Communication, Renegotiation, and the Scope for Collusion

David J. Cooper
Florida State University

Kai-Uwe Kühn
University of Michigan and CEPR

September 5, 2011

Abstract: We experimentally study the effect of communication in a game where cooperation is consistent with equilibrium play *if players share an understanding that cheating will be punished*. Consistent with communication acting as a coordinating device, pre-play messages including a credible threat to punish cheating are the most effective type of message for improving collusion. Sending promises to collude also improves cooperation. Credible threats do not emerge in a limited message space treatment that allows for threats of punishment, suggesting that communication is inherently different with only a limited message space. Contrary to theory, allowing for renegotiation increases collusion.

Acknowledgments: We would like to thank the National Science Foundation (SES-0720993) for funding these experiments. Angelo Benedetti, E. Glenn Dutcher, John Jensenius, Cortney Rodet, Micah Sanders, Elena Spatoulos, and Evan Starr provided valuable research assistance on this project. We would like to thank Tim Cason, Gary Charness, Martin Dufwenberg, Guillaume Frechette, Larry Samuelson, Steve Salant, Yossi Spiegel, Roberto Weber, Mike Whinston, and seminar participants at Florida State University, IESE, University of Cologne, University of East Anglia, University of Innsbruck, the University of Erlangen-Nuremberg, University of Michigan, the EWEBE meetings in Innsbruck, the IO Society sessions at the ASSA meetings in San Francisco, the Conference on Cartels and Tacit Collusion at Norwich, the IIOC conference in Arlington, the 2009 CEPR Applied IO conference in Mannheim, and the FSU Workshop on Experimental Game Theory for their helpful feedback. The authors are solely responsible for any errors in this manuscript.

"The Finns will respect the Spanish dominance in Spain if ENCE really increase their prices in other countries: If Fincell learn about prices below US $360 also in the future, they will reconsider their policy as to sales in Spain!"

The preceding quote is from a document found in the European Woodpulp case (Decision 85/202/EEC, 1985, published in OJ L85/1). Finncell, the joint sales organization for Finnish producers, promises to abide by a collusive agreement if ENCE, the leading Spanish producer of wood pulp, does so as well, but threatens to punish ENCE in the future if it departs from the agreement. Communications containing explicit threats and promises of this sort are intuitively understood to be at the heart of illegal cartel activities. The per-se prohibitions on price fixing under the Sherman Act in the US and Article 81 of the Treaty of Rome in the EU are effectively prohibitions of having such conversations and enforcement focuses on discovering evidence of communication and explicit agreements. Despite this emphasis by anti-trust authorities, incriminating communication commonly occurs within cartels, as is well documented in case and field evidence (e.g. Genesove and Mullin, 2001). The sheer frequency of inter-firm communication about collusion in the face of large fines suggests that it must be a valuable tool for effectively establishing collusive outcomes. However, it is not well understood why this is the case and what exactly needs to be said to make collusion successful.

Price collusion between firms is only one of many cases where repeated interaction between economic agents makes cooperation consistent with equilibrium play even though agents have incentives to behave non-cooperatively in the short run. A common element across such environments is that cooperation can be enforced in equilibrium by the threat to switch from cooperative behavior to non-cooperative behavior after a deviation is observed. We frame the discussion in this paper in terms of "price collusion", but the insights we gain about the role of communication in improving cooperation are equally applicable to other cases where cooperative equilibria have this structure. This includes examples where cooperation is socially desirable such as implicit contracting, team production, and the provision of local public goods.

This intuition has been formalized in the theory of infinitely repeated games (Abreu, 1988; Abreu, Pearce, Stacchetti, 1990). The theory does not address communication, but implies that collusion relies on firms solving a difficult coordination problem: cooperation can only be supported if players know that deviations will be punished by switching from a high payoff to a low payoff continuation equilibrium. The literature typically assumes that coordination on the necessary contingent strategies is easily achieved. However, experimental evidence from related coordination games shows that subjects often fail to reach a Pareto optimal equilibrium in the absence of an explicit coordination device (Van Huyck, Battalio, and Beil, 1990). Communication dramatically increases the likelihood of coordination on an efficient equilibrium (Cooper, De Jong, Forsythe, Ross, 1992; Blume and Ortmann, 2007; Brandts and Cooper, 2007). Coordination on a collusive equilibrium is far more complex, requiring agreements not just on a single action but on entire contingent plans. It is not obvious that communication can be as effective a coordination device in such settings.

At the same time experimental evidence suggests that communication may have a role beyond coordination, since it can lead to more cooperative outcomes in one-shot (or finitely repeated) games where cooperation is not an equilibrium outcome (e.g. Dawes, MacTavish, and Shaklee, 1977; Isaac and Walker, 1988; Cason and Mui, 2009; Charness and Dufwenberg, 2006). There exist numerous explanations for why communication makes subjects more willing to cooperate in the absence of monetary incentives, including increased group identity, aversion to lying (or guilt aversion, which is slightly different)[1], and improved understanding of the mutual benefits of cooperation. If communication primarily improves collusion through such channels, the mechanism of rewards and punishment in the continuation game may be unnecessary to sustain cooperative outcomes. Such a finding would reduce the practical relevance of the theory of repeated games which explains cooperation as an equilibrium phenomenon that relies on this mechanism.

We explore these different explanations for how communication might foster collusion through a series of experiments that vary the type of communication available. Subjects play a sequence of two period collusion games with random rematching between games. Collusion (mutual choice of a high price) can be supported in the first period as part of a subgame perfect equilibrium, but only if deviations are punished by play of a Pareto inferior equilibrium in the second period. The two period collusion game therefore captures the main strategic features of infinitely repeated collusion games while easing a number of methodological problems occurring for experiments on communication in indefinitely repeated game.

[1] These concepts are closely related to results from the psychology literature finding that communication allows players to make promises which are then binding due to strong norms, both internal and social, against violating commitments (Kerr and Kaufman-Gilliland, 1994).

In an initial phase without communication, first period play quickly collapses to the non-cooperative equilibrium.[2] Communication is then added for the second phase of the experiment. If communication is limited to pre-game statements of intent to collude in Period 1 with no possibility of specifying a punishment scheme for deviations, an initial increase in collusion is followed by a collapse back to the non-cooperative equilibrium. This is in line with results reported by Holt and Davis (1990). Adding the possibility of specifying a punishment scheme without otherwise expanding the message space does not improve matters, as play still collapses back to the non-cooperative equilibrium. When a rich pre-game message space is used – subjects have access to a chat window and can send and receive unlimited messages – there is again an initial burst of collusive behavior followed by gradual deterioration. Unlike the treatments with limited message spaces, this decline slows and is eventually reversed in the pre-play chat treatment. By the end of the experiment, collusive behavior returns to its initial high levels. When renegotiation is allowed by adding chat between periods of the game, collusion is even more common and never exhibits a decline. This contradicts the unambiguous theoretical prediction for the game we implement: renegotiation should eliminate all collusion by making it impossible to credibly commit to punish cheating.

Using detailed analysis of the chat content we identify two channels by which communication leads to greater collusion in the treatment with pre-play chat: credible threats and promises. Subjects who either send or receive credible threats (i.e. a threat of punishment that, if believed, makes it incentive compatible to abide by a collusive agreement) that non-collusive play will be punished are significantly less likely to cheat on collusive agreements. The effect is large. Sending a credible threat is estimated to lower the probability of cheating by 40% and receiving a credible threat lowers the probability of cheating by 26%. Credible threats are by far the most effective type of communication for bolstering collusion in the treatment with pre-play communication. Underlying the effectiveness of credible threats are changes in the incentive to collude: Generally we observe higher payoffs for cheating on a collusive agreement than complying with it, but this reverses when a credible threat is sent or received.

The powerful effect of credible threats supports the role of communication as a coordination device to achieve a collusive equilibrium consistent with the standard theory. If collusion was

[2] The payoffs in our game are designed to make it likely that collusion will fail in the absence of communication, but this is not a universal feature of collusion games. See Dal Bo and Frechette (2011) for a study of which features of the payoff table make it likely that tacit collusion will occur.

observed but such punishment was rarely mentioned or if it played a smaller role in fostering collusion than other types of communication, this would have been a cause for skepticism about the theory itself. Our result complements earlier experimental work on collusion by Dal Bo (2005) which demonstrated that collusive play is more likely when players operate "under the shadow of the future." We show that the "shadow of the future" becomes much more important when *explicit* threats of punishment can be made.

Given the importance of credible threats in the pre-play chat treatment, it is surprising that collusion is so low in the treatment with a limited message space that allows for threats of punishment. Analysis of messages sent in this treatment indicates that the problem is not a failure to use threats of punishment, but instead a failure to specify sufficiently harsh punishments for cheating to be unprofitable. This result provides an important methodological insight: Limited message treatments are not a good substitute for more natural conversations that take place with open chat. Not only do limited message treatments run the risk of missing the types of messages that actually matter, the available messages are used differently than they would be within the context of a natural conversation.

The second channel through which communication boosts collusion in the treatment with pre-play chat is promises of trustworthy behavior. Specifically, *sending* promises of trustworthy behavior is associated with a significant decrease in cheating by the sender on collusive agreements. The marginal effect is less than half of the effect of sending a credible threat, reducing cheating by 19%. On the surface this lines up with the results of Charness and Dufwenberg (2006), as well as numerous papers from the psychology literature (see Kerr and Kaufman-Gilliland, 1994, for a summary) finding that promises increase cooperation. However, in our setting neither an aversion to lying or guilt aversion is necessary to explain the self-commitment effect of promises: subjects who send explicit promises face sufficiently strong punishment for cheating that collusion becomes incentive compatible. Surprisingly, this self-commitment is not valuable to the subject sending the promise since it does not reduce cheating by the individual *receiving* the promise.

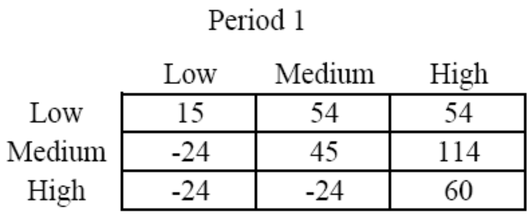
Turning to the treatment allowing for renegotiation, analysis of the second period chat provides a clear explanation for why collusion is more successful than in any other treatment. Consistent with renegotiation theory, players try to avoid going through with a punishment following cheating and these attempts have some success. As a result, average monetary punishments after deviations from collusive agreements are the weakest of all communication treatments. However, this is counteracted by a second important effect of allowing chat between the two periods of the game: individuals who are cheated can reproach those who cheated them. They seize upon this opportunity with high frequency and great enthusiasm. The availability of an inexpensive and effective form of punishment in the treatment with renegotiation provides a good explanation for the high and stable levels of collusion achieved in this treatment.[3]

The paper is organized as follows. In section 2 we discuss collusion theory and the theory of communication in games. Section 3 describes the experimental design in detail. In section 4 we present the experimental results. We first analyze the benchmark behavior when there is no communication possible and then look at the short and long run treatment effects of various communication treatments. We then go into more detail on the content analysis of the pre-play chat treatment and the renegotiation treatment. Section 5 concludes the paper and discusses the relationship between our work and field studies of communication and collusion.

1. Communication and the Theory of Collusion: The standard theoretical approach to price collusion is quite simple conceptually: collusion can be supported at a price greater than the competitive price if the short-run gain from undercutting is less than the long run losses induced by future punishment involving a switch from collusive to competitive behavior. What is critical for the argument is that both the promise of future collusion as a reward for past collusive behavior and the threat of future competitive behavior as a punishment for a past deviation are credible in the sense that they involve equilibrium play. A credible threat therefore requires a coordinated switch between different equilibria of the continuation game. A central question in our study is how communication helps players coordinate on the contingent play necessary for a collusive equilibrium. This section develops predictions on what kind of communication we should observe if cheap talk can be used as a coordination device in collusion games.

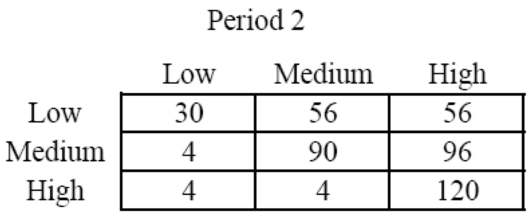
To fix ideas we discuss these issues using the specific two-period game employed in our experiments. The following two matrices (with player 1's strategies being the rows and player 2's strategies the columns) show the payoffs for the game for player 1 in Period 1 and Period 2 respectively.

