Published as:
"'I'm just a soul whose intentions are good': The role of communication in noisy repeated games."
Antonio A. Arechar, Anna Dreber, Drew Fudenberg, and David G. Rand (2017). Games and Economic Behavior. 104 726-743. DOI: 10.1016/j.geb.2017.06.013

"I'm just a soul whose intentions are good": The role of communication in noisy repeated games

By ANTONIO A. ARECHAR, ANNA DREBER, DREW FUDENBERG, AND DAVID G. RAND[*]

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 messages were honest (although the majority of participants lied at least occasionally). 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 lower return to cooperation, communication reduces overall cooperation. These results suggest that cheap talk communication can promote cooperation in repeated games, but only when there is already a self-interested motivation to cooperate.

First version: 3/16/2016

This version: 3/2/2017

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

JEL codes: C7, C9, D00

[*] Arechar: Department of Psychology, Yale University, Sheffield-Sterling-Strathcona Hall, Room 206, 1 Prospect Street, New Haven, CT 06511 (e-mail: antonio.alonso@yale.edu); Dreber: Department of Economics, Stockholm School of Economics, Box 6501, 113 83 Stockholm, Sweden (e-mail: anna.dreber@hhs.se); Fudenberg: Department of Economics, MIT, 50 Memoiral Drive, Cambridge, MA 02139, (e-mail: drew.fudenberg@gmail.com); Rand: Department of Psychology, Yale University, Sheffield-Sterling-Strathcona Hall, Room 204, 1 Prospect Street, New Haven, CT 06511 (e-mail: david.rand@yale.edu). 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 Ola Andersson, Tore Ellingsen, Daniele Nosenzo, Chris Starmer and Emmanuel Vespa for helpful conversations and comments.

I. 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 2011, 2015a, 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 interactions, such as relationships between colleagues, neighbors, friends, or romantic partners, are able to communicate with each other, and do so regularly. Communication may help to solve a major problem for repeated interactions, namely that actions are typically implemented with noise: while good intentions often lead to good outcomes, with randomness this is not always the case, so intentions cannot thus always be inferred from outcomes. One example is interactions between romantic partners, where most actions are (presumably) well-intended, but a friendly gesture might be misinterpreted as hostile. The industrial cartels discussed by Levenstein and Suslow (2015) and Harrington and Skrypcacz (2011) are another setting where cheap-talk communication has been thought to help participants cooperate (which here means to collude). In these settings, cost and market demand are often stochastic, which creates a form of imperfect monitoring, and cartels often use self-reported costs or sales as part of their collusive agreement.

To assess the role of communication in a noisy repeated environment, we conduct an experiment where we vary whether participants can communicate their intentions in the context of an infinitely repeated prisoner's dilemma with imperfect or "noisy" public monitoring of intended actions.[1] 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 one of two possible messages indicating the action they intended to play. This minimal form of communication has two advantages relative to richer forms (e.g. free-form messages). First, our approach makes it straightforward to analyze the content and truthfulness of messages. Second, it provides a conservative test of the effect of communication, as previous work in other domains (e.g. Charness and Dufwenberg 2010) suggests that free-form messages often have larger effects compared to the type of impersonal messages we used here.

The messages were transmitted without error, but there was a constant probability (known to the participants) that the *action* they intended was not the one that was realized. The message and the realized action were displayed simultaneously for the receiver, *after* the receiver had chosen her own message and action. The payoffs at each stage depended only on the realized 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

[1] Noise reduces cooperation when intentions cannot be observed, both theoretically (Kandori 1992) and in the lab (Aoyagi and Frechette 2009, Fudenberg et al. 2012).

[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).

equilibrium outcome when it would not be otherwise. In addition, there is experimental evidence that players in noisy repeated games attempt to infer their partners' 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 promote cooperation.

Given the lack of clear theoretical predictions but the evident importance of communication in daily life, we sought to collect empirical data exploring how communication is used in noisy repeated games, and what impact this has on cooperation levels. Our hope is to identify systematic ways in which communication affects behavior in order to inspire future theoretical work.

To this end, 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 supergame 47% percent of the time in the absence of communication, but 60% in the treatment with communication, so communication had a 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 had 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 by specifying a finite set of strategies and then using maximum likelihood estimation to estimate the share of participants choosing each strategy. 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.

Finally, we note that our design lets us study how the honesty of participants unfolds over the course of a given supergame. We find that participants are less likely to deceive their partners as the game develops. In particular, this is driven by subjects becoming more likely to admit to defection (perhaps because they dislike lying, or because they want it to be clear that their defection is a punishment).

Our tentative interpretation of these various findings is the following. First, the reason for the relatively low amount of 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 learns to cooperate, which benefits them. However, in the *low*

[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 findings in section 3.

treatment, the message “I meant to cooperate” isn’t credible, because cooperation isn’t supported by a reasonably simple equilibrium. Here, communication reduces cooperation, 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.

II. Literature review

Our paper is not the first to study some sort of communication in some sort of repeated cooperation game. The most closely related paper is that of Embrey et al. (2013), who explore an infinitely repeated partnership game with imperfect monitoring, where participants choose effort levels that determine the probability of the joint project being successful or not. There are some important differences between our study and Embrey et al. While Embrey et al. explore a binary public signal (“success” or “failure”), we have four possible signals each period (2 actions*2 messages). Embrey et al. let the participants sequentially exchange structured messages about their future plans, while in our setting participants send a message while they choose an intended action. While we set out to explore the role of communication in the noisy game in relation to revealing intentions, and to assess how honest communication is in different contexts, Embrey et al. study whether communication leads participants to play a renegotiation-proof equilibrium.

Most other experiments on communication in cooperation games look at a finite time horizon. In finitely repeated prisoner’s dilemmas, 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. There are also examples of infinitely repeated cooperation games with communication. For example, Andersson and Wengström (2007) explore communication in an infinitely repeated Bertrand game with perfectly observed actions and find that while costless communication can be detrimental for cooperation, costly communication enhances it. However, they do not explore the role of noise.

In treatments with cooperative equilibria, the infinitely repeated prisoner's dilemma has some features of a coordination game, and experimental evidence from coordination games suggests that communication sometimes but not always increases equilibrium play (see e.g., Cooper et al. (1992), Charness (2000), Andersson and Wengström (2012), Charness (2012), Embrey et al. (2013), and Cooper and Kühn (2014)).[4] There is also a substantial literature on communication in bargaining games; see, e.g. Charness (2012). Typically but not always, non-binding and costless communication in these games leads to more efficient outcomes, even in instances where the theoretical predictions say that communication should not matter.

Blume and Ortmann (2007) study the effect of communicating intentions in 9-player, 7-action versions of the median-effort and minimum-effort coordination games introduced by Van Huyck et al. (1990, 1991). Subjects played a single iteration of an 8-period game (so once again it is hard to tell which behaviors would be robust to learning) and were given vague instructions about how many periods the game would last. As in the "high" treatment in our study, they find that

[4] For example, Andersson and Wengström (2012) find that communication between a cooperation game and a subsequent coordination game is detrimental for cooperation, while Cooper and Kühn (2014) find the opposite. The difference in results may depend on the communication protocol, which is structured in Andersson and Wengström (2012) and free-form in Cooper and Kühn (2014).

communication leads to more efficient outcomes, which they attribute to a reduction in strategic uncertainty. In both games the majority of messages are both efficient and truthful, but in two of the 8 minimum-game sessions play “unraveled” to low/inefficient payoffs and a substantial number of untruthful messages.

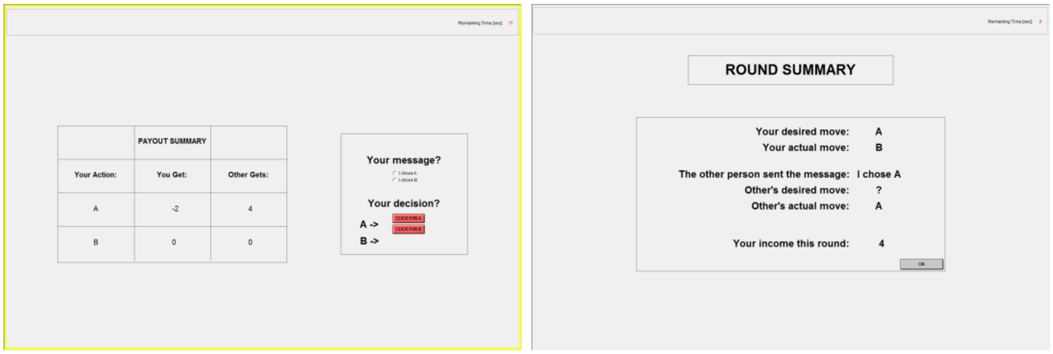
There is also a related literature on honesty or deception. For example, Gneezy (2005) explores a single one-shot interaction where one party can send a truthful or deceitful message, and a second party can choose whether or not to believe the message. His results suggest that a substantial fraction of participants have an aversion to lying and thus act honestly, and that participants were sensitive to both their own gains from lying and the costs imposed on others. Charness and Dufwenberg (2011) instead explore honesty in a principal-agent setup where agents can be of high or low ability, the agent type is unknown to the principal, and the payoff if the principal hires the agent depends on the agent’s type and choice, with high ability agents being safer to hire. Participants can send anonymous free-form messages to each other. The results from their small sample suggest that when low ability agents have the possibility to participate in a Pareto-improving outcome for both the agent and the principal, communication increases the rate at which this option is chosen from 40% to 78%. Moreover, many claims are actually honest: claims of being of low ability but with the intention to be nice to the principal are true in 15 out of 15 cases in this treatment (and tend to be trusted). However, there are also lies: some low ability participants claim to be of high ability (6/28), and interestingly, some choose to be silent (7/28). Another example that is relevant to our study is Vespa and Wilson (2016) who study honesty in an infinitely repeated sender-receiver game where there is uncertainty about a payoff-relevant state, the two players have misaligned interests and the sender can be honest or not. When participants are allowed to communicate freely before each supergame, efficiency and honesty increase.

III. Experimental design

We study infinitely repeated prisoner dilemmas with a constant continuation probability of δ=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-δ that the particular game ends and participants are re-matched to play another supergame.

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

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 (see Figure 1).[5]



	PAYOUT SUMMARY	
Your Action:	You Get:	Other Gets:
A	-2	4
B	0	0

Your message?
○ I chose A
○ I chose B

Your decision?
A ->
B ->

ROUND SUMMARY

Your desired move: A
Your actual move: B

The other person sent the message: I chose A
Other's desired move: ?
Other's actual move: A

Your income this round: 4

FIGURE 1. SCREENSHOT OF A SAMPLE DECISION SCREEN AND PERIOD SUMMARY SCREEN

[5] We use this two-message design rather than free-form communication or a design with the option not to send a message since one of our goals was to explore the strategies used. Allowing a third "no message" action would have made the strategy space much larger and more complex. Free-form communication would have made it even harder to analyze the strategies used; it would also have made it harder to see when people were communicating honestly.

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.[6] See Figure 2 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.

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 2. 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 2 (albeit with neutral language), but not the expected payoff matrix.

