Round 1	All round
	(1)	(2)
Trustee's pledge	0.405***	0.242***
	(0.100)	(0.046)
Trustee's pledge*More trustworthy trustee	-0.214*	-0.101*
	(0.101)	(0.050)
Time trend	–	0.823***
		(0.099)
Constant	8.73***	8.53***
	(0.823)	(0.591)
Obs.	135	630
Cluster		135

Note: Table 4 displays the OLS coefficients of the linear regression (1) for round 1 and the GLS coefficients of the random effects regression (2) for all rounds. Standard errors in parentheses. The unit of observation is at the group level. Stars indicates significant differences from the Hidden*Promise broken. *p <0.05, **p <0.01, ***p <0.001.

Result 3 (Pledge Effect) Controlling for a trustee's pledge, trustors withdraw on average fewer tokens when the trustee has been revealed as more trustworthy, both in the first round and across all rounds.

Result 3 has important ramifications for our thinking about the pledge stage in social dilemmas. In our experiment, group members' pledges affect withdrawals in the maintenance game, despite being cheap talk. On average, higher (less cooperative) pledges by others make group members choose more (less cooperative) withdrawal levels. This is evidence that the structured communication provided by the pledge stage has a systematic impact on takings decisions and, hence, efficiency. Trustworthiness information, when available, also impacts on efficiency because the statistical association between trustors' takings decisions and a trustee's pledges differs when the trustee is more trustworthy. Specifically, withdrawal levels remain more cooperative (increase less) given the trustee's pledge when the trustee is more trustworthy, thus increasing efficiency.

6 Conclusion

In social dilemmas, pledges are used by parties to publicly announce their future cooperative behavior. The notion that providing for such structured pre-play communication can enhance cooperation has increasingly informed how international climate cooperation is conducted under the framework of the UNFCCC. In preparation for the 2015 Paris Accord, a pledge stage was introduced for the first time and countries were mandated to make pledges called 'Intended Nationally Defined Contributions'. Since then, INDCs are supposed to convey information about countries' intended future cuts in greenhouse gas emissions. However, as has been pointed out, not all countries' INDCs are likely to be taken equally seriously.

In the present paper, we examined how differences in trustworthiness affect outcomes in social dilemmas with a pledge stage, as foreseen in the post-Paris framework. Using the repeated Maintenance Game as a persuasive abstraction of the problem of global climate change, we manipulated the composition of three-player groups by inserting a group member of more or less trustworthiness, as determined by trustee behavior in a preceding Promise-Keeping Game, and varied whether the other group members, all trustors in a preceding PKG, could observe that trustworthiness or not.

Based on data from 795 participants of an online implementation of the experiment, we first confirmed that our manipulation succeeded in seeding the groups with behaviorally distinct types of trustees in terms of cooperativeness and pledge credibility. Testing our three predictions against the experimental data, we then found that two of our three predictions were borne out by the experimental data. The first prediction, the Composition Effect, was not confirmed by the data. While cooperation in groups with a more trustworthy trustee as the third group member tended to be higher than cooperation in groups with a less trustworthy trustee, the difference was not significant when trustworthiness was not observable. We interpret this finding as indicating that the behavioral differences between more and less trustworthy trustee need favorable circumstances such as many rounds of interaction in order to shift outcomes in the social dilemma alone. This implies that without additional information, the inherent trustworthiness characteristics of participating parties struggle to affect outcomes in international climate cooperation with pledges.

The second prediction, the Information Effect, was borne out by the experimental data, but in an asymmetric fashion. In contrast to the earlier literature, we found that a social dilemma will be resolved more efficiently when group members are knowingly matched with a more trustworthy co-player. This would suggest the climate cooperation greatly benefits from countries' awareness that the other countries have honored their promises on other matters of international concern in the past. Efforts to assess and publicize trustworthy behavior (such as the LSE folks) are therefore conducive to the

efficiency with which the social dilemma is resolved. The converse is not true, however: We found no evidence that knowingly interacting with a less trustworthy co-player decreased cooperation compared to unknowingly interacting with one. Reports of good behavior are therefore the only way of influencing the efficiency of cooperation outcomes.

The third prediction, the Pledge Effect, also established itself in the experiment: Group members behaved more cooperatively toward the pledges of more trustworthy co-players. This suggests that part of the Information Effect comes from the synergies between observable trustworthiness and the pledge stage: Group members expect more credible pledges from more trustworthy co-players and express this belief in the form of cooperating more. The presence of the Pledge Effect could also explain why the Information Effect so clearly asserted itself in our data in contrast to previous experiments without a pledge stage. This highlights one reason why the concept of INDCs introduced in the Paris Accord could have helped improve outcomes in international climate cooperation.

References

Averchenkova, A. and Bassi, S. (2016). Beyond the targets: assessing the political credibility of pledges for the paris agreement. *Policy Brief, Feb. 2016, Grantham Research Center, LSE*.