[3] This is similar to the effect of non-pecuniary punishments (disapproval points) in public goods games (Masclet, Noussair, Tucker, and Villeval, 2003). The effect here is more persistent, possibly due to the richer set of verbal punishments available. Likewise, Xiao and Houser (2005) find that the possibility of verbal punishment reduces rejection rates in ultimatum games.



Period 1

	Low	Medium	High
Low	15	54	54
Medium	-24	45	114
High	-24	-24	60



Period 2

	Low	Medium	High
Low	30	56	56
Medium	4	90	96
High	4	4	120

The Period 1 stage game can be interpreted as a standard Bertrand game in which firms have a choice between 3 possible prices and have a sunk cost of 24. The unique Nash equilibrium of the Period 1 stage game played in isolation is (L,L). The Period 2 stage game, which is based on the continuation profits of an infinitely repeated version of the Period 1 stage game,[4] is a coordination game in which there are three pure strategy equilibria, (L,L), (M,M), and (H,H). These equilibria are Pareto ranked with (H,H) being the Pareto dominant equilibrium.

We refer to the full two-period game as the Two Period Bertrand Game (TPBG). The necessary incentive conditions are satisfied so that (L,L), (M,M), or (H,H) can all be sustained in the first period of a subgame perfect equilibrium if players play (L,L) in Period 2 after any deviation in Period 1 and (H,H) otherwise. Asymmetric Period 1 outcomes (H,L) and (M,L) (and their permutations) can also be sustained in this way but not (H,M).[5] All collusive equilibria of the TPBG, defined as equilibria that yield an outcome other than (L,L) for Period 1, require that players use the first period outcome as a coordination device for play in the second period. This captures the essential structure of all theories of collusion based on infinitely repeated games (Abreu 1988, Abreu, Pearce, and Stacchetti, 1990) or finitely repeated games (Benoit and Krishna, 1985).

For pre-play communication to have a systematic impact on outcomes, players must take messages to have meaning. But players cannot believe just any message because opponents may have an incentive to lie in order to induce favorable behavior. We assume that two features of communication are necessary for credible communication. First, for a message to be credible each player should have an incentive to do what he proposes as long as he expects the proposal to be believed by the other player.[6] Second, a pair of proposals from two different players

[4] See Appendix C for the details.

[5] Note that (H,H) can only be sustained in period 1 if (L,L) is played in Period 2 after a deviation and (H,H) otherwise. All other Period 1 subgame perfect equilibria can also be sustained by playing (L,L) after a deviation and (M,M) otherwise in Period 2. The only Period 1 outcome that can be sustained by playing (M,M) after a deviation and (H,H) otherwise is (M,M) in Period 1.

[6] This condition corresponds to the concept of "self-commitment in Aumann (1990). Aumann also requires that messages are "self-signaling": the sender only wants the message to be believed if he is telling the truth. This

should be considered credible only if the proposals are compatible. These two conditions are equivalent to requiring agreement on a Nash equilibrium of the game. Since all Nash equilibria in the TPBG that support Period 1 prices above L involve contingent play in Period 2 (i.e. play in Period 2 varies depending on the Period 1 outcome), these assumptions about credibility predict that communication can facilitate collusion only if messages specifying contingent strategies are available and used. Going further, only messages that threaten to punish cheating with play of Low in Period 2 provide credible support for a proposal to play (H,H) in Period 1.

If communication is used as a coordination device prior to Period 1, we should also expect subjects to go further and coordinate on an equilibrium *whenever* they can communicate. This raises the issue of renegotiation prior to Period 2. Renegotiation theory is built on the assumption that players will use an unspecified coordination device to achieve a Pareto undominated equilibrium after any history of play.[7] Suppose players in the TPBG can communicate and hence coordinate before *both* periods. Since continuation equilibria in the TPBG are Pareto ranked, the intuition underlying renegotiation theory implies players will agree on and play (H,H) in Period 2 regardless of the Period 1 outcome. Messages prior to Period 1 that specify a different equilibrium for Period 2 should therefore be ignored because rational individuals will anticipate the impact of renegotiation prior to Period 2. It follows that messages prior to Period 1 cannot credibly specify a contingent strategy for Period 2. No collusion should therefore occur in Period 1 of the TPBG if we allow for communication as a coordination device between Period 1 and Period 2 play.

It is not a universal property of collusion games that allowing renegotiation eliminates the possibility to collude. This occurs in the TPBG because the equilibria of the Period 2 game can be Pareto ranked. The scope for collusion under renegotiation is significantly wider when there are multiple asymmetric continuation equilibria that cannot be Pareto ranked. The TPBG is designed to make the predicted effect of renegotiation as stark as possible, yielding a clean test of the main idea underlying all models of renegotiation: the ability to renegotiate will eliminate

condition is much more controversial. As Farrell and Rabin (1996) point out, this condition leads to unlikely predictions about the effectiveness of communication in stag-hunt games and cannot be satisfied in collusion games.

[7] See Bernheim, Peleg, and Whinston's (1987), Bernheim and Ray (1989), and Benoit and Krishna (1993) for finitely repeated games and Van Damme (1989), Farrell and Maskin (1989), and Abreu, Pearce, Stacchetti (1993) for supergames. Note that the assumption that a Pareto undominated equilibrium will be chosen is not generally accepted in the cheap talk literature for cases, unlike the TPBG, with conflicts of interest about the equilibrium to be chosen (Farrell and Rabin 1996).

the use of Pareto dominated continuation equilibria and therefore limit the outcomes that can be supported as equilibria in the full game.

Several other features of the TPBG are designed to make the results easier to interpret. The equilibrium (M,M) is risk dominant (Harsanyi and Selten, 1988) in the second period game. This makes it relatively likely that M will be chosen in the absence of an explicit coordination device. If, in contrast, (H,H) were both Pareto and Risk dominant, H would be a natural default choice in period 2. Players could then easily coordinate on (H,H) without an explicit coordination device and it would be more difficult to identify the effect of communication as a coordination device when renegotiation is allowed. Making (M,M) risk dominant also helps to identify contingent behavior. Since collusion at (H,H) in period 1 can only be supported by a switch from (H,H) to (L,L) in period 2, more effective communication should move period 2 play away from (M,M) to the extremes. Note that (L,L) in period 2 is unlikely to occur due to coordination failure rather than a conscious decision to punish cheating because it is both Pareto and Risk dominated.

More generally, we designed the TPBG to provide strong incentives in the sense that collusion at H is highly beneficial, strong punishments are needed to maintain collusion in equilibrium, and the loss from cheating and being punished in equilibrium is large. In line with these goals, the payoff from play of (H,H) is 33% larger in each period than the payoff from play of (M,M), collusion at H in Period 1 can only be supported as an equilibrium outcome via reversion to (L,L) in Period 2 which reduces payoffs by 75% compared to play of (H,H), and the payoff from colluding at (H,H) in both periods is 25% greater than the payoff from defection to M in Period 1 followed by reversion to (L,L) in Period 2 (180 vs. 144). Subjects are given strong incentives to reach and abide by collusive agreements, but the punishment called for by the equilibrium is sufficiently harsh for both players that it is unlikely to be undertaken lightly.

3. Experimental Design

A. General Design: Subjects play twenty rounds of the TPBG in all treatments. A "round" refers to an entire play of the TPBG while a "period" refers to one of the two games played within a single round of the TPBG. Subjects are randomly matched with a new opponent in each round. Sessions are sufficiently large (minimum of twenty subjects) that it is unlikely that there are repeated game effects between rounds.

For the first ten rounds in all treatments subjects play the TPBG without any communication. This is sufficient for subjects to understand the experimental interface, payoff tables, and main strategic issues in the TPBG before attempting to master use of communication. Having ten rounds of play before introducing communication also allows play to converge to the one shot Nash equilibrium in the first period, making the task facing subjects in Rounds 11 - 20 more challenging. Treatments vary by the type of communication available in Rounds 11 – 20.

B. Use of the TPBG: By using the TPBG our design differs from earlier work on collusion that used indefinitely repeated games as proxies for supergames (e.g. Roth and Murnigham, 1978; dal Bo 2005; dal Bo and Frechette 2011; Duffy and Ochs, 2009). The simplicity of a two period game has significant methodological advantages for a study that focuses on communication and learning. The most important is speed of play and learning. Subjects go through a lengthy learning process, and our conclusions would differ substantially if this process lacked adequate time to converge. Generating the requisite experience is non-trivial since chat slows down play significantly and we limited ourselves to two hour sessions, including instructions, to avoid subject fatigue. The TPBG is sufficiently simple that most sessions were completed within two hours, even with chat, and play has largely stabilized by the end of Round 20. Another advantage provided by the TPBG's simplicity is that we can make sharp theoretical predictions, as described in Section 1, about what type of proposals for play of (H,H) in Period 1 are credible and what equilibria are renegotiation proof. Finally, using a relatively simple game helps with the statistical evaluation of the outcomes. Analyzing the content of messages is a daunting task even when the number of types of relevant messages that can be sent is small. If we expand the strategy space, and by extension increase the number of relevant message types, the analysis quickly becomes intractable.[8]

C. The Communication Treatments: Our experimental design compares play in five communication treatments: No Communication, Period 1 Limited Communication, Period 1 Limited Communication with Contingencies, Pre-play Chat, and Renegotiation. The following subsection briefly describes each of these treatments.

1) No Communication (N Treatment): The rules for Rounds 11 – 20 are identical to those in Rounds 1 – 10, with no communication between players allowed. The treatments with communication have a pause following Round 10 while new instructions are read. To exclude the possibility that collusion in Round 11 is caused by a restart effect, the N treatment is included

[8] In a follow-up paper (Cooper and Kühn, 2010), we compare play in two and three period Bertrand games. Having a three period game allows for the use of asymmetric punishments as well as making the connection between the first two periods more transparent. No differences in play are observed that would affect our conclusions.

as a control. It also has a pause following Round 10 for an announcement that the games for Rounds 11 – 20 will use the same rules as Rounds 1 – 10.

2) Period 1 Limited Communication (P1 Treatment): The P1 treatment gives subjects the opportunity to send a message prior to the beginning of Period 1. The message space is limited to suggesting actions for Period 1. Specifically, subjects are given the prompt, "I think we should choose the following in Period 1." They are asked to choose between "Low", "Medium", "High", or "No Response" for both "My Choice" and "Your Choice." Messages are chosen simultaneously and each player is shown both parts of both players' messages at the same time as choices are made for Period 1. The feedback at the end of Period 1 reiterates the messages as well as reporting the outcome for Period 1. Subjects cannot send any messages about their intent for Period 2. The purpose of this treatment is to establish that simply calling for collusive behavior in Period 1 is not sufficient to generate Period 1 collusion.

3) Period 1 Limited Communication with Contingencies (P1C Treatment): In addition to specifying what actions should be chosen for Period 1, subjects in the P1C treatment also indicate what actions should be chosen for Period 2 subject to the outcome for Period 1. The set of possible messages about Period 2 is limited: subjects are prompted "[i]f we choose the preceding [actions I think should be chosen] in Period 1, I think we should choose the following in Period 2" and "[i]f we DO NOT choose the preceding [actions I think should be chosen] in Period 1, I think we should choose the following in Period 2." Limiting the message space simplifies the problem facing subjects while still making it possible to send a credible message that supports collusion in Period 1. Subjects are shown both players' messages, in full, at the same time as choices are made for Period 1. The messages are displayed again as part of the feedback for Period 1. The purpose of the P1C treatment is to establish that the possibility of sending contingent messages is not sufficient to generate Period 1 collusion in a limited message space treatment.

4) Pre-Play Chat (PChat Treatment): At the heart of our experimental design are the two chat treatments. Starting in Round 11, the PChat treatment allows players to communicate using the chat option in version 3.1 of z-tree (Fischbacher, 2007). This is very similar to using an IM program. Continuous back-and-forth communication is possible until one of the players makes a decision for the period. Subjects are given no guidance on how the chat should be used or what they might say, although it is fairly obvious that it is meant for discussing the game. The purpose of the PChat treatment is to study whether and how pre-play chat is used to facilitate Period 1 collusion.

