# Communicating Intentions in Noisy Repeated Games*

First complete version: 3/16/2016

This version: 4/1/16

Antonio A. Arechar[a], Anna Dreber[b], Drew Fudenberg[c] and David G. Rand[d]

[a]Department of Psychology, Yale University
[b]Department of Economics, Stockholm School of Economics
[c]Departments of Economics, Harvard University and Yonsei University
[d]Department of Psychology, Department of Economics, Cognitive Science Program, School of Management, Yale University

**Abstract**

To explore the role of communication in promoting cooperation, we let participants indicate their intended action in a repeated game experiment where actions are implemented with errors. Even though communication is cheap talk, we find that the majority of participants communicate honestly. As a result, communication has a positive effect on cooperation when the payoff matrix makes the returns to cooperation high. When the payoff matrix gives a low return to cooperation, conversely, there is a negative effect of communication on cooperation. These results suggest that cheap talk communication, which is a common feature of daily life, can promote cooperation in repeated games, but only when there is already a self-interested motivation to cooperate.

Keywords: cooperation, communication, prisoner's dilemma, repeated games, intentions

JEL codes: C7, C9, D00

* We thank the Jan Wallander and Tom Hedelius Foundation (Svenska Handelsbankens Forskningsstiftelser), the Knut and Alice Wallenberg Foundation, the John Templeton Foundation, and National Science Foundation grant SES- 1258665 for financial support, and Tore Ellingsen and Emmanuel Vespa for helpful conversations and comments.

## 1. Introduction

Understanding when and how repeated interaction leads to cooperation in social dilemmas is a key issue for economics and other social sciences. The existing theory of repeated games is of only partial use for understanding this cooperation, as repeating a game never eliminates any of the static equilibria. Moreover, experiments show that although cooperation tends not to be a long-run outcome when it cannot be supported by equilibrium, it is not true that people always cooperate when cooperation *can* be one of the equilibrium outcomes (Dal Bó 2005, Dal Bó and Frechette 2012, 2015, Fudenberg et al. 2012, Rand and Nowak 2013). It is thus important to develop a richer and more detailed body of experimental results about when cooperation does arise.

A central element of cooperation in repeated games outside the laboratory is communication: participants in most real-world repeated interaction settings such as relationships between colleagues, neighbors, friends, or romantic partners are able to communicate with each other, and do so regularly. Yet this issue has received little prior attention in the experimental literature on infinitely repeated games.[1] Thus we conduct an experiment to investigate how communication affects cooperation in the context of an infinitely repeated prisoner's dilemma with imperfect or "noisy" public monitoring of intended actions.

We focus on games with noise as they involve a natural topic about which to communicate, namely the intended actions. Many interactions outside of the lab have some sort of noise or random events that prevent the players' intentions from being fully inferred from their actions: Bad outcomes can occur despite high effort, and friends may be too busy or sick to help. Noise thus leads actions to sometimes differ from intentions, so that any reciprocally cooperative strategy must sometimes punish accidental defections in order to provide any incentive at all for others to cooperate. As a result, noise reduces cooperation when intentions cannot be observed, both theoretically (Kandori 1992) and in the lab (Aoyagi and Frechette 2009, Fudenberg et al. 2012).

[1] Previous experiments on communication in repeated games have only considered finitely repeated games without noise. In these games cooperation is not an equilibrium, and in the absence of communication it eventually unravels (Embrey et al. 2014). Bochet et al. (2006) find that verbal communication in chat rooms is almost as efficient as face-to-face communication when it comes to increasing contributions in a 10-period public goods game, while numerical communication (via computer terminals) has no effect on contributions, but as their participants only played one iteration of the ten-period game it is hard to know if the observed behavior would be robust to feedback and learning.

In our experiment, participants played an infinitely repeated prisoner's dilemma with noise and communication. Specifically, in each period participants chose both their intended action and a binary message indicating the action they intended to play. The messages were transmitted without error, but there was a constant probability (known to the participants) that the *action* they chose was not the one that was implemented. The payoffs at each stage depended only on the implemented actions - the messages were a form of "cheap talk" with no direct payoff consequences. In this game, allowing for communication does not change the set of pure-strategy equilibria.[2] If, however, participants rarely lie, and believe that others also rarely lie, then communication can transform a game with imperfect monitoring into one where intentions are perfectly observed, which can permit cooperation to be an equilibrium outcome when it would not be otherwise. In addition, there is experimental evidence that players in noisy repeated games attempt to infer their partner's intentions based on the past history of play (Rand et al. 2015). This suggests that the restricted form of communication allowed by our protocol will be salient to most participants, and (if enough of the messages are truthful) could help to promote cooperation.

Because play in some repeated game experiments systematically changes over the course of the session (e.g., Dal Bó and Frechette 2015, Embrey et al. 2014), we let participants play at least 8 iterations of the repeated game. It turns out that there was little apparent change in play over the course of a session, but our design also lets us study how the honesty of participants unfolds over the course of a supergame. Here we find that participants are less likely to deceive their partners as the game develops, and in particular become more likely to admit to defection.

We test the impact of communication under two different payoff treatments, where we vary the rewards to cooperation by using two different payoff matrices. In the *"high*" treatment the payoff matrix and other parameters (error rate and continuation probability) are such that there are cooperative equilibria that use simple strongly symmetric strategies such as "Grim," which says to start out cooperating but defect forever once one defect is observed. Importantly, though, in this treatment "Always Defect" risk-dominates "Grim", meaning that Always Defect is the best response to a population in which half the players use one of these strategies and the rest use the other. Here participants cooperated in the first period of a new interaction 47% percent of the time in the absence of communication, but 60% in the

[2] More generally it has no impact on the set of perfect public equilibria (Fudenberg, Levine and Maskin (1994)); its effect on the larger set of mixed equilibria in private strategies is not currently known. In contrast, communication is known to enlarge the set of equilibrium outcomes in repeated games with imperfect *private* monitoring, see Compte (1998), Kandori and Matsushima (1998), and Fudenberg and Levine (2007).

treatment with communication, so communication had a substantial positive effect on cooperation. Moreover, in this treatment most of the participants sent honest messages.

In the "*low*" treatment, the return to joint cooperation is low enough that cooperation cannot be supported by strongly symmetric strategies such as Grim, though it could be if intentions were perfectly observed. Here there is only 39% first-period cooperation without communication, while introducing communication leads this cooperation rate to drop to 28%. Thus, unlike in the *high* treatment, here communication has, if anything, a negative effect on cooperation. Furthermore, in this treatment, participants sent dishonest messages more often.

We also apply the "structural frequency estimation method" (SFEM) introduced by Dal Bó & Frechette (2011) to our data. The SFEM results also suggest high shares of honest behavior and that participants played strategies that conditioned on messages, particularly in the *high* treatment; these findings are reinforced by our descriptive analyses of the data.

Our tentative interpretation of these findings is the following. First of all, the reason that there is relatively little cooperation in the *high* treatment without communication is the strategic uncertainty faced by the players: even though it would be the best response to use a conditionally cooperative strategy if all other players did, the loss incurred when meeting a non-cooperator is too large to make cooperation worthwhile when only half of the population is willing to cooperate.[3] In this treatment communication helps because it has the potential to increase long-run payoffs by facilitating coordination on the cooperative equilibrium: players tended to be honest, which makes cooperative arrangements more rewarding and so makes players more willing to risk initial cooperation. As a result, a substantial fraction of players learn to cooperate, which benefits them. However, in the *low* treatment the message "I meant to cooperate" isn't credible, because cooperation isn't supported by an equilibrium. Here not only does communication not help support cooperation, it reduces it, perhaps because it makes the players more suspicious of one another. Regardless of whether our explanation is correct, the data shows an interesting connection between strategic incentives and honesty.

Our past work on the role of intentions in noisy repeated games (Rand et al. 2015) shows that when both the partner's intended and actual action are revealed, most people condition only on intentions and ignore the realized action, and moreover that this conditioning leads to higher cooperation rates in settings where cooperative equilibria exist. Our results here show that cheap talk about intentions gets some of this benefit, but not all of it: We find that

[3] This is consistent with the theoretical model of Blonski et al (2011) on cooperation in repeated games with observed actions. We discuss related experimental finding in section 3.

communication is only effective in raising cooperation levels in the *high* treatment where cooperative equilibria exist even without revealed intentions (the overall cooperation with communication is 44%, compared to 33% without it). However, in the *low* treatment without cooperative equilibria, when intentions are hidden by noise, adding communication does not help, in contrast to the observed-intentions treatment of Rand et al. (2015): here the overall cooperation rate was 21% with communication and 25% without it.

## 2. Experimental design

We study infinitely repeated prisoner dilemmas with a constant continuation probability of $\delta=7/8$. This means that in each period of each supergame, there is a probability of 7/8 that the particular supergame continues, and a probability of $1-\delta$ that the particular game ends and participants are rematched to play another supergame.[4]

In all treatments, there is a known constant error probability of $E=1/8$ that an intended action is not implemented but changed to the opposite action. Participants are not informed about the intended action of the other player but only the realized action and whether their own intended action was implemented or not.

We used a 2x2 design to test the impact of communication under two different treatments. First, we varied whether or not communication is possible. In our communication treatments, participants had to send a message indicating their intended action (on the same screen in which they make their actual choice). We used a stage game where cooperation and defection take the "benefit/cost" (b/c) form, where cooperation means paying a cost c to give a benefit b to the other player, while defection gives 0 to each player.[5] See Figure 1 where payoffs are denoted in points. We used neutral language, with cooperation denoted as "action A" and defection labelled "action B"; in the communication treatments, participants chose between sending messages "I chose A" or "I chose B", but for clarity we will refer to these as C and D in our analyses. In the control treatments, there were no such messages to be sent.

[4] Sessions consisted of up to 20 supergames with lengths pre-generated according to the specified geometric distribution, such that in each session every sequence of interactions had similar lengths, i.e.: 7, 6, 11, 5, 8, 1, 19, 12, 3, 5, 10, 4, 15, 5, 7, 14, 1, 10, 7, and 2. This allows us to avoid cross-treatment noise introduced by stochastic variation in game lengths between treatments. In our 7th and 10th sessions, however, one of the games was accidentally skipped; we find no evidence that this affects any of our results.

[5] The prisoner's dilemma is of course more general than this, but the b/c setup fulfills the criteria of having the short-run gain to playing D instead of C being independent of the other player's action.

**Realized payoffs**

*Low* (b/c = 1.5)

| | C | D |
|---|---|---|
| C | 1,1 | -2,3 |
| D | 3,-2 | 0,0 |

*High* (b/c = 2)

| | C | D |
|---|---|---|
| C | 2,2 | -2,4 |
| D | 4,-2 | 0,0 |

**Expected payoffs**

*Low* (b/c = 1.5, E = 1/8)

| | C | D |
|---|---|---|
| C | 0.875, 0.875 | -1.375, 2.375 |
| D | 2.375, -1.375 | 0.125, 0.125 |

*High* (b/c = 2, E = 1/8)

| | C | D |
|---|---|---|
| C | 1.75, 1.75 | -1.25, 3.25 |
| D | 3.25, -1.25 | 0.25, 0.25 |

***Figure 1.*** *Payoff matrices for each treatment. Payoffs are in points.*

We study two different payoff matrices with varying rewards to cooperation. In the *low* treatment the b/c ratio is 1.5, whereas in the *high* treatment this ratio is 2. As in prior work (Fudenberg et al. 2012, Rand et al. 2015), participants were presented with both the b/c representation of the game and the resulting pre-error payoff matrix as in Figure 1 (albeit with neutral language), but not the expected payoff matrix.

For each treatment we performed three sessions. Within a session, a single treatment was implemented. Participants were informed about the specifics of their treatment but were unaware of the existence of other treatments. This leaves us with 12 sessions and a total sample size of 312 participants. See Table 1 for more details.

All sessions took place in the computer laboratory of the Centre for Decision Research & Experimental Economics (CeDEx) at the University of Nottingham from March to May 2015. The game was computerized and programmed in the experimental software z-Tree (Fischbacher 2007). Participants were invited by e-mail using the recruiting software ORSEE (Greiner, 2004).

At the start of each session, participants drew a ticket from a bag containing 30 numbers. The number determined their cubicle in the laboratory. Once all participants were seated, they received a copy of the instructions for the experiment. The instructions were read out loud to the participants by the same experimenter through all the sessions and they were given the opportunity to ask questions individually. Finally, participants' understanding of the game was tested by having them individually answer a series of comprehension questions. The experimental part of the session ended when all the participants completed the series of

repeated prisoner's dilemmas. Afterwards, participants completed a questionnaire about their socio-demographics and the strategies they used.[6]

Participants received a show-up fee of £10 plus the total number of units earned throughout the experiment, converted at the exchange rate of 30 units = £1. Since stage-game payoffs could be negative, participants started the experiment with an initial endowment of 50 units.[7] Including the show-up fee, participants were paid an average of £14.42 privately in cash at the end of the session, with a range from £11 to £23. The average session length was 90 minutes.[8]

***Table 1.*** *Summary statistics by treatment.*

| | No Communication | | Communication | |
|---|---|---|---|---|
| | *Low* | *High* | *Low* | *High* |
| Number of sessions | 3 | 3 | 3 | 3 |
| Number of participants | 78 | 76 | 80 | 78 |
| Average number of supergames | 12.3 | 15.4 | 11.9 | 12.4 |
| Average number of periods per supergame | 7.9 | 7.9 | 8.1 | 7.9 |

## 3. Questions

In this section we introduce four questions about play in repeated games with errors and communication that we explore using our experiments. For each question, we consider how the answer varies with game payoffs and history of play.

QUESTION 1: Does the ability to communicate increase cooperation levels?

Since previous experimental work on communication in cooperation games has focused on finitely repeated games without errors, we have little empirical basis for hypotheses in our infinitely repeated game with errors. However, previous experiments using the type of minimal communication protocol used in our design suggest that this type of communication may not have any effect on cooperation. And since cheap-talk communication does not enlarge the set of perfect public equilibria, the standard approach of using the most efficient such equilibrium to generate predictions suggests that communication here will not have an effect on cooperation in either treatment. However, the infinitely repeated prisoner's dilemma resembles a coordination game, and experimental evidence from coordination games suggests

[6] In particular, we asked them to describe their strategies, the number of periods of past play considered, and in treatments with communication whether they paid attention to messages, actions, or both.
[7] No participant ever had less than 7 units, and only 2 out of 312 participants ever dropped below 30 units.
[8] Participants had to make choices within 30 seconds, and were told that after 30 seconds choices would be randomized. The average decision time was 1.8 seconds and just 51, or 0.16%, of the choices were random.

that communication sometimes but not always increases equilibrium play (see e.g., Cooper et al. (1992), Andersson and Wengström (2012), and Cooper and Kühn (2014)). It is thus not clear *a priori* how communication will affect play in our experiment.

QUESTION 2: How honestly do participants communicate their intentions?

We would expect that communication is most likely to promote cooperation when a substantial fraction of participants honestly communicate their intentions. And past work gives reason to expect that at least some of our participants *will* be honest; for example Gneezy (2005) finds that in one-shot interactions, a substantial fraction of participants have an aversion to lying, and thus act honestly. Furthermore, in past work participants were found to be sensitive to both their own gains from lying and the costs imposed on others. As this work was done in substantially different settings, however, we have little evidence to guide a quantitative prediction about what fraction of participants will be always or mostly honest and whether this will be sufficient to allow communication to impact cooperation, or how the level of honesty will vary with the payoffs (although the honesty observed in one-shot games, where there are no cooperative equilibria, suggest that at least some participants will be honest even in our lower-returns treatment).