Barrett, S. and Dannenberg, A. (2016). An experimental investigation into 'pledge and review'in climate negotiations. *Climatic Change*, 138(1):339–351.

Bauer, N., Bertram, C., Schultes, A., Klein, D., Luderer, G., Kriegler, E., Popp, A., and Edenhofer, O. (2020). Quantification of an efficiency–sovereignty trade-off in climate policy. *Nature*, 588(7837):261–266.

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.

Bolton, G. E., Katok, E., and Ockenfels, A. (2005). Cooperation among strangers with limited information about reputation. *Journal of Public Economics*, 89(8):1457–1468.

Burlando, R. M. and Guala, F. (2005). Heterogeneous agents in public goods experiments. *Experimental Economics*, 8(1):35–54.

Cagala, T., Glogowsky, U., Grimm, V., and Rincke, J. (2019). Public goods provision with rent-extracting administrators. *The Economic Journal*, 129(620):1593–1617.

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

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

Chen, X.-P. and Komorita, S. S. (1994). The effects of communication and commitment in a public goods social dilemma. *Organizational Behavior and Human Decision Processes*, 60(3):367–386.

Cherry, T. L., Kallbekken, S., Sælen, H., and Aakre, S. (2021). Can the paris agreement deliver ambitious climate cooperation? an experimental investigation of the effectiveness of pledge-and-review and targeting short-lived climate pollutants. *Environmental Science & Policy*, 123:35–43.

Dariel, A. and Nikiforakis, N. (2014). Cooperators and reciprocators: A within-subject analysis of pro-social behavior. *Economics Letters*, 122(2):163–166.

De Oliveira, A. C., Croson, R. T., and Eckel, C. (2015). One bad apple? heterogeneity and information in public good provision. *Experimental Economics*, 18(1):116–135.

Diekmann, A. (1985). Volunteer's dilemma. *Journal of conflict resolution*, 29(4):605–610.

Dreber, A., Fudenberg, D., and Rand, D. G. (2014). Who cooperates in repeated games: The role of altruism, inequity aversion, and demographics. *Journal of Economic Behavior & Organization*, 98:41–55.

Falkner, R. (2016). The paris agreement and the new logic of international climate politics. *International Affairs*, 92(5):1107–1125.

Fischbacher, U. and Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, 100(1):541–556.

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

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. and Thöni, C. (2005). Social learning and voluntary cooperation among like-minded people. *Journal of the European Economic Association*, 3(2-3):303–314.

Goeschl, T. and Jarke, J. (2017). Trust, but verify? monitoring, inspection costs, and opportunism under limited observability. *Journal of Economic Behavior & Organization*, 142:320–330.

Guido, A., Robbett, A., and Romaniuc, R. (2019). Group formation and cooperation in social dilemmas: A survey and meta-analytic evidence. *Journal of Economic Behavior & Organization*, 159:192–209.

Harstad, B. (2023). Pledge-and-review bargaining: From kyoto to paris. *The Economic Journal*, 133(651):1181–1216.

Heinz, M. and Schumacher, H. (2017). Signaling cooperation. *European Economic Review*, 98:199–216.

Höhne, N., Kuramochi, T., Warnecke, C., Röser, F., Fekete, H., Hagemann, M., Day, T., Tewari, R., Kurdziel, M., Sterl, S., et al. (2017). The paris agreement: resolving the inconsistency between global goals and national contributions. *Climate Policy*, 17(1):16–32.

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

Koessler, A.-K., Page, L., and Dulleck, U. (2021). Public cooperation statements. *Journal of Economic Interaction and Coordination*, pages 1–21.

Lippert, S. and Tremewan, J. (2021). Pledge-and-review in the laboratory. *Games and Economic Behavior*, 130:179–195.

McEvoy, D. M., Haller, T., and Blanco, E. (2022). The role of non-binding pledges in social dilemmas with mitigation and adaptation. *Environmental and Resource Economics*, 81(4):685–710.

Pauw, W. P. and Klein, R. J. (2021). *Making Climate Action More Effective: Lessons Learned from the First Nationally Determined Contributions (NDCs)*. Routledge.

Pogrebna, G., Krantz, D. H., Schade, C., and Keser, C. (2011). Words versus actions as a means to influence cooperation in social dilemma situations. *Theory and Decision*, 71(4):473–502.

Reischmann, A. and Oechssler, J. (2018). The binary conditional contribution mechanism for public good provision in dynamic settings—theory and experimental evidence. *Journal of*

Public Economics, 159:104–115.

Rogelj, J., Den Elzen, M., Höhne, N., Fransen, T., Fekete, H., Winkler, H., Schaeffer, R., Sha, F., Riahi, K., and Meinshausen, M. (2016). Paris agreement climate proposals need a boost to keep warming well below 2 c. *Nature*, 534(7609):631–639.