5) Renegotiation (RChat Treatment): Communication before Period 1 occurs in the RChat treatment in exactly the same way as in the PChat treatment. The RChat treatment only differs from the PChat treatment by also allowing communication through a chat window after first period actions are observed and before second period actions are chosen, making renegotiation possible. The purpose of this treatment is to test whether the intuition underlying theories of renegotiation is empirically valid.

Initial Hypotheses: Our initial hypotheses for Period 1 play are derived from the discussion of communication and collusion in Section 2. While it is possible to achieve tacit collusion in an experimental collusion game without communication, the TPBG is designed to make this unlikely in the N treatment.[9] The limited message space available in the P1 treatment makes it impossible to reach a credible agreement to collude in Period 1. We therefore did not expect long term collusion in this treatment. Consistent with earlier results of Holt and Davis (1990), we expected an initial increase in collusion when communication was introduced since it takes some time to learn whether messages calling for Period 1 collusion are credible. The P1C treatment allows for credible agreements to collude, so higher and more stable collusion was expected than in the P1 treatment. For the same reason, the PChat treatment was also expected to yield higher and more persistent collusion than the P1 treatment. The rich communication available in the PChat treatment also provides more scope for behavioral factors such as trust and guilt to have an effect, so we anticipated higher and more stable collusion in the PChat treatment than in the P1C treatment.[10] The intuition underlying theories of renegotiation predicts that allowing chat between Periods 1 and 2 should undermine collusion in Period 1 and lead to coordination on H in Period 2. We therefore expected to see less collusion in the RChat treatment than in the PChat treatment.

[9] The TPBG is designed so collusion at H is not risk-dominant. Dal Bo and Frechette (2008) present evidence that making collusion at H risk-dominant is necessary but not sufficient to guarantee successful collusion. We therefore did not expect collusion at H in the absence of communication.

[10] Charness and Dufwenberg (2010) compare rich and limited communication in a one-shot trust game and find that rich messages makes subjects more trusting and somewhat more trustworthy. Our game is different because cooperation is consistent with equilibrium and messages about contingent strategies have an important role to play. As will be seen, our results are both stronger and have a different interpretation. Rather than revolving around differences in what can be said in the two treatments, the the P1C and PChat treatments differ in the usage of a common set of messages.

E. Procedures: The experiments were conducted at Case Western Reserve University in Fall 2006 and Spring 2007 using subjects recruited via emails sent to all undergraduates. Sessions were run in a computerized laboratory using z-Tree (Fischbacher, 2007), and took between 1½ and 2 hours. Average earnings were slightly more than twenty dollars, including a six dollar show-up fee. Subjects were paid their total earnings from all twenty rounds of the TPBG. Payoffs were denominated in experimental currency units (ECUs), converted to dollars at a rate of 130 ECUs equal $1. Table 1 summarizes the number of subjects and sessions for each treatment. There are three sessions for each treatment with at least twenty subjects per session.

Table 1: Summary of Treatments

	N	P1	P1C	PChat	RChat
Number of Sessions	3	3	3	3	3
Number of Subjects	64	68	74	64	76
First Period Limited Messages					
Second Period Limited Messages					
First Period Chat					
First and Second Period Chat					

The instructions were read to the subjects, and were also shown on the subjects' computer screens. Several times the payoff tables were projected on an overhead screen for examples, making the payoffs common knowledge. The matching for this experiment is relatively complex (fixed matching within a round, random re-matching between rounds), so this point was emphasized. Following the instructions, subjects took a quiz testing their ability to read the payoff tables and their understanding of the instructions.[11]

The experimental materials are framed using abstract language. For example, the materials do not refer to prices. In Period 1 subjects choose between "A", "B", and "C" and in Period 2 they choose between "D", "E", and "F", with the three labels in each period corresponding to low, medium, and high prices. To ease exposition, the terms "Low", "Medium", and "High" (or L, M, and H) are used throughout this paper even though these are not the labels seen by subjects.

Subjects knew they would be playing a total of twenty rounds of the TPBG. They also knew that the first ten rounds would be played without communication and followed by a pause

[11] Instructions are available at [add url].

for additional instructions. The possibility of communication was only introduced at this intermediate point. To maintain parallelism there was a pause prior to the 11th round in the N treatment to announce that none of the rules would change.

In the P1 and P1C treatments, the instructions prior to round 11 described in detail what messages could be sent, including the option to send "no response", but provided no guidance about why any particular message should be sent. We stressed that the messages are cheap talk.

Subjects in the two chat treatments received extensive instructions, largely focused on the mechanics of using the chat program. The instructions gave the subjects no guidance on what types of messages should be sent other than (i) requesting that they not identify themselves and (ii) asking them to avoid offensive language. These instructions also stressed that the messages are cheap talk with no direct effect on payoffs.

Subjects had printed copies of the payoff tables for *both* periods available whenever they made a decision. When choosing a price for either period, the interface showed subjects any messages or chat from either player for the current period as well as a summary of outcomes (prices and payoffs) for all previous rounds. The interface automatically showed the summary for the three most recent rounds with a scroll bar that could be used to see earlier rounds. When choosing a price in Period 2, subjects could see the prices and payoffs for both players in Period 1, but could not see any communication from Period 1. At the end of each period subjects received a summary of the prices chosen by both players as well as both players' payoffs for the period. Period 2 feedback also included the sum of payoffs across both periods for both players.

The interface (with one exception) did not include identifying information about a subjects' opponent to limit the possibility of repeated game effects across rounds. To make it possible for subjects to tell whether a message had been sent by themselves or their opponent, messages in the chat window were tagged with a randomly generated ten digit "chat id". At the time these sessions were run, we were unable to generate new chat ids across rounds or to use non-identifying tags. Subjects were not allowed to have any writing implements during the experiment to prevent them from writing down the other players' chat ids, and it seems unlikely that they remembered long random numbers across multiple rounds. With one exception, the content of the chat contains no evidence that subjects knew when they had played an opponent previously.[12] In follow-up experiments at FSU we have run sessions with both a “chat id” and non-identifiable tags (“mine” and “other”). The type of tag has no discernable effect on messages or behavior.

Sessions were automatically ended at the two hour mark to avoid subject fatigue (this was *not* announced to subjects in advance). Due to this rule, one session of the RChat treatment only had sixteen rounds and another only had eighteen rounds.

4. Results: Collusion in the TPBG is defined in terms of Period 1 play. When we discuss collusion in the results section, this refers to Period 1 choices. Our analysis of Period 2 choices focuses on how these depend on Period 1 outcomes, a central point of interest for understanding the relationship between communication and collusion. See Appendix D for tables giving frequencies of all actions in both periods for all rounds, broken down by treatment.

A. Behavior in Rounds 1 -10: Initially modest levels of collusion collapse over ten rounds without communication. In Round 1, 44% of the subjects choose Low in Period 1 compared with only 12% choosing High. By Round 10, play converges to the Nash equilibrium for the Period 1 game with 86% of subjects choosing Low for Period 1 and only 4% choosing High. There is no statistically significant difference between treatments for the first ten rounds. Underlying the collapse of collusion in Rounds 1 – 10 are poor incentives to collude. In Round 1 the average payoffs for the entire round (i.e. the sum of payoffs for Periods 1 and 2) following Period 1 choices of High, Medium, and Low are 34, 82, and 98 ECUs respectively. Subjects earn almost three times as much if Low is chosen in Period 1 rather than High. Similarly bad incentives exist throughout Rounds 1 – 10.

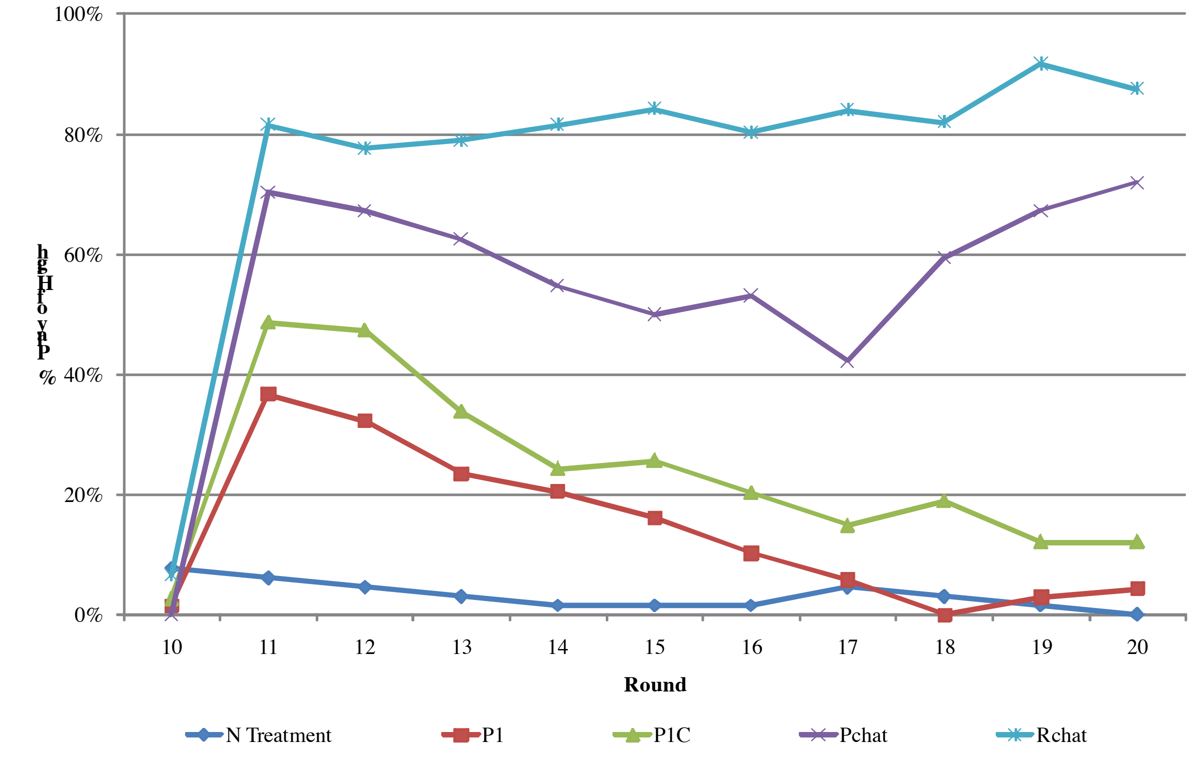
Period 2 choices are more diffuse than Period 1 choices, but by Round 10 a clear mode has emerged at Medium with 53% of all Period 2 choices. Period 2 choices depend significantly on Period 1 choices. Aggregating across Rounds 1 – 10, if the other player’s Period 1 choice is High or Medium then High is chosen in Period 2 for 40% of the observations and Low for 16%. However, if the other player chooses Low in Period 1, the distribution of Period 2 choices shifts to 27% choosing High and 26% choosing Low. The effect of Period 1 play on Period 2 choices strengthens over time, but is never sufficiently large to prevent Low from being the profit maximizing choice for Period 1. The degree to which Period 2 behavior is contingent on Period

[12] One subject in the RChat treatment tried to pass on the identity of a subject who had cheated him to two other subjects (he didn’t exactly remember the chat ID of the offending party, but came close). None of our conclusions change if the affected observations are dropped.

1 outcomes suggests that the type (but not the magnitude) of contingent behavior needed for tacit collusion is present in the data before communication becomes available.

B. Prices in Rounds 11 – 20: Figure 1 plots the frequency of High being played in Period 1 across Rounds 10 – 20 for each treatment. For the N treatment, there is no discernable restart as play continues the pattern from Rounds 1 – 10 with steady movement toward 100% play of Low in Period 1. In contrast, all four communication treatments show a large initial increase in collusion as choice of High in Period 1 rises in Round 11 when communication becomes available. After this initial burst the evolution of play through Rounds 11 – 20 strongly differs across the four communications treatments.

Figure 1: Play of High in Period 1



In the two limited message space treatments, P1 and P1C, an initial increase in collusion is followed by a steady collapse. The P1 treatment shows the least initial increase in collusion of the communication treatments. Medium rather than High is the modal choice in Round 11. There follows a sharp decline in the use of High, which almost dies out by round 20, as well as a steady drop in use of Medium. Play converges toward the non-cooperative outcome. The P1C treatment does little better, in spite of the availability of contingent messages. The initial increase in use of High is somewhat stronger than in P1, with High being the clear modal strategy in Round 11, but a similar collapse follows. Period 1 play has not converged by Round 20, but with choice of High having become quite rare (12%) and a mode of choosing Low (51%) having emerged, it seems clear where things are headed.