QUESTION 3: To what extent do participants condition on the intentions communicated by their partner?

To assess whether (and in which ways) one's play is affected by the partner's communicated intentions, we examine how players' probability of cooperating, and of signaling cooperation, depend on both their partner's prior actions and signals (and how this varies with payoff treatment). We imagine that communication may improve cooperation by making players more likely to be lenient after a partner's defection if the partner signaled that they intended to cooperate, but we would expect repeated instances of mismatch between communicated intention and actual action to undermine a participant's faith in her partner's communication.

QUESTION 4: What predicts which participants are more likely to condition on their partner's communicated intentions, and which participants are more likely to communicate honestly?

To the extent that prosociality drives both cooperation and honesty, one might expect cooperators to be more honest than defectors. A similar prediction arises from the possibility that people cooperate (and tell the truth) in an effort to maximize payoffs (particularly when cooperative equilibria exist, as in Dreber et al. (2014)). There may also be gender differences in honesty. There is some evidence suggesting that if anything, men tend to lie more when there are material benefits for them doing so (e.g., Dreber and Johannesson 2008, Erat and Gneezy 2012), although the evidence here is mixed, and several studies suggest that there are no gender differences in deception (e.g., Childs 2012).

## 4. Strategy Frequency Estimation

Before answering these specific questions, we present a general assessment of the strategies played by participants in our experiments using the SFEM introduced by Dal Bó & Frechette (2011), in which a finite set of strategies is specified, and the probability of participants choosing each strategy (along with a probability of mental error) is estimated from the data.[9]

To use this method, it is necessary to choose which strategies to include, because it is not possible to include all of the infinitely many pure strategies of the repeated game. To construct this list of strategies, we restrict our attention to a set of strategies that look no further back than the last three periods of play, as in prior work on repeated games with errors (Fudenberg et al., 2012; Rand et al., 2015).[10]

The simplest strategies we consider either unconditionally cooperate all the time (ALLC) or defect all the time (ALLD). For treatments with communication, we look at three unconditional strategies that always either: cooperate and send the C message (ALLC(C)), defect and send the C message (ALLD(C)), or defect and send the D message (ALLD(D)).[11]

In the treatments without communication, we also consider the conditional strategies Grim (GRIM1) and tit-for-tat (TFT) which depend only on the previous period's outcome;

[9] This method estimates the frequency of each strategies in based on the histories of play. It relies on MLE and assumes that participants play the same strategy throughout the session, but make mental mistakes in implementing that strategy and chose an action that is not recommended.

[10] Unlike prior work, however, our treatments with communication require strategies that specify messages as well as actions.

[11] We do not look at strategies with intended move C(D) because they occurs so rarely in our dataset (0.73%) that is not possible to make meaningful inference about them.

GRIM2, 2TFT, TF2T, which look back two periods; and GRIM3, 3TFT, TF3T, 2TF2T, which look back 3 periods.[12]

In treatments with communication, conditional strategies must specify which combinations of moves are considered "defection" (and therefore cause the strategy to trigger). We therefore include versions of each of the above strategies that: *ignore messages* and treat both D(C) and D(D) as defection; *trust messages* and treat both C(D) and D(D) as defection; are *punitive* and treat anything other than C(C) as defection; or are *tolerant* and treat only D(D) as defection. For GRIM2, TF2T, and 2TF2T (lenient strategies that wait for two defections in a row before triggering), we also include versions that are lenient as described except when they observe D(D), in which case they trigger immediately; and for GRIM3 and TF3T (lenient strategies that wait for three defections in a row before triggering), we include versions that trigger immediately upon observing D(D); and that trigger after observing two periods in a row of D(D).

In treatments with communication, there is also the question of which actions a strategy uses when in non-punishing and punishing states. Thus we included strategies that either played: C(C) when non-punishing and D(D) when punishing; C(C) when non-punishing and D(C) when punishing; or D(C) when non-punishing and D(D) when punishing.

For both treatments with and without messages, we also include additional versions of each possible strategy that start with a move different from the one outlined, for example C-ALLD starts by cooperating in the first period and then switches to ALLD for the rest of the interaction. For treatments without communication the starting move could be either C or D, whereas for treatments with communication the starting move could be either C(C), D(C), or D(D).

As a product of these variations, our full set of possible strategies contains a total of 21 strategies for treatments without communication, and 541 strategies for treatments with communication.[13] To determine which of these possible strategies are most useful in describing the play of participants in our experiments, we use the following procedure. First,

[12] As in prior work, we assume that defections by either player will trigger Grim strategies.

[13] The total number of strategies with communication was actually 607, but there were certain combinations that were excluded beforehand because their similarity made it seem hard to disentangle them in the data. In particular, for treatments without communication we excluded D-GRIM1 because it is identical to ALLD except when a player mistakenly cooperates in the first period, and the other player also cooperates (in this case ALLD would defect and D-GRIM1 would cooperate). In similar fashion, for treatments with communication we excluded Grim strategies that start with D(D) and trigger defection when observing D(D), or that start with D(C) and trigger when observing D(C).

for each participant, we determine which strategy correctly predicts the highest fraction of that participant's moves (in the event of ties, we use the simplest strategy in terms of memory).[14] We then removed all strategies that were not best predictors for at least two participants. Using this reduced set, we performed the SFEM procedure as described in Dal Bó & Frechette (2011) to estimate the frequency of each strategy. We then further eliminated strategies whose estimated frequency was not significantly greater than zero (at the 10% significance level, based on bootstrapped standard errors).[15] Finally, using the surviving strategies, we again performed SFEM to arrive at a final estimate of strategy frequencies, which are presented in Table 2 below.[16]

***Table 2.*** *SFEM results for treatments with and without communication*

| Strategy | *Low* | *High* |
|---|---|---|
| ***Treatments without communication*** | | |
| ALLD | 0.40*** (0.06) | 0.37*** (0.06) |
| GRIM1 | 0.13*** (0.04) | |
| TFT | 0.08** (0.03) | 0.11** (0.04) |
| D-TFT | 0.07** (0.03) | 0.06* (0.03) |
| 2TFT | | 0.08** (0.04) |
| D-2TFT | 0.14*** (0.04) | 0.09*** (0.03) |
| TF2T | 0.06** (0.03) | 0.11*** (0.04) |
| GRIM3 | 0.07** (0.03) | 0.05* (0.03) |
| 3TFT | 0.07** (0.03) | |
| 2T2T | | 0.14*** (0.05) |
| Mental error | 0.13 | 0.13 |
| ***Treatments with communication*** | | |
| ALLD(C) | 0.17*** (0.04) | 0.06** (0.03) |
| ALLD(D) | 0.30*** (0.06) | 0.21*** (0.05) |
| GRIM1 that believes messages | | 0.05* (0.03) |
| TFT that ignores messages | | 0.05* (0.03) |
| D(C)-TFT that is punitive | 0.08** (0.04) | |
| D(C)-TFT that is punitive and defects using D(C) | 0.07** (0.03) | |

[14] Our approach includes all moves within a session because we find no evidence of learning (as described below). Moreover, in Table B6 of the Appendix we find qualitatively similar results when restricting to the last four supergames.

[15] None of the strategies deleted in the first stage had a frequency greater than 0.04, and none of the remaining strategies in the second stage increased its frequency by more than 0.05.

[16] Compared to treatments without communication (and prior work without messages), our treatments with communication have substantially higher rates of mental errors (calculated as the probability that the chosen action is not the one recommended by the strategy). This is not surprising, given that the strategy set is much more complicated, and there are three ways to make a mistake rather than just one. For the treatments with communication, we tested the validity of the estimation procedure on simulated data. We assigned strategies to 80 computer agents in *low*, and 78 computer agents in *high*, according to the estimated strategy frequency distribution. We then performed SFEM on the simulated histories of play of a total of 12 supergames (agents were randomly paired and played games of lengths similar to the ones induced experimentally). Results in Table A9 of the Appendix reveal consistency between actual and simulated frequencies.

| | | |
|---|---|---|
| D(C)-2TFT that ignores messages and cooperates using D(C) | 0.07** (0.03) | |
| TF2T that is punitive | | 0.12** (0.05) |
| TF2T that is punitive and defects using D(C) | | 0.09*** (0.03) |
| TF2T that is punitive and immediately punishes D(D) | 0.31*** (0.06) | |
| GRIM3 that ignores messages and immediately punishes after two periods of D(D) | | 0.08** (0.03) |
| TF3T that is punitive and immediately punishes D(D) | | 0.24*** (0.05) |
| 2TF2T that is punitive, immediately punishes D(D), and defects using D(C) | | 0.11*** (0.04) |
| Mental error | 0.29 | 0.24 |

*Notes:* The prefix of a conditional strategy indicates the opening move; if no prefix is given, the opening move is C(C). *Punitive* refers to strategies that treat any move other than C(C) as defection. Unless otherwise specified, strategies cooperate using C(C) and defect using D(D) (i.e. play C(C) when un-triggered, and by D(D) when triggered). Mental error is calculated as the probability that the chosen action is not the one recommended by the strategy. Bootstrapped standard errors (shown in parentheses) used to calculate p-values. ***p<0.01, **p<0.05, *p<0.10.

## 5. Main Results

We start by evaluating how much participants appear to learn and adjust their play over the course of a session. In most of our treatments, the percentage of people cooperating in the first period of each supergame did not vary over the course of the experiment (Figure 2), nor did the frequency of cooperation over all periods or the frequency of messages indicating cooperative intent.[17] Given this, we base our analysis on decisions from all supergames to maximize the amount of data available.[18] Figure 2 also indicates substantial differences in cooperation levels across treatments. First period cooperation rates vary between 28% and 60% depending on the treatment, and overall cooperation rates vary between 21% and 44%.[19]

We now turn to our experimental questions.

[17] This is confirmed by treatment-specific linear regressions that control for the supergame played and are clustered on both participant and supergame pair (p=0.153, Table A1 of the Appendix). We use linear models rather than *logit* or *probit* because the coefficients produced are more interpretable. Our conclusions are the same regardless of the approach used. We find a similar lack of learning when considering cooperation over all periods (p=0.261, Table A2 of the Appendix) and likelihood of indicating cooperative intent using messages over all periods (p=0.358, Table A3 of the Appendix).

[18] Moreover, we find qualitatively similar results if we restrict our attention to the last 4 super games played (Appendix B).

[19] We note that there is substantially less cooperation in the *high* treatment without communication here compared to what was observed previously in Fudenberg et al. (2012). Given that the experimental setup is identical between the two papers, it seems likely that this difference reflects differences in participant pool (Nottingham vs Harvard), particularly given Camerer et al. (2016)'s nearly exact replication of the Fudenberg et al. (2012) results using a CalTech participant pool.

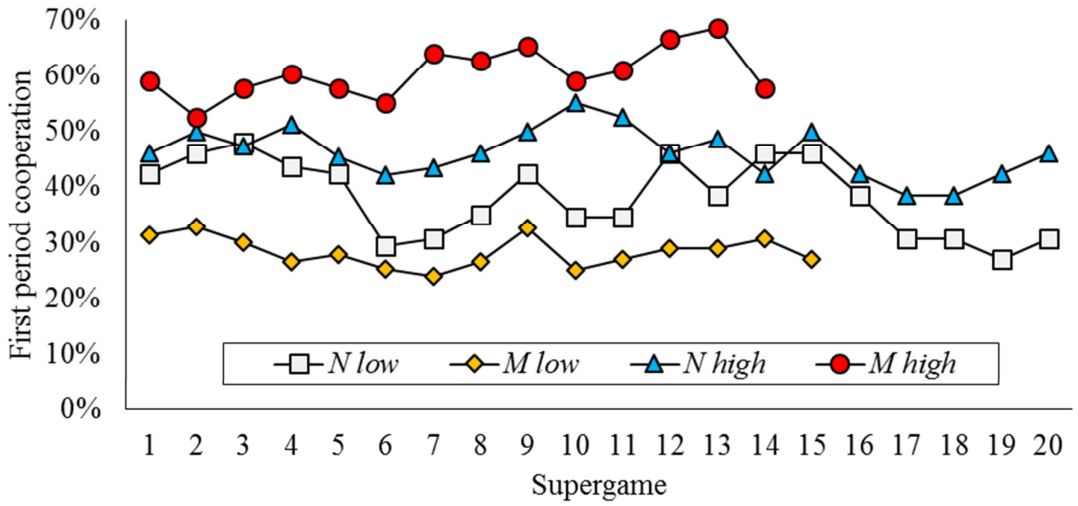


***Figure 2.*** *First period cooperation over the course of the session, by treatment.*

QUESTION 1: *Does the ability to communicate increase cooperation levels?*

In contrast to predictions based on the most efficient equilibria, which predict full cooperation in the *high* treatment even without communication, Figure 3 reveals that the ability to communicate increases cooperation levels (first period cooperation: no messages 47%, messages 60%, $p$=0.044; overall cooperation: no messages 33%, messages 44%, $p$=0.012).[20] Interestingly, allowing for communication results in a marginally significant decrease in first period cooperation in the *low* treatment (first period cooperation: no messages 39%, messages 28%, $p$=0.063; overall cooperation: no messages 25%, messages 21%, $p$=0.313).

[20] We report pairwise comparisons based on the results of linear regressions with a treatment value dummy as the independent observation, errors clustered on both participant and supergame pair.

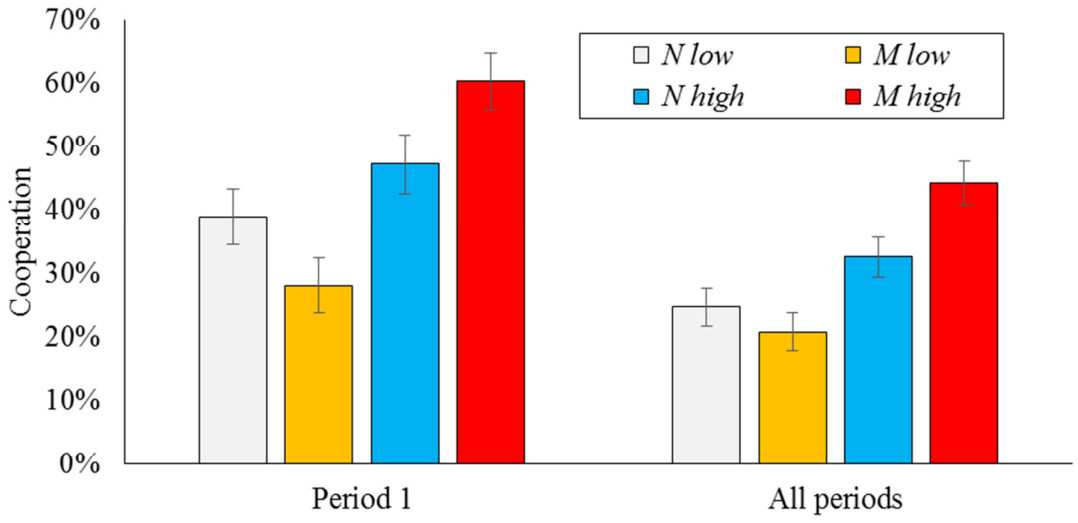


***Figure 3.*** *First period and overall cooperation, by treatment.*

These results suggest that participants treat communicated intentions differently in the two payoff treatments, an issue which we explore more thoroughly below in Question 3.