Vanberg, C. (2008). Why do people keep their promises? an experimental test of two explanations 1. *Econometrica*, 76(6):1467–1480.

Williamson, O. E. (1983). Credible commitments: Using hostages to support exchange. *The American economic review*, 73(4):519–540.

Yamagishi, T., Mifune, N., Li, Y., Shinada, M., Hashimoto, H., Horita, Y., Miura, A., Inukai, K., Tanida, S., Kiyonari, T., et al. (2013). Is behavioral pro-sociality game-specific? pro-social preference and expectations of pro-sociality. *Organizational Behavior and Human Decision Processes*, 120(2):260–271.

Appendix

A-1 Additional Tables

A-1.1 Balance check

Table A-1 summarizes participants demographics by treatments. Using two-sided Fisher exact tests, we find no significant difference between treatments in the percentage of female (p=0.825) and the percentage of participants who indicated English as their native language (p=0.207). Using two-sided Mann-Whitney tests and Fligner-Policello tests of means, we find no significant difference between treatments in age ($p = 0.305$ and $p = 0.306$, respectively), and perceived clarity of the instructions ($p = 0.248$ and $p = 0.350$, respectively). These results suggest that our randomization was successful.

Table A-1: Balance check

	Hidden	Observable
female (%)	45.14	46.03
mean age	38.87	39.86
	(0.621)	(0.570)
English (%)	89.03	91.84
mean clarity	2.64	2.58
	(0.034)	(0.032)
Obs.	344	441

Note: Table A-1 displays the number of participants, the percentage of female, the mean age, the percentage of English native speakers, and the mean perceived clarity of the instructions, by treatments. Standard errors in parentheses.

A-1.2 Manipulation check

Table A-2: Trustees' withdrawal decisions.

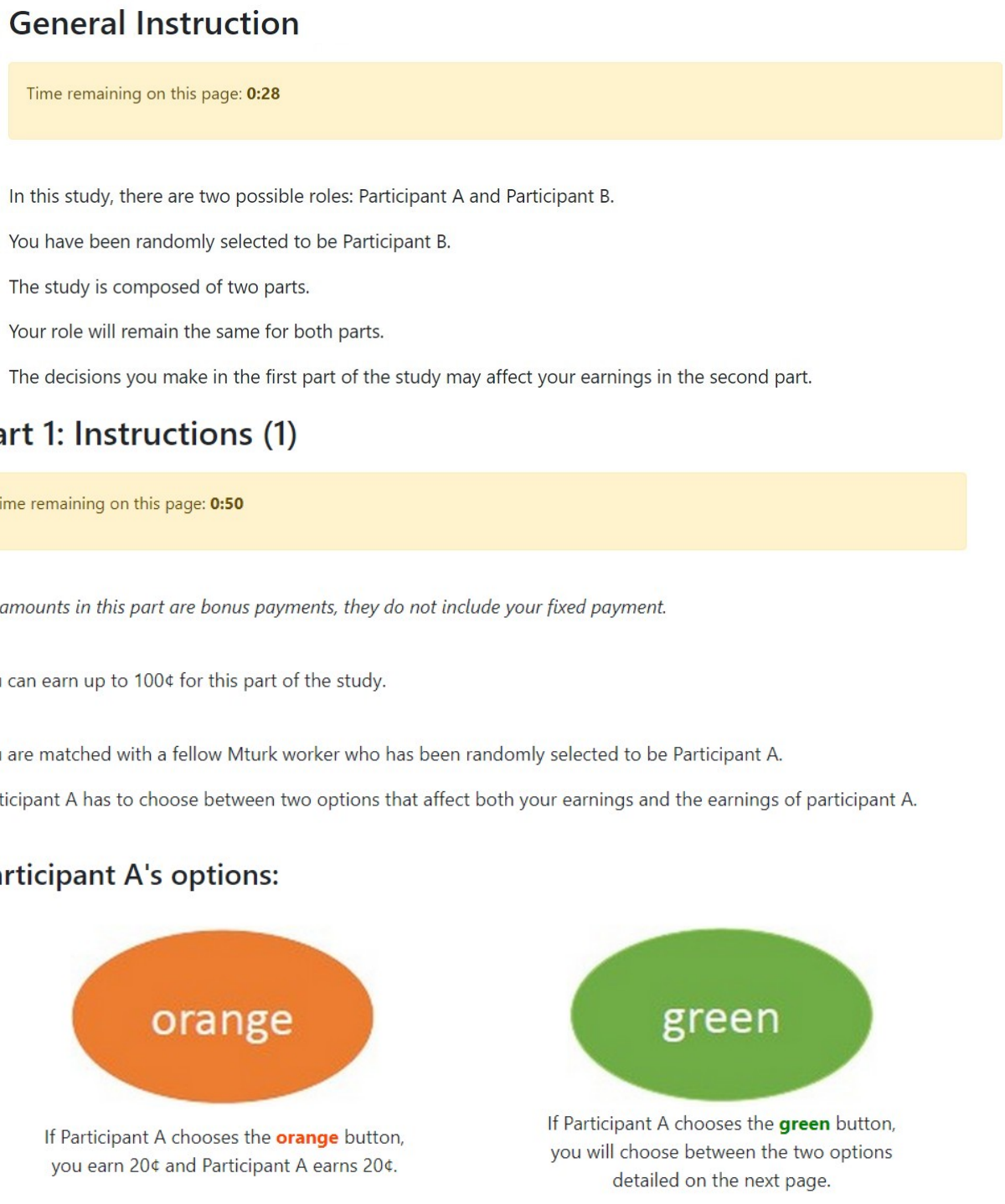
	Untrustworthy Trustee			Trustworthy Trustee		
	Round 1	All rounds	Last 2 rounds	Round 1	All rounds	Last 2 rounds
Takings	14.03	16.06	–	9.98***	12.71***	–
	(0.886)	(0.617)		(0.636)	(0.533)	
Credibility	-2.34	-3.06	-4.07	-1.31	-1.74	-1.85*
	(1.063)	(0.746)	(0.933)	(0.514)	(0.341)	(0.419)
Obs.	61	61	61	167	167	167