The initial spike in play of High is much stronger in both chat treatments than in either limited message space treatment. In Round 11, the modal Period 1 choice is High for both PChat and RChat with play of Low being almost non-existent (5% in both treatments). The subsequent dynamics differ dramatically. In the PChat treatment use of High falls sharply until Round 17, similar to the declines observed in the P1 and P1C treatments. Unlike the limited message space treatments this decline reverses in Rounds 17 – 20. Round 20 choices are largely indistinguishable from those made for Round 11 with 72% of the subjects choosing High. There is a clear trend towards collusion in the long run.

In the RChat treatment, no decline in collusion is observed. The distribution of Period 1 actions is essentially constant over Rounds 11 – 20 with negligible use of Low. The small shifts upward in Rounds 17 and 19 are artifacts driven by RChat sessions ending in different rounds. Breaking the data down by sessions, little change is observed over Rounds 11 – 20. In Round 16, the last round where all subjects played, 80% of subjects chose High in Period 1.

Appendix A reports the results of two regressions that provide statistical backing for the preceding observations about Period 1 prices in Rounds 11 – 20. The first model finds that almost all of the differences *between* treatments are statistically significant. The second model looks at changes *within* treatments over time. It confirms that Period 1 prices decline and then rebound significantly in the PChat treatment. The dip and recovery in the PChat treatment are unlikely to have occurred by chance.

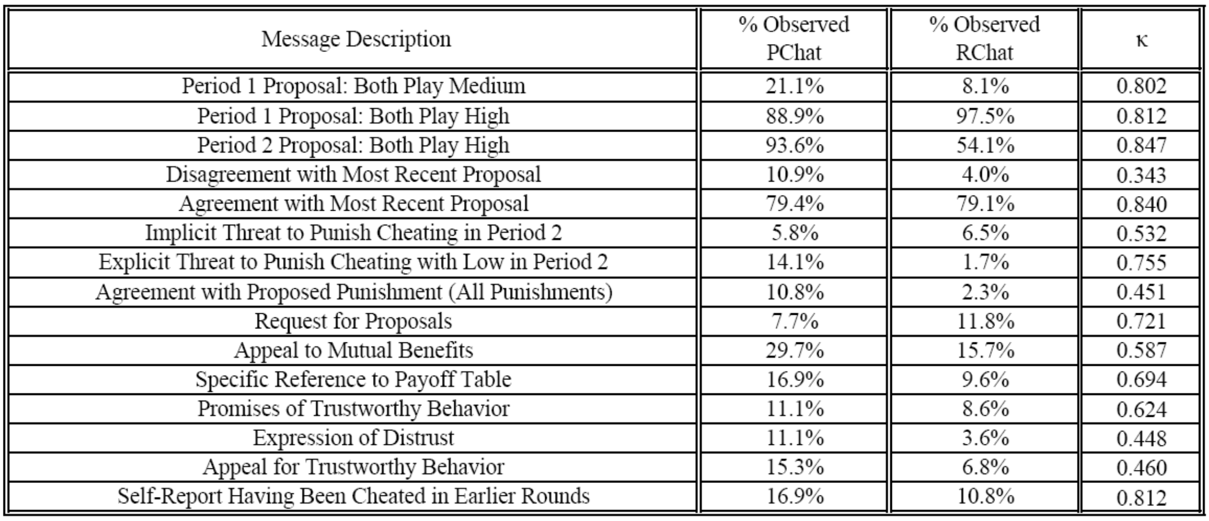
C. Content Analysis: The results of the P1 and P1C treatments show that the possibility of communicating an intent to collude, even if backed with the possibility of sending Period 2 messages contingent on Period 1 outcomes, is insufficient to generate stable collusion. Something contained in the rich exchanges of the chat treatments must be leading to high rates of collusion. To systematically study how the content of chat affects play we quantified content by coding all of the dialogues. The goal was to be as comprehensive as possible, including a category for any type of message that might conceivably be relevant for play of the game, rather than only coding categories we thought likely, *ex ante*, to be important in generating collusion. See Appendix B for a full list of categories.

Two research assistants independently coded all messages. No effort was made to force agreement among coders – the goal was to have two independent codings so errors were uncorrelated. At no point in the coding process was either RA informed about any hypotheses that were to be tested. The RAs were repeatedly told that their job was to capture what had been said rather than why it was said or what effect it had. Coding was binary – a message was coded as a 1 if it was deemed to contain the relevant category of content and zero otherwise. We had no requirement on the number of codes for a message – a coder could check as many or few categories as he deemed appropriate.

Table 2 summarizes the frequency of the most common categories, defined as any category that was coded in at least 10% of the dialogues (averaging across coders) for one of the chat treatments. Implicit threats are also included as a category of particular interest, even though this category does not reach the 10% threshold in either chat treatment. The measure of frequency reported in Table 2 is the percentage of dialogues where the category was coded. For messages prior to Period 2 from the RChat treatment, observations are broken down by whether both players colluded in Period 1 (both players choose High) or one player cheated (deviated from an agreement to choose High). Observations where both players cheated are not included. The final column of each panel shows • , a common measure of inter-coder agreement (Cohen, 1960). The • 's generally show substantial agreement between the coders, especially since the number of possible codes was large and no attempt was made to force agreement between coders.

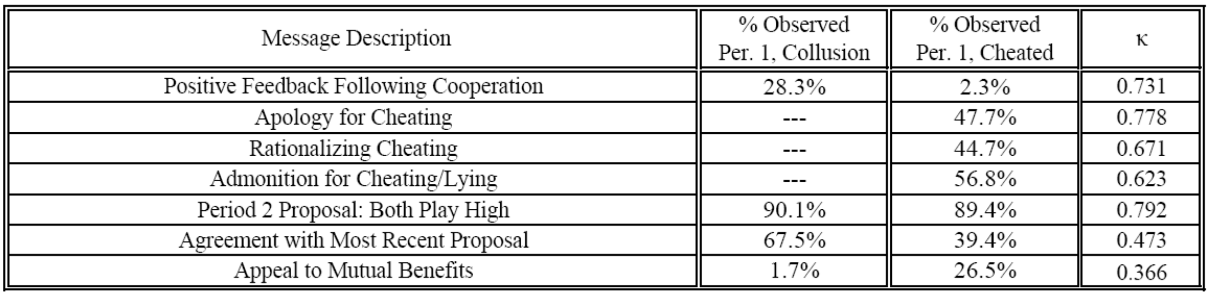
Subjects almost always engaged in communication relevant to the game. Period 1 prices were proposed in 618 of 622 dialogues, and only one dialogue lacked substantive discussion of how the game would be played. The content of the dialogues differed substantially between the PChat and RChat treatments. Some of these differences are to be expected (e.g. subjects are less likely to propose a Period 2 action prior to Period 1 in the RChat treatment, since they have another opportunity to discuss Period 2 prices prior to Period 2), but others are more striking. Subjects in the PChat treatment were more likely to make explicit threats to punish cheating with play of Low in Period 2, while implicit threats were more common in the RChat treatment. There was generally more substantive conversation prior to Period 1 in the PChat treatment. Subjects not only behave differently in the two chat treatments, they also communicate differently. We therefore do *not* pool data from the two chat treatments in the content analysis.

Table 2
Use of Period 1 Messages in Chat Treatments



Message Description	% Observed PChat	% Observed RChat	κ
Period 1 Proposal: Both Play Medium	21.1%	8.1%	0.802
Period 1 Proposal: Both Play High	88.9%	97.5%	0.812
Period 2 Proposal: Both Play High	93.6%	54.1%	0.847
Disagreement with Most Recent Proposal	10.9%	4.0%	0.343
Agreement with Most Recent Proposal	79.4%	79.1%	0.840
Implicit Threat to Punish Cheating in Period 2	5.8%	6.5%	0.532
Explicit Threat to Punish Cheating with Low in Period 2	14.1%	1.7%	0.755
Agreement with Proposed Punishment (All Punishments)	10.8%	2.3%	0.451
Request for Proposals	7.7%	11.8%	0.721
Appeal to Mutual Benefits	29.7%	15.7%	0.587
Specific Reference to Payoff Table	16.9%	9.6%	0.694
Promises of Trustworthy Behavior	11.1%	8.6%	0.624
Expression of Distrust	11.1%	3.6%	0.448
Appeal for Trustworthy Behavior	15.3%	6.8%	0.460
Self-Report Having Been Cheated in Earlier Rounds	16.9%	10.8%	0.812

Use of Period 2 Messages in RChat Treatment



Message Description	% Observed Per. 1, Collusion	% Observed Per. 1, Cheated	κ
Positive Feedback Following Cooperation	28.3%	2.3%	0.731
Apology for Cheating	---	47.7%	0.778
Rationalizing Cheating	---	44.7%	0.671
Admonition for Cheating/Lying	---	56.8%	0.623
Period 2 Proposal: Both Play High	90.1%	89.4%	0.792
Agreement with Most Recent Proposal	67.5%	39.4%	0.473
Appeal to Mutual Benefits	1.7%	26.5%	0.366

i. Agreements and Cheating: In all four communication treatments we can identify agreements between the players on what price to set in Period 1. For the P1 and P1C treatments, we define the players as having come to an agreement if their messages prior to Period 1 suggest the same Period 1 prices. In the PChat and RChat treatments, the players are defined as having an agreement if a proposal for Period 1 prices is made by one player and then accepted by the other. The agreed price is given by the *last* proposal that was agreed upon.[13]

Pairs almost always reach an agreement on Period 1 prices in the PChat (98% reach agreement) and RChat treatments (100% reach agreement), but often do not reach an agreement in the P1 (51% reach agreement) and P1C treatments (55% reach agreements). It is relatively hard to reach an agreement in the limited message space treatments since subjects cannot revise proposals when their initial proposals do not agree. Subject to reaching an agreement, players in

[13] The same Period 1 price is specified for both players in all of the agreements we identify in the data.

all three treatments with communication overwhelmingly agree on choosing High in Period 1 (92% in P1, 97% in P1C, 85% in PChat, and 94% in RChat).

We define cheating as choosing Low or Medium in Period 1 following an agreement to choose High in Period 1. The probability of cheating in the P1 and P1C treatments is substantial even in early rounds, 57% and 34% respectively for Rounds 11 – 12, and rises over time, reaching 93% and 78% respectively for Rounds 19 – 20. This high frequency of cheating indicates that the cause of declining collusion in the P1 and P1C treatments is *not* a failure to reach agreements on play of High in Period 1. Even if all pairs reached collusive agreements, it is unlikely that collusion could survive such pervasive cheating.

For the chat treatments, cheating follows a matching pattern to that observed for play of High in Figure 1. Cheating in the PChat treatment is initially only slightly higher than in the RChat treatment (15% vs. 8% in Round 11), but rises steadily to a peak of 46% in Round 17. It then declines back to almost its initial level (23% in Round 20). Cheating in the RChat treatment is low and steady throughout.

The standard theory of collusion suggests that the decision to cheat on an agreement to collude is driven by whether cheating is likely to be punished. Can we therefore explain the low levels of cheating in the two chat treatments by increased punishment? Define *unilateral* cheating as the case where one player cheats on an agreement to choose High in Period 1 while the other player does not. Unilateral cheating is frequently punished in Rounds 11 – 20 of all four treatments with communication. Surprisingly, punishment is strongest in the P1 treatment. The proportion of non-cheaters (subjects who do not cheat on an agreement to choose High) choosing High in Period 2 drops from 90% to 9% when their opponent cheats. While not quite as extreme, non-cheaters also punish strongly in the P1C treatment, with the proportion choosing High dropping from 99% to 33% when their opponent cheats. The type of contingent behavior needed to support collusion is present in the P1 and P1C treatments among those who comply with agreements to play High in Period 1. Comparing the two chat treatments, punishment is significantly more frequent in the PChat treatment.[14] The proportion of non-cheaters that choose High in Period 2 changes from 97% to 45% when their opponent cheats in the PChat treatment, as opposed to 100% and 65% in the RChat treatment. This difference between the chat

[14] Punishment of unilateral cheating is significantly more likely in the PChat treatment at the 5% level for Rounds 11 – 20 and at the 1% level for Rounds 16 – 20 based on ordered probit regressions that control for round effects and individual effects.

treatments becomes larger over time. In Rounds 16 - 20 cheating lowers the proportion of Period 2 choices of High by non-cheaters from 98% to 39% in the PChat treatment, compared with a change from 100% to 75% in the RChat treatment. Consistent with the theory of renegotiation, punishment of cheating is signficantly weaker in the RChat treatment. *Nonetheless, collusion is higher*.