QUESTION 2: *How honestly do participants communicate their intentions?*

Figure 4 shows that participants are honest much of the time when communicating their intentions.[21] In the *high* treatment, 78% of actions across all periods were consistent with their corresponding messages. The corresponding number for the *low* treatment was also high, 68%, but significantly lower ($p$=0.005). Furthermore, honesty has different flavors across treatments. Candid cooperation occurred significantly more often in the *high* treatment than in the *low* treatment (44% versus 20%, $p$=0.001), whereas for honest defections the opposite is true (48% versus 34%, $p$=0.002). Not surprisingly, in both treatments virtually all lying involved defecting while claiming to have intended cooperation. This intended deception was more prevalent in the *low* treatment: only 8% of realized D(C) outcomes in the *low* treatment were actual cases of accidental defection, compared to 24% in *high*.

[21] For brevity, in this section we only discuss results when considering all periods of play; the results for first period play (displayed in Figure 4A) are qualitatively equivalent.

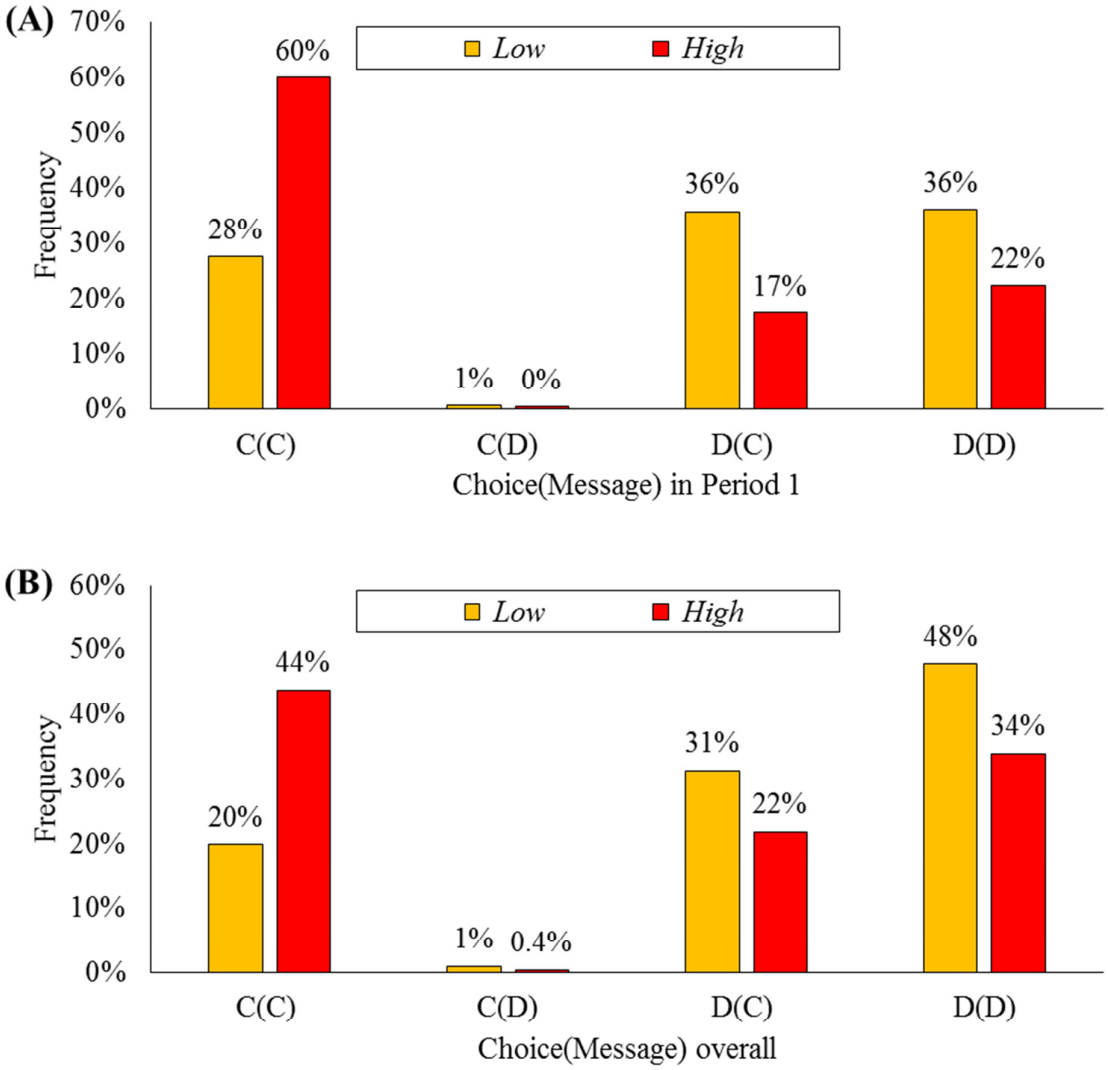


***Figure 4.*** *Frequency of intended actions in the message treatments. A) Period 1; B) Overall*

We therefore focus our subsequent discussion of honesty on cases where the intended action was D. In particular, we calculate an "honest-defection" as theratio D(D)/[D(D)+D(C)]. Using this measure, we find 60% honesty in the *low* treatment and 61% honesty in the *high* treatment. Interestingly, the first time the participant played defection, this value drops to 48% in the *low* treatment and 45% in the *high* treatment. Hence, it appears that the longer a player waits to defect, the more likely she is to be honest about it. Furthermore, Figure 5 below reveals that the fraction of honest reports tends increase over the course of a supergame in both treatments.

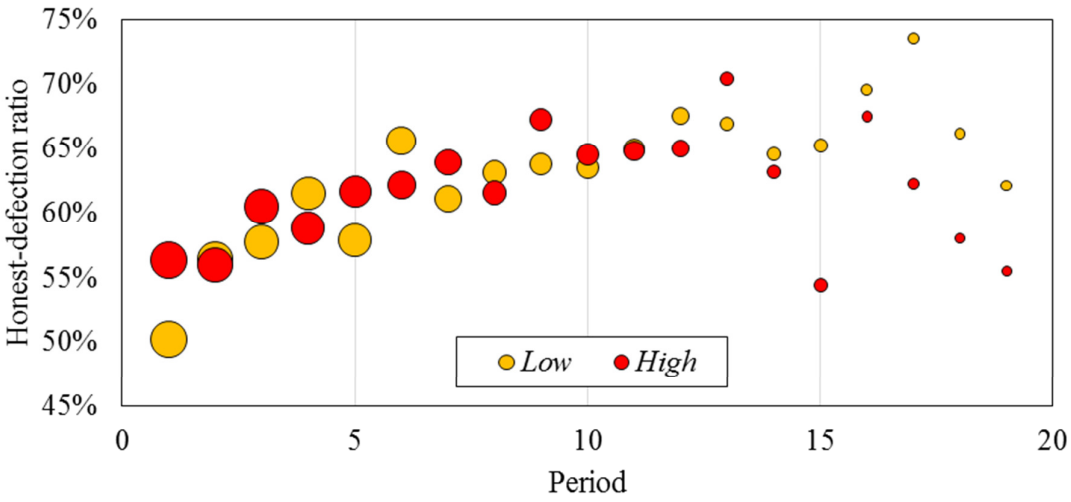


***Figure 5.*** *Honest-defection ratio (D(D)/[D(D)+D(C)]) by period; dot size is proportional to the number of observations in each period.*

Although the overall level of honesty is high, some participants were more honest than others. Figure 6 displays the frequency of people who lied a given number of times. As can be seen, a large majority of the participants sent dishonest messages at some point throughout the course of the session. In the *low* and *high* treatments respectively, 78 (98%) and 72 (92%) participants were *not* honest at least once. Moreover, most of the participants lied sparsely and sent dishonest messages 30% of the time or less: 43 participants (54%) in the *low* treatment and 50 participants (64%) in the *high* treatment.

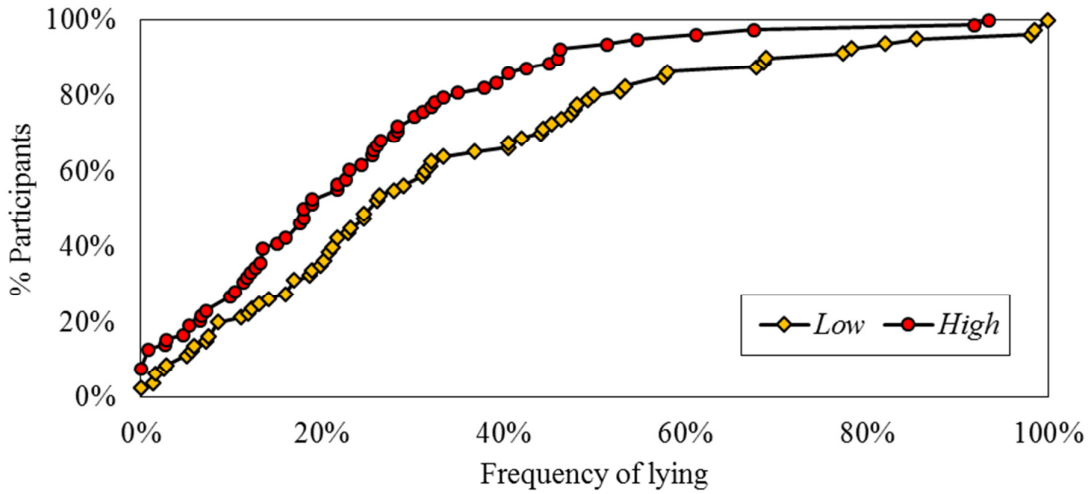


***Figure 6.*** *Cumulative distribution of participants by how often they lied.*

As with our descriptive analysis, the SFEM results (shown in Table 2) indicate that participants were honest much of the time in both treatments, and that there was more honesty in the *high* treatment. Specifically, in the *low* treatment, strategies that never lied (i.e. never played D(C)) had total probability of 61%; in the *high* treatment, 74%. Thus we find convergent evidence in support of a high level of honesty among our participants.

QUESTION 3: *To what extent do participants condition on the intentions communicated by their partner?*

We begin by taking a descriptive approach to answering this question. We find that a large proportion of the participants conditioned their responses on what their partner communicated. Figure 7 reports intended responses to the message and observed action of the other player in the immediately previous period. When participants saw that their partner both cooperated and signaled cooperation, 71% of the participants in the *high* treatment both cooperated and reported cooperation. The corresponding number for the *low* treatment is significantly lower, 53% ($p$=0.001). Moreover, in the event that the partner defected but signaled cooperation, participants in the *high* treatment were more than twice as lenient as those in the *low* treatment: they cooperated and sent the honest message 33% of the time versus only 14% of the time ($p$=0.001).[22]

[22] To support this observation, we report the result of linear regressions that predict cooperation based on the partner's message and observed action in the previous period (Table A7 of the Appendix). In both treatments, we find significant positive effects of cooperative messages and cooperative actions, as well as a significant positive interaction between the two ($p<0.001$ for all).

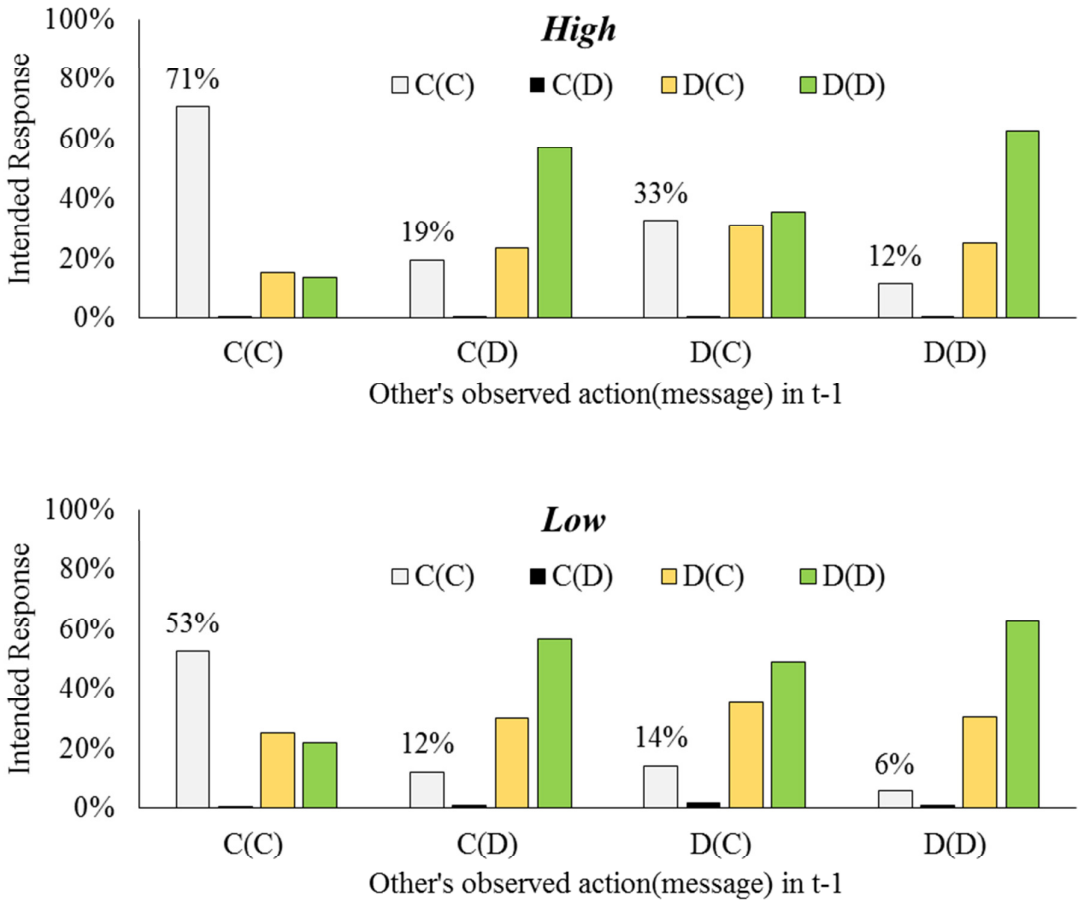


***Figure 7.*** *Intended response to other's observed action and message in the previous period.*

Note that Figure 7 implicitly assumes that participants ignored all of the history of the interaction except for what happened in the previous period. Yet there is evidence that people use strategies that look back more than one period, especially in games with noise (Fudenberg et al 2012, Rand et al 2015). This appears to be the case with our data too. Figure 8 shows that most participants in the message treatments (75% in *low*, 87% in *high*) as well as participants in the no-message treatments (65% in *low*, 87% in *high*) reported that they considered more than just the last period. Further evidence comes from the SFEM results shown above in Table 2, where strategies that looked back more than one period had probability weights of 33% in low and 66% in high with messages; and 20% in low and 33% in high without messages. These results suggest that messages facilitate the use of longer memory lenient strategies in the high treatment.