Note: Table A-2 displays the mean takings and mean credibility of untrustworthy and trustworthy trustees. Standard errors in parentheses. Stars indicate significant differences between the behavior of untrustworthy and trustworthy trustees. *p <0.05, **p <0.01, ***p <0.001.

A-2 Instructions

A-2.1 Trustees Instructions (Observable Treatment)

We present the screens for a trustee who made the promise to choose the cooperative option and did choose the cooperative option. Screens for a trustee who made the promise to choose the cooperative option and did not choose the cooperative option are the same except that the 'group composition' screen reads 'You promised to choose the purple button in Part 1 and you did **not** do so.' instead of 'You promised to choose the purple button in Part 1 and you did so.'



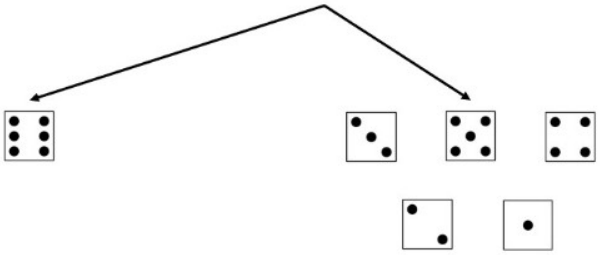
Part 1: Instructions (2)

Time remaining on this page: **0:53**

Your options if Participant A chooses the **green** button:

If you choose the **yellow** button, you earn 100¢ and Participant A earns 10¢.

If you choose the **purple** button, you will roll a virtual die.



If the die lands on a **6**, you earn 50¢ and Participant A earns 10¢.

If the die lands on **any other number**, you earn 50¢ and Participant A earns 100¢.

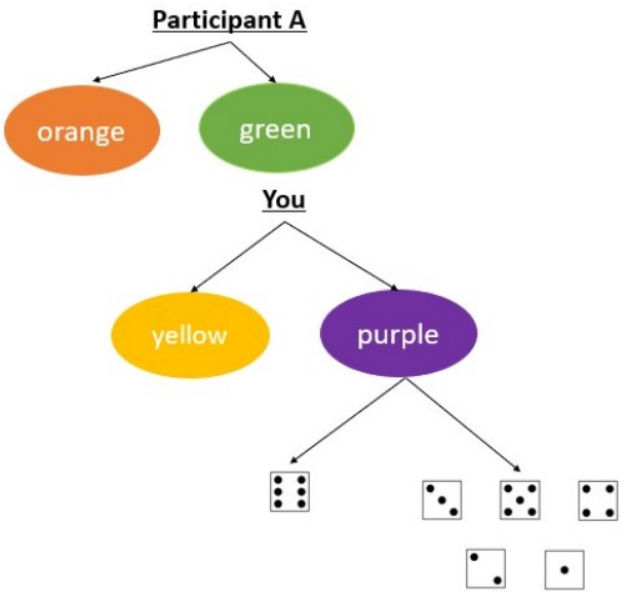
Note that your decision will only be implemented if Participant A chooses the **green** button.

Since you do not know the decision of Participant A when making your decision, we ask you to make your decision for the eventuality that Participant A chooses the **green** button.

Part 1: Instructions (3)

Time remaining on this page: **0:57**

Here is an overview of all the possible action paths and their outcomes.