[6] 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.

For each treatment we performed three sessions. Within a session, a single treatment was implemented and participants played from 8 to 20 supergames, with most of the variation coming from how quickly participants made their decisions.[7] After each supergame, participants were randomly re-matched with another person in the room for a new supergame. 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.

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

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, 2015).

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, which are included in the online Appendix. The instructions were read out loud to

[7] Because play in some repeated game experiments systematically changes over the course of the session (e.g., Dal Bó and Frechette 2015b, Embrey et al. 2014), we let participants play at least eight supergames. The game lengths used in each supergame were 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.

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.[8]

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.[9] 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.[10]

Finally, we introduce the following terminological conventions. Intended cooperation is referred to as C and intended defection as D; realized actions have a circumflex (hat) added; and messages come immediately after their corresponding action and are indicated in parentheses (e.g. D(C) for intended defection paired with the message "I played C"; or $\hat{C}$(D) for realized cooperation paired with the message "I played D"). In treatments with communication, a player's pairing of action and message is referred to as a "response."

[8] 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.

[9] No participant ever had less than 7 units, and only 2 out of 312 participants ever dropped below 30 units.

[10] 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 of the 32,256 choices were random.

IV. 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 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. Moreover, previous experimental studies of communication in repeated games have had mixed results. 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 communicates their intentions. And past work gives reason to expect that at least some of our participants *will* be honest. As most of 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; 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, suggests 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 partners?*

To assess whether (and in which ways) play is affected by the partner's communicated intentions, we examine how players' likelihood of cooperating, and of signaling cooperation, depend on both their partners' prior actions and signaled intentions. 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 realized action to undermine a participant's faith in her partner's communication.

QUESTION 4: *What additional insight do we gain from the strategies outlined by the SFEM?*

We assess the strategies used 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.[11] As in previous work (Fudenberg et al 2012, Rand et al 2015), we believe that the strategies obtained with this method could further inform us regarding how participants use messages, are lenient, and condition on their partner's choices.

V. 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 intending to cooperate in the first period of each supergame did not vary over the course of the experiment (Figure 3), nor did the frequency of intended cooperation

[11] This method estimates the frequency of each strategy based on the histories of play. It relies on maximum likelihood estimation and assumes that all participants select a strategy from a common distribution and stick to it, but make mental errors in implementing that strategy, meaning that they sometimes choose an action other than what is prescribed by that strategy. In the games with messages, the method does not make any specific assumptions about the structure of the mental errors. Instead, it just scores each strategy's prediction for each action/message pair as "correct" or "incorrect" and the resulting mental error rate is the fraction of times the strategy is incorrect.

over all periods or the frequency of messages indicating cooperative intent.[12] Given this, we base our main analyses on decisions from all supergames to maximize the amount of data available (and report results in the online Appendix restricting to the last four supergames played, which look qualitatively equivalent).

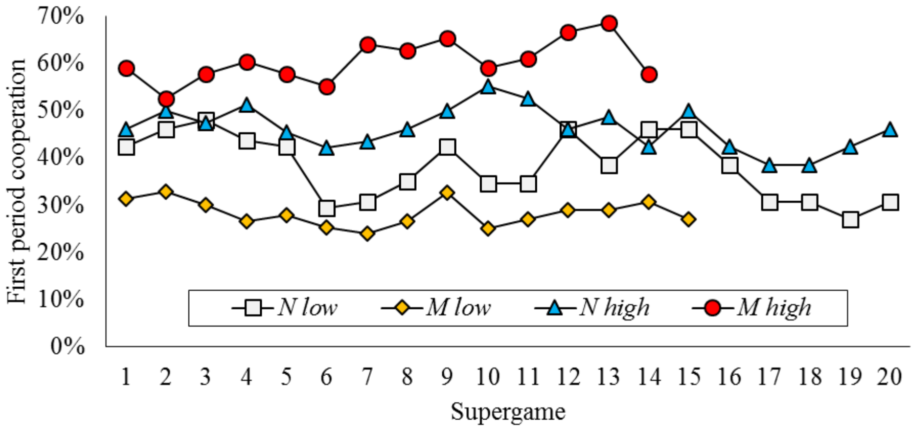


FIGURE 3. FIRST PERIOD INTENDED COOPERATION OVER THE COURSE OF THE SESSION, BY TREATMENT. "N" AND "M" DENOTE THE ABSENCE OR PRESENCE OF MESSAGES

Figure 3 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%.[13]

We now turn to our experimental questions.

[12] This is confirmed by treatment-specific linear regressions that control for the supergame played and are clustered on both participant and supergame pair. We use linear models rather than *logit* or *probit* because the coefficients produced are more interpretable, and note that our conclusions are the same regardless of the approach used. We find no change with supergame in intended first period cooperation (p*s*>0.153, Table A1 of the Appendix), intended cooperation over all periods (p*s*>0.261, Table A2 of the Appendix) or likelihood of indicating cooperative intent using messages over all periods (p*s*>0.358, Table A3 of the Appendix).

[13] 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) or in the sequence of game lengths implemented, particularly given Camerer et al. (2016)'s nearly exact replication of the Fudenberg et al. (2012) results using a CalTech participant pool and the same game lengths as in the original paper.

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 4 reveals that the ability to communicate increases cooperation levels (first period intended cooperation: no messages 47%, messages 60%, *p*=0.044; overall intended cooperation: no messages 33%, messages 44%, *p*=0.012).[14] Interestingly, allowing for communication results in a marginally significant *decrease* in first period cooperation in the *low* treatment (first period intended cooperation: no messages 39%, messages 28%, *p*=0.063; overall intended cooperation: no messages 25%, messages 21%, *p*=0.313).

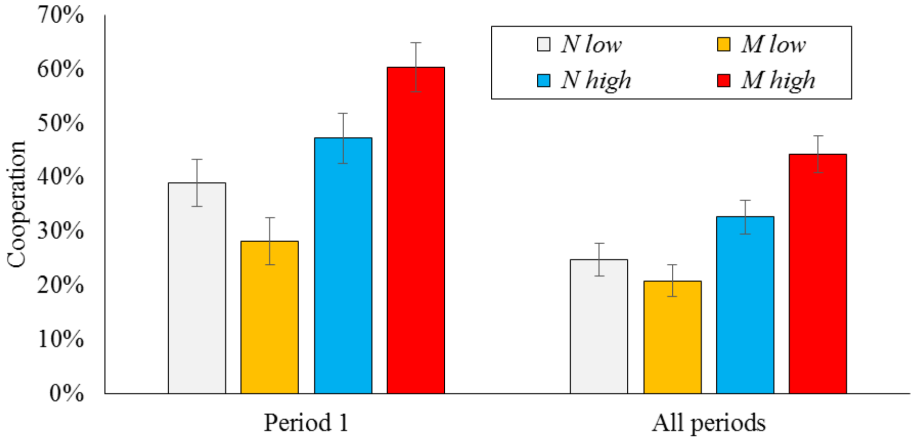


FIGURE 4. FIRST PERIOD AND OVERALL INTENDED COOPERATION, BY TREATMENT

These results show that cheap talk messages influence cooperation levels, and that they do so differently in the two payoff treatments. We explore this issue more thoroughly below, in Question 3.

[14] 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.

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

Figure 5 shows that participants are honest much of the time when communicating their intentions.[15] 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. C(C) occurred significantly more often in the *high* treatment than in the *low* treatment (44% versus 20%, *p*=0.001), whereas for D(D) 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 $\hat{\text{D}}$(C) outcomes in the *low* treatment were actual cases of accidental defection, compared to 24% in *high*.

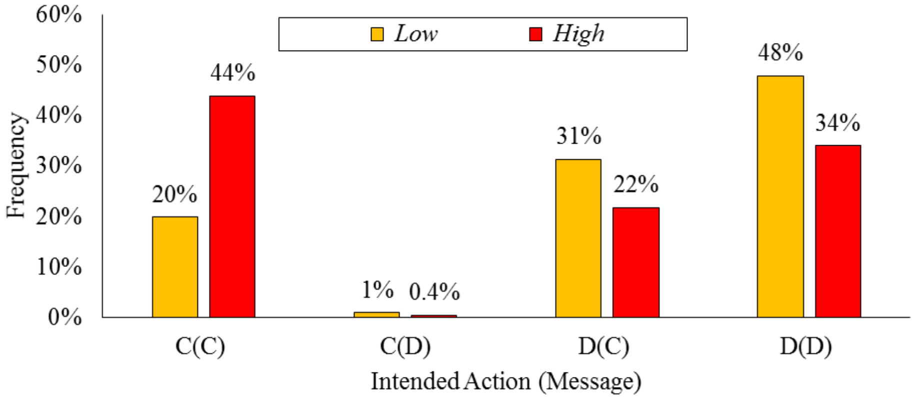


FIGURE 5. OVERALL FREQUENCY OF INTENDED ACTIONS IN TREATMENTS WITH COMMUNICATION

We therefore focus our subsequent discussion of honesty on cases where the intended action was D. In particular, we calculate an "honest-defection" index as the ratio D(D)/[D(D)+D(C)]. Using this measure, we find 60% honesty in the *low*

[15] For brevity, in this section we only discuss results when considering all periods of play; the results for first period play are qualitatively equivalent.

treatment and 61% honesty in the *high* treatment. This reveals that the greater overall honesty in the high treatment is driven by a greater level of cooperation, rather than reflecting an actual decrease in lying conditional on defecting.

Participants are less honest the first time they defect in a given supergame: 49% honest in the *low* treatment and 45% in the *high* treatment. Perhaps, after cooperation has broken down and there is no possibility of deceiving the partner, people switch to honest defection (suggesting an aversion to lying or that participants are investing in credibility for the future).

Furthermore, Figure 6 below reveals that the fraction of honest defections tends to increase over the course of a supergame in both treatments overall, and that the effect is mainly driven by instances where the other person played $\hat{D}$(D) in the previous period. This pattern may reflect defections later in the supergame being more likely to be used as punishment rather than attempted exploitation.