The theory focuses on punishment as a deterrent for cheating, but the frequency of punishment is not the only factor determining the incentives to cheat. The more likely a subject believes it is that his opponent will cheat, the greater his own incentives to cheat. Comparing total payoffs for the round (summing over both periods) when a subject cheats on a collusive agreement with total payoffs when he does not cheat, it becomes obvious why cheating is so prevalent in the limited message space treatments. In the P1 and P1C treatments, the average increases in total payoff from cheating are 19 and 26 ECUs respectively in Rounds 11 – 12, increasing to 36 and 43 ECUs for Rounds 19 – 20. As a point of comparison, a subject gains 54 ECUs by cheating on a collusive agreement if the other player does not cheat and Period 2 actions are unaffected. These strong incentives to cheat are not driven by a lack of punishment but reflect frequent cheating by others.

In the PChat treatment, it is initially more profitable to cheat on an agreement than to follow it with an average gain of 31 ECUs in Rounds 11 – 12, but over time the incentive to cheat shrinks to an average of 12 ECUs in Rounds 19 – 20. This reflects both an increase in punishment and a decreasing threat over Rounds 17 – 20 of being cheated. The incentives to cheat are in line with the initial decrease in collusion, but do not explain why collusion recovers in later rounds since cheating continues to pay (albeit, by less). The opposite trend emerges in the RChat treatment. The average gain from cheating is only 10 ECUs over Rounds 11 – 12, but this jumps to 35 ECUs for Rounds 15 – 16 (the last two rounds before the first session ended). There is little danger of getting cheated in the RChat treatment, but punishment is sufficiently weak in the later rounds that compliance with collusive agreements is clearly not incentive compatible. Nonetheless, collusion remains stable.

The analysis of cheating and punishment leaves us with a pair of puzzling observations. In neither chat treatment are the average gains from cheating negative at any point, yet both treatments show high levels of collusion. Likewise, the incentives to cheat are strong in the long run for the RChat treatment, particularly in comparison to the PChat treatment, yet this is the treatment with the highest and most stable levels of collusion. These observations only make sense if something in the message content counteracts the generally poor incentives to honor collusive agreements. We therefore turn to the effects of specific types of messages.

ii. Credible Messages and Punishment: As discussed in Section 1, a collusive agreement to play High is only credible if accompanied by a proposal that mutual choice of High in Period 1 leads to play of High in Period 2 while any other outcome in Period 1 will be punished by play of Low. In this subsection we examine the incidence and effect of messages that include threats to punish cheating.

Our coding scheme distinguishes between two different types of proposed punishment schemes, explicit and implicit. Explicit punishment refers to cases where subjects state that failure to collude will lead to use of Low rather than High in Period 2. Often times an explanation is given for why the threat of punishment makes collusion incentive compatible. The following dialogue from the PChat treatment is a nice example in which the subjects successfully colluded. The leading numbers indicate which player is talking. This is a verbatim transcript, typos included, except that we have replaced the abstract labels used in the experimental materials with the more descriptive terminology used elsewhere in this paper:

> 1 both go [Period 1 High] and [Period 2 High]?
> 1 do i have you word?
> 2 if you dont pick [Period 1 High] im going [Period 2 Low] and well both make less... you can do the same for that way there is no incentive to screw you over the first round
> 1 ok fine [Period 1 High] than [Period 1 High]

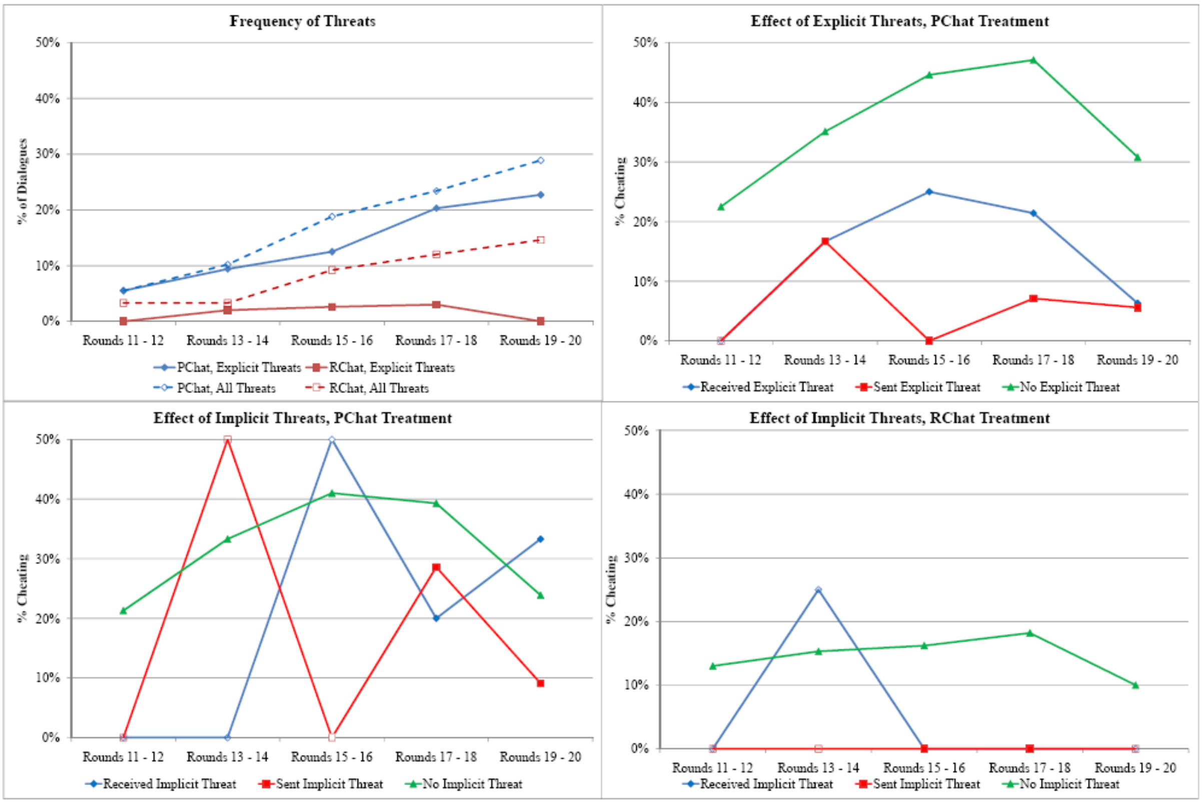
Implicit punishment refers to cases where a subject threatened to punish non-collusion but did not specify how this would work, as in the following example:

> 1 [Period 1 High]/[Period 2 High]?
> 2 lets call it a truce: [mutual play of Period 1 High], if you play me all bets are off, and vice versa
> 2 [mutual play of Period 1 High], [mutual play of Period 2 High]
> 1 sounds good

There is a clear threat to punish here, but it remains unclear what action the players will choose if one of them cheats. Coordination on a specific punishment is not achieved. Nevertheless subjects did collude in this case.

The upper left panel of Figure 2 shows the frequency of threats in the two chat treatments. Data is grouped into two round chunks to reduce the noise. Solid lines show the percentage of dialogues (averaged across the two coders) where an explicit threat was observed, and dashed lines show the percentage of dialogues where either type of threat was observed. The difference between the solid and dashed lines for each treatment shows the percentage of dialogues where *only* an implicit threat was made.[15] The use of threats rises over time in both treatments. The frequency of threats is always higher in the PChat treatment than in the RChat treatment, and most threats are explicit in the PChat treatment but not in the RChat treatment.[16]

Figure 2: Use and Effect of Threats



The remaining panels of Figure 2 examine the effect of threats on the incidence of cheating on collusive agreements (mutual play of High in Period 1). *Hollow markers indicate cells with less than five observations for Figure 2 as well as all subsequent figures*. Only implicit threats are considered for the RChat treatment, since too few explicit threats are made in this treatment to draw any conclusion about their effect.

[15] Coding of both types of threats in a dialogue largely comes from cases where a subject clarified an implicit threat by subsequently making an explicit threat.

[16] The difference in frequency of threats (explicit or implicit) between the PChat and RChat treatments is significant at the 5% level in Rounds 16 – 20. For explicit threats, the difference is significant at the 10% level for Rounds 11 – 15 and at the 1% level for Rounds 16 – 20. These statements are based on ordered probits that control for round and individual effects.

The top right panel plots the percentage of subjects in the PChat treatment who cheat on collusive agreements as a function of whether they received (and did not send) or sent an *explicit* threat. A message is categorized as including an explicit threat if this was coded by *either* coder. ***This rule is followed in all figures where the presence or absence of a code is forced to be a binary variable.*** Cheating on an agreement is less frequent when an explicit threat is either received or sent, especially in later rounds.

The bottom panels show the percentage of subjects in the PChat and RChat treatments cheating as a function of whether they sent or received (and did not send) an *implicit* threat. The effect of receiving an implicit threats differs across the two chat treatments. Focusing on the later rounds when implicit threats are reasonably frequent, receiving an implicit threat has no obvious impact on the probability of cheating in the PChat treatment, but clearly reduces cheating in the RChat treatment. Sending an implicit threat is associated with a lower frequency of cheating in both treatments.

iii. Promises and Appeals to Mutual Benefits: The theories of collusion and cheap talk make threats a natural focus for our analysis of chat content. However, the psychology and economics literatures both suggest that promises can play an important role in increasing cooperation (Kerr and Kaufman-Gilliland, 1994; Charness and Dufwenberg, 2006). Communication can also reinforce group identity and norms of maximizing joint benefits (Dawes, Orbell, and van de Kragt, 1988), providing an explanation for decreased cheating in the last few rounds of the PChat treatment even if threats are not used.

We focus on three common coding categories that relate to these alternative explanations: promises of trustworthy behavior, appeals for trustworthy behavior by the other player, and appeals to the mutual benefits of collusion. The following is a good example of a dialogue coded for the first two categories:

1 [Period 1 High] and [Period 2 High] work?
2 Let's put [Period 1 High] and [Period 2 High]
2 Yeah
1 ok
1 ppl have been so tricky
1 i mean honestly
2 YEah
1 so PROMISE?
2 I swear
1 haha not that we could kno anyway

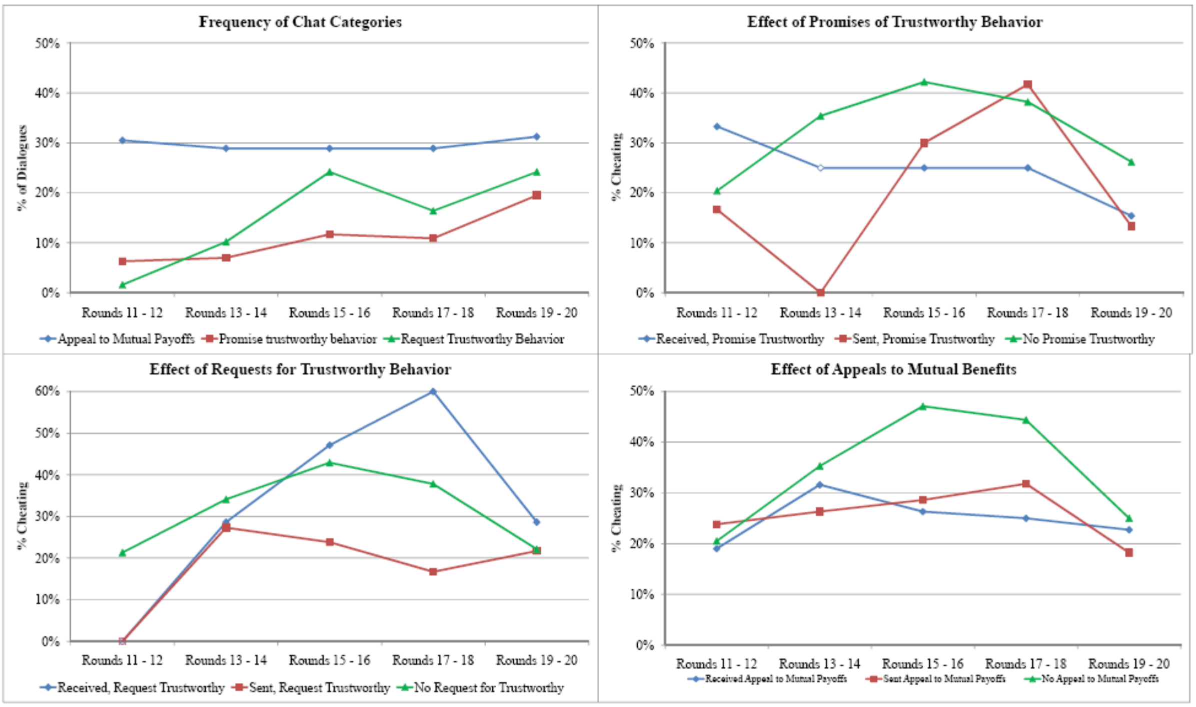
1 but it's like ppl don't have consciences...