	Your earnings	Participant A earnings
Participant A chooses **orange**	20¢	20¢
Participant A chooses **green** and you choose **yellow**.	100¢	10¢
Participant A chooses **green**, you choose **purple** and the die lands on a 6.	50¢	10¢
Participant A chooses **green**, you choose **purple** and the die does not land on a 6.	50¢	100¢

You will be informed of the decision of participant A at the time of payment.

Part 1: Instructions (4)

Time remaining on this page: **0:35**

Before Participant A makes a decision between the **green** and the **orange** button, you can make a promise to choose the **purple** button if Participant A chooses the **green** button.

Promises are not binding, which means that even if you decide to make a promise, you are still free to choose either the **purple** button or the **yellow** button.

If you choose to make a promise, Participant A will see the following message:
*« I promise that I will choose the **purple** button ».*

If you choose not to make a promise, Participant A will see the following message:
*« I do not promise that I will choose the **purple** button».*

Participant A will be informed about your promise before making a decision.

Part 1: Comprehension Questionnaire

Time remaining on this page: **2:01**

You need to answer all the questions correctly before moving on to the next page.

If you need help, use the "REMIND ME OF THE GAME" button at the bottom of this page.

Suppose that Participant A chooses the **orange** button:

1) How much do you earn?

○ You earn 10¢.

○ You earn 20¢.

○ You earn 100¢.

2) How much does Participant A earn?

○ Participant A earns 20¢.

○ Participant A earns 50¢.

○ Participant A earns 100¢.

Suppose that Participant A chooses the **green** button and that you choose the **purple** button:

3) How much do you earn?

○ You earn 50¢.

○ You earn 100¢.

○ It depends on the outcome of the die roll.

4) How much does Participant A earn?

○ Participant A earns 10¢.

○ Participant A earns 100¢.

○ It depends on the outcome of the die roll.

5) If you promise to choose the **purple** button, you will have to choose the **purple** button for sure.

○ True

○ False

CHECK MY ANSWERS | REMIND ME OF THE GAME

Part 1: Promise

Time remaining on this page: **0:48**

Select the message that you want to send to Participant A:

○ I promise that I will choose the **purple** button.

○ I do not promise that I will choose the **purple** button.

Your message will be displayed on Participant A's screen before Participant A makes a decision.

REMIND ME OF THE GAME

Part 1: Decision

Time remaining on this page: **0:31**

Select one of the buttons below to make a decision:

Note that your decision will only be implemented if Participant A chooses the **green** button.

Since you do not know the decision of Participant A at this point, we ask you to make your decision for the eventuality that Participant A chose the **green** button.

REMIND ME OF THE GAME

Part 2: Instructions (1)

Time remaining on this page: **2:30**

All amounts in this part are bonus payments, they do not include your fixed payment.

You can earn up to 400¢ in this part of the study.

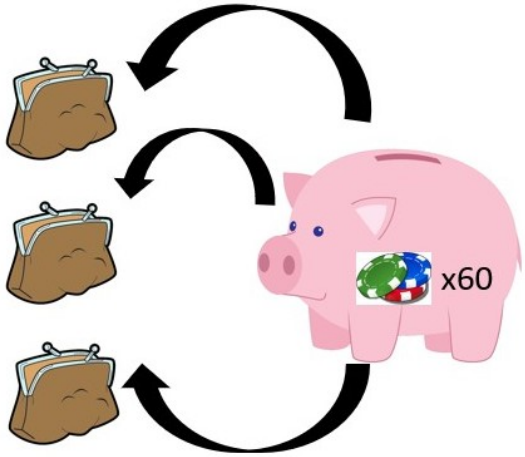
You are now matched with two other Mturk workers to form a group of three. Your group receives a piggy bank containing 60 tokens.

Each group member can decide to leave the tokens in the piggy bank or to take some tokens for his or her own wallet.

Each token left in the piggy bank yields 5¢ to each group member.

Each token taken by a group member yields 10¢ to him-/herself, and nothing to the other group members.

Each group member can take up to 20 tokens from the piggy bank.



Your earnings:

Your earnings for this part are computed as follows:
10¢ x [Number of tokens you take for yourself] + 5¢ x [sum of tokens left in the piggy bank].

Part 2: Instructions (2)

Time remaining on this page: **1:04**

You will play this game for 5 rounds with the same group members.

Unfolding of a round:

A round is composed of 4 stages:

1. In the **Intention Stage**, you have to indicate how many tokens you intend to take. Intentions are not binding, which means that you and your fellow group members are free to change their mind.

2. In the **Announcement Stage**, intentions are made public.

3. In the **Decision Stage**, you decide how many tokens to actually take from the piggy bank.

4. In the **Feedback Stage**, you will receive information about the number of tokens taken by each group member, as well as their earnings for the current round.

Payment:

At the end of the study, one out of the five rounds will be selected at random for payment. You will be paid according to your decision, and the decisions of your fellow group members in this particular round.