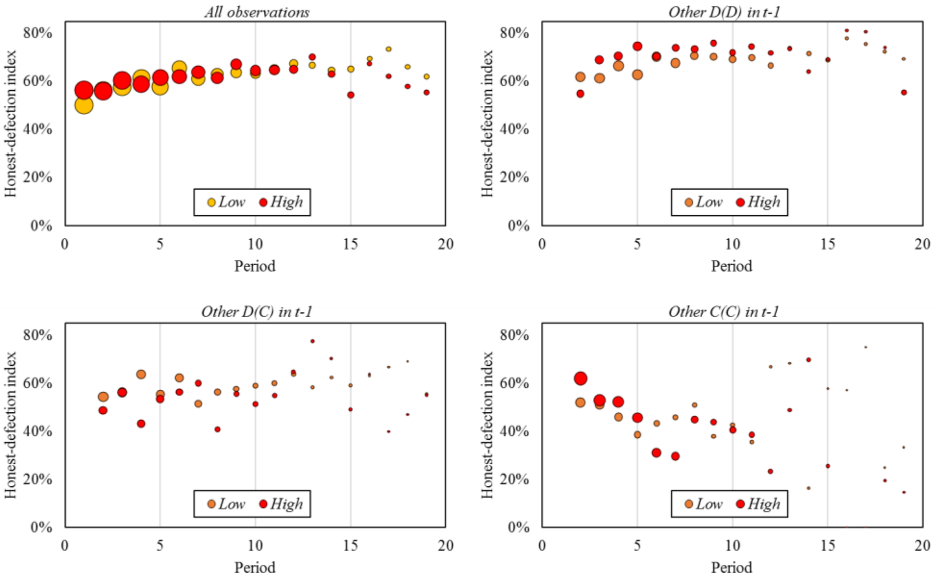


FIGURE 6. HONEST-DEFECTION INDEX (D(D)/[D(D)+D(C)]) BY PERIOD; DOT SIZE IS PROPORTIONAL TO THE NUMBER OF OBSERVATIONS IN EACH PERIOD

Figure 7 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.[16]

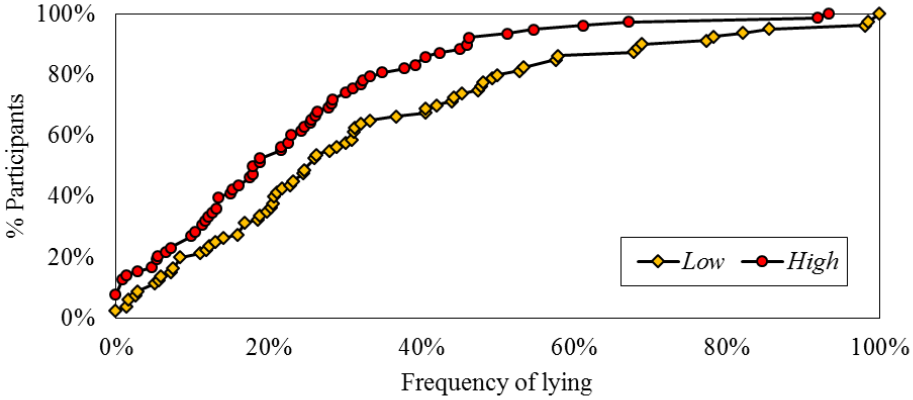


FIGURE 7. CUMULATIVE DISTRIBUTION OF PARTICIPANTS BY HOW OFTEN THEY LIED. TREATMENT MARKERS SHOW THE PERCENTAGE OF INDIVIDUALS THAT HAVE LIED LESS THAN X% OF THE TIME

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

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 partners communicated. Figure 8 reports intended responses to the message and realized action of the other player in the previous period. When participants saw

[16] 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 D" against her own overall cooperation, gender and age. We find significant positive effects for female gender and age (coeff = 0.131, p<0.014 and coeff = 0.018, p<0.031, respectively). 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. 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).

that their partner played $\hat{C}$(C), 71% of the participants in the *high* treatment responded in kind with C(C). 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 (played $\hat{D}$(C)), participants in the *high* treatment were more than twice as lenient as those in the *low* treatment: they responded with C(C) 33% of the time, versus only 14% of the time in *low* (*p*=0.001).[17]

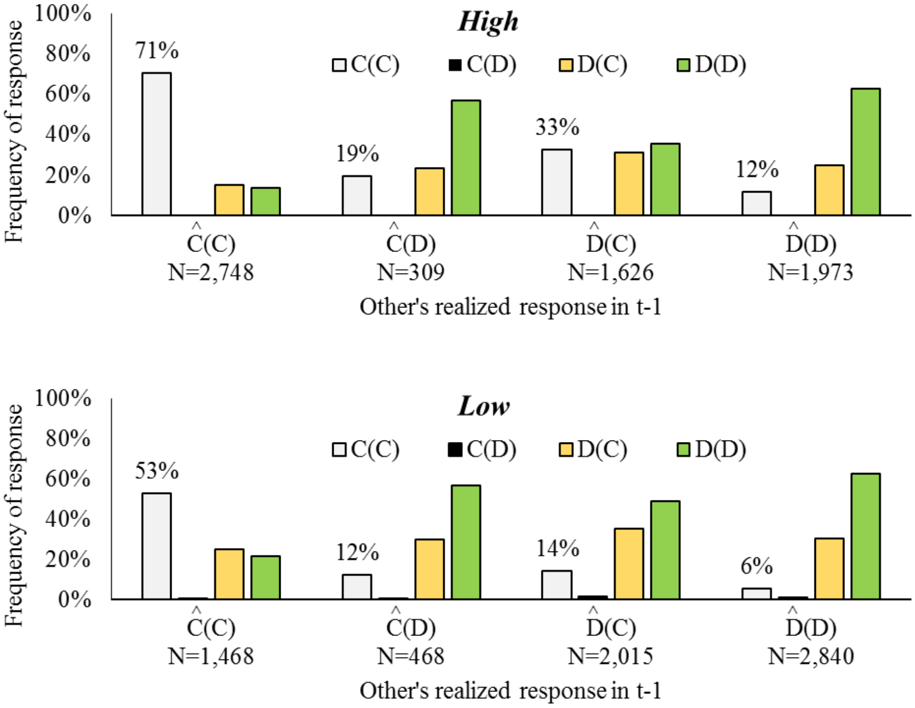


FIGURE 8. INTENDED RESPONSE TO OTHER'S REALIZED RESPONSE IN THE PREVIOUS PERIOD

Note that Figure 8 implicitly assumes that participants ignored all of the history of the supergame except for what happened in the previous period. Yet there is evidence that people use strategies that look back more than one period, especially

[17] Table A4 of the Appendix supports this observation by reporting linear regressions that predict cooperation in Period 2 of each supergame based on the partner's message and realized action in Period 1, including session fixed effects. In both treatments, we find significant positive effects of cooperative messages and cooperative actions (p<0.092 for all).

in games with noise (Fudenberg et al 2012, Rand et al 2015). This appears to be the case with our data too. Figure 9 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.

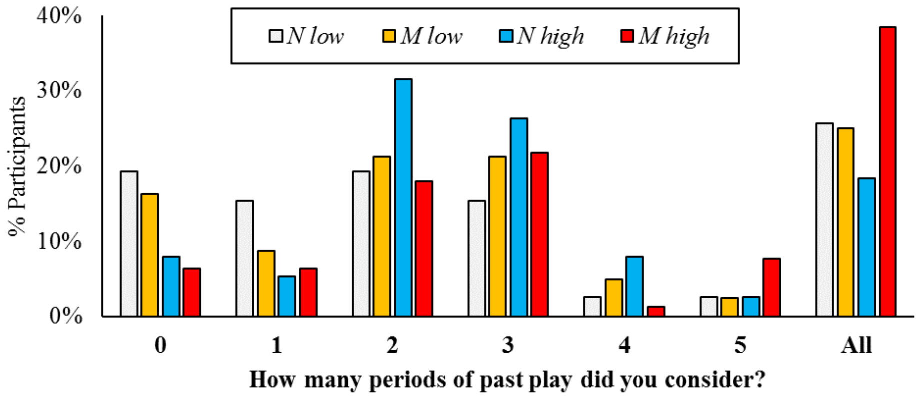


FIGURE 9. NUMBER OF PERIODS BACK THAT PARTICIPANTS SELF-REPORTED CONSIDERING

Thus we consider the extent to which participants conditioned on play two periods ago. In particular, we focus on the case where the partner defected but communicated the intention to cooperate (played $\hat{D}$(C)) one period ago (Figure 10). We see that in the *high* treatment, if the partner played $\hat{C}$(C) two periods ago, the $\hat{D}$(C) of one period ago was forgiven 52% of the time; compared to only 16% of the time if the partner also played $\hat{D}$(C) two periods ago. A similar pattern (but lower overall level of cooperation) is seen in the *low* treatment, with 31% cooperation if the partner played $\hat{C}$(C) two periods ago compared to 7% after two periods of $\hat{D}$(C). These results are confirmed statistically in Table A5 of the Appendix.

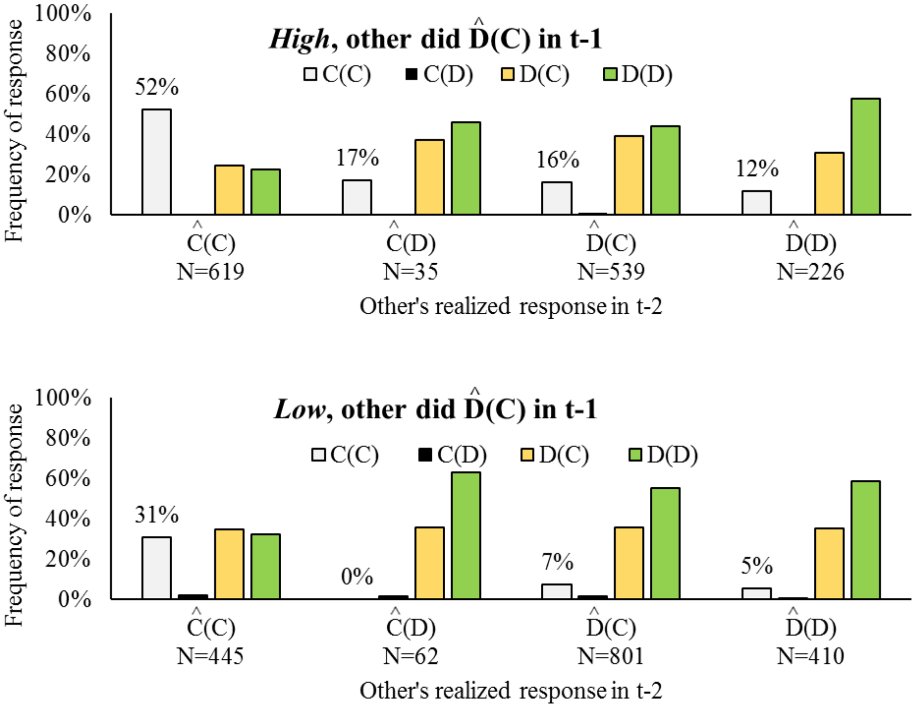


FIGURE 10. INTENDED RESPONSE TO OBSERVING OTHER'S DEFECTION AND MESSAGE "I CHOOSE C."

Interestingly, the impact of play two periods ago is different in the case where the partner played D̂(D) one period ago. Here, we see substantially less dependency on play two periods ago compared to when the partner played D̂(C) one period ago: people are less likely to cooperate after the partner plays D̂(D), even if two periods ago the partner played Ĉ(C): 32% cooperation in the *high* treatment (vs 52% above), 14% in the *low* (vs 31% above). (Figure A1 of the Appendix). This difference between D̂(C) and D̂(D) provides evidence that the signal part of D̂(C) had a substantial impact on play, promoting leniency.[18]

[18] 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 action in period *t*-1 (in histories where these were different). We find that in 34% of cases, the participant's current decision matched the

QUESTION 4: *What additional insight do we gain from the strategies outlined by the SFEM?*

To use the SFEM, 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. 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).[19]

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: always cooperate and send the C message (ALLC(C)), always defect and send the C message (ALLD(C)), or always defect and send the D message (ALLD(D)).[20]

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; GRIM2, 2TFT, TF2T and apologetic TFT (ATFT) which look back two periods; and GRIM3, 3TFT, TF3T, 2TF2T, which look back 3 periods.[21]

In treatments with communication, conditional strategies must specify which combinations of moves are considered "defection" (and therefore cause the strategy

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.