To explicitly test whether participants are indeed conditioning on more than only their

observation of the most recent period, we conduct a linear regression with correlated random effects, regressing own decision in period $t$ against own play in period $t$ - *1*, other's play in $t$ – *1*, own play in $t$ – *2* and other's play in $t$ – *2*; we also include own average frequency of first period cooperation and overall cooperation to help reduce possible heterogeneity bias.[23] Consistent with the use of longer memories, we find a significant effect of other's play two periods ago when pooling all treatments (coeff = 0.09, $p < 0.001$) as well as for each treatment separately.[24]

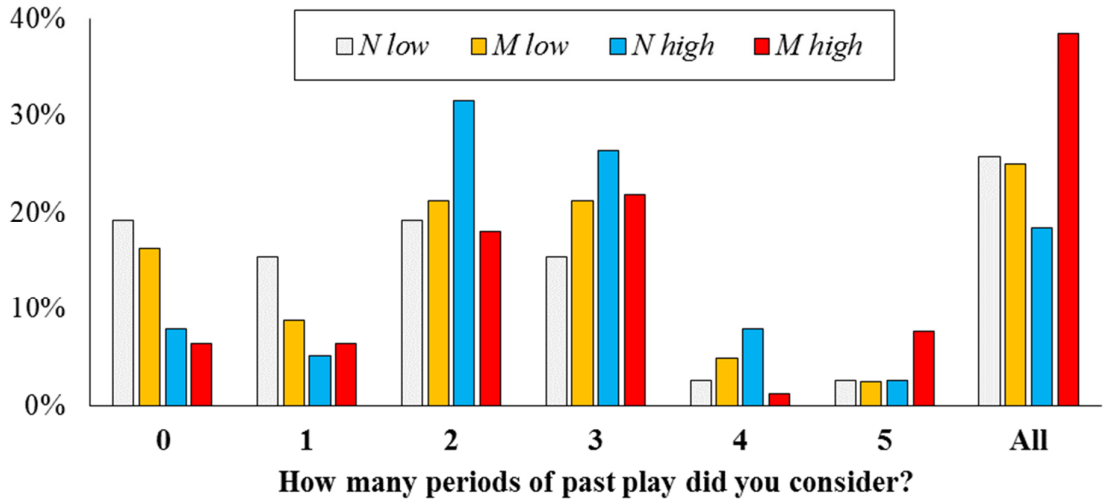


***Figure 8.*** *Number of periods back that participants self-reported considering.*

To further analyze the role of communicated intentions, we also examine dependence on play two periods ago (Figure 9). In particular, we focus on the case where the partner defected but communicated the intention to cooperate one period ago. We see that in the *high* treatment, if the partner had cooperated and signaled cooperation two periods ago, the defection of one period ago was forgiven 52% of the time; compared to only 16% of the time if the partner also defected but signaled cooperation two periods ago. A similar pattern (but lower overall level of cooperation) is seen in the *low* treatment, with 31% cooperation if the partner cooperated and signaled cooperation two periods ago compared to 7% after two

[23] Bias introduced by heterogeneity presents a potential challenge for this approach: the other's play two periods ago interacts with own play two periods ago to determine the history in the previous period, so other's play two periods ago could have a spuriously significant coefficient in a heterogeneous population of participants all of whom use memory-1 strategies. To control for bias introduced by heterogeneity, we include controls for the type of the player making the decision, as in Aoyagi and Frechette (2009) and Fudenberg et al. (2012).

[24] This effect is similar across treatments: *N low*, coeff=0.07; *N high*, coeff=0.10; *M low*, coeff=0.08; *M high*, coeff=0.10; $p<0.001$ for all.

periods of the partner defecting but signaling cooperation. These results are confirmed statistically in Table A8 of the Appendix. Interestingly, if we consider cases where the partner defected and communicated the intention to defect one period ago (Figure A7 of the Appendix), we see substantially less leniency compared to defection and signaled intent to cooperate, even if two periods ago the partner cooperated (32% cooperation in the *high* treatment (vs 52% above), 14% in the *low* (vs 31% above)). This provides evidence that the signal had a substantial impact on play, promoting leniency.[25]

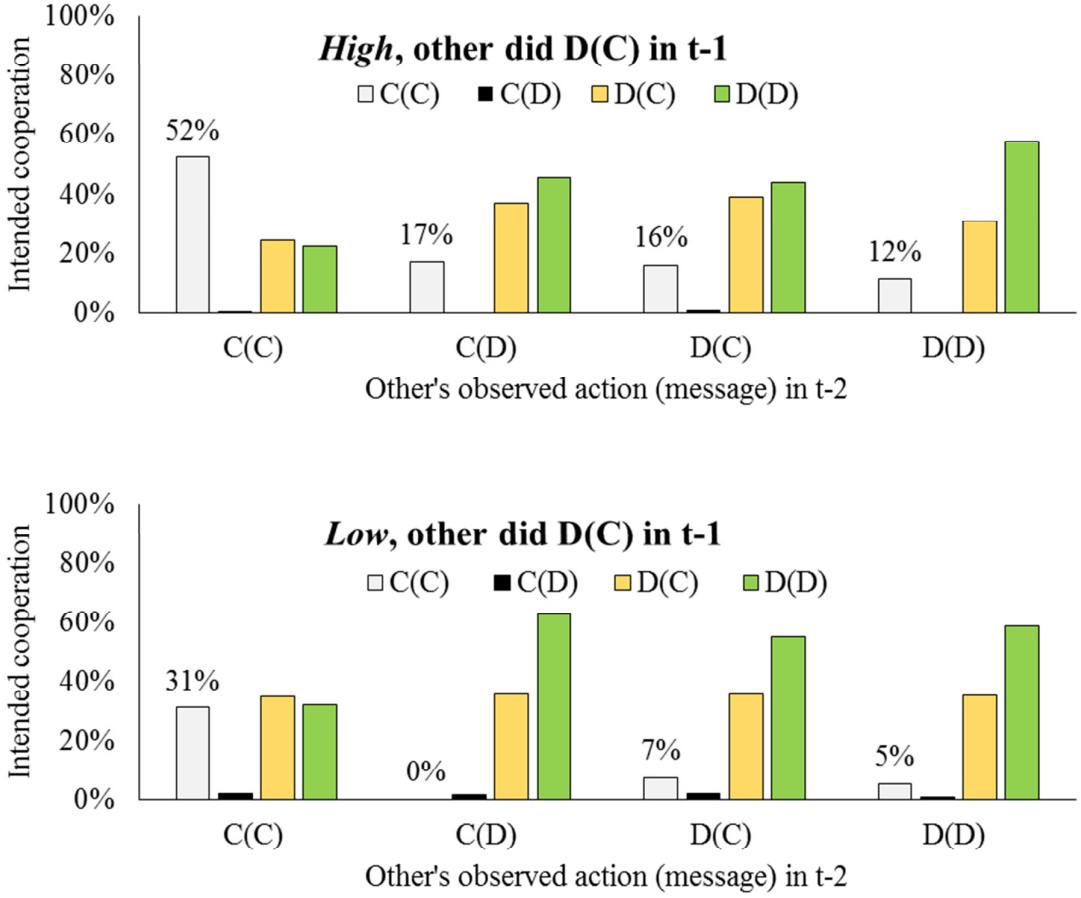


***Figure 9.*** *Intended response to observing other's defection and message "I choose A".*

Next we compliment this descriptive approach by examining the results of the SFEM

[25] To provide additional evidence that participants attended to the messages, and to provide some quantitative sense of how much this was true, we ask how often a player's move in period *t* matched the partner's message in period *t*-1 as opposed to the partner's message in period *t*-1 (in histories where these were different). We find that in 34% of cases, the participant's current decision matched the partner's previous message rather than the partner's previous action (41% in high and 29% in low); while in the remaining 66% of cases (59% in high, 71% in low), the participant's current decision matched the partner's previous action.

shown in Table 2 above. Consistent with the descriptive results, the SFEM assigns substantial weight to strategies that condition on messages. Such strategies account for 46% of the total probability weight in the *low* treatment, and 61% in the *high* treatment. Furthermore, many of these strategies look back longer than the previous period in their assessment of messages: 31% in the low treatment, and 56% in the high treatment, are lenient, such that they do not initially punish when a defecting partner sends a cooperate message, but eventually switch to punishing after two or three such occurrences. Thus, both the SFEM results and the descriptive analyses suggest that many players took messages seriously, but that repeated inconsistency between message and action undermined the credibility of the messages.

QUESTION 4: *What predicts which participants are more likely to condition on their partner's communicated intentions, and which participants are more likely to communicate honestly?*

Figure 10 suggests that the magnitude of the rewards to cooperation influences what people report paying attention to. In the *low* treatment, the majority of the participants focused on the outcome of the play, whereas in the *high* treatment, the majority focused jointly on both outcome and message. We also see that the more participants reported paying attention to messages, the more they tended to cooperate (in both treatments).

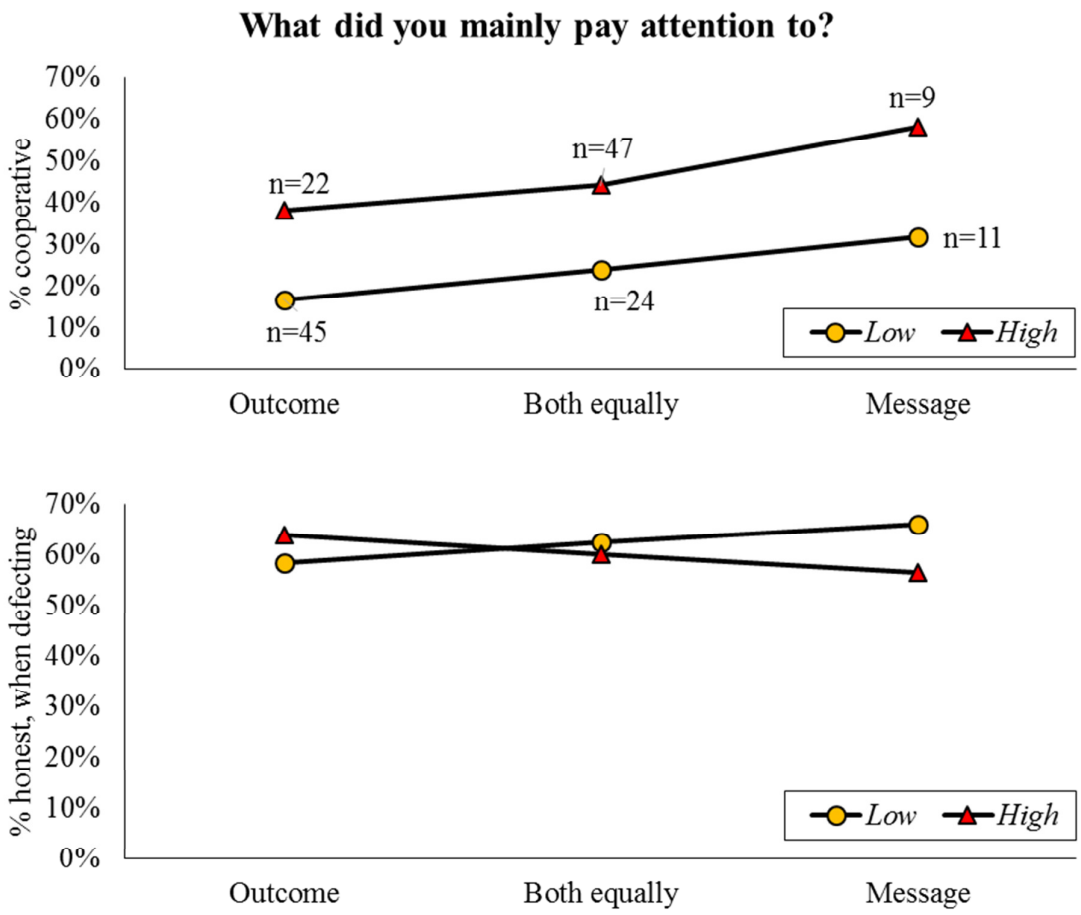


***Figure 10.*** *Frequency of honesty by treatment and the type of information participants reported paying more attention to (messages, outcome, or both).*

With regards to average payoffs, participants in the *high* treatment who reported looking at messages earned the most (0.97 MUs), whereas participants who reported looking at outcomes earned 0.09 MUs less, and participants who reported looking at messages and outcomes equally earned 0.07 MUs less. In the *low* treatment, participants who reported looking at messages also earned the most (0.30 MUs), but the difference between them and participants who reported looking at outcomes (messages and outcomes equally) was of just 0.01 (0.05) MUs.

We also explore how cooperativeness and demographic variables predict honestly signaling defection. We regress the likelihood that a participant who defects chooses the message "I choose B" against own overall cooperation, gender and age. We find significant

positive effects for female gender[26] and age (coeff = 0.131, $p<0.014$ and coeff = 0.018, $p<0.031$, respectively).[27]

**Discussion**

We now ask which behaviors were most successful by examining how participants' payoffs relate to their willingness to cooperate, and to believe their partner's messages. To do so, we investigate how participants' realized payoffs depend on their play. From the outcomes in the analogous no-message treatment of Fudenberg et al. (2012), we expect payoffs to be decreasing in the extent of cooperation in our no-message *low* treatment. In the *high* treatment, Grim is an equilibrium, but ALLD is still risk-dominant over Grim. Furthermore, in the analogous no-message treatment of Fudenberg et al. (2012), various cooperative strategies earned roughly equal payoffs to ALLD. Thus, we might expect the same in our no-message *high* treatment, but hope that communication would allow cooperators to out-earn defectors in the presence of messages.

With this in mind, we now examine how participants' payoffs in the experiment varied with their strategy. We begin by asking which of the strategies listed in Table 2 most accurately predicts each participant's play. For each strategy, we then calculate the average payoff per period over all participants identified with that strategy. This provides a (noisy) measure of actually experienced payoffs for people assigned to each strategy, which (because it depends on who they were matched with and the realizations of the monitoring errors) is in turn a noisy estimate of their expected payoff against a randomly drawn member of the participant pool.

We also use simulations to compare how these imputed payoffs compare to the payoffs expected if the SFEM strategy frequencies are correct. To do so, we compute the average payoff when each pair of strategies interact by averaging over 100,000 simulated supergames, and then calculate expected payoff for each strategy by weighting these payoffs based on the observed strategy frequencies in Table 2. Table 3 shows the estimated frequency of each strategy, along with the experimentally observed and expected payoffs.

[26] Our gender finding is in line with some but not all previous results on dishonesty (see, e.g., Dreber and Johannesson 2008, Childs 2012, Erat and Gneezy 2012).

[27] See table A8 of the Appendix for details. Our results are robust to two alternative cooperation measures: first period cooperation and whether the participant played C or D on the very first move of the whole session.

***Table 3.*** *Observed frequencies and resulting payoffs for each strategy with messages*

| | *Low* | | *High* | |
|---|---|---|---|---|
| *Strategy* | *Frequency* | *Observed (expected) payoff* | *Frequency* | *Observed (expected) payoff* |
| ALLD(C) | 0.17 | 0.39 (0.45) | 0.06 | 0.99 (1.29) |
| ALLD(D) | 0.30 | 0.25 (0.22) | 0.21 | 0.79 (0.68) |
| GRIM1 that believes messages | | | 0.05 | 0.82 (1.15) |
| TFT that ignores messages | | | 0.05 | 0.80 (1.07) |
| D(C)-TFT that is punitive | 0.08 | 0.26 (0.34) | | |
| D(C)-TFT that is punitive and defects using D(C) | 0.07 | 0.35 (0.45) | | |
| D(C)-2TFT that ignores messages and cooperates using D(C) | 0.07 | 0.25 (0.32) | | |
| TF2T that is punitive | | | 0.12 | 1.14 (1.17) |
| TF2T that is punitive and defects using D(C) | | | 0.09 | 0.99 (1.20) |
| TF2T that is punitive and immediately punishes D(D) | 0.31 | 0.22 (0.17) | | |
| GRIM3 that ignores messages and immediately punishes after two periods of D(D) | | | 0.08 | 0.82 (1.14) |
| TF3T that is punitive and immediately punishes D(D) | | | 0.24 | 0.90 (1.19) |
| 2TF2T that is punitive, immediately punishes D(D), and defects using D(C) | | | 0.11 | 1.05 (1.17) |

*Notes:* The prefix of a strategy indicates the opening move. If absent, participants start with C(C). *Punitive* refers to strategies that only treat C(C) as cooperation; unless otherwise specified, participants cooperate using C(C) and defect using D(D).