Part 2: Instructions (3)

Time remaining on this page: **0:57**

Group formation:

The composition of your group is based on your role and the role of the two other group members in Part 1 of this study.

Your group is composed of one participant B (you) and two participants A (different from the ones you encountered in Part 1).

At the beginning of Part 2, the two other participants in your group will receive information about your decision in Part 1.

More specifically, the two other participants in your group will be informed about whether you made a promise to choose the **purple** button in Part 1 and whether you kept your promise or not.

Part 1: Comprehension Questionnaire

Time remaining on this page: **3:55**

You need to answer all the questions correctly before moving on to the next page.

If you need help, use the "REMIND ME OF THE GAME" button at the bottom of this page.

1) How much do you earn if you take 0 token from the piggy bank and the two other members in your group also take 0 token each?

¢.

2) How much do you earn if you take 20 tokens from the piggy bank and the two other members in your group also take 20 tokens each?

¢.

3) How much do you earn if you take 10 tokens from the piggy bank, another group member takes 15 tokens and the last group member takes 5 tokens?

¢.

4) My group members will remain the same throughout the 5 rounds.

○ True

○ False

CHECK MY ANSWERS | REMIND ME OF THE GAME

Part 2: Instruction (4)

Time remaining on this page: **0:56**

Each member of your group has been given a unique label: P1, P2 or P3. These labels will remain the same throughout the 5 rounds.

Group Composition

You are **P1**. You promised to choose the **purple** button in Part 1 and you did so.

P2 was a participant A in Part 1.

P3 was a participant A in Part 1.

Part 2

Time remaining on this page: **1:57**

Each member of your group has been given a unique label: P1, P2 or P3. These labels will remain the same throughout the 5 rounds.

Group Composition

You are **P1**. You promised to choose the **purple** button in Part 1 and you did so.

P2 was a participant A in Part 1.

P3 was a participant A in Part 1.

Intention Stage (1/5)

In this stage, you have to indicate how many tokens (out of 20) you intend to take from the piggy bank.

Your statement will be made public to your fellow group members in the next stage.

Intentions are not binding, which means that you will be free to change your mind in the Decision Stage.

I intend to take [ ] **tokens** from the piggy bank.

REMIND ME OF THE GAME NEXT

Part 2

Time remaining on this page: **0:26**

Each member of your group has been given a unique label: P1, P2 or P3. These labels will remain the same throughout the 5 rounds.

Group Composition

You are **P1**. You promised to choose the **purple** button in Part 1 and you did so.

P2 was a participant A in Part 1.

P3 was a participant A in Part 1.

Announcement Stage (1/5)

You intend to take 0 token(s).

P2 intend to take 0 token(s).

P3 intend to take 0 token(s).

Part 2

Time remaining on this page: **1:14**

Each member of your group has been given a unique label: P1, P2 or P3. These labels will remain the same throughout the 5 rounds.

Group Composition

You are **P1**. You promised to choose the **purple** button in Part 1 and you did so.

P2 was a participant A in Part 1.

P3 was a participant A in Part 1.

Decision Stage (1/5)

Reminder intentions:
You intend to take 0 token(s).

P2 intend to take 0 token(s).

P3 intend to take 0 token(s).

In this stage, you have to decide how many tokens (out of 20) to actually take from the piggy bank.

I take [ ] **tokens** from the piggy bank.

REMIND ME OF THE GAME NEXT

Part 2

Time remaining on this page: **0:43**

Each member of your group has been given a unique label: P1, P2 or P3. These labels will remain the same throughout the 5 rounds.

Group Composition

You are **P1**. You promised to choose the **purple** button in Part 1 and you did so.

P2 was a participant A in Part 1.

P3 was a participant A in Part 1.

Feedback Stage (1/5)

	Intent to take (in tokens)	Decision to take (in tokens)	Earnings for this round (in ¢)
You	0	0	200
P2	0	10	300
P3	0	10	300

If this round is selected for payment, you will earn 200¢ for this part of the study.

A-2.2 Trustees Instructions (Hidden Treatment)

Since the instructions for Part 1 are the same in both the Observable and the Hidden treatment, we only reproduce the screen for Part 2 below.

Part 2: Instructions (2)

Time remaining on this page: **1:06**

You will play this game for 5 rounds with the same group members.

Unfolding of a round:

A round is composed of 4 stages:

1. In the **Intention Stage**, you have to indicate how many tokens you intend to take. Intentions are not binding, which means that you and your fellow group members are free to change their mind.

2. In the **Announcement Stage**, intentions are made public.

3. In the **Decision Stage**, you decide how many tokens to actually take from the piggy bank.

4. In the **Feedback Stage**, you will receive information about the number of tokens taken by each group member, as well as their earnings for the current round.

Payment:

At the end of the study, one out of the five rounds will be selected at random for payment. You will be paid according to your decision, and the decisions of your fellow group members in this particular round.