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

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

[21] As defined in prior work, GRIMX and TFXT are versions of GRIM/TFT with delayed triggers, and XTFT are versions of TFT that punish for X periods. Apologetic TFT plays TFT unless two periods ago, it accidentally defected while the partner cooperated, in which case it forgives a defection by the partner one period ago – this strategy is similar in spirit to Boyd (1989)'s "Contrite TFT". As in prior work, we assume that Grim strategies will be triggered by either player's defection.

to trigger). We therefore include versions of each of the above strategies that: *ignore messages* and treat both $\hat{D}(C)$ and $\hat{D}(D)$ as defection; *trust messages* and treat both $\hat{C}(D)$ and $\hat{D}(D)$ as defection; are *punitive* and treat anything other than $\hat{C}(C)$ as defection; or are *tolerant* and treat only $\hat{D}(D)$ 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 $\hat{D}(D)$, 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 $\hat{D}(D)$; or that trigger after observing two periods in a row of $\hat{D}(D)$).

In treatments with communication, there is also the question of which actions a strategy uses when the analogous no-communication strategy plays C versus D. Thus we included strategies that used C(C) for C and D(D) for D; C(C) for C and D(C) for D; or D(C) for C and D(D) for D.

In all treatments, we also include additional versions of each possible strategy that differ in their starting move: for example, C-ALLD starts by cooperating in the first period and then switches to ALLD for the rest of the supergame.[22] We also let strategies condition their response to a defection on their own actual play in the previous period – they can choose to tolerate a realized defection in the previous period (treat it as cooperation) if their own realized action was defection as a result of an error. For treatments with communication, this exception can apply to either $\hat{D}(C)$ and $\hat{D}(D)$ cases or to $\hat{D}(C)$ only cases.

[22] For treatments without communication the starting move could be either C or D; 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 43 strategies for treatments without communication, and 1713 strategies for treatments with communication.[23] To determine which of these possible strategies are most useful in describing the play of participants in our experiments, we use the following procedure.[24] First, 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).[25] We then removed all strategies that were not best predictors for at least two participants. Using this reduced set, we performed the SFEM 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).[26] Using the surviving strategies, we calculated the posterior probability of each strategy for each subject, and only kept strategies that were the most likely for at least one subject.[27] Finally,

[23] The total number of strategies with communication was actually 1845, but some strategies were excluded beforehand because their similarity with others in the full set 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 $\hat{D}$(D), or that start with D(C) and trigger when observing $\hat{D}$(C).

[24] The results in the text consider all supergames within each session when estimating strategy frequencies, because we find no evidence of learning (as described below). Moreover, in Table C6 of the online Appendix we find qualitatively similar results when restricting to the last four supergames.

[25] In the event of ties, we use the simplest strategy in terms of memory-rank, defined as follows: GRIM1> TFT> GRIM2> 2TFT> TF2T> GRIM3> 3TFT> TF3T> 2T2T.

[26] None of the strategies deleted in the first stage had a frequency greater than 0.07, and none of the remaining strategies in the second stage increased its frequency by more than 0.15.

[27] Using the estimated distribution *on* the surviving strategies, we use Bayes rule in our data to compute the posterior probabilities that a given strategy was used. We then eliminate strategies that were not the most likely for any subject (*M high:* from 8 to 3 strategies; *M low:* from 5 to 3 strategies; *N low:* from 4 to 2 strategies; *N high:* from 8 to 3 strategies), and finally combine the resulting sets in each communication condition (5 strategies using messages and 4 strategies not using messages).

we again performed SFEM on the reduced sets of strategies to arrive at a final estimate of strategy frequencies, which are presented in Tables 2 and 3 below.[28]

TABLE 2—SFEM RESULTS FOR TREATMENTS WITH MESSAGES

Strategy	*Low*	*High*
ALLD(C)	0.20	0.06
ALLD(D)	0.41	0.22
TFT that ignores messages, defects using D(C), and treats other's $\hat{D}$(C) or $\hat{D}$(D) in t-1 as C(C) if in period t-1 the subject accidentally defected	0.06	0.17
TF2T that immediately punishes $\hat{D}$(D), but waits for two periods of $\hat{D}$(C) or $\hat{C}$(D) before punishing	0.18	0.30
TF2T that is punitive, and treats other's $\hat{D}$(C) in t-1 as C(C) if in period t-1 the subject accidentally defected	0.16	0.25
Mental error	0.30	0.26

Notes: Punitive refers to strategies that treat any move other than $\hat{C}$ (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 D(D) when triggered). Mental error is calculated as the probability that the chosen action is not the one recommended by the strategy.

TABLE 3—SFEM RESULTS FOR TREATMENTS WITHOUT MESSAGES

Strategy	*Low*	*High*
ALLD	0.63	0.44
ATFT	0.19	0.33
2TF2T	0.04	0.13
2TF2T that treats other’s $\hat{D}$ in t-1 as C if in t-1 the subject accidentally defected	0.14	0.11
Mental error	0.15	0.14

Note: Mental error is calculated as the probability that the chosen action is not the one recommended by the strategy.

[28] For the treatments with communication, we tested the validity of the estimation procedure on simulated data. To this end, we employed a similar approach to earlier work without messages (Fudenberg et al. 2012) in that we assigned strategies to 80 computer agents in *low*, and 78 computer agents in *high* (i.e. we used the same numbers of experimental subjects). We then assigned strategies to the computer agents based on the estimated strategy frequency distribution shown in Table 2. Next, we took these 80 (or 78) agents and generated simulated histories of play of a total of 12 supergames, with agents being randomly paired for each supergame and the set of game lengths being similar to the ones induced experimentally. For each game, we recorded the message, intended action, and actual action for each agent in each period. As in the experiment, our agents also faced a 1/8 probability that their intended action would not be implemented. We then used the SFEM on the simulated data using the same approach as in our analysis of the experimental data (do the SFEM, eliminate strategies not significant at 10% level and redo the SFEM, keep only strategies with posterior probabilities that were most likely for at least one subject and redo the SFEM again). Results, shown in in Table A6 of the Appendix, reveal good consistency between the strategy distribution programmed into the agents and the strategy distribution estimated by the SFEM.

Compared to treatments without communication (and prior repeated PDs without messages), our treatments with communication have substantially higher rates of "mental errors" (that is, the probability that the chosen action is not the one recommended by the strategy); these errors can be viewed as a measure of how well the specified strategy set fits the data. 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.[29] Thus, the few surviving strategies in our estimation might best be thought of as stand-ins for classes of similar strategies (as it is unclear how well the method can make fine distinctions). We also note that adding back in further strategies only causes a very minimal decrease in the estimated error rate, on the order of 1 or 2 percent.

Note that in line with Fudenberg et al. (2012), we find that a substantial proportion of participants play strategies that always defect (ALLD(D) and ALLD(C)), and that the cooperative strategies are lenient and forgiving.[30]

Like the descriptive results used to answer our second question, the SFEM indicates that participants were honest much of the time in both treatments, and that there was more honesty (driven by higher cooperation rates) in the *high* treatment: in the *low* treatment, the always-lying strategy ALLD(C) and a version of TFT that punished using D(C) had a combined probability of 26%; while in the *high* treatment, this combined probability was 23%. All other strategies never lied. Thus we find convergent evidence in support of a high level of honesty among our

[29] Indeed, the improvement in metal error rate compared to random choice is roughly the same for the games with and without messages: from 50% error to 14% or 15% error in games without messages, and from to 66% error (assuming that C(D) is never chosen, so there are 3 possible options) to 26% or 30% in the games with messages. Embrey et al. (2013) observe a similar mental error rate as in our communication treatment in a game with three choices.

[30] Not surprisingly, although the overall pattern is similar, the exact frequencies reported here differ from Fudenberg et al. (2012). For a more direct comparison, see Table A7 in the Appendix that uses the strategy set from Fudenberg et al. (2012). Including these additional strategies results in only a very small reduction in mental error (from 0.15 to 0.12 in *Low,* and from 0.14 to 0.13 in *High*) at the expense of substantial increase in the number of strategies (from 4 to 11).

participants.

The SFEM results are also consistent with the answer to our third question, showing that participants considered more than just the last period when making their decisions. In particular, strategies that looked back more than one period had probability weights of 34% in *low* and 55% in *high* with messages; and 37% in *low* and 56% in *high* without messages.

Consistent with the descriptive results, we also note that the strategies that condition on messages look back two periods in their assessment of messages: they do not initially punish when a defecting partner sends a cooperate message, but switch to punishing after two 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.

VI. Discussion

We now ask which behaviors were most successful by examining how participants' payoffs relate to their willingness to cooperate, and to believe their partners' messages. From the outcomes in the analogous no-message treatment of Fudenberg et al. (2012), we expect in the no-message *low* treatment that more cooperative strategies will earn lower payoffs. 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 estimated strategy. We begin by using participants' intended first period actions as a rough proxy for their strategy. In particular, we compare payoffs between participants classified into three types based on their period 1 play: consistent defectors (who intended to cooperate in 25% or fewer of supergame period 1s), intermediate (who intended to cooperate in more than 25% but fewer than 75% of supergame period 1s), and consistent cooperators (who intended to cooperate in 75% or more of supergame period 1s). We believe that these differences in opening moves are a reasonable proxy for strategies more generally because, as shown in Table 4, participants who open with C 25% or less rarely cooperate in any other periods, so their play resembles the ALLD strategy. Conversely, participants who open with C 75% or more are more cooperative overall than intermediate participants.

TABLE 4—OVERALL COOPERATION RATES (EXCLUDING PERIOD 1) FOR PARTICIPANTS BY PERIOD 1 ACTION: AT LEAST 75% D, A MIX OF D AND C, AT LEAST 75% C

		Period 1 choice	
	D ≥75%	*Mixed*	*C ≥75%*
N low	0.10 (N=38)	0.27 (N=21)	0.42 (N=19)
N high	0.11 (N=34)	0.33 (N=12)	0.54 (N=30)
M low	0.09 (N=47)	0.32 (N=22)	0.39 (N=11)
M high	0.11 (N=23)	0.43 (N=15)	0.58 (N=40)

Using this classification system to examine payoffs, Table 5 shows that consistent defectors tend to out-earn more cooperative participants in the *low* treatments. In the *high* treatments, the opposite is true: consistent cooperators earn the highest payoffs, although the difference does not reach statistical significance.[31]

TABLE 5—OVERALL PAYOFF FOR PARTICIPANTS BY PERIOD 1 ACTION: AT LEAST 75% D, A MIX OF D AND C, AT LEAST 75% C

	D ≥75%	*Mixed*	*C ≥75%*
N low	0.39 (N=38)	0.26 (N=21)	0.21 (N=19)
N high	0.76 (N=34)	0.62 (N=12)	0.79 (N=30)
M low	0.34 (N=47)	0.17 (N=22)	0.24 (N=11)
M high	0.87 (N=23)	0.82 (N=15)	0.95 (N=40)

[31] Linear regression with robust standard errors and session fixed effects predicting overall payoff shows a significant positive effect for a "D ≥75%" dummy when analyzing the data from the *low* treatments (b=0.155; p<0.001), and a non-significant positive effect for a "C ≥75%" dummy when analyzing the data from the *high* treatments (b=.065; p=0.112).