In line with Fudenberg et al. (2012), we find a high proportion of participants playing unconditional strategies that always defect (ALLD(D) and ALLD(C)). Interestingly, while ALLD(D) performs very poorly, the consistently dishonest ALLD(C) is actually the best performing strategy in *low*, and one of the best in *high* because it is capable of exploiting strategies that trust messages. Yet ALL(D) is substantially more common than ALLD(C) in both treatments – perhaps because lying is psychologically costly, as suggested by the one-shot experiments of Gneezy (2005).

Also in line with Fudenberg et al. (2012), we find that lenient strategies perform well in the *high* treatment, but not in the *low* treatment where dishonest strategies (i.e. strategies that sometimes play D(C)) perform better. In fact, participants who earned the most in the *high* treatment played a version of TF2T that defects only after their partner defects (either in

actions or messages) for two consecutive periods; whereas participants who earned the most in the *low* treatment played ALLD(C).

For treatments without messages we perform a similar analysis and report the results in Table 4. We note that in both treatments a substantial proportion of participants always defects, but doing so only earns high payoffs in the *low* treatment. Consistent with our findings regarding treatments with messages, we find that strategies that are cooperative and lenient are more frequent in the *high* treatment. However, without messages, such strategies do not perform well even in the *high* treatment. This suggests that the ability to send messages improves the relative performance of lenient cooperative strategies.

Overall, then, we see that cooperative strategies (and in particular, longer memory cooperative strategies) perform well in the *high* treatment with messages, but not elsewhere – in all other treatments, non-cooperative strategies are the highest earners.

***Table 4.*** *Observed frequencies and resulting payoffs for each strategy without messages*

| | *Low* | | *High* | |
|---|---|---|---|---|
| *Strategy* | *Frequency* | *Expected (actual) payoff* | *Frequency* | *Expected (actual) payoff* |
| ALLD | 0.40 | 0.36 (0.37) | 0.37 | 0.72 (0.76) |
| GRIM1 | 0.13 | 0.26 (0.31) | | |
| TFT | 0.08 | 0.15 (0.2) | 0.11 | 0.75 (0.81) |
| D-TFT | 0.07 | 0.22 (0.36) | 0.06 | 0.80 (0.65) |
| 2TFT | | | 0.08 | 0.75 (0.58) |
| D-2TFT | 0.14 | 0.31 (0.3) | 0.09 | 0.80 (0.74) |
| TF2T | 0.06 | 0.01 (0) | 0.11 | 0.71 (0.81) |
| GRIM3 | 0.07 | 0.06 (0.22) | 0.05 | 0.71 (0.69) |
| 3TFT | 0.07 | 0.25 (0.38) | | |
| 2T2T | | | 0.14 | 0.73 (0.86) |

*Note:* These strategies all start with C except for D-TFT and D-2TFT.

To provide further evidence regarding payoffs that does not rely on SFEM, we use participants' intended first period decisions as a rough proxy for their strategy. In particular, we compare participants who always chose defection in period 1 with those who sometimes chose defection and sometimes chose cooperation in period 1, and those who always chose cooperation in period 1. We believe that these differences in opening moves are a reasonable proxy for strategies more generally because, as shown in Table 5, participants who always open with D virtually never cooperate, so their play resembles the ALLD strategy, while participants who always open with C are more cooperative than intermediate participants.

***Table 5.*** *Overall cooperation rates (excluding period 1) for participants by period 1 choice: always defection, a mix of defection and cooperation, and always cooperation..*

| | *Period 1 choice* | | |
|---|---|---|---|
| | *Always D* | *Mixed* | *Always C* |
| *N low* | 0.05 (N=21) | 0.26 (N=53) | 0.72 (N=4) |
| *N high* | 0.03 (N=14) | 0.33 (N=50) | 0.56 (N=12) |
| *M low* | 0.07 (N=34) | 0.26 (N=42) | 0.51 (N=4) |
| *M high* | 0.08 (N=12) | 0.42 (N=46) | 0.59 (N=20) |

Moreover, Table 6 shows that participants who always open with defection tend to out-earn more cooperative participants in every treatment except for the *high* treatment with communication. There, the opposite is true: participants who always open with cooperation earn the highest payoffs.

***Table 6.*** *Average payoff for always open D(D), intermediate, always open C(C)*

| | *Participant (P) who always opens with D* | *Intermediate* | *Participant (P) who always opens with C* |
|---|---|---|---|
| *N low* | 0.46 (N=21) | 0.26 (N=53) | 0.16 (N=4) |
| *N high* | 0.80 (N=14) | 0.74 (N=50) | 0.70 (N=12) |
| *M low* | 0.37 (N=34) | 0.21 (N=42) | 0.27 (N=4) |
| *M high* | 0.86 (N=12) | 0.87 (N=46) | 0.99 (N=20) |

Next, we examine the payoff consequences of communication by comparing the average payoff of participants based on their combination of opening action and message. In the *low* treatment, no more than 5% of participants always opened with C(C), C(D), or D(D). However, the 14% of participants who always opened by lying (i.e. playing D(C)) earned substantially more per period (0.45 MUs) than other participants (0.25 MUs). In our *high* treatment, conversely, it was participants who always opened with C(D), D(C), or D(D) that were rare, and the 24% of the participants who always opened with C(C)[28] out-earned (0.99 MUs) other participants without a consistent opening move (0.87 MUs).[29]

[28] In addition to being more initially cooperative, we find that participants who always open with C(C) are more lenient in the *high* treatment: when their partner's realized outcome is D(C) in period 1, participants who always open with C(C) are substantially likely to cooperate in period 2 (70%C) compared to other participants (49%C). Furthermore, this leniency is specifically driven by sensitivity to the message: when the partner opened with D(D), participants who always open with C(C) are not any more likely to cooperate (31%C) than other participants (38%C).

[29] Note that this approach provides evidence that is in line with the one from the SFEM. In particular, in Table 3 only ALLD(D) and ALLD(C) do not start with C(C), and on average they had a payoff of 0.83, which is not far from the 0.87 here observed.

This suggests that persistent dishonest defection paid off when the returns to cooperation were low, whereas honest cooperation paid off when the returns to cooperation were high. To try and understand why this might be, we examine how payoffs vary based on the partner's opening move (Table 7). We see that in the *low* treatment, participants that always opened with D(C) out-earned others regardless of the partner's opening move. In the *high* treatment, participants who always open with C(C) out-earned others when they were matched with partners who also opened with C(C), but when matched with D(C) partners they were out-earned by participants who always opened with D(C).

***Table 7.*** *Average payoff by type of partner's opening in period 1. Shown in parentheses is the number of interactions in which each combination of participant's strategy and partner's opening move occurred.*

| | *Low* | | | *High* | | |
|---|---|---|---|---|---|---|
| *Other's Realized Period 1 Play* | *P who always opens with D(C)* | *P who always opens with C(C)* | *All other openings* | *P who always opens with D(C)* | *P who always opens with C(C)* | *All other openings* |
| *C(C)* | 0.83 (31) | 0.68 (15) | 0.63 (230) | 1.33 (10) | 1.46 (138) | 1.25 (376) |
| *C(D)* | 0.84 (4) | 0.31 (5) | 0.45 (33) | - (0) | 0.14 (8) | 0.66 (25) |
| *D(C)* | 0.34 (32) | 0.11 (16) | 0.11 (279) | 0.89 (5) | 0.60 (53) | 0.59 (166) |
| *D(D)* | 0.18 (45) | 0.10 (13) | 0.05 (249) | 0.73 (5) | 0.21 (38) | 0.24 (144) |

To learn more about why those who always open with C(C) do not fare poorly against partners who open with D(C), in Table 8 we compare earnings in cases where the partner's D(C) was either a true accident (i.e. partner chose C(C) and noise changed the action to D) or dishonesty (i.e. partner choice D(C)). We see that in the *high* treatment (but not the *low* treatment), participants who always open with C(C) earn 0.18 MUs less per period than others when the partner's D(C) was intentional, but earn 0.30 MUs *more* per period than others when the partner's D(C) was accidental. Thus, the risk of exploitation by liars is balanced out by the gains from being more cooperative when one's partner makes an honest mistake.

***Table 8.*** *Average payoff of players when other opens with D(C). Shown in parentheses is the number of interactions in which each combination of participant's strategy and partner's opening move occurred.*

| | *Low* | | | *High* | | |
|---|---|---|---|---|---|---|
| | *P who always opens with D(C)* | *P who always opens with C(C)* | *All other openings* | *P who always opens with D(C)* | *P who always opens with C(C)* | *All other openings* |
| *D(C) intended* | 0.29 (30) | 0.08 (14) | 0.09 (258) | 0.40 (3) | 0.19 (33) | 0.37 (107) |
| *D(C) unintended* | 1.06 (2) | 0.23 (2) | 0.48 (21) | 1.00 (2) | 1.28 (20) | 0.98 (59) |

## Conclusion

In many real-world repeated interactions, participants can communicate with each other, making promises, excuses, and threats. In this paper we studied the impact of a very limited communication protocol, namely announcements of the intended action, on cooperation in an indefinitely repeated prisoner's dilemma. We found that even though most participants are mostly honest, communication only led to higher cooperation rates in the treatment with relatively higher gains from cooperation. In this treatment honest cooperation also maximized the participants' earnings: even though these cooperators could be exploited by liars, they could also reap the benefits from future cooperation after having trusted an honest mistake. In the other treatments, where honesty did not maximize payoff, it was much less common.

This paper used a very restrictive communication protocol, to keep the strategy space from being too complex and to make the data easier to analyze. It would be interesting to explore the effects of other sorts of communication protocols, though designing richer modes of communication that still provide analyzable data is a challenge for future work.

**References**


Andersson, Ola and Erik Wengström. 2012. Credible Communication and Cooperation: Experimental Evidence from Multi-Stage Games. *Journal of Economic Behavior & Organization*, 81: 207-219.

Aoyagi, Masaki, V. Bhaskar, and Guillaume Frechette. 2013. The Impact of Monitoring in Infinitely Repeated Games: Perfect, Public, and Private. *Manuscript in preparation*.

Aoyagi, Masaki, and Guillaume Frechette. 2009. Collusion as Public Monitoring Becomes Noisy: Experimental Evidence. *Journal of Economic Theory* 144 (3): 1135–65.

Bigoni, Maria, Jan Potters, and Giancarlo Spagnolo. 2012. Flexibility and Collusion with Imperfect Monitoring. *Working Paper*.

Blonski, Matthias, Peter Ockenfels, and Giancarlo Spagnolo. 2011. Equilibrium Selection in the Repeated Prisoner's Dilemma: Axiomatic Approach and Experimental Evidence. *American Economic Journal: Microeconomics* 3 (3): 164–92.

Bochet, Oliver, Talbot Page and Louis Putterman. 2006. Communication and Punishment in Voluntary Contribution Experiments. *Journal of Economic Behavior & Organization*, 60, 11-26.

Camera, Gabriele, Marco Casari, and Maria Bigoni. 2013. Binding Promises and Cooperation Among Strangers. *Economics Letters*, 118(3): 459-461.

Camerer, Collin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen & Hang Wu. 2016. Evaluating Replicability of Laboratory Experiments in Economics. *Science*.

Childs, Jason. 2012. Gender differences in lying. *Economics Letters*, 114(2): 147-149.

Compte, Olivier. 1998. Communication in Repeated Games with Imperfect Private Monitoring. *Econometrica*, 66(3), 597–626.

Cooper, David J., and Kai-Uwe Kühn. 2014. Communication, Renegotiation, and the Scope for Collusion. *American Economic Journal: Microeconomics*, 6(2): 247-78.

Cooper, Russell, Douglas V. DeJong, Robert Forsythe, and Thomas W. Ross. 1992. Communication in Coordination Games. *The Quarterly Journal of Economics* 107 (2): 739-771.

Dal Bó, Pedro. 2005. Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games. *American Economic Review* 95 (5): 1591–1604.

Dal Bó, Pedro, and Guillaume R. Frechette. 2011. The Evolution of Cooperation in Infinitely Repeated Games: Experimental Evidence. *American Economic Review* 101 (1): 411–29.

Dal Bó, Pedro, and Guillaume R. Frechette. 2012. Strategy Choice In The Infinitely Repeated Prisoners Dilemma. *Working paper.*

Dal Bó, Pedro, and Guillaume R. Frechette. 2015. On the Determinants of Cooperation in Infinitely Repeated Games: A Survey. Forthcoming in the *Journal of Economic Literature.*

Dreber, Anna, and Magnus Johannesson. 2008. Gender Differences in Deception. *Economics Letters*, 99(1), 197-199.

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

Embrey, Matthew, Guillaume R. Fréchette, and Sevgi Yuksel. 2014. Cooperation in the finitely repeated prisoner's dilemma. *Working Paper.*

Erat, Sanjiv, and Uri Gneezy. 2012. White Lies. *Management Science*, 58 (4), 723-733.

Fischbacher, Urs. 2007. Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments. *Experimental Economics* 10 (2): 171–78.

Fudenberg, Drew, and David K. Levine. 2007. The Nash-threats folk theorem with communication and approximate common knowledge in two player games, *Journal of Economic Theory*, 132(1), 461–473.

Fudenberg, Drew, David K. Levine, and Eric Maskin. 1994. The Folk Theorem in Repeated Games with Imperfect Public Information. *Econometrica* 62, 997-1039.

Fudenberg, Drew, David G. Rand, and Anna Dreber. 2012. Slow to Anger and Fast to Forgive: Cooperation in an Uncertain World. *American Economic Review* 102 (2), 720-749.

Gneezy, Uri. 2005. Deception: The Role of Consequences, *American Economic Review*,95(1), 384-394.

Greiner, Ben. 2004. An online Recruitment System for Economic Experiments. In: Kurt Kremer, Volker Macho (Eds.), Forschung und wissenschaftliches Rechnen 2003. GWDG Bericht 63: 79-93.

Kandori, Michihiro. 1992. The Use of Information in Repeated Games with Imperfect Monitoring, *Review of Economic Studies*, 59, 581-594.

Kandori, Michihiro, and Hitoshi Matsushima. 1998. Private Observation, Communication and Collusion, *Econometrica*, 66(3), 627–652.

Rand, David G., Drew Fudenberg and Anna Dreber. 2015. It's the Thought that Counts: The Role of Intentions in Noisy Repeated Games. *Journal of Economic Behavior and Organization*,116: 481-499.

Rand, David G., and Martin A. Nowak. 2013. Human cooperation. *Trends in Cognitive Sciences* 17: 413-425.