Part 2: Instructions (3)

Time remaining on this page: **0:57**

Group formation:

The composition of your group is based on your role and the role of the two other group members in Part 1 of this study.

Each group is composed of one participant B (you) and two participants A (different from the ones you encountered in Part 1).

Part 2: Instruction (4)

Time remaining on this page: **0:58**

Each member of your group has been given a unique label: P1, P2 or P3. These labels will remain the same throughout the 5 rounds.

Group Composition

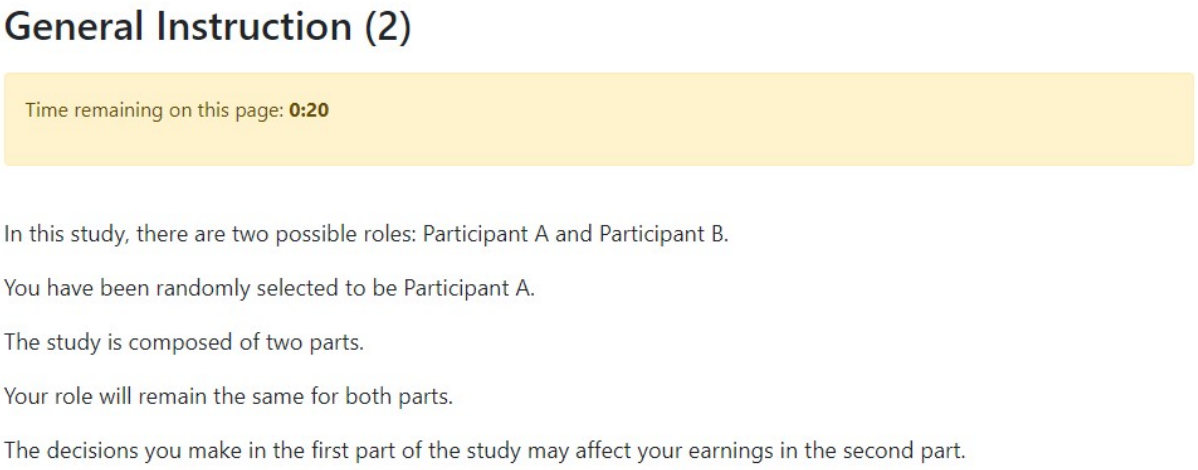
You are **P1**.

P2 was a participant A in Part 1.

P3 was a participant A in Part 1.

A-2.3 Trustors Instructions (Observable Treatment)

Below are the instructions provided to the trustors when they differ from the trustees' instructions. As before, we provide instructions for Part 2 for a trustor who has been matched with a trustee who made the promise to choose the cooperative option and did choose the cooperative option in Part 1. Screens for a participant matched with a trustee who made the promise to choose the cooperative option and did not choose the cooperative option are the same except that the 'group composition' screen reads 'He/she promised to choose the purple button and he/she did **not** do so.' instead of 'He/she promised to choose the purple button and he/she did so.'



Part 1: Instructions (1)

Time remaining on this page: **0:55**

All amounts in this part are bonus payments, they do not include your fixed payment.

You can earn up to 100¢ for this part of the study.

You are matched with a fellow Mturk worker who has been randomly selected to be Participant B.

You have to choose between two options that affect both your earnings and the earnings of participant B.

Your options:

If you choose the **orange** button,
you earn 20¢ and Participant B earns 20¢.

If you choose the **green** button,
Participant B will choose between the two options
detailed on the next page.

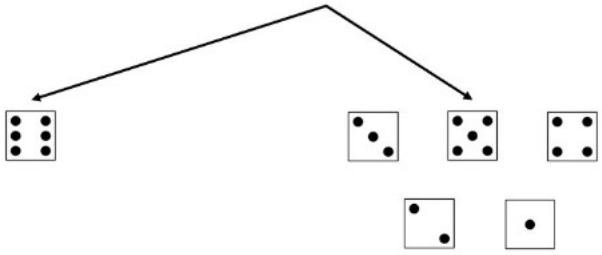
Part 1: Instructions (2)

Time remaining on this page: **0:51**

Options of Participant B if you choose the green button:

If Participant B chooses the **yellow** button,
you earn 10¢ and Participant B earns 100¢.

If Participant B chooses the **purple** button,
he/she will roll a virtual die.