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, 20% of participants opened at least 75% of the time by lying (i.e. playing D(C)), and earned substantially more per period (0.43 MUs) than other participants (0.24 MUs). In the *high* treatment, 51% of the participants opened at least 75% of the time with C(C), and out-earned (0.95 MUs) other participants combined (0.85 MUs).[32]

This suggests that persistent dishonest defection paid off when the returns to cooperation were low, while honest cooperation paid off when the returns to cooperation were high. To try to understand why this might be, we examine how payoffs vary based on the partner's opening move (Table 6).

TABLE 6—AVERAGE PAYOFF BY PARTNER'S REALIZED FIRST PERIOD OUTCOME

	Low		*High*	
Other's Realized Period 1 Play	*Opens with D(C)≥75%*	*Opens with C(C)≥75%*	*Opens with D(C)≥75%*	*Opens with C(C)≥75%*
$\hat{C}(C)$	0.80	0.64	1.32	1.37
	(56)	(40)	(53)	(282)
$\hat{C}(D)$	0.75	0.39	0.98	0.22
	(5)	(7)	(4)	(13)
$\hat{D}(C)$	0.29	0.06	0.71	0.63
	(52)	(50)	(15)	(118)
$\hat{D}(D)$	0.14	0.02	0.34	0.24
	(61)	(37)	(23)	(97)

Notes: Shown in parentheses is the number of supergames in which each combination of participant's strategy and partner's opening move occurred.

We see that in the *low* treatment, participants that usually opened with D(C) out-earned others regardless of the partner's opening move. In the *high* treatment,

[32] In addition to being more initially cooperative, we find that participants who open with C(C) at least 75% of the time are more lenient of D(C) in the *high* treatment: when their partner's realized outcome is D(C) in period 1, participants who consistently open with C(C) are substantially likely to cooperate in period 2 (60%C) compared to other participants (48%C). Furthermore, this leniency is specifically driven by sensitivity to the message: when the partner opened with D(D), participants who usually open with C(C) are not any more likely to cooperate (35%C) than other participants (36%C).

participants who usually opened with C(C) out-earned others when they were matched with partners who opened with $\hat{C}$(C), but when matched with partners that opened with $\hat{D}$(C) they were out-earned by participants who usually opened with D(C).

Finally, we complement these analyses based on first period cooperation with an analysis of payoffs based on the SFEM results. To do so, we first ask which of the strategies listed in Tables 2 and 3 had the highest posterior likelihood for each subject. For each strategy, we then calculate the average payoff per period over all participants identified with that strategy. Because these payoffs depend on who they were matched with and the realizations of the monitoring errors, this is a noisy estimate of their expected payoff against a randomly drawn member of the participant pool.

We also compute the pairwise payoffs for each combination of the SFEM strategies by averaging over 100,000 simulated supergames; we then calculate expected payoff for each strategy by weighting these payoffs based on the estimated strategy frequencies.[33] Table 7 shows the estimated frequency of each strategy, along with the observed and expected payoffs.

Interestingly, while ALLD(D) performs poorly, the consistently dishonest ALLD(C) is actually the best performing strategy in *low* because it is capable of exploiting the strategy that trusts messages. Yet ALLD(D) is substantially more common than ALLD(C) – perhaps because lying is psychologically costly, as suggested by e.g. the one-shot experiments of Gneezy (2005), or because the ALLD(D) players failed to learn that ALLD(C) was more profitable. Similarly, in *high*, ALLD(C) performs much better than ALLD(D) (and performs roughly as well as the cooperative strategies), but is almost never played. And unlike in *low*,

[33] As in prior work, in the simulations strategies were implemented without mental error.

in *high* the various cooperative strategies all out-perform ALLD(D): in *high*, the average cooperative player earned an observed (expected) average payoff per round of 0.91 (1.13), compared to 0.84 (0.77) for the average non-cooperative player; conversely, in *low* the average cooperative player earned 0.25 (0.16) compared to 0.30 (0.36) for the average non-cooperative player.[34]

TABLE 7—STRATEGY FREQUENCIES AND TWO MEASURES OF THEIR PAYOFFS, TREATMENTS WITH MESSAGES

	Low		*High*	
Strategy	*Frequency*	*Observed (expected) payoff*	*Frequency*	*Observed (expected) payoff*
ALLD(C)	0.20	0.39 (0.49)	0.06	0.99 (1.09)
ALLD(D)	0.41	0.26 (0.30)	0.22	0.80 (0.68)
TFT that ignores messages, defects using D(C), and treats other's $\hat{D}$(C) and $\hat{D}$(D) in t-1 as C(C) if in period t-1 the subject accidentally defected.	0.06	0.34 (0.17)	0.17	1.01 (1.10)
TF2T that immediately punishes $\hat{D}$(D), but waits for two periods of $\hat{D}$(C) or $\hat{C}$(D) before punishing	0.18	0.25 (0.19)	0.30	0.84 (1.14)
TF2T that is punitive, and treats other's $\hat{D}$(C) in t-1 as C(C) if in period t-1 the subject accidentally defected	0.16	0.21 (0.12)	0.25	0.94 (1.14)

Notes: Punitive refers to strategies that only treat $\hat{C}$(C) as cooperation; unless otherwise specified, participants cooperate using C(C) and defect using D(D) (i.e. play C(C) when un-triggered, and D(D) when triggered).

For treatments without messages, we perform a similar analysis and report the results in Table 8. In the *low* treatment, ALLD is by far the most common strategy,

[34] Average payoffs are computed by averaging the payoffs of the three cooperative strategies (or two non-cooperative strategies), weighted by the estimated frequency of each strategy.

and substantially out-performs the other more cooperative strategies. Consistent with our findings regarding treatments with messages, we find that cooperative and lenient strategies are more frequent in the *high* treatment. However, without messages, such strategies do not substantially out-perform ALLD even in the *high* treatment: in *low* the average cooperative player earned an observed (expected) average payoff per round of 0.20 (0.08), compared to 0.36 (0.39) for the average non-cooperative player, and in *high* the average cooperative player earned 0.74 (0.81) compared to 0.75 (0.74) for the average non-cooperative player.

TABLE 8—STRATEGY FREQUENCIES AND TWO MEASURES OF THEIR PAYOFFS, TREATMENTS WITHOUT MESSAGES

	Low		*High*	
Strategy	*Frequency*	*Observed (expected) payoff*	*Frequency*	*Observed (expected) payoff*
ALLD	0.63	0.36 (0.39)	0.44	0.75(0.74)
ATFT	0.19	0.30 (0.17)	0.33	0.75(0.86)
2TF2T	0.04	0.22 (0.02)	0.13	0.79(0.74)
2TF2T that treats other's $\hat{D}$ in t-1 as C if in t-1 the subject accidentally defected	0.14	0.07 (-0.03)	0.11	0.63(0.72)

Taken together, these observations suggest that the ability to send messages improves the performance of cooperative strategies relative to non-cooperative strategies: the average cooperative player substantially out-earns the average non-cooperative player in the *high* treatment with messages, but not elsewhere.

VII. 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 (but almost all participants are sometimes dishonest), 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.

Our past work on the role of intentions in noisy repeated games (Rand et al. 2015) shows that when the partner’s intended and actual actions are both 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 communication is only effective in raising cooperation levels in the *high* treatment where cooperative equilibria exist even without revealed intentions. 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).

Our 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.

APPENDIX

TABLE A1—INTENDED COOPERATION IN THE FIRST PERIOD 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

Notes: The dependent variable is the intended cooperation in the first period of each supergame, per treatment. We report standard errors clustered on both participant and supergame pair.

*** Significant at the 1 percent level.
** Significant at the 5 percent level.
* Significant at the 10 percent level.

TABLE A2—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

Notes: The dependent variable is the overall intended cooperation, per treatment. We report standard errors clustered on both participant and supergame pair.

*** Significant at the 1 percent level.
** Significant at the 5 percent level.
* Significant at the 10 percent level.

TABLE A3—COOPERATIVE MESSAGES IN THE FIRST PERIOD OF A SUPERGAME, AND OVERALL

	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

Notes: The dependent variable is the fraction of cooperative messages, per treatment. We report standard errors clustered on both participant and supergame pair.

*** Significant at the 1 percent level.
** Significant at the 5 percent level.
* Significant at the 10 percent level.

Table A4—The role of actions and intentions communicated in period 1 for cooperation in Period 2

	Low (L)		*High (H)*		*L & H*
	(1)	(2)	(3)	(4)	(5)
Partner’s realized	0.212***	0.164*	0.213***	0.157*	0.164*
action in t-1 (*A*)	(0.048)	(0.091)	(0.048)	(0.093)	(0.091)
Partner’s message	0.093**	0.082**	0.188***	0.172***	0.082**
in t-1 (*M*)	(0.037)	(0.038)	(0.049)	(0.057)	(0.038)
A x M		0.059		0.066	0.059
		(0.103)		(0.107)	(0.102)
High (*H*)					0.014
					(0.087)
H x A					-0.007
					(0.130)
H x M					0.090
					(0.068)
H x A x M					0.007
					(0.148)
Constant	0.210***	0.216***	0.252***	0.260***	0.216***
	(0.058)	(0.059)	(0.063)	(0.062)	(0.050)
Session f.e.	Yes	Yes	Yes	Yes	Yes
Observations	870	870	888	888	1758
R^2	0.084	0.085	0.101	0.101	0.139

We report standard errors clustered on both participant and supergame pair.
*** Significant at the 1 percent level.
** Significant at the 5 percent level.
* Significant at the 10 percent level.

Table A5—The role of actions and intentions communicated in t-2 for cooperation if the other decided to defect and sent the message “I choose C” in Period 3

	Low (L)		*High (H)*		*L & H*
	(1)	(2)	(3)	(4)	(5)
Partner’s realized	0.113***	0.012	0.122***	0.099	0.012
action in t-2 (*A*)	(0.034)	(0.066)	(0.038)	(0.080)	(0.066)
Partner’s message in	0.130***	0.106***	0.274***	0.268***	0.106***
t-2 (*M*)	(0.036)	(0.037)	(0.047)	(0.056)	(0.036)
A x M		0.126		0.026	0.126*
		(0.077)		(0.103)	(0.077)
High (*H*)					0.013
					(0.080)
H x A					0.087
					(0.104)
H x M					0.162**
					(0.066)
H x A x M					-0.100
					(0.128)
Constant	0.142***	0.156***	0.247***	0.250***	0.156***
	(0.044)	(0.045)	(0.055)	(0.057)	(0.045)
Session f.e.	Yes	Yes	Yes	Yes	Yes
Observations	870	870	865	865	1735
R^2	0.050	0.053	0.098	0.099	0.135

We report standard errors clustered on both participant and supergame pair.
*** Significant at the 1 percent level.
** Significant at the 5 percent level.
* Significant at the 10 percent level.