**Appendix A**

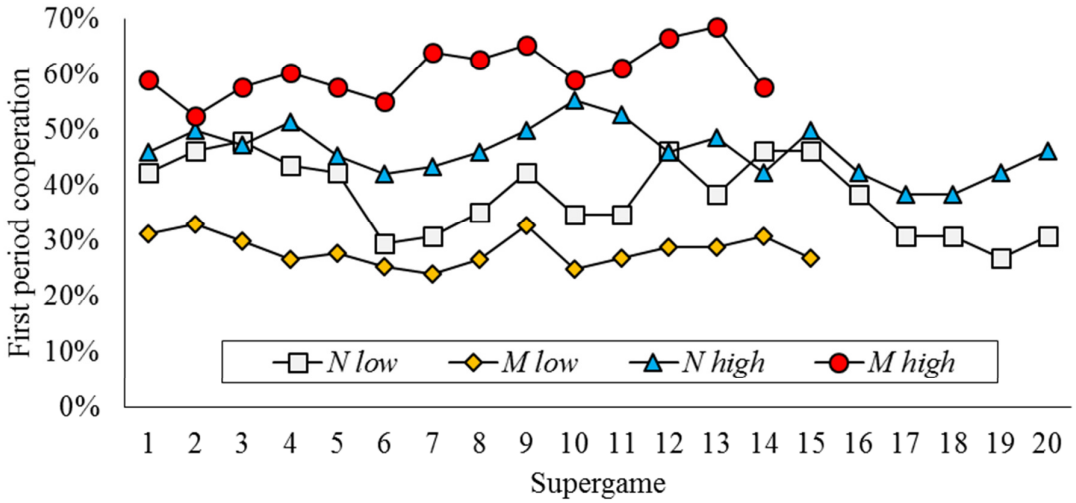


***Figure A1 (Figure 2).*** *First period cooperation over the course of the session, by treatment.*

**Table A1.** Intended cooperation in the first period of each supergame in all treatments. Standard errors clustered on both participant and supergame pair; * p<0.1, ** p<0.05, *** p<0.01.

| *Dependent variable: Intended Cooperation in the first interaction of each supergame* | | | | |
|---|---|---|---|---|
| | *N low* | *M low* | *N high* | *M high* |
| Supergame | -0.006 | -0.002 | -0.002 | 0.007 |
| | (0.005) | (0.006) | (0.005) | (0.005) |
| Constant | 0.436*** | 0.296*** | 0.486*** | 0.553*** |
| | (0.048) | (0.047) | (0.052) | (0.052) |
| Observations | 960 | 946 | 1169 | 968 |
| $R^2$ | 0.004 | 0.001 | 0.001 | 0.003 |

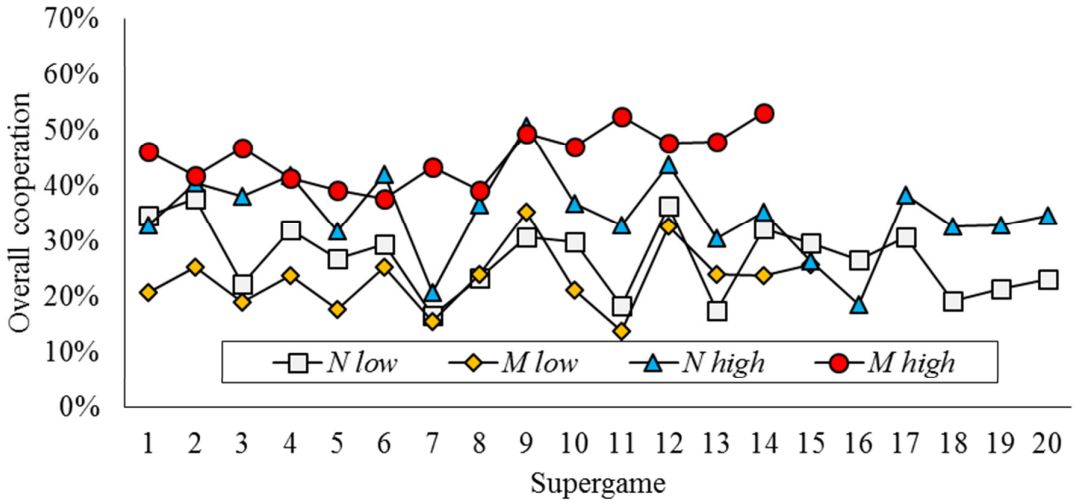

***Figure A2.*** *Overall cooperation over the course of the session, by treatment.*

**Table A2.** Overall Intended cooperation in all treatments. Standard errors clustered on both participant and supergame pair; * p<0.1, ** p<0.05, *** p<0.01.

| | *Dependent variable: Overall Intended Cooperation* | | | |
|---|---|---|---|---|
| | *N low* | *M low* | *N high* | *M high* |
| Supergame | -0.004 | 0.002 | -0.004 | 0.005 |
| | (0.004) | (0.004) | (0.003) | (0.005) |
| Constant | 0.282*** | 0.192*** | 0.357*** | 0.405*** |
| | (0.032) | (0.028) | (0.039) | (0.041) |
| Observations | 7597 | 7737 | 9247 | 7624 |
| $R^2$ | 0.003 | 0.001 | 0.001 | 0.002 |

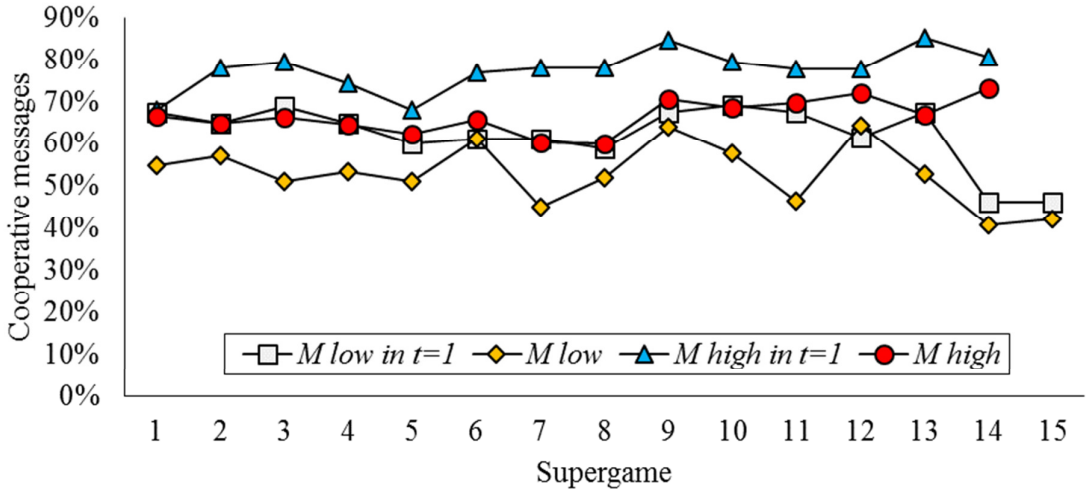


***Figure A3.*** *Fraction of "C" messages in the first period of a supergame, and overall, by treatment.*

**Table A3.** Cooperative messages in the first period of a supergame, and overall. Standard errors clustered on both participant and supergame pair; * p<0.1, ** p<0.05, *** p<0.01.

| | *Dependent variable: Cooperative messages* | | | |
|---|---|---|---|---|
| | *First period* | | *Overall* | |
| | *Low* | *High* | *Low* | *High* |
| Supergame | -0.006 | 0.008** | -0.003 | 0.004 |
| | (0.006) | (0.004) | (0.004) | (0.004) |
| Constant | 0.671*** | 0.719*** | 0.531*** | 0.628*** |
| | (0.471) | (0.044) | (0.034) | (0.042) |
| Observations | 952 | 968 | 7756 | 7630 |
| $R^2$ | 0.002 | 0.005 | 0.001 | 0.001 |

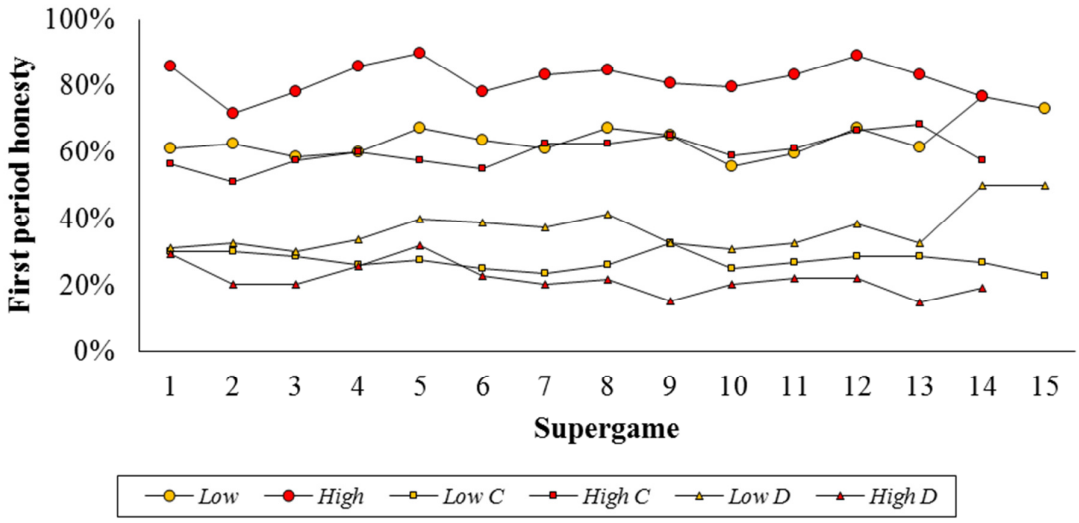


***Figure A4.*** *First period honesty by supergame played. The letter next to the treatment represents the combination of intended action and message.*

**Table A4.** First period honesty in all treatments with messages. Standard errors clustered on both participant and supergame pair; * p<0.1, ** p<0.05, *** p<0.01.

| | *Dependent variable: First period honesty* | | | |
|---|---|---|---|---|
| | *Cooperation, C(C)* | | *Defection, D(D)* | |
| | *Low* | *High* | *Low* | *High* |
| Supergame | -0.002 | 0.009* | 0.006 | -0.007* |
| | (0.006) | (0.005) | (0.006) | (0.004) |
| Constant | 0.288*** | 0.541*** | 0.321*** | 0.268*** |
| | (0.045) | (0.052) | (0.047) | (0.043) |
| Observations | 946 | 968 | 946 | 968 |
| $R^2$ | 0.001 | 0.004 | 0.002 | 0.004 |

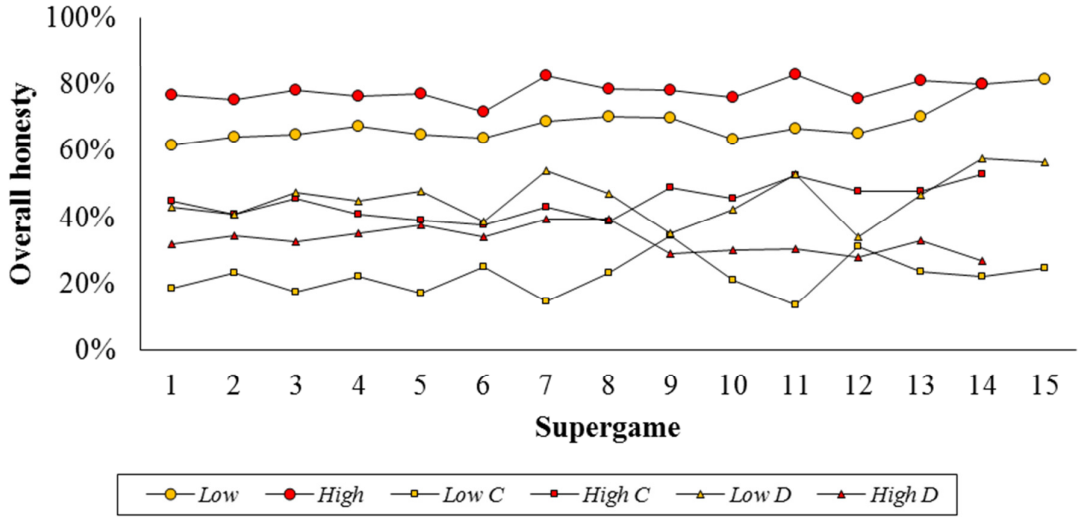


***Figure A5.*** *Overall honesty by supergame played. The letter next to the treatment represents*

*the combination of intended action and message.*

**Table A5.** Overall honesty in all treatments with messages. Standard errors clustered on both participant and supergame pair; * p<0.1, ** p<0.05, *** p<0.01.

| | *Dependent variable: Overall Honesty* | | | |
|---|---|---|---|---|
| | *Cooperation, C(C)* | | *Defection, D(D)* | |
| | *Low* | *High* | *Low* | *High* |
| Supergame | 0.003 | 0.006 | 0.004 | -0.003 |
| | (0.004) | (0.005) | (0.004) | (0.004) |
| Constant | 0.175*** | 0.396*** | 0.453*** | 0.363*** |
| | (0.027) | (0.041) | (0.034) | (0.041) |
| Observations | 7737 | 7624 | 7737 | 7624 |
| $R^2$ | 0.001 | 0.002 | 0.001 | 0.001 |

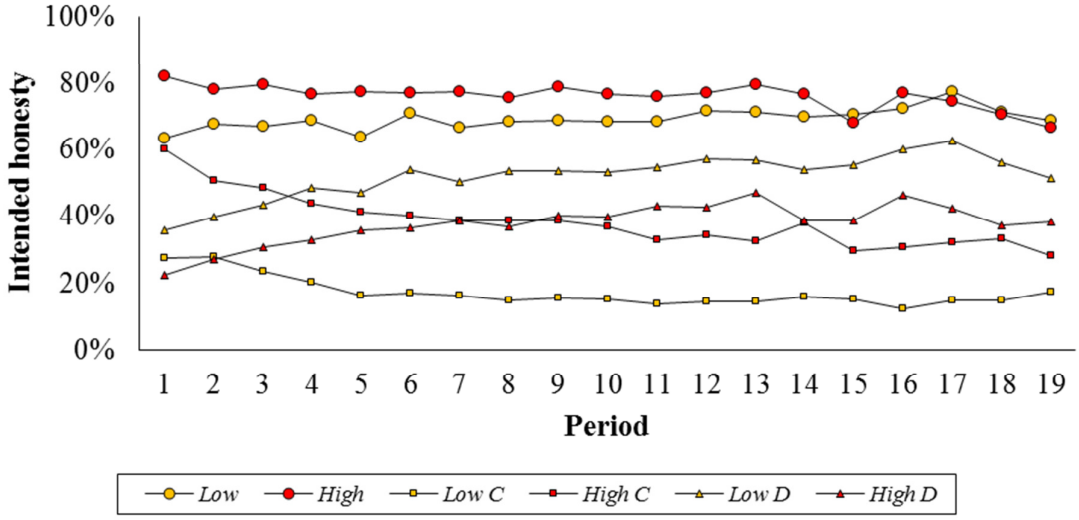


***Figure A6.*** *Mean honesty by period. The letter next to the treatment represents the combination of intended action and message.*

***Table A6.*** *The role of actions and intentions communicated in previous period for cooperation.*

| | *Low* | | *High* | | *Low & High* |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |

| | | | | | |
|---|---|---|---|---|---|
| Partner's action in t-1 (*A*) | 0.274*** | 0.067*** | 0.315*** | 0.075*** | 0.067*** |
| | (0.031) | (0.021) | (0.021) | (0.024) | (0.021) |
| Partner's message in t-1 (*M*) | 0.163*** | 0.093*** | 0.281*** | 0.209*** | 0.093*** |
| | (0.019) | (0.015) | (0.027) | (0.028) | (0.015) |
| A x M | | 0.305*** | | 0.303*** | 0.305*** |
| | | (0.035) | | (0.031) | (0.035) |
| High (*H*) | | | | | 0.057*** |
| | | | | | (0.021) |
| H x A | | | | | 0.008 |
| | | | | | (0.032) |
| H x M | | | | | 0.116*** |
| | | | | | (0.032) |
| H x A x M | | | | | -0.002 |
| | | | | | (0.047) |
| Constant | 0.036*** | 0.065*** | 0.090*** | 0.123*** | 0.067*** |
| | (0.010) | (0.010) | (0.018) | (0.018) | (0.021) |
| Observations | 6791 | 6791 | 6656 | 6656 | 13447 |
| $R^2$ | 0.178 | 0.201 | 0.255 | 0.267 | 0.284 |