Player 1's distrust was well-founded since Player 2 cheated. In this example Player 2 made an explicit promise, but the category for promises of trustworthy behavior also includes messages where subjects indicate that they should be trusted (e.g. "uve just gotta trust me" and "well, just trust me"). Even if an explicit promise is not made in messages like these, an implicit promise of trustworthy behavior is clear.

The following dialogue features an explicit appeal to mutual benefits:

1 which letter do you want to pick?
2 [Period 1 High]
2 We both make the most money this way
2 Right?
1 k, i will too

Figure 3: Use and Effect of Promises and Appeals to Mutual Benefits, PChat Treatment



The upper left panel of Figure 3 shows the percentage of dialogues (averaged across coders) in the PChat treatment where these three categories were observed. Appeals to the mutual benefits of collusion are far more common than explicit threats, while requests for and promises of trustworthy behavior are roughly as frequent as explicit threats. Use of requests for and promises of trustworthy behavior increases over time.

The remaining three panels illustrate the effectiveness of these three categories in the PChat treatment, showing the likelihood of cheating on a collusive agreement subject to whether a message from the category was received (but not sent) or sent. Receiving a promise of trustworthy behavior reduces cheating, with the effect shrinking in the later rounds, while receiving a request for trustworthy behavior is associated with more cheating. Sending either type of message is associated with less cheating, but the effect weakens in the later rounds. Either receiving or sending appeals to the mutual benefits of cooperation leads to a reduction in cheating, particularly for the middle rounds, but this positive effect largely vanishes by the end of the experiment. None of these categories have the large and growing impact that explicit threats have in the PChat treatment.

Figure 4: Use and Effect of Promises and Appeals to Mutual Benefits, RChat Treatment

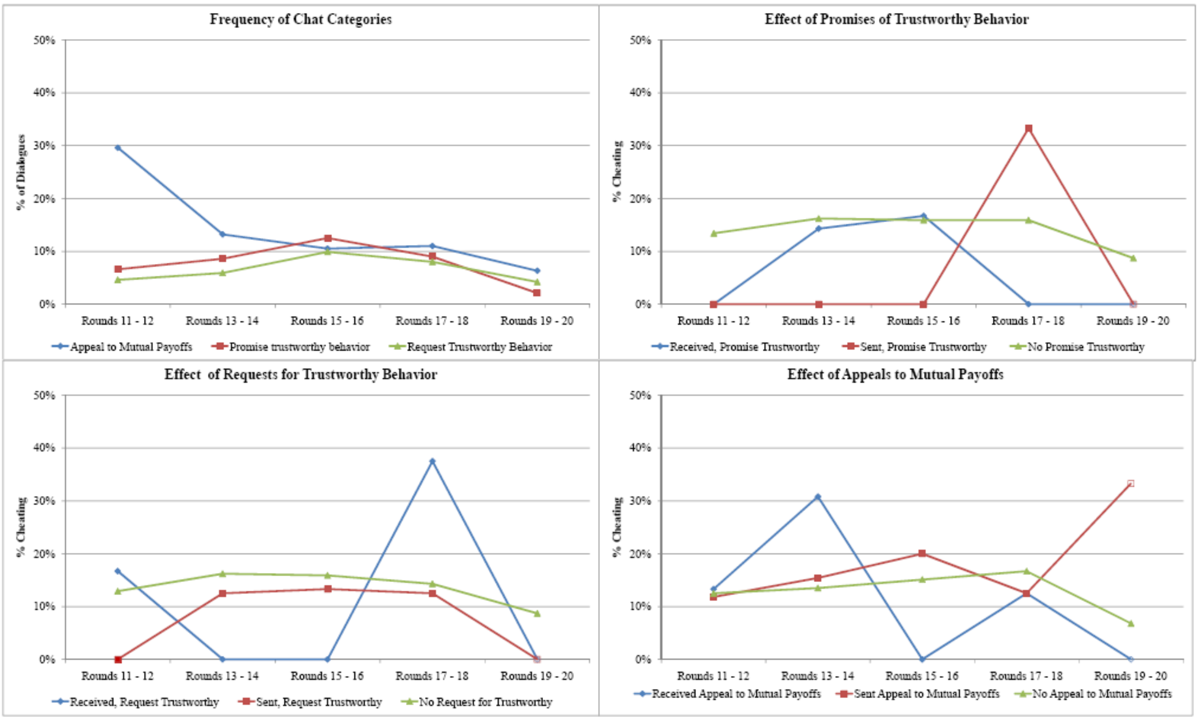


Figure 4 shows data from the RChat treatment, paralleling Figure 3. All three of these categories are coded less frequently in the RChat treatment than in the PChat treatment, and, unlike the PChat treatment, use of these categories is flat or declining over time.[17] None have a clear effect on cheating.

[17] For Rounds 16 – 20, based on ordered probit regressions controlling for round and individual effects, the differences between PChat and RChat is significant at the 1% level for the frequency of appeals to the mutual benefits of collusion and requests for trustworthy behavior and at the 10% level for promises of trustworth behavior.

iv. Regression Analysis: The preceding discussion suggests what types of pre-play communication lead to collusion, but we should not read too much into the raw data. First, most dialogues include many codes, including codes that we have not thus far discussed. Since there is correlation between the coding of various categories, the effects described above may reflect a failure to control for the effects of other types of messages. Second, Figures 2 – 4 exclude data from observations where subjects do not agree on play of High for Period 1, thus eliminating 12% of the data. Selection could therefore play a role in the apparent effects of chat. Finally, the data includes multiple observations from the same individuals. Individual effects are certainly present in the data, potentially distorting the observed effects of chat. We can deal with all of these issues using regression analysis.

Table 3 reports the results for regressions in which the dependent variable is a dummy for whether the subject cheated on an agreement. Observations where an agreement was not reached (1% of the dataset) are dropped. Since cheating is a binary variable, the regressions use a probit specification. Standard errors are corrected for clustering at the subject level. See Appendix A for the equation being estimated.

The independent variables of primary interest measure whether a comment from a specific category was received or sent by the individual in the dialogue prior to Period 1. These variables average across the two coders and therefore have three possible values: 0, ½, and 1. With minor exceptions, the regressions incorporate all categories that were coded in at least 10% of dialogues (averaging across coders) for the treatment in question.[18] Because we are particularly interested in the effect of threats, the categories for implicit and explicit threats were included even when the 10% threshold was not reached. Reducing the threshold for including a category from a 10% coding rate to 5% has little effect on the results reported below beyond cluttering the regression tables. The discussion below notes the exceptions where including more categories affects the results.

Table 3: Probit Regressions on Effect of Chat Categories

	Model 1	Model 2
Data Set	PChat	RChat
Number of Observations/Subjects	626/64	602/76

[18] Dummies for the agreed upon price are included as independent variables rather than the messages proposing prices. These are highly correlated but not identical. The category for agreeing to a punishment scheme is highly correlated with sending a threat, and is excluded to avoid colinearity.

Received Implicit Threat	.462 (.444)	-1.223*** (.463)
Received Explicit Threat	-.816*** (.307)	.379 (.784)
Received Request for Proposals		.462 (.286)
Received Appeal to Mutual Payoffs	-.287 (.193)	-.018 (.308)
Received Specific Reference to Payoff Table	-.125 (.268)	
Received Promise of Trustworthy Behavior	-.183 (.301)	
Received Expression of Distrust	.180 (.284)	
Received Appeal for Trustworthy Behavior	.369 (.260)	
Received Self-Report Being Cheated Earlier	-.179 (.231)	.157 (.266)
Sent Implicit Threat	-.430 (.477)	Perfectly Predicts No Cheating
Sent Explicit Threat	-1.232*** (.380)	Perfectly Predicts No Cheating
Sent Request for Proposals		.233 (.313)
Sent Appeal to Mutual Payoffs	-.214 (.246)	-.158 (.336)
Sent Specific Reference to Payoff Table	.258 (.289)	
Sent Promise of Trustworthy Behavior	-.599** (.248)	
Sent Expresion of Distrust	.323 (.380)	
Sent Appeal for Trustworthy Behavior	-.447 (.380)	
Sent Self-Report Being Cheated Earlier	-.268 (.242)	-.502* (.303)
Log Likelihood	-325.77	-229.49

Note: Three (***), two (**), and one (*) stars indicate statistical significance at the 1%, 5%, and 10% respectively.

As additional controls, the regressions include dummies for the agreed upon price, the current round, the subject's average Period 1 price in Rounds 1 – 10, and the average Period 1 price of their opponent in Rounds 1 – 10. The latter two variables address different potential sources of omitted variable bias. The subject's own history of cooperation in Rounds 1 – 10 provides a control for individual effects beyond the correction for clustering. If individuals who have a propensity to cooperate also tend to make certain types of comments and no control is included for being a cooperative type, the estimated effects of chat could reflect uncontrolled individual effects rather than a causal relationship. The opponent's history of cooperation in Rounds 1 – 10 cannot capture a direct effect on decisions to cheat, since subjects do not observe their opponent's history. Instead, individuals who are inherently more cooperative may also communicate in subtly different ways not captured by the coding scheme. If these differences correlate with use of the included coding categories, then omitted variable bias will result. Including the opponent's history for Rounds 1 – 10 controls for such effects.

Omitted variable bias can also arise due to the interactive nature of dialogues. There is correlation between the types of messages that are sent and the types that are received. This can make it appear that receipt of a message causes cooperative action, when the effect is actually driven by a related message the subject sent. Including controls for sent messages eliminates this source of omitted variable bias as the impact of sent messages is directly accounted for. As it turns out, none of our conclusions about the effects of receiving messages change when controls for sent messages are included. Interpreting the parameter estimates for the *sent* messages as identifying causal relationships with cheating is problematic since sent messages and cheating could be jointly determined by a common unobserved individual effect. We use the within subject variation of sent messages to pin down whether unobserved individual effects cause the observed associations between sent messages and cheating.

Model 1 uses data from the PChat treatment. The parameter estimates for the agreement dummies, round dummies, and controls for Period 1 prices in Rounds 1 – 10 are not of direct interest and are not reported to save space.[19] Looking at the effect of *received* messages, only explicit threats have a significant effect on the likelihood of cheating. The marginal effect is large, with an estimated 26% reduction in the probability of cheating. Receiving an appeal to the mutual benefits of collusion, a promise of trustworthy behavior, or an appeal for trustworthy behavior all fail to have a significant effect on the likelihood of cheating. *Model 1 confirms that explicit threats are by far the type of message whose receipt has the strongest effect on collusion in Period 1.*

[19] Copies of the full regression output are available from the authors upon request.

Turning to *sent* messages, subjects who send an explicit threat are significantly less likely to cheat. The marginal effect is large, with an estimated 40% reduction in cheating. Subjects who promise to be trustworthy are also significantly less likely to cheat on an agreement. The marginal effect is large at 19%, but less than half of the marginal effect of sending an explicit threat. To clarify whether these two effects are due to uncontrolled individual effects, we created a new variable measuring whether the subject had *ever* sent a message of that sort for both types of message. We then reran Model 1 with these two new variables included. If the effects of sending an explicit threat or promising to be trustworthy reflect uncontrolled individual effects, the new variables should be statistically significant and the variables measuring when these messages are actually sent should not be. In other words, what should matter is whether a subject is a type who would send one of these messages, not whether he sent one in the current round. The estimated parameters for the two new variables are small and not close to statistical significance. The parameter estimates for sending these two types of messages are virtually unchanged in magnitude and remain statistically significant at the 5% level.[20] It is therefore unlikely that the significant effects of sending explicit threats or promises of trustworthy behavior reflect uncontrolled individual effects.