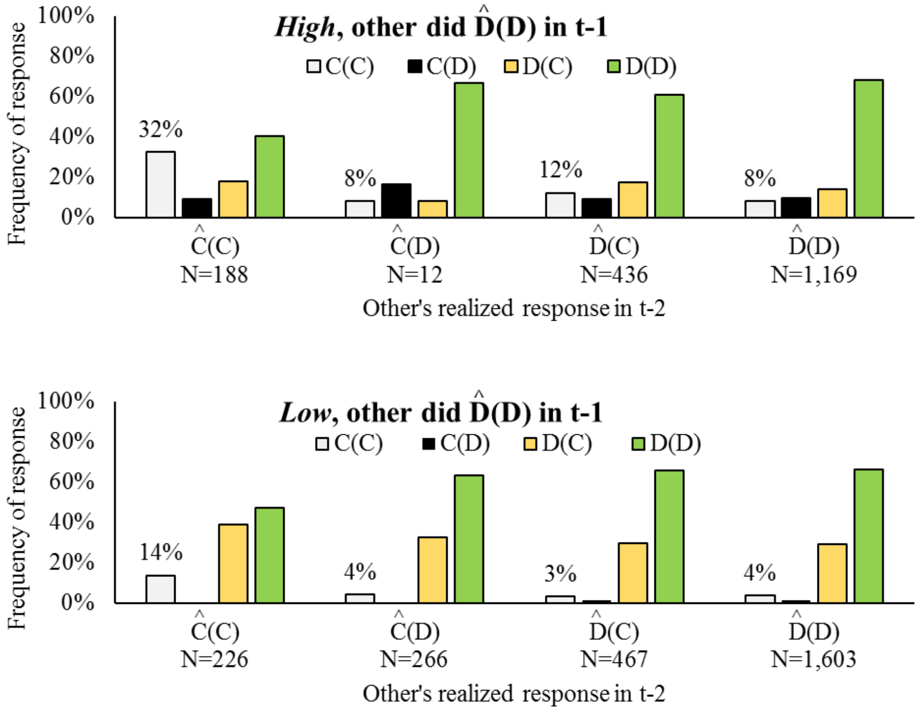


FIGURE A1. INTENDED RESPONSE TO OBSERVING OTHER'S DEFECTION AND MESSAGE "I CHOOSE D"

TABLE A6— MAXIMUM LIKELIHOOD ESTIMATES FOR SIMULATED HISTORIES, TREATMENTS WITH MESSAGES

Strategy	*Low*	*High*
ALLD(C)	0.20	0.06
ALLD(D)	0.41	0.22
TFT that ignores messages, defects using D(C), and treats other's $\hat{D}$(C) or $\hat{D}$(D) in t-1 as C(C) if in period t-1 the subject accidentally defected	0.05	0.17
TF2T that immediately punishes $\hat{D}$(D), but waits for two periods of $\hat{D}$(C) or $\hat{C}$(D) before punishing	0.18	0.31
TF2T that is punitive, and treats other's $\hat{D}$(C) in t-1 as C(C) if in period t-1 the subject accidentally defected	0.16	0.24
Mental error	0.00	0.00

Notes: Punitive refers to strategies that treat any move other than $\hat{C}$ (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 D(D) when triggered). Mental error is calculated as the probability that the chosen action is not the one recommended by the strategy.

TABLE A7— SFEM COMPARISON FOR TREATMENTS WITHOUT MESSAGES USING THE STRATEGY SET IN FUDENBERG ET AL. (2012)

	Treatments without messages		*Fudenberg et al. 2012*	
	Low	*High*	*b/c=1.5*	*b/c=2*
ALLC	0.01	0.01	0	0.03
TFT	0.01	0.10***	0.19***	0.06
TF2T	0.02	0.09**	0.05	0
TF3T	0.01	0	0.01	0.03
2TFT	0.12***	0.10**	0.06	0.07*
2T2T	0.06*	0.13***	0	0.11**
GRIM1	0.13***	0.02	0.14***	0.07
GRIM2	0.03	0.06**	0.06*	0.18***
GRIM3	0.05*	0	0.06	0.28***
ALLD	0.44***	0.37***	0.29***	0.17***
D-TFT	0.10***	0.12***	0.14***	0
Mental error	0.12	0.13	0.10	0.12

*** Significant at the 1% level.

** Significant at the 5%.

* Significant at the 10% level.

REFERENCES

Andersson, O., Wengström, E., 2007. Do antitrust laws facilitate collusion: Experimental evidence on costly communication in duopolies. Scandinavian Journal of Economics 109(2), 321-339.

Andersson, O., Wengström, E., 2012. Credible communication and cooperation: Experimental evidence from multi-stage games. Journal of Economic Behavior & Organization, 81, 207-219.

Aoyagi, M., Frechette, G., 2009. Collusion as public monitoring becomes noisy: Experimental evidence. Journal of Economic Theory 144**(**3**)**, 1135–65.

Blonski, M., Ockenfels, P., Spagnolo, G., 2011. Equilibrium selection in the repeated prisoner's dilemma: Axiomatic approach and experimental evidence. American Economic Journal: Microeconomics 3**(**3**)**, 164–92.

Blume, A., Ortmann, A., 2007. The effects of costless pre-play communication: Experimental evidence from games with Pareto-ranked equilibria. Journal of Economic Theory 132(1), 274-290.

Bochet, O., Page, T., Putterman, L., 2006. Communication and punishment in voluntary contribution experiments. Journal of Economic Behavior & Organization, 60, 11-26.

Boyd, R., 1989. Mistakes allow evolutionary stability in the repeated prisoner's dilemma. Journal of Theoretical Biology, 136, 47-56.

Camerer, C.F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen M., Wu, H., 2016. Evaluating replicability of laboratory experiments in economics. Science. 351(6280), 1433-1436.

Charness, G., 2000. Self-serving cheap talk: A test of Aumann's Conjecture. Games and Economic Behavior, 38, 177-194.

Charness, G., 2012. Communication in bargaining experiments. The Oxford _Handbook of Economic Conflict Resolution. Editors: Rachel Croson and Gary E. Bolton.

Charness, G., Dufwenberg, M., 2010. Bare promises. Economics Letters 107, 281-283.

Charness, G., Dufwenberg, M., 2011. Participation. American Economic Review, 101, 1211-1283.

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

Compte, O., 1998. Communication in repeated games with imperfect private monitoring. Econometrica, 66(3), 597–626.

Cooper, D.J., Kühn, K.-U., 2014. Communication, renegotiation, and the scope for collusion. American Economic Journal: Microeconomics, 6(2), 247-78.

Cooper, R., DeJong, D.V., Forsythe, R., Ross, T.W., 1992. Communication in coordination games. The Quarterly Journal of Economics 107(2), 739-771.

Dal Bó, P., 2005. Cooperation under the shadow of the future: Experimental evidence from infinitely repeated games. American Economic Review 95**(**5**)**, 1591–1604.

Dal Bó, P., Frechette, G.R., 2011. The evolution of cooperation in infinitely repeated games: Experimental evidence. American Economic Review 101**(**1**)**, 411–29.

Dal Bó, P., Frechette, G.R., 2015a. On the determinants of cooperation in infinitely repeated games: A survey. Forthcoming in the Journal of Economic Literature.

Dal Bó, P., Frechette, G.R., 2015b. Strategy choice in the infinitely repeated prisoners dilemma. Working paper.

Dreber, A., Johannesson, M., 2008. Gender differences in deception. Economics Letters, 99(1), 197-199.

Embrey, M., Fréchette, G.R., Stacchetti, E., 2013. An experimental study of imperfect public monitoring: Efficiency versus renegotiation-proofness. Working Paper.

Embrey, M., Fréchette, G.R., Yuksel, S. 2014. Cooperation in the finitely repeated prisoner's dilemma. Working Paper.

Erat, S., Gneezy, U., 2012. White lies. Management Science, 58(4), 723-733.

Fischbacher, U., 2007. Z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10**(**2**)**, 171–78.

Fudenberg, D., Levine, D.K., 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, D., Levine, D.K., Maskin, E., 1994. The folk theorem in repeated games with imperfect public information. Econometrica, 62, 997-1039.

Fudenberg, D., Rand, D.G., Dreber, A., 2012. Slow to anger and fast to forgive: Cooperation in an uncertain world. American Economic Review 102(2), 720-749.

Gneezy, U., 2005. Deception: The role of consequences, American Economic Review, 95(1), 384-394.

Greiner, B., 2015. Subject pool recruitment procedures: organizing experiments with ORSEE. Journal of the Economic Science Association, 1(1), 114-125.

Harrington, J.E., Skrzypacz, A., 2011. Private monitoring and communication in cartels: Explaining recent collusive practices. American Economic Review, 101(6), 2425-2449.

Kandori, M., 1992. The use of information in repeated games with imperfect monitoring, Review of Economic Studies, 59: 581-594.

Kandori, M., Matsushima, H., 1998. Private observation, communication and collusion, Econometrica, 66(3), 627–652.

Levenstein, M.C., Suslow, V.Y., 2015. Cartels and collusion - empirical evidence. The Oxford Handbook of International Antitrust Economics, vol. 2, R. Blair and D. Sokol, eds., Oxford University Press.

Rand, D.G., Fudenberg, D., Dreber, A., 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, D.G., Nowak, M.A., 2013. Human cooperation. Trends in Cognitive Sciences 17, 413-425.

Van Huyck, J.B., Battalio, R.C., Beil, R.O., 1990. Tacit coordination games, strategic uncertainty, and coordination failure. American Economic Review, 80(1), 234-48.

Van Huyck, J.B., Battalio, R.C., Beil, R.O., 1991. Strategic uncertainty, equilibrium selection, and coordination failure in average opinion games. Quarterly Journal of Economics, 106(3), 885-910.

Vespa, E., Wilson, A.J., 2016. Information transmission under the shadow of the future: An experiment. Working Paper.

Online Appendix

“I’m just a soul whose intentions are good”: The role of communication in noisy repeated games

Appendix B — Sample Instructions

Here we provide a sample copy of the experimental instructions used in our treatment “M High”. The instructions for the other treatments were adapted accordingly.

Instructions:

Thank you for participating in this experiment.

Please read the following instructions carefully. If you have any questions, do not hesitate to ask us. Aside from this, no communication is allowed during the experiment.

This experiment is about decision making. You will be randomly matched with other people in the room. None of you will ever know the identity of the others. Everyone will receive a fixed show-up amount of £10 for participating in the experiment. In addition, you will be able to earn more money based on the decisions you and others make in the experiment. Everything will be paid to you in cash immediately after the experiment.

You begin the session with 50 units in your account. Units are then added and/or subtracted to that amount over the course of the session as described below. At the end of the session, the total number of units in your account will be converted into cash at an exchange rate of 30 units = £1.

The Session

The session is divided into a series of interactions between you and other participants in the room.