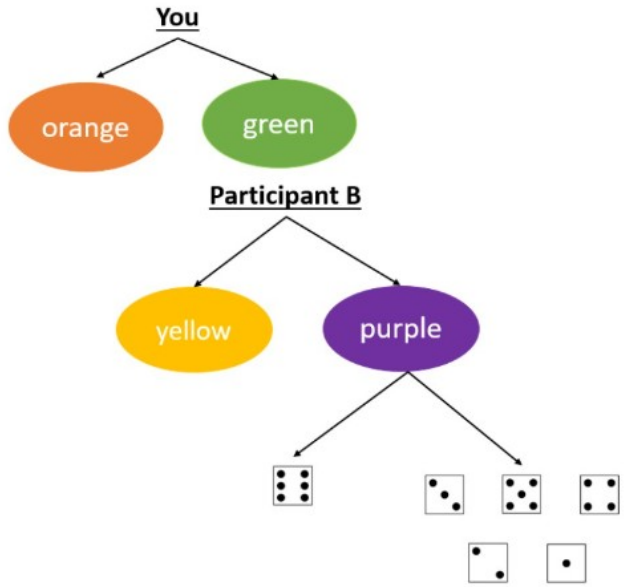
If the die lands on a **6**, you earn 10¢ and Participant B earns 50¢.

If the die lands on **any other number**, you earn 100¢ and Participant B earns 50¢.

Part 1: Instructions (3)

Time remaining on this page: **0:57**

Here is an overview of all the possible action paths and their outcomes.



	Your earnings	Participant B earnings
You choose **orange**	20¢	20¢
You choose **green** and Participant B chooses **yellow**.	10¢	100¢
You choose **green**, Participant B chooses **purple** and the die lands on a 6.	10¢	50¢
You choose **green**, Participant B chooses **purple** and the die does not land on a 6.	100¢	50¢

If you choose the **green** button, you will be informed of Participant B's decision at the time of payment.

Part 1: Instructions (4)

Time remaining on this page: **0:36**

Before you make a decision between the **green** and the **orange** button, Participant B can make a promise to choose the **purple** button if you choose the **green** button.

Promises are not binding, which means that even if Participant B decides to make a promise, he/she is still free to choose either the **purple** button or the **yellow** button.

If Participant B chooses to make a promise, you will see the following message:
*« I promise that I will choose the **purple** button ».*

If Participant B chooses not to make a promise, you will see the following message:
*« I do not promise that I will choose the **purple** button».*

You will be informed about Participant B's promise before making a decision.

Part 1: Comprehension Questionnaire

Time remaining on this page: **2:02**

You need to answer all the questions correctly before moving on to the next page.

If you need help, use the "REMIND ME OF THE GAME" button at the bottom of this page.

Suppose that you choose the **orange** button:

1) How much do you earn?

○ You earn 10¢.

○ You earn 20¢.

○ You earn 100¢.

2) How much does Participant B earn?

○ Participant B earns 20¢.

○ Participant B earns 50¢.

○ Participant B earns 100¢.

Suppose that you choose the **green** button and that Participant B chooses the **purple** button:

3) How much do you earn?

○ You earn 10¢.

○ You earn 100¢.

○ It depends on the outcome of the die roll.

4) How much does Participant B earn?

○ Participant B earns 50¢.

○ Participant B earns 100¢.

○ It depends on the outcome of the die roll.

5) If participant B promises to choose the **purple** button, participant B will choose the **purple** button for sure.

○ True

○ False

CHECK MY ANSWERS | REMIND ME OF THE GAME

Part 1: Decision

Time remaining on this page: **0:35**

Participant B's message:

*I do not promise that I will choose **purple**.*

Select one of the buttons below to make a decision:

REMIND ME OF THE GAME

Part 1: Decision

Time remaining on this page: **0:27**

Participant B's message:

*I promise that I will choose **purple**.*

Select one of the buttons below to make a decision:

REMIND ME OF THE GAME

Part 2: Instructions (3)

Time remaining on this page: **0:58**

Group formation:

The composition of your group is based on your role and the role of the two other group members in Part 1 of this study.

Your group is composed of one participant B (different from the one you encountered in Part 1) and two participants A (including you).

At the beginning of Part 2, you will receive information about the decision of your group member who was allocated the role of participant B in Part 1.

More specifically, you will be informed about whether participant B made a promise to choose the **purple** button in Part 1 and whether participant B kept his/her promise or not.

Part 2: Instruction (4)

Time remaining on this page: **0:36**

Each member of your group has been given a unique label: P1, P2 or P3. These labels will remain the same throughout the 5 rounds.

Group Composition

P1 was a participant B in Part 1. He/she promised to choose the **purple** button and he/she did so.

You are **P2**.

P3 was a participant A in Part 1.

A-2.4 Trustors Instructions (Hidden Treatment)

Part 2: Instructions (3)

Time remaining on this page: **0:58**

Group formation:

The composition of your group is based on your role and the role of the two other group members in Part 1 of this study.

Each group is composed of one participant B (different from the one you encountered in Part 1) and two participants A (including you).

Part 2: Instruction (4)

Time remaining on this page: **0:48**

Each member of your group has been given a unique label: P1, P2 or P3. These labels will remain the same throughout the 5 rounds.

Group Composition

P1 was a participant B in Part 1.

P2 was a participant A in Part 1.

You are **P3**.