***Table A7.*** *The role of actions and intentions communicated in t-2 for cooperation if the other decided to defect and sent the message "I choose A".*

| | *Low* | | *High* | | *Low & High* |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Partner's action in t-1 (*A*) | 0.204*** | 0.008 | 0.258*** | 0.035 | 0.008 |
| | (0.028) | (0.016) | (0.017) | (0.022) | (0.016) |
| Partner's message in t-1 (*M*) | 0.155*** | 0.089*** | 0.298*** | 0.232*** | 0.089*** |
| | (0.019) | (0.016) | (0.026) | (0.025) | (0.016) |
| A x M | | 0.287*** | | 0.281*** | 0.287*** |
| | | (0.029) | | (0.031) | (0.029) |
| High (*H*) | | | | | 0.048** |
| | | | | | (0.020) |
| H x A | | | | | 0.027 |
| | | | | | (0.027) |
| H x M | | | | | 0.142*** |
| | | | | | (0.029) |
| H x A x M | | | | | -0.006 |
| | | | | | (0.042) |
| Constant | 0.045*** | 0.072*** | 0.090*** | 0.120*** | 0.072 |
| | (0.010) | (0.011) | (0.017) | (0.017) | (0.011) |
| Observations | 5921 | 5921 | 5768 | 5768 | 11689 |
| $R^2$ | 0.128 | 0.150 | 0.222 | 0.232 | 0.247 |

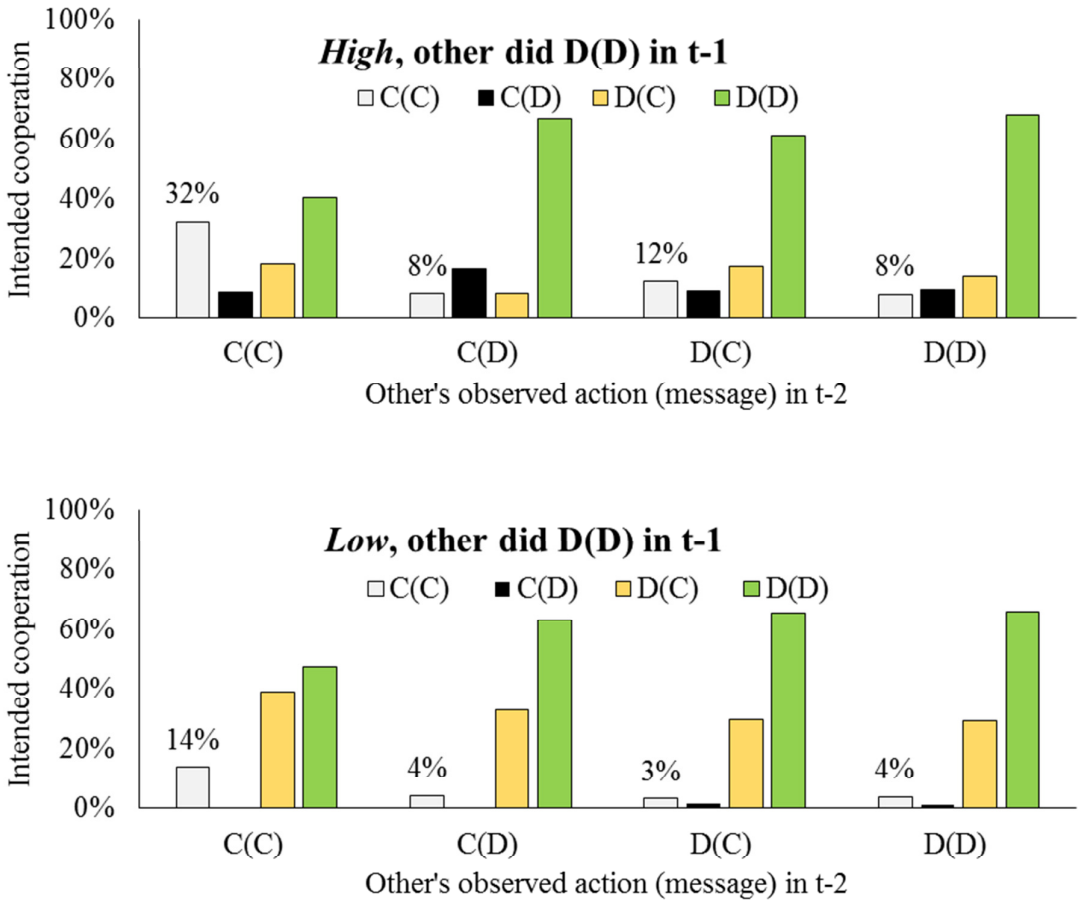


***Figure A7.*** *Intended response to observing other's defection and message "I choose B".*

***Table A8.*** *The role of the history of play for honest choices of player choosing D.*

| | *Low* (1) | *High* (2) | *Low & High* (3) |
|---|---|---|---|
| Average contribution | 0.217 | -0.109 | 0.008 |
| | (0.183) | (0.126) | (0.105) |
| Female | 0.095 | 0.182** | 0.131** |
| | (0.079) | (0.071) | (0.053) |
| Age | 0.004 | 0.036** | 0.018** |
| | (0.010) | (0.015) | (0.008) |
| High | | | -0.029 |
| | | | (0.050) |
| Constant | 0.410* | -0.252 | 0.146 |
| | (0.238) | (0.332) | (0.188) |
| N | 80 | 78 | 158 |
| $R^2$ | 0.036 | 0.160 | 0.063 |

***Table A9.*** *Maximum likelihood estimates for simulated histories*

| Strategy | *Actual* | *Estimated* |
|---|---|---|
| ***Low*** | | |
| ALLD(C) | 0.17 | 0.18*** (0.04) |
| ALLD(D) | 0.30 | 0.30*** (0.05) |
| D(C)-TFT that is punitive | 0.08 | 0.08*** (0.03) |
| D(C)-TFT that is punitive and defects using D(C) | 0.07 | 0.08*** (0.002) |
| D(C)-2TFT that ignores messages and cooperates using D(C) | 0.07 | 0.06** (0.03) |
| TF2T that is punitive and immediately punishes D(D) | 0.31 | 0.31*** (0.06) |
| Mental error | 0.29 | 0.00 |
| ***High*** | | |
| ALLD(C) | 0.06 | 0.06 (0.06) |
| ALLD(D) | 0.21 | 0.21*** (0.06) |
| GRIM1 that believes messages | 0.05 | 0.04* (0.03) |
| TFT that ignores messages | 0.05 | 0.05* (0.03) |
| TF2T that is punitive | 0.12 | 0.12*** (0.04) |
| TF2T that is punitive and defects using D(C) | 0.09 | 0.09** (0.04) |
| GRIM3 that ignores messages and immediately punishes after two periods of D(D) | 0.08 | 0.06** (0.02) |
| TF3T that is punitive and immediately punishes D(D) | 0.24 | 0.27*** (0.07) |
| 2TF2T that is punitive, immediately punishes D(D), and defects using D(C) | 0.11 | 0.10*** (0.03) |
| Mental error | 0.24 | 0.00 |

*Notes:* The prefix of a conditional strategy indicates the opening move; if no prefix is given, the opening move is C(C). *Punitive* refers to strategies that treat any move other than C(C) as defection. *Tolerant* refers to strategies that only treat D(D) as defection. Unless otherwise specified, strategies cooperate using C(C) and defect using D(D) (i.e. play C(C) when un-triggered, and by D(D) when triggered). Mental error is calculated as the probability that the chosen action is not the one recommended by the strategy. Bootstrapped standard errors (shown in parentheses) used to calculate p-values. ***p<0.01, **p<0.05, *p<0.10.

**Appendix B: Analysis restricted to the last four supergames played.**

*Here we present the results of our restricted analysis of the last four super games played.*

*Question 1.* We find similar differences in cooperation levels across treatments. Overall cooperation rates vary between 21% and 46% depending on the treatment; cooperation in the first period of each supergame vary between 26% and 63%. Figure B1 reveals that the ability to communicate increases cooperation levels, but only in the first period when there are cooperative equilibria (first period cooperation: *high*, $p$=0.088; *low*, $p$=0.281; overall cooperation: *high*, $p$=0.113; *low*, $p$=0.952).

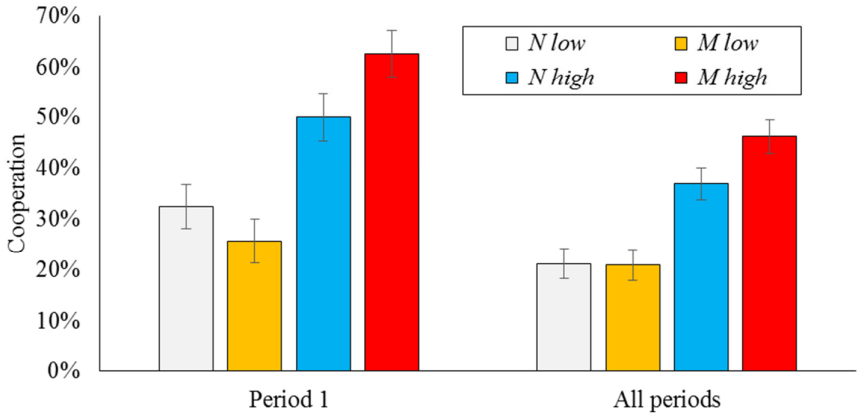


***Figure B1.*** *First period and overall cooperation, by treatment, averaged over the last four supergames of each session*

*Question 2.* Figure B2 is remarkably similar to Figure 4, the only notable differences is that candid cooperation in the *high* treatment occurs slightly more often (46% versus 44%) and honest defections slightly less (32% versus 34%).

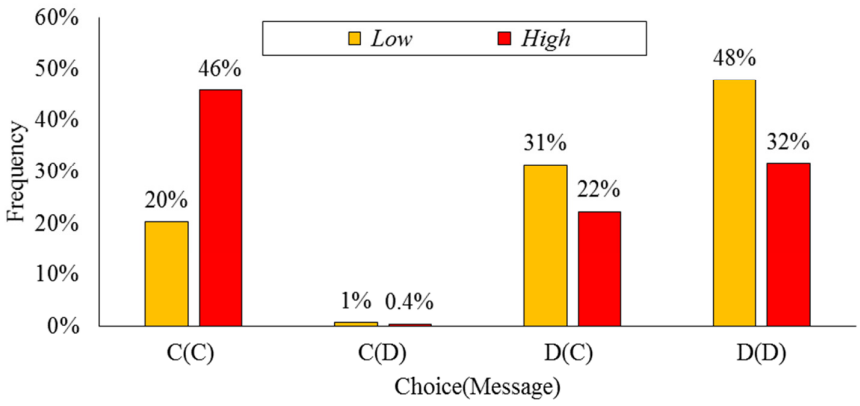

***Figure B2.*** *Frequency of intended actions in the message treatments, averaged over the last four supergames of each session*

Results on honest defections in the restricted dataset are also very similar to the ones found in the extended dataset: 60% overall in *low* and 59% in *high*; 51% in *low* and 51% in *high*, if we restrict our attention to the first period of each supergame. Also, Figure B3 shows a similar trend as before.

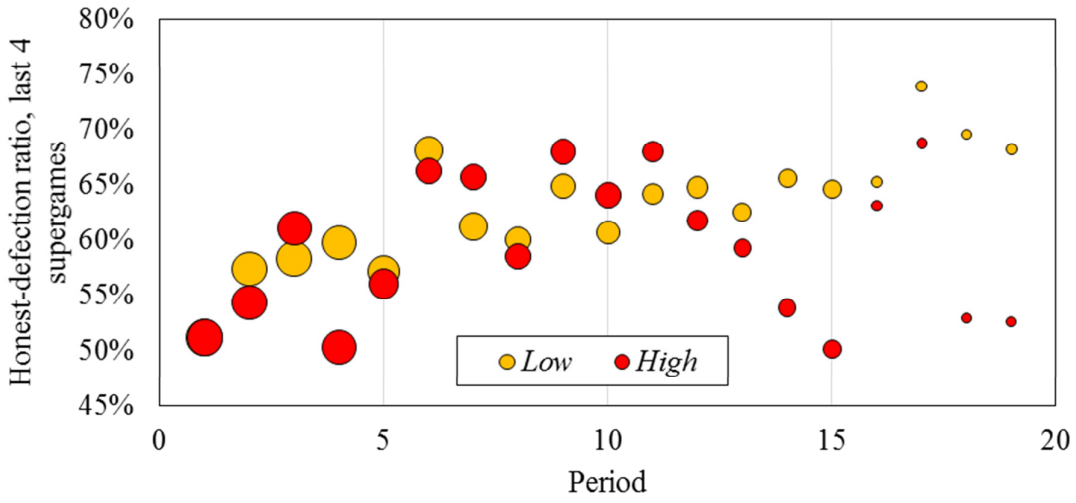


***Figure B3.*** *Honest-defection ratio (D(D)/[D(D)+D(C)]) over Period in the last 4 supergames; dot size is proportional to the number of observations in each period.*

Figure B4 reveals that in the last four supergames participants became slightly more honest. In the *low* and *high* treatments respectively, 71 (89%) and 65 (83%) participants are not honest at least once.

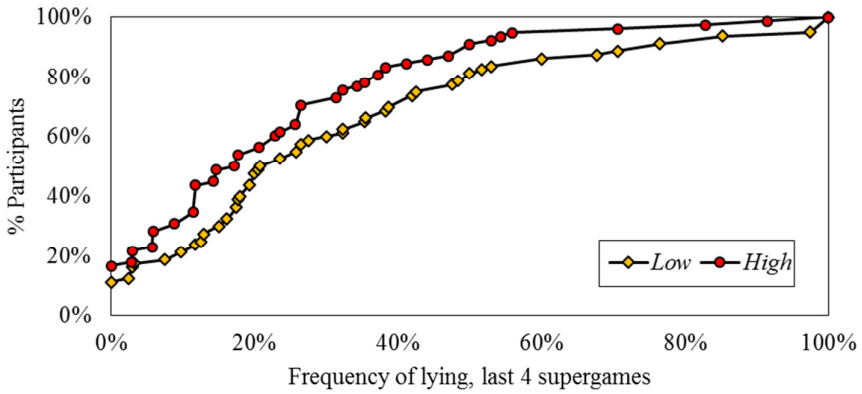

***Figure B4.*** *Cumulative distribution of participants who lied a determined number of times, averaged over the last four supergames of each session*

*Question 3.* We also find that a large proportion of the participants conditioned their responses on what their partner communicated. Figure B5 shows that when participants see that their partner both cooperated and signaled cooperation, 72% of the participants in *high* both cooperate and report cooperation. The corresponding number for *low* is significantly lower, 60% (p=0.051). In the event that the partner defected but sent the non-matching signaling indicating intended cooperation, participants in *high* cooperate candidly 37% of the time versus only 15% of the time in *high* (p=0.001).