In each interaction, you play a random number of rounds with another person. In each round you and the person you are interacting with can choose one of two options. In each round, you and the other person also send a message to each other about the action you each chose. Once the interaction ends, you get randomly re-matched with another person in the room to play another interaction.

The setup will now be explained in more detail.

The round

In each round of the experiment, the same two possible options are available to both you and the other person you interact with: A or B.

The payoffs of the options (in units)

Option	You will get	The other person will get
A:	−2	+4
B:	0	0

If your move is A then you will get −2 units, and the other person will get +4 units.

If you move is B then you will get 0 units, and the other person will get 0 units.

Calculation of your income in each round:

Your income in each round is the sum of two components:
- the number of units you get from the move you played
- the number of units you get from the move played by the other person.

Your round-total income for each possible action by you and the other player is thus

Other person

		A	B
You	A	+2	-2
	B	+4	0

For example:
If you play A and the other person plays A, you would both get +2 units.
If you play A and the other person plays B, you would get -2 units, and they would get +4 units.
If you play B and the other person plays A, you would get +4 units, and they would get -2 units.
If you play B and the other person plays B, you would both get 0 units.

Your income for each round will be calculated and presented to you on your computer screen.

The total number of units you have at the end of the session will determine how much money you earn, at an exchange rate of 30 units = £1.

Each round you must enter your choice within 30 seconds, or a random choice will be made.

A chance that the your choice is changed

There is a 7/8 probability that the move you choose actually occurs. But with probability 1/8, your move is changed to the opposite of what you picked. That is:

When you choose A, there is a 7/8 chance that you will actually play A, and 1/8 chance that instead you play B. The same is true for the other player.

When you choose B, there is a 7/8 chance that you will actually play B, and 1/8 chance that instead you play A. The same is true for the other player.

Both players are informed of the moves which actually occur. Neither player is informed of the move *chosen* by the other. Thus with 1/8 probability, an error in execution occurs, and you never know whether the other person's action was what they chose, or an error.

For example, if you choose A and the other player chooses B then:

- With probability (7/8)*(7/8)=0.766, no changes occur. You will both be told that your move is A and the other person's move is B. You will get -2 units, and the other player will get +4 units.

- With probability (7/8)*(1/8)=0.109, the other person's move is changed. You will both be told that your move is A and the other person's move is A. You both will get +2 units.

- With probability (1/8)*(7/8)=0.109, your move is changed. You will both be told that your move is B and the other person's move is B. You will both get +0 units.

- With probability (1/8)*(1/8)=0.016, both your move and the other person's moves are changed. You will both be told that your move is B and the other person's move is A. You will get +4 units and the other person will get -2 units.

Ability to send a message in each round

When choosing a move, you will also choose a message that will be sent to the other person.

In each round of the experiment, the same two possible messages are available to both you and the other person you interact with.

The messages are:

I chose A

I chose B

To send a message, first select your message then click the move you want to play. You will not be able to select a move without first selecting a message.

After you both make your selections, you will both be shown the move that actually occurred for you and for the other person, as well as the message that you and the other person sent. (Unlike your actions, there is NOT a chance that your message will be changed – messages are always shown exactly as chosen.)

Each round you must send a message within 30 seconds, or a random message will be chosen.

Random number of rounds in each interaction

A random number generator has determined how many rounds each interaction will have. After each round, the random number generator placed 7/8 probability on the interaction continuing for at least one more round, and 1/8 probability on the interaction ending. After each interaction, you will be randomly re-matched with another person in the room for a new interaction. Each interaction has the same setup. You will play a number of such interactions with other people.

Summary

To summarize, every interaction you have with another person in the experiment includes a random number of rounds. After every round, a random number generator has placed 7/8 probability on the interaction continuing for another round. There will be a number of such interactions, and your behavior has no effect on the number of rounds or the number of interactions.

There is a 1/8 probability that the option you choose will not happen and the opposite option occurs instead, and the same is true for the person you interact with. You will be told which moves actually occur, but you will not know what move the other person actually chose. When choosing the action, you and the other person will also send each other a message.

At the beginning of the session, you have 50 units in your account. At the end of the session, you will receive £1 for every 30 units in your account.

You will now take a very short quiz to make sure you understand the setup.

The session will then begin with one practice round. This round will not count towards your final payoff.

Appendix C — Analysis restricted to the last four supergames played

Here we present the results of our restricted analysis of the last four supergames 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 varies between 26% and 63%. Figure C1 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).

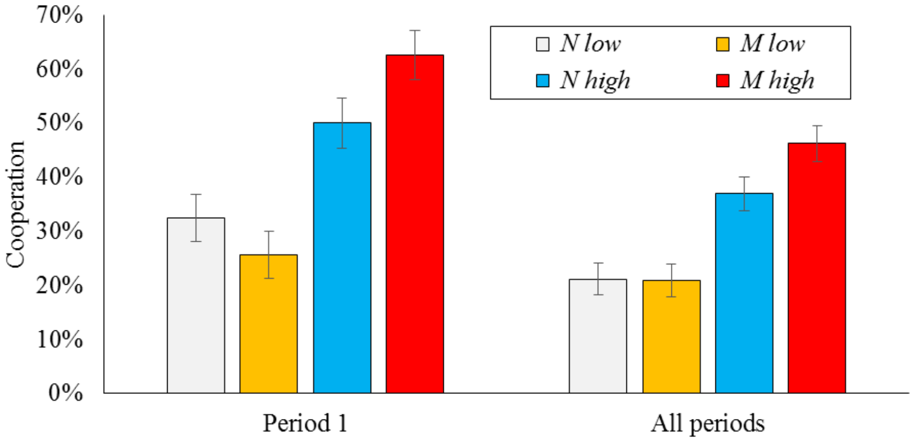


FIGURE C1. FIRST PERIOD AND OVERALL COOPERATION, BY TREATMENT, AVERAGED OVER THE LAST FOUR SUPERGAMES OF EACH SESSION

Question 2. Figure C2 is remarkably similar to Figure 5; the only notable differences is that candid cooperation in the *high* treatment occurs slightly more often (46% versus 44%) and honest defection slightly less (32% versus 34%).

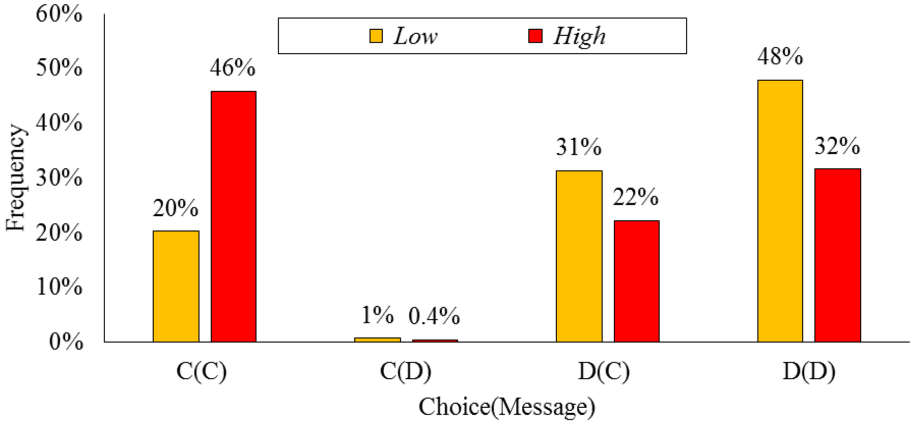


FIGURE C2. FREQUENCY OF INTENDED ACTIONS IN THE MESSAGE TREATMENTS, AVERAGED OVER THE LAST FOUR SUPERGAMES OF EACH SESSION

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*; 48% in *low* and 39% in *high*, if we restrict our attention to the first they defect. Also, Figure C3 shows a similar trend as before.

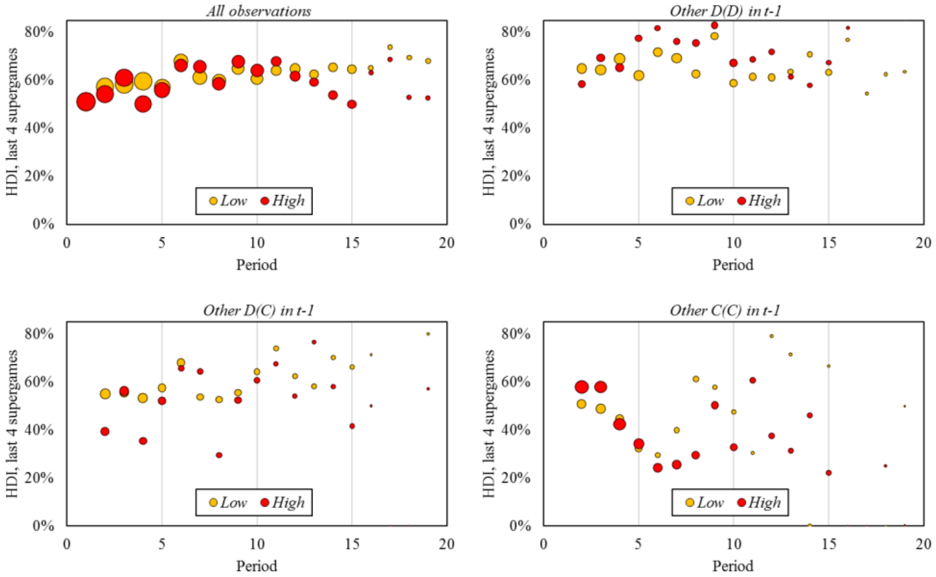


FIGURE C3. HONEST-DEFECTION INDEX (HDI; D(D)/[D(D)+D(C)]) OVER PERIOD IN THE LAST 4 SUPERGAMES; DOT SIZE IS PROPORTIONAL TO THE NUMBER OF OBSERVATIONS IN EACH PERIOD

Figure C4 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.

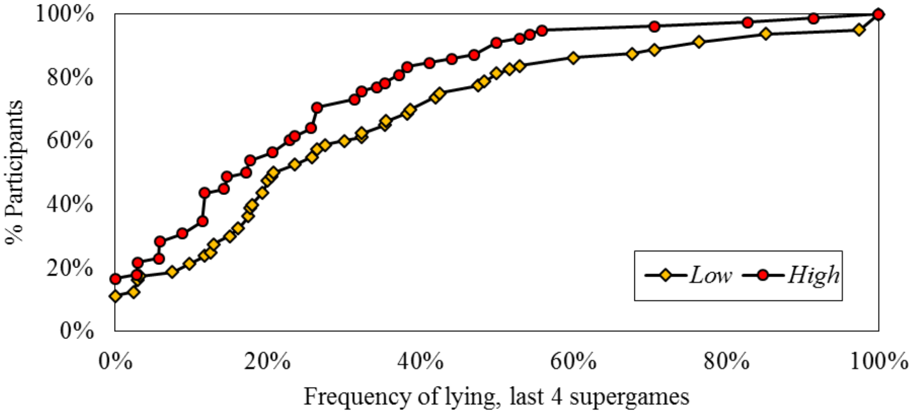


FIGURE C4. CUMULATIVE DISTRIBUTION OF PARTICIPANTS WHO LIED A DETERMINED NUMBER OF TIMES, AVERAGED OVER THE LAST FOUR SUPERGAMES OF EACH SESSION

Question 3. We also find that a large proportion of the participants conditioned their responses on what their partner communicated. Figure C5 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 *low* (p=0.001). Indeed, Table C1 confirms a significant main effect of the partner’s message across treatments.