Model 2 confirms that Period 1 communication in the RChat treatment has very different impact than in the PChat Treatment. Receiving an implicit threat led to significantly less cheating in the RChat treatment. The parameter estimate for receiving an implicit threat is larger than the estimate for receiving explicit threats in the PChat treatment, but the estimated marginal effects are about the same (27% vs. 26%). Receiving an explicit threat has a small (and statistically insignificant) positive effect on cheating. *Model 2 confirms that it is receipt of implicit threats rather than explicit threats that reduces cheating in the RChat treatment.*

Sending either an implicit or explicit threat has a significant negative effect on cheating in the RChat treatment. Sending a message reporting to have been cheated earlier is also marginally significant. Neither promises of trustworthy behavior nor requests for trustworthy behavior are included in Model 2 since both categories are below the 10% threshold. If these are included, sending a promise of trustworthy behavior causes a significant (at the 1% level) reduction in cheating.

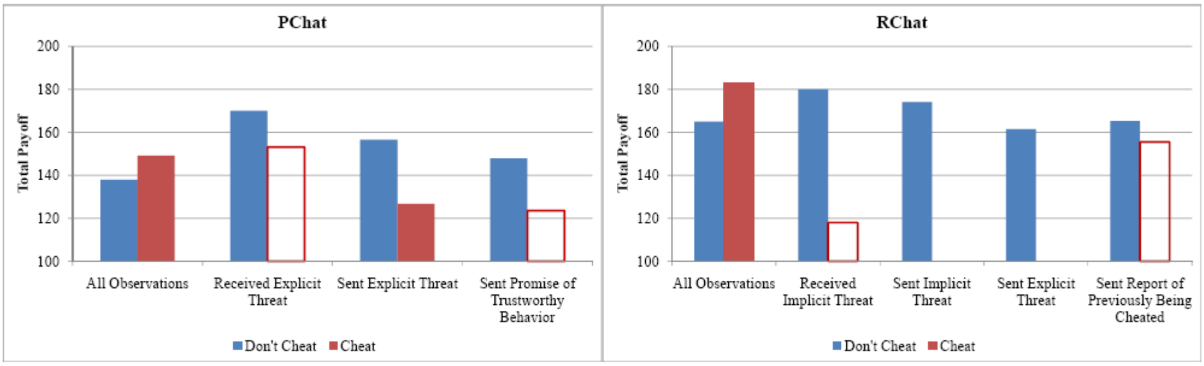
[20] The parameter estimates for the two new variables are -.180 and -.082 with standard errors of .345 and .283 respectively. The parameter estimates for whether these two categories are sent in the current round become -1.084 and -.569 with standard errors of .432 and .254.

To check whether the significant effects of sent messages are due to uncontrolled individual effects, we also modified Model 2 to include variables measuring whether a subject was *ever* coded for sending an implicit threat, an explicit threat, or a message reporting to have been cheated earlier. All three estimates are small and far from statistical significance.[21] The estimated effect of sending a message reporting to have been cheated is reduced by about a third and is no longer statistically significant.

We have explored a wide variety of alternative specifications for Models 1 and 2. Appendix A discusses some of these alternative specifications. Regressions with fixed effects provide additional confirmation that our results cannot be attributed to uncontrolled individual effects, and an instrumental variable approach provides additional evidence of the strong causal relationship between receiving an explicit threat and reduced cheating.

v. Chat and the Incentives to Cheat: We now can resolve the puzzles that subjects in the chat treatments cheat less than expected given the weak incentives to collude and that collusion rebounds in the last periods of the PChat treatment. The key insight is that what matters to subjects is not the *average* incentive to cheat, but instead the incentive to cheat given the *specific* messages they have sent and received. Figure 5 shows that subjects who send or receive messages associated with reducing cheating face lower incentives to cheat.

Figure 5: Gains from Cheating and Chat



For each case where we found a statistically significant effect on cheating from sending and/or receiving a type of message, Figure 5 displays the relationship between sending and/or receiving this type of message and the incentives to cheat. Data is only included from observations where the subjects agreed to play High in Period 1. The average total payoffs for the entire round are plotted conditional on whether or not the player cheated on the agreement.

[21] The parameter estimates are .300, .073, and -.178 with standard errors of .366, .614, and .264.

Hollow bars indicate cells with less than five observations. Two cells had no observations and therefore are missing. As a point of comparison we also plot the corresponding average total payoff for all observations with agreements to play High in Period 1.

Types of communication that reduce cheating generally reduce the incentives to cheat. This occurs primarily through an increase in the payoff when subjects do *not* cheat (blue columns). The messages with the greatest impact on cheating, explicit threats in PChat and implicit threats in RChat, also have the most dramatic effect on payoffs from not cheating. Subjects sending such threats almost never cheat, so subjects receiving these threats have less of an incentive to cheat themselves. This is mutually reinforcing, so sending explicit threats in the PChat treatment or implicit threats in the RChat treatment improves the payoffs when the sender does not cheat (blue columns).

The effect of these messages on payoffs following cheating is less clear given that the relevant sample is small in most cases. The most obvious effect is from sending a promise of trustworthy behavior in the PChat treatment which clearly reduces the payoff after cheating (red column). Subjects who promise not to cheat and then renege are punished harshly. Although monetary incentives explain well why subjects do not cheat after sending a promise, they do not explain why opponents punish more often when promises are broken. This feature of the data has an obvious behavioral explanation that subjects are especially angry about being cheated if they feel that the cheater went out of their way to seem trustworthy.

The cumulative effect of effective messages is large. While the average incentives to not cheat are poor, these incentives are quite sensitive in the presence of effective messages. In the PChat treatment, subjects who send or receive an explicit threat or send a promise to not cheat lose an average of 28 ECUs by cheating on a collusive agreement while other subjects gain an average of 24 ECUs by cheating. In line with these incentives, there is only a 14% chance of cheating by subjects who send or receive an explicit threat or send a promise as compared with 36% for others. Over time, the likelihood of effective messages skyrockets and the impact of effective messages on incentives to cheat and the likelihood of cheating increase slightly. Working together, these changes create the conditions for a return to collusion.

Effective messages also create strong local incentives in the RChat treatment. Subjects who send or receive an implicit threat, send an explicit threat, or report having been cheated lose an average of 24 ECUs by cheating on a collusive agreement while other subjects gain an average of 22 ECUs by cheating.

vi. P1C vs. PChat: We have established for the PChat treatment that explicit threats are the only type of message that significantly lowers cheating by the recipient and that incentives created by sending and receiving explicit threats play a central role in the late increase in collusion. Threats to punish cheating with Period 2 play of Low are also available in the P1C treatment, yet collusion shows no sign of recovering over time in this treatment. In some sense it should be easier for subjects to learn to use threats of punishment in the P1C treatment, since the design suggests the possibility of contingent strategies to the subjects and there are no other types of messages to distract the subjects. *Explicit* calls for punishment for deviating from a collusive agreement are indeed more common in the P1C treatment (24% of pairs) than in the PChat treatment (18% of pairs).[22] Why do punishments then not lead to more collusion in the P1C treatment?

The reason does not seem to be that the single round of communication makes it more difficult to come to an agreement than in PChat. In follow up experiments where subjects could take multiple sequential turns sending messages, leading to higher agreement rates, collusion was no better than in the original P1C treatment (Cooper and Kühn 2010).

Instead, the difference between the P1C and PChat treatment appears to lie in the types of punishments proposed. In the PChat treatment, all messages coded as explicit threats call for play of Low in Period 2 if the other player cheats. Subjects never threaten to punish cheating with play of Medium. In contrast, 90% of all explicit threats in the P1C treatment call for cheating to be punished with play of Medium in Period 2. Subjects in the P1C treatment often use threats, but they almost never use threats which make collusion incentive compatible. For subjects who receive an explicit threat in the P1C treatment, the average loss from cheating on an agreement is negligible (0.3 ECUs). Unlike the PChat treatment where threats are associated with lower payoffs from cheating, both in theory and in practice, in the P1C treatment threats are simply not very threatening. Not surprisingly, receiving an explicit threat makes subject in the P1C treatment no more likely to abide by a collusive agreement: 65% of subjects receiving

[22] As in PChat, some P1C subjects send ambiguous messages about Period 2. In 15% of the pairs with collusive agreements, a message is sent calling for play of High in Period 2 if the agreement is followed and "No Response" if somebody cheats. These messages might represent an implicit threat or may simply be expressions of uncertainty. At the very least, the recipient is given reason to be less certain that High will be played following cheating.

explicit threats cheat on collusive agreements versus 56% for all others. In the PChat treatment threats are effective and become more common over time, but in the P1C treatment threats are ineffective and subjects abandon their use.

Our study's focus on analyzing open chat is necessary because communication is fundamentally different when subjects participate in a natural conversation rather than using a limited message space. This is not just because more types of messages are available and these can be combined and sequenced in more ways. In our experiments, subjects use the available messages less effectively with a limited message space than when messages are embedded within the natural environment of a conversation.

viii. The Effect of Renegotiation: As shown in Section 4.6, effective messages prior to Period 1 improve local incentives in both chat treatments. Unlike the PChat treatment, these improved incentives do not provide a satisfactory explanation for consistently high collusion in the RChat treatment. Effective messages are uncommon until the later rounds but collusion is high throughout. Moreover, cheating is rare (16%) even when effective messages are not present. It is also puzzling that implicit threats are effective in the RChat treatment given that they are not in the PChat treatment and the theory of cheap talk gives no reason why non-specific threats should be useful.

The obvious difference between the PChat and RChat treatments is the addition of a communication phase between the Period 1 and 2 decisions. Consistent with the theory of renegotiation, attempts at renegotiation following unilateral cheating are frequent and reasonably successful. For games with unilateral cheating on collusive agreements, mutual play of High in Period 2 is suggested by at least one of the subjects in 89% of the Period 2 dialogues. When such suggestions occur, 64% of the pairs successfully coordinate on mutual play of High in Period 2. This compares to 14% of pairs coordinating on mutual play of High in Period 2 for the (admittedly infrequent) cases where play of High is not suggested. Successful renegotiation causes the financial consequences of cheating to be weak in the RChat treatment, creating strong incentives to cheat in Period 1

The following Period 2 dialogue gives a sense of why cheating is nevertheless infrequent in the RChat treatment. These subjects had agreed to choose High in Period 1 with minimal discussion. Player 1 chose High, but Player 2 cheated and chose Medium. Following the conversation below, Player 2 chose High in Period 2 while Player 1 chose Low.

1 YOU MEANIE!!!!!!!!!!!
2 its in both our interests to choose [Period 2 High]
1 no
1 i wont help you
2 but you hurt yourself in the process
1 you already hurt me :(

Beyond financially harming Player 2, Player 1 makes it clear that he disapproves of Player 2's actions. This type of negative verbal response to cheating is quite common. In cases of unilateral cheating on a collusive agreement, 73% of the subjects who were cheated admonished the other player for cheating and/or lying. This percentage jumps from 46% in Rounds 11 – 12 to 86% in Rounds 13 – 14 and remains stable thereafter. The messages were often quite emotional, could be very personal in nature, and frequently ignored the instructions to avoid cursing. Examples include "good job, [expletive deleted]," "you are a bad person . . . i hope someone [expletive deleted] you over as well", and (our favorite) "you know, they shoot you for that in Texas." Messages of this type are best interpreted as non-pecuniary punishments in the spirit of Masclet *et al*. (2003). Verbal punishment provides a cheap way to punish cheating. If subjects dislike being told off, less cheating occurs than when only conventional punishment (choosing Low in Period 2) is possible.