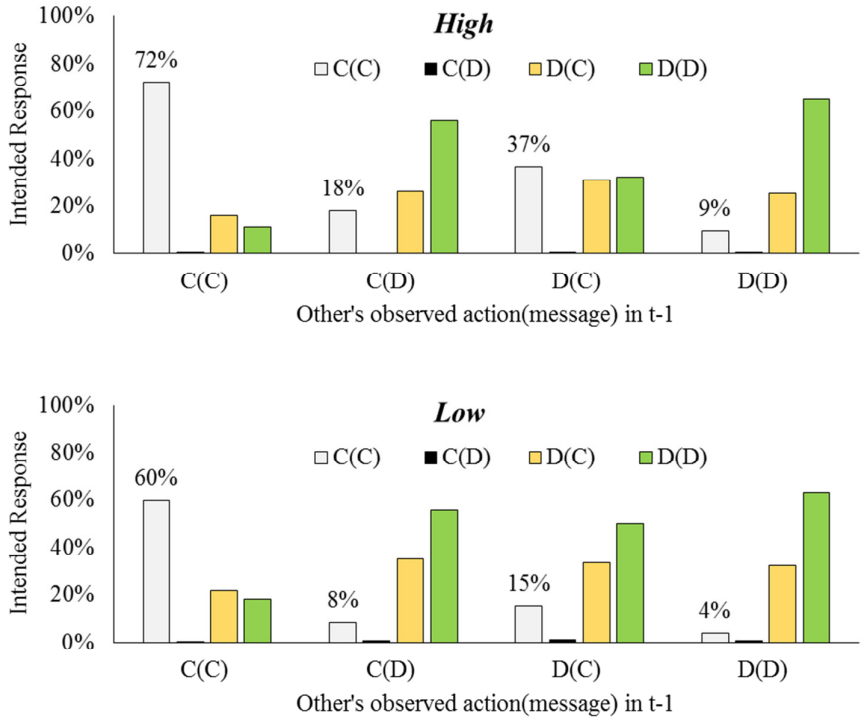


***Figure B5.*** *Intended response to other's observed action and message in the previous period, averaged over the last four supergames of each session*

Table B1 confirms a significant main effect of treatment and an interaction between treatment and partner's message. Interestingly, when including the interaction between action and message, the significance of partner's action in t-1 disappears.

***Table B1.*** *The role of actions and intentions communicated in previous period for cooperation, averaged over the last four supergames of each session*

| | *Low* | | *High* | | *Low & High* |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Partner's action in t-1 (*A*) | 0.318*** | 0.044 | 0.302*** | 0.083* | 0.044 |
| | (0.041) | (0.028) | (0.027) | (0.042) | (0.027) |
| Partner's message in t-1 (*M*) | 0.201*** | 0.117*** | 0.341*** | 0.73*** | 0.117*** |
| | (0.027) | (0.025) | (0.041) | (0.043) | (0.024) |
| A x M | | 0.392*** | | 0.276*** | 0.392*** |
| | | (0.044) | | (0.053) | (0.044) |
| High (*H*) | | | | | 0.051** |
| | | | | | (0.022) |
| H x A | | | | | 0.038 |
| | | | | | (0.050) |
| H x M | | | | | 0.156*** |
| | | | | | (0.050) |
| H x A x M | | | | | -0.116* |
| | | | | | (0.069) |
| Constant | 0.012 | 0.046*** | 0.065*** | 0.098*** | 0.046*** |
| | (0.011) | (0.011) | (0.021) | (0.019) | (0.010) |
| Observations | 2481 | 2481 | 2362 | 2362 | 4843 |
| $R^2$ | 0.245 | 0.281 | 0.283 | 0.293 | 0.335 |

A visual comparison between Figures 7 and B6 confirms that participants react similarly in the last four super games and during the whole session. If anything, we observe more cooperative players in Figure B5 in response to their partner cooperating and sending the message "I choose B". This is mainly due to the reduced number of observations (17), though.

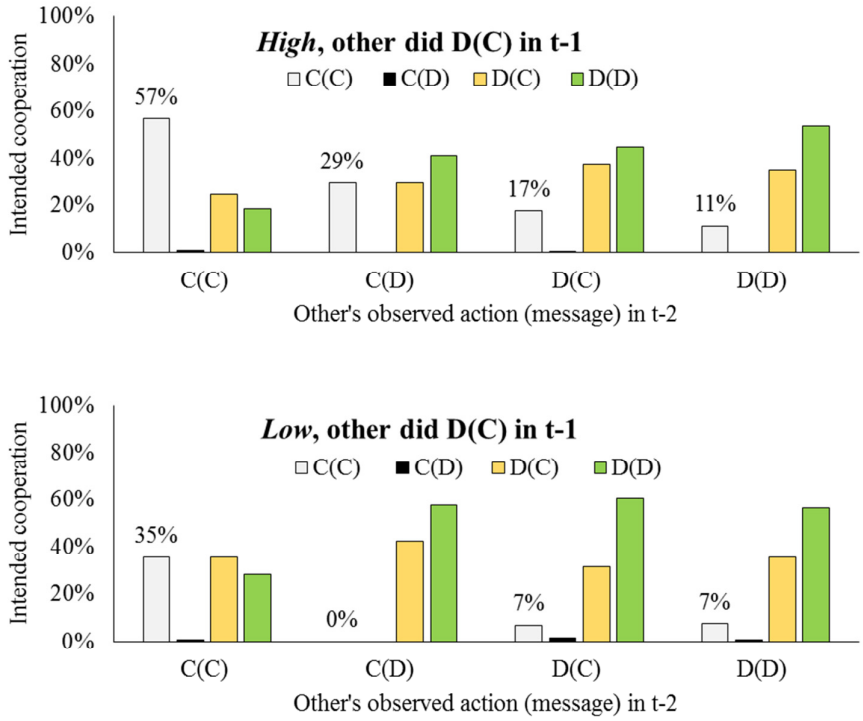


***Figure B6.*** *Intended response to observing other's defection and message "I choose A", averaged over the last four supergames of each session*

Not surprisingly, Table B2 shows that the number of participants who either always chose defection or always choose cooperation in period 1 increases when we restrict out attention the last four interactions. Most importantly, this Table also shows that participants who always open with D virtually never cooperate, so their play resembles the ALLD strategy, while participants who always open with C are more cooperative than intermediate participants.

***Table B2.*** *Overall cooperation for participants period 1 choice is always defection, a mix of defection and cooperation, and always cooperation.*

| | *Period 1 choice* | | |
|---|---|---|---|
| | *Always D* | *Mixed* | *Always C* |
| *N low* | 0.05 (N=42) | 0.32 (N=21) | 0.50 (N=15) |
| *N high* | 0.07 (N=31) | 0.45 (N=15) | 0.63 (N=30) |
| *M low* | 0.10 (N=49) | 0.31 (N=18) | 0.49 (N=13) |
| *M high* | 0.10 (N=23) | 0.47 (N=13) | 0.66 (N=42) |

We calculate payoffs as the average earned by each participants in each period played. As shown in Table B3, participants who always open with defection now only out-earn more cooperative participants in treatments with *low* payoffs.

***Table B3.*** *Average payoff for always open D(D), intermediate, always open C(C)*

| | *Participant (P) who always opens with D* | *Intermediate* | *Participant (P) who always opens with C* |
|---|---|---|---|
| *N low* | 0.38 (N=42) | 0.17 (N=21) | 0.11 (N=15) |
| *N high* | 0.75 (N=31) | 0.65 (N=15) | 0.91 (N=30) |
| *M low* | 0.34 (N=49) | 0.16 (N=18) | 0.23 (N=13) |
| *M high* | 0.94 (N=23) | 0.92 (N=13) | 0.96 (N=42) |

In Table B4 we also see that in the *low* treatment, participants that always opened with D(C) out-earned others regardless of the partner's opening move. In the *high* treatment, we see that the success of participants who always open with C(C) is driven by productive interactions with partners who also opened with C(C) too. Interestingly, we see that participants who always open with C(C) are not actually at a disadvantage relative to others when meeting a partner who opens with D(C).

***Table B4.*** *Average payoff by type of partner's opening in period 1. Shown in parentheses is the number of interactions in which each combination of participant's strategy and partner's opening move occurred.*

| | *Low* | | | *High* | | |
|---|---|---|---|---|---|---|
| *Other's Realized Period 1 Play* | *P who always opens with D(C)* | *P who always opens with C(C)* | *All other openings* | *P who always opens with D(C)* | *P who always opens with C(C)* | *All other openings* |
| *C(C)* | 0.73 (26) | 0.55 (16) | 0.56 (44) | 1.33 (19) | 1.37 (97) | 1.35 (59) |
| *C(D)* | 0.64 (2) | -0.10 (4) | 0.53 (11) | 1.17 (1) | -0.10 (4) | 0.14 (5) |
| *D(C)* | 0.36 (24) | 0.34 (14) | 0.07 (73) | 1.29 (14) | 0.58 (46) | 0.55 (27) |
| *D(D)* | 0.19 (24) | -0.09 (14) | 0.10 (68) | 0.63 (5) | 0.17 (21) | 0.18 (25) |

Table B5 confirms that also in our restricted dataset the risk of exploitation by liars is balanced out by the gains from being more cooperative when one's partner makes an honest mistake.

***Table B5.*** *Average payoff of players when other opens with D(C). Shown in parentheses is the number of interactions in which each combination of participant's strategy and partner's*

*opening move occurred.*

| | *Low* | | | *High* | | |
|---|---|---|---|---|---|---|
| | *P who always opens with D(C)* | *P who always opens with C(C)* | *All other openings* | *P who always opens with D(C)* | *P who always opens with C(C)* | *All other openings* |
| *D(C) intended* | 0.28 (22) | 0.25 (12) | 0.08 (70) | 0.00 (2) | 0.29 (30) | 0.38 (18) |
| *D(C) unintended* | 1.00 (2) | 0.74 (2) | -0.23 (3) | 1.50 (1) | 1.17 (16) | 0.86 (9) |

Table B6 shows that the results of an SFEM restricted to the last 4 supergames shows qualitatively similar results. That is, unconditional strategies are heavily used, and lenient strategies are found more often in treatments with high payoffs. Moreover, the mental errors slightly decrease in all treatments, which would suggest that participants err slightly less as the session nears its end.

***Table B6.*** *SFEM results for treatments with and without communication (last 4 supergames)*

| Strategy | *Low* | *High* |
|---|---|---|
| ***Treatments without communication*** | | |
| ALLC | 0.05* (0.03) | |
| ALLD | 0.52*** (0.06) | 0.38*** (0.07) |
| GRIM1 | 0.11*** (0.04) | |
| TFT | | 0.09** (0.04) |
| D-TFT | 0.06* (0.03) | 0.09** (0.04) |
| 2TFT | | 0.11** (0.04) |
| D-2TFT | 0.12** (0.05) | |
| TF2T | | 0.24*** (0.06) |
| GRIM3 | 0.13*** (0.04) | 0.09** (0.04) |
| Mental error | 0.09 | 0.10 |
| ***Treatments with communication*** | | |
| ALLD(C) | 0.15*** (0.04) | 0.07** (0.03) |
| D(C)-ALLD(D) | 0.08** (0.04) | |
| ALLD(D) | 0.35*** (0.06) | 0.20*** (0.04) |
| TFT that ignores messages | 0.19*** (0.06) | |
| D(C)-TFT that is punitive and defects using D(C) | 0.10** (0.04) | |
| D(D)-TFT that believes messages and cooperates using D(C) | | 0.05* (0.03) |
| GRIM2 that ignores messages | 0.12*** (0.05) | |
| GRIM2 that believes messages | | 0.16** (0.07) |
| TF2T that is punitive and immediately punishes D(D) | | 0.35*** (0.07) |
| TF3T that ignores messages, defects with D(C), and immediately punishes after two consecutive D(D) | | 0.18*** (0.05) |
| Mental error | 0.25 | 0.22 |

*Notes:* The prefix of a conditional strategy indicates the opening move; if no prefix is given, the opening move is C(C). *Punitive* refers to strategies that treat any move other than C(C) as defection. Unless otherwise specified, strategies cooperate using C(C) and defect using D(D) (i.e. play C(C) when un-triggered, and by D(D) when triggered). Mental error is calculated as the probability that the chosen action is not the one recommended by the

strategy. Bootstrapped standard errors (shown in parentheses) used to calculate p-values. ***p<0.01, **p<0.05, *p<0.10.

Table B7 and B8 complete our SFEM results with a look at the payoffs earned in the last four supergames. Similar to what we previously found, a large fraction of participants still chose unconditionally defective strategies. This type of strategies are heavily used in the *low* treatment without communication, and indeed the observed payoffs were the highest with this strategy. Moreover, in treatments with communication, lenient strategies pay off particularly well in *high,* whereas deceptive strategies earn the most in *low*.

***Table B7.*** *Observed frequencies and resulting payoffs for each strategy with messages (last 4 supergames)*

| | *Low* | | *High* | |
|---|---|---|---|---|
| *Strategy* | *Frequency* | *Observed (expected) payoff* | *Frequency* | *Observed (expected) payoff* |
| ALLD(C) | 0.15 | 0.35 (0.33) | 0.07 | 1.08 (1.44) |
| D(C)-ALLD(D) | 0.08 | 0.26 (0.30) | | |
| ALLD(D) | 0.35 | 0.24 (0.30) | 0.20 | 0.77 (0.72) |
| TFT that ignores messages | 0.19 | 0.27 (0.08) | | |
| D(C)-TFT that is punitive and defects using D(C) | 0.10 | 0.32 (0.30) | | |
| D(D)-TFT that believes messages and cooperates using D(C) | | | 0.05 | 0.95 (1.17) |
| GRIM2 that ignores messages | 0.12 | 0.29 (0.07) | | |
| GRIM2 that believes messages | | | 0.16 | 0.97 (1.00) |
| TF2T that is punitive and immediately punishes D(D) | | | 0.35 | 1.10 (1.13) |
| TF3T that ignores messages, defects with D(C), and immediately punishes after two consecutive D(D) | | | 0.18 | 1.08 (1.03) |

*Notes:* The prefix of a strategy indicates the opening move. If absent, participants start with C(C). *Punitive* refers to strategies that only treat C(C) as cooperation; unless otherwise specified, participants cooperate using C(C) and defect using D(D).

***Table B8.*** *Observed frequencies and resulting payoffs for each strategy without messages (last 4 supergames)*

| | *Low* | | *High* | |
|---|---|---|---|---|
| *Strategy* | *Frequency* | *Expected (actual)* | *Frequency* | *Expected (actual)* |

| | | *payoff* | | *payoff* |
|---|---|---|---|---|
| ALLC | 0.05 | -.24 (-.59) | | |
| ALLD | 0.52 | 0.35 (0.41) | 0.38 | 0.71 (0.80) |
| GRIM1 | 0.11 | 0.14 (0.25) | | |
| TFT | | | 0.09 | 0.97 (0.82) |
| D-TFT | 0.06 | 0.21 (0.22) | 0.09 | 0.64 (0.89) |
| 2TFT | | | 0.11 | 0.73 (0.82) |
| D-2TFT | 0.12 | 0.16 (0.31) | | |
| TF2T | | | 0.24 | 0.99 (0.75) |
| GRIM3 | 0.13 | 0.18 (-.02) | 0.09 | 0.81 (0.75) |

*Notes:* The prefix of a strategy indicates the opening move. If absent, participants start with C.