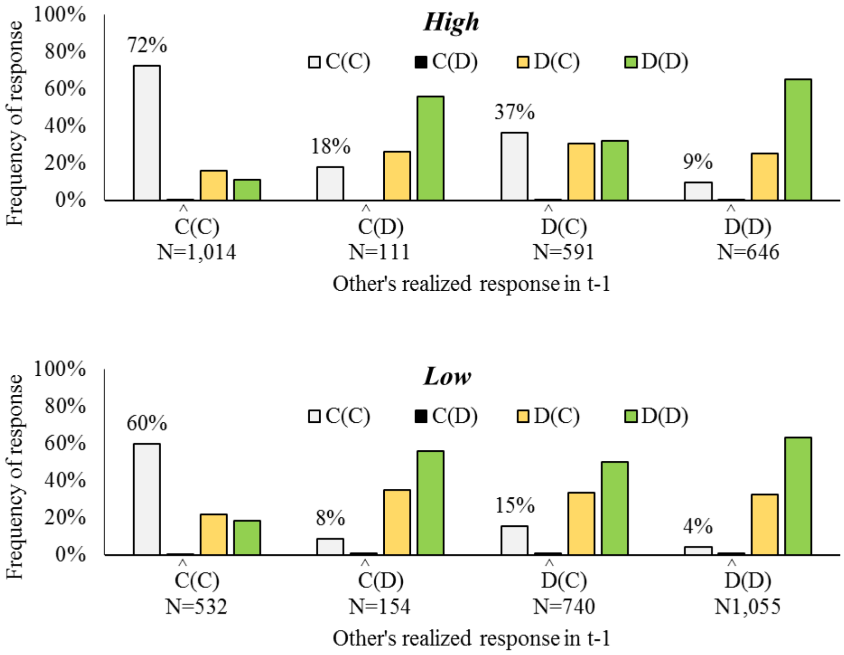


FIGURE C5. INTENDED RESPONSE TO OTHER'S REALIZED RESPONSE IN THE PREVIOUS PERIOD, AVERAGED OVER THE LAST FOUR SUPERGAMES OF EACH SESSION

TABLE C1—THE ROLE OF ACTIONS AND INTENTIONS COMMUNICATED IN PERIOD 1 FOR COOPERATION IN PERIOD 2, AVERAGED OVER THE LAST 4 SUPERGAMES OF EACH SESSION

	Low (L)		*High (H)*		*L & H*
	(1)	(2)	(3)	(4)	(5)
Partner's action in	0.195**	0.201	0.185***	0.222	0.201
t-1 (*A*)	(0.078)	(0.130)	(0.070)	(0.200)	(0.130)
Partner's message	0.150**	0.151**	0.250***	0.260***	0.151**
in t-1 (*M*)	(0.066)	(0.074)	(0.081)	(0.095)	(0.074)
A x M		-0.007		-0.044	-0.007
		(0.164)		(0.224)	(0.164)
High (*H*)					0.137
					(0.118)
H x A					0.022
					(0.238)
H x M					0.108
					(0.121)
H x A x M					-0.036
					(0.277)
Constant	0.163***	0.162**	0.250**	0.244**	0.058
	(0.060)	(0.064)	(0.097)	(0.099)	(0.066)
Session f.e.	Yes	Yes	Yes	Yes	Yes
Observations	291	291	288	288	579
R^2	0.100	0.100	0.108	0.108	0.169

Notes: We report standard errors clustered on both participant and supergame pair.
*** Sig. at the 1% level; ** Sig. at the 5% level; * Sig. at the 10% level.

A visual comparison between Figures 10 and C6 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 C5 in response to their partner cooperating and sending the message "I choose D." This is mainly due to the reduced number of observations (17), though.

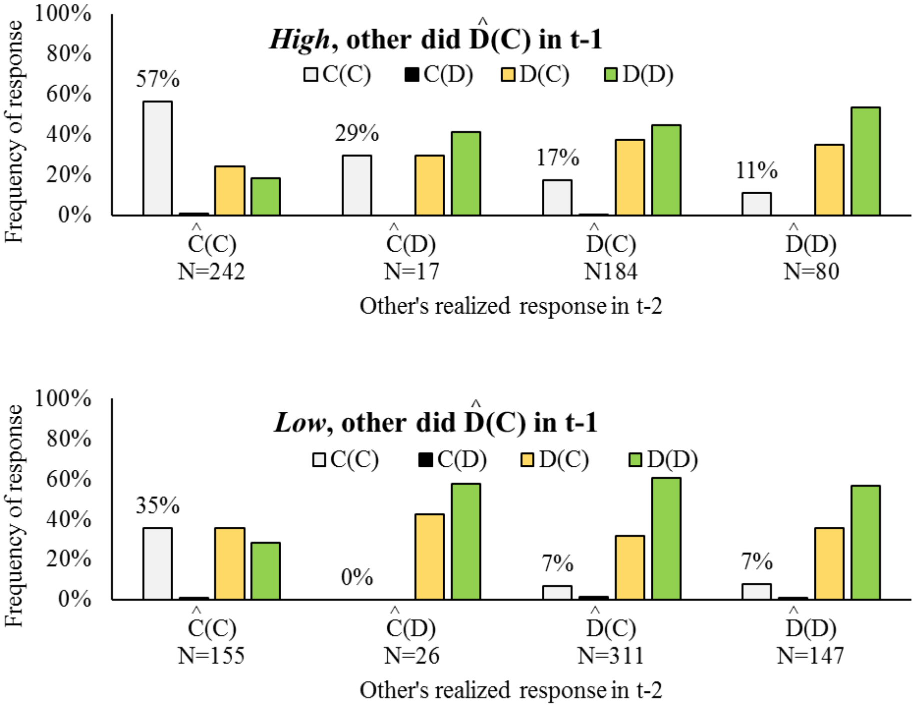


FIGURE C6. INTENDED RESPONSE TO OBSERVING OTHER'S DEFECTION AND MESSAGE "I CHOOSE A," AVERAGED OVER THE LAST FOUR SUPERGAMES OF EACH SESSION

Not surprisingly, Table C2 shows that the number of participants who either choose defection at least 75% of the time or choose cooperation at least 75% of the time in period 1 increases when we restrict out attention to the last four interactions. Most importantly, this Table also shows that participants who usually open with D virtually never cooperate, so their play resembles the ALLD strategy, while participants who usually open with C are more cooperative than intermediate participants.

TABLE C2—OVERALL COOPERATION RATES (EXCL. PERIOD 1) FOR PARTICIPANTS BY PERIOD 1 CHOICE: D 75% OF THE TIMES, A MIX OF D AND C, AND C 75% OF THE TIMES.

		Period 1 choice	
	D 75%	*Mixed*	*C 75%*
N low	0.07 (N=49)	0.31 (N=8)	0.42 (N=21)
N high	0.11 (N=35)	0.33 (N=5)	0.57 (N=36)
M low	0.13 (N=58)	0.33 (N=4)	0.45 (N=18)
M high	0.15 (N=27)	0.52 (N=4)	0.60 (N=47)

We calculate payoffs as the average earned by each participants in each period played. As before, participants who usually open with defection out-earn more cooperative participants in treatments with *low* payoffs; whereas participants who consistently cooperate in the *high* treatment are not out-earned by others.

In Table C3 we see that in the *low* treatment, participants that usually opened with D(C) out-earned other participants who usually opened with C(C) when they were matched with partners who opened with C(C) or D(D), but when matched with partners that opened with D(C) they were slightly out-earned by participants who usually opened with C(C). In the *high* treatment, we see that the success of participants who usually open with C(C) is driven by productive interactions with partners who opened with C(C).

TABLE C3—AVERAGE PAYOFF BY PARTNER'S REALIZED FIRST PERIOD OUTCOME

	Low		*High*	
Other's Realized Period 1 Play	*75% opens with D(C)*	*75% opens with C(C)*	*75% opens with D(C)*	*75% opens with C(C)*
$\hat{C}(C)$	0.66 (34)	0.59 (20)	1.36 (29)	1.37 (107)
$\hat{C}(D)$	0.32 (5)	-0.10 (4)	1.17 (1)	-0.16 (5)
$\hat{D}(C)$	0.24 (33)	0.25 (24)	0.85 (5)	0.60 (50)
$\hat{D}(D)$	0.15 (28)	-0.01 (20)	0.45 (9)	0.19 (26)

Notes: Shown in parentheses is the number of supergames in which each combination of participant's strategy and partner's opening move occurred.

Table C4 shows that the results of an SFEM restricted to the last four supergames show 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 C4—SFEM results for treatments with and without communication

Strategy	*Low*	*High*
Treatments with communication		
ALLD(C)	0.21	0.09
ALLD(D)	0.41	0.22
TFT that ignores messages	0.01	0.01
TFT that believes messages, and defects with D(C)	0.03	0.20
TF2T that is punitive and immediately punishes $\hat{D}$(D)	0.34	0.47
Mental error	*0.26*	*0.24*
Treatments without communication		
ALLD	0.52	0.36
GRIM1	0.14	0.09
ATFT	0.04	0.21
D-2TFT	0.17	0.10
TF2T	0.13	0.24
Mental error	*0.10*	*0.10*

Notes: Punitive refers to strategies that only treat $\hat{C}$(C) as cooperation; unless otherwise specified, participants 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.

Table C5 completes 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. Yet, various cooperative strategies out-perform the strategy ALLD (ALLD(D), for treatments with communication).

TABLE C5—STRATEGY FREQUENCIES AND TWO MEASURES OF THEIR PAYOFFS

	Low		*High*	
Strategy	*Frequency*	*Observed (expected) payoff*	*Frequency*	*Observed (expected) payoff*
Treatments with communication				
ALLD(C)	0.21	0.34 (0.51)	0.09	1.12 (1.43)
ALLD(D)	0.41	0.23 (0.24)	0.22	0.75 (0.52)
TFT that ignores messages	0.01	0.27 (0.14)	0.01	0.74 (0.95)
TFT that believes messages, and defects with D(C)	0.03	0.43 (0.02)	0.20	1.08 (1.08)
TF2T that is punitive and immediately punishes $\hat{D}$ (D)	0.34	0.28 (0.17)	0.47	0.96 (1.11)
Treatments without communication				
ALLD	0.52	0.34 (0.32)	0.36	0.69 (0.74)
GRIM1	0.14	0.09 (0.19)	0.09	0.73 (0.75)
ATFT	0.04	0.39 (0.07)	0.21	0.87 (0.80)
D-2TFT	0.17	0.37 (0.28)	0.10	0.74 (0.82)
TF2T	0.13	0.01(-0.09)	0.24	0.94 (0.69)

Notes: Punitive refers to strategies that only treat $\hat{C}$(C) as cooperation; 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.