Subjects who were cheated and sent a verbal punishment were more likely to engage in conventional punishment as well: 42% of subjects who were cheated and admonished their opponents also chose a price other than High, compared with only 22% of subjects who were cheated and did not admonish their opponent. While the subjects who admonish cheating presumably believe they are punishing the recipients, the recipients may not care. We can divide subjects into those who *never* received an admonishment and those who *ever* received an admonishment. For convenience, we label these groups as "non-cheaters" and "cheaters." Most of the subjects (56/76) are non-cheaters. Non-cheaters only cheat on 2% of collusive agreements and frequently use verbal punishment, admonishing cheating/lying in 80% of observations where they followed a collusive agreement and their opponent cheated. We cannot provide direct evidence that non-cheaters dislike receiving verbal punishment since they do not discuss the matter and, by definition, never receive verbal punishment. However, their frequent use of verbal punishment certainly suggests they may think receiving verbal punishment is bad. The effectiveness of implicit threats, which may be seen as threats of non-pecuniary punishment, and the low usage of explicit threats can also be seen as indirect evidence that receiving verbal punishments is considered bad.[23]

Cheaters cheat on 49% of collusive agreements, with the probability of cheating remaining stable over time. In 80% of the observations in which cheaters cheat, they are admonished. If they disliked getting admonished, we would expect them to cheat less following an admonishment. Instead, the rate of cheating by cheaters who have previously been admonished rises slightly to 52%. When cheaters do not cheat but get cheated by their opponents, they are far less likely to admonish (42%) than non-cheaters.

Given that our population consists of two types with relatively stable (if very different) behavior, stable cooperation in the RChat treatment occurs because the vast majority of our subjects were non-cheaters who use verbal punishment and presumably fear receiving it. We can make an educated guess that non-cheaters do not cheat, in spite of poor monetary incentives, to avoid verbal punishment. Cheaters do not seem to be inhibited by verbal punishment, but there are too few of them to destabilize the social norm of cooperation.

5. Conclusions: The primary purpose of our experiment was to study how communication facilitates the development of stable collusion. Our results indicate that allowing subjects to communicate an intent to collude is insufficient to generate persistent collusion even if subjects can send contingent messages calling for punishment of cheating, but allowing a rich message space leads to persistent collusion. As suggested by theory, the use of explicit threats to punish deviation from collusive agreements is the most effective type of message for promoting collusion when only pre-game communication is allowed. Collusion is also promoted by sending a message promising trustworthy behavior. In sessions where renegotiation is allowed, high levels of collusion occur contrary to standard theories of renegotiation. While attempts at renegotiation occur and are reasonably successful, as predicted by theory, the effect of renegotiation is overwhelmed by the impact of verbal punishment of cheating which provides an inexpensive and easily understood means of supporting collusion. Pre-play use of implicit threats is quite effective in the renegotiation treatment, presumably because implicit threats raise the specter that cheating will be met by verbal punishment.

[23] We cannot test whether subjects who make implicit threats are more likely to use verbal punishment because there are almost no observations where a subject makes an implicit threat and is cheated.

In our experiments, monetary incentives dominate behavioral explanations for collusive behavior. Sending promises of trustworthy behavior leads to reduced cheating, in line with the results of Charness and Dufwenberg (2006), but the fact that a lie about trustworthy behavior is severely punished gives subjects a very concrete reason not to renege on promises, eliminating the need to rely on more subtle psychological factors like guilt aversion. Receiving a promise of trustworthy behavior has no measurable effect on the likelihood of cheating.

Given the stress on promises in the economics and psychology literatures, readers may be surprised by the lack of an effect in our data. We conjecture that the critical feature of our experiments is that cooperation is consistent with equilibrium play, so monetary incentives are sufficient to support collusion without relying on behavioral factors.

The most surprising result of our paper is that renegotiation facilitates collusion even though the basic logic of renegotiation theory finds support in the data. This suggests that personal interactions play an important and theoretically underappreciated role in collusion. However, a natural question to ask is when results that rely so heavily on an emotional reaction from subjects will extend to other settings, particularly decisions made in a corporate setting. We suspect that this depends on two factors. First, given that there are two distinct types in our population, differences may occur between populations because of a differing mix of types. Beyond this, holding the mix of types constant, differences between settings are likely to be driven by the social context. For example, almost all business decisions are made in a group context and within an organizational hierarchy. Individuals in these settings may feel more responsible to other people within their corporation than to a person in another firm and hence be less sensitive to being admonished for cheating. These issues are central to our ongoing research.

Our experimental results have serious methodological implications. We have simplified the study of communication and cooperation in repeated game settings by using a finite game, but have also made it more complex by using chat rather than a limited message space. Comparing the P1C and PChat treatments, using a limited message space treatment appear to limit subjects' ability to effectively use a given message space even if we do not eliminate vital components of communication such as verbal punishment in the RChat treatment. In a follow-up experiment, we have modified the P1C treatment to allow for renegotiation, allowing each player to send a message about their intended action for Period 2 after the outcome for Period 1 has been observed. This has a more modest effect than RChat has relative to PChat, neither increasing nor decreasing collusion relative to P1C.[24] The difference in results illustrates another problem with using limited message space to study communication, as the verbal punishments that play such a critical role in RChat are not available with the limited message space. Given the two preceding observations, we believe that using limited message space treatments to study the role of communication in fostering cooperation risks missing important features of how communication functions in the more natural context of a conversation.

We view our work as a complement to field work studying the transcripts of communication between colluding firms (e.g. Genesove and Mullin 2001). Field data on communication and collusion comes from firms that were sufficiently successful at colluding to warrant prosecution and sufficiently indiscrete (or possibly unlucky) to get caught, but in the lab we observe the full population, including firms who try to collude and fail. The controlled environment of the lab also allows us to manipulate what types of communication are available to our subjects, for example turning the ability to renegotiate on or off. Players in the field are presumably fairly experienced at the game being played, but in the lab we get to see the learning process as subjects gain experience by playing the game repeatedly. Lab experiments cannot match the verisimilitude of field data, but the richness of the data and the ability to study non-naturally occurring communication structures make them a valuable tool for understanding the relationship between communication and collusion.

On a broad level our results highlight the importance of making a sharp distinction between explicit and tacit collusion in anti-trust policy as suggested by Whinston (2006). It is possible to achieve tacit collusion in the lab, but direct communication makes collusion easier. More importantly, our experiments point to the types of communication that should be of particular concern in anti-trust enforcement. Calls for collusion in isolation may not be terribly effective. What truly matters is laying out a punishment for failure to stick to a collusive agreement.

[24] In a similar vein, Andersson and Wengström (2010) report evidence in favor of renegotiation theory from an experiment with limited message spaces. They actually find somewhat lower cooperation with renegotiation. The game studied in their paper differs from ours, allowing only two actions rather than three among other things, and this may responsible for the slightly different findings.

References

1. **Abreu, Dilip.** “On the Theory of Infinitely Repeated Games with Discounting.” *Econometrica*, 1988, 56(2), pp. 383-396.
2. **Abreu, Dilip; Pearce, David and Stacchetti, Ennio.** “Toward a Theory of Discounted Repeated Games with Imperfect Monitoring.” *Econometrica*, 1990, 58(5), pp. 1041-1063.
3. **Abreu, Dilip; Pearce, David, and Stacchetti, Ennio.** "Renegotiation and Symmetry in Repeated Games." *Journal of Economic Theory*, 1993, 60(2): 217-240.
4. **Andersson, Ola and Wengström, Erik.** “More communication, less cooperation: Experimental evidence from multi-stage games,” Lund University and University of Copenhagen, working paper, 2010.
5. **Aumann, Robert.** "Nash Equilibria are not Self-Enforcing." In Gabszewicz, J. J., J.-F. Richard, and L. A. Wolsey, eds., Economic Decision-Making: Games, Econometrics and Optimisation. Amsterdam: Elsevier, 1990, pp. 201-6.
6. **Bernheim, Douglas B. and Ray, Debraj.** “Collective Dynamic Consistency in Repeated Games.” Games and Economic Behavior, 1989, 1(4), pp. 295-326.
7. **Bernheim, B. Douglas; Peleg, Bezalel; and Whinston, Michael D.** ‘‘Coalition-Proof Nash Equilibrium: I. Concepts.’’ *Journal of Economic Theory*, 1987, 42(1), pp. 1-12.
8. **Benoit, Jean-Pierre and Krishna, Vijay.** “Finitely Repeated Games.” *Econometrica*, 1985, 53(4), pp. 905-922.
9. **Benoit, Jean-Pierre and Krishna, Vijay.** “Renegotiation in Finitely Repeated Games.” *Econometrica*, 1993, 61(2), pp. 303-323.
10. **Brandts, Jordi and Cooper, David J.** “It's What You Say, Not What You Pay: An Experimental Study of Manager–Employee Relationships in Overcoming Coordination Failure.” *Journal of the European Economic Association*, 2007, 5(6), pp. 1223-1268.
11. **Blume, Andreas and Ortmann, Andreas.** “The effects of costless pre-play communication: Experimental evidence from games with Pareto-ranked equilibria.” *Journal of Economic Theory*, 2007, 132(1), pp. 274-290.
12. **Cason, Tim and Vai-Lam Mui,** “Coordinating Resistance through Communication and Repeated Interaction,” Purdue and Monash Universities, working paper, 2009.
13. **Charness, Gary and Dufwenberg, Martin.** “Promises and Partnership.” *Econometrica*, 2006, 74(6), pp. 1579-1601.
14. **Charness, Gary and Dufwenberg, Martin.** “Bare Promises,” *Economic Letters,* 2010, forthcoming.
15. **Cohen, J.**, “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, 1960, 20(1), pp.37–46.
16. **Cooper, Russell; DeJong, Douglas V.; Forsythe, Robert and Ross, Thomas W.** “Communication in Coordination Games.” *The Quarterly Journal of Economics*, 1992, 107(2), pp. 739-771.
17. **Cooper, David J. and Kühn, Kai-Uwe.** “The Effect of Limited vs. Rich Message Spaces on Collusion,” Florida State University and University of Michigan, working paper, 2010
18. **Dal Bó, Pedro.** “Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games.” *The American Economic Review*, 2005, 95(5), pp. 1591-1604.
19. **Dal Bó, Pedro and Fréchette, Guillaume R.** “The Evolution of Cooperation in Infinitely Repeated Games.” *The American Economic Review,* 2011, 101(1), pp. 411–29. .

20. Dawes, Robyn M.; McTavish, Jeanne and Shaklee, Harriet. “Behavior, Communication, and Assumptions about other People's Behavior in a Commons Dilemma Situation.” *Journal of Personality and Social Psychology*, 1977, 35(1), pp. 1-11.

21. Dawes, Robyn, Orbell, John, & van de Kragt, Alphons. “Not me or thee but we: The importance of group identity in eliciting cooperation in dilemma situations.” *Acta Psychologica*, 1988, 68, 83-97.

22. Duffy, John and Ochs, Jack. “Cooperative Behavior and the Frequency of Social Interaction.” *Games and Economic Behavior*, 2009, 66(2), pp. 785-812.

23. Farrell, Joseph and Maskin, Eric. “Renegotiation in repeated games.” *Games and Economic Behavior*, 1989, 1(4), pp. 327-360.

24. Farrell, Joseph and Rabin, Matthew. “Cheap Talk.” *The Journal of Economic Perspectives*, 1996, 10(3), pp. 103-118.

25. Fischbacher, Urs (2007): z-Tree: Zurich Toolbox for Ready-made Economic Experiments, Experimental Economics 10(2), 171-178.

26. Genesove, David and Mullin, Wallace P. “Rules, Communication, and Collusion: Narrative Evidence from the Sugar Institute Case.” *The American Economic Review*, 2001, 91(3), pp. 379-98.

27. Harsanyi, John and Selten, Reinhard. A General Theory of Equilibrium Selection in Games, Cambridge, MIT Press, 1988.

28. Holt, Charles and Davis, Douglas. “The Effects of Non-binding Price Announcements on Posted Offer Markets.” *Economics Letters*, 1990, 34, pp. 307–310

29. Isaac, Mark R. and Walker, James M. “Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism.” *The Quarterly Journal of Economics*, 1988, 103(1), pp. 179-199.

30. Kerr, Norbert L., and Kaufman-Gilliland, Cindi M. “Communication, Commitment, and Cooperation in Social Dilemma.” *Journal of Personality and Social Psychology,* 1994, 66, pp. 513-529.

31. Masclet, David; Noussair, Charles; Tucker, Steven and Villeval, Marie-Claire. “Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism.” *The American Economic Review*, 2003, 93(1), pp. 366-380.

32. Roth, Alvin E. and Murnighan, J. Keith. “Equilibrium behavior and repeated play in prisoners' dilemma games.” *Journal of Mathematical Psychology*, 1978, 17, pp. 189-198.

33. Van Damme, Eric. “Renegotiation-proof Equilibria in Repeated Prisoners' Dilemma.” *Journal of Economic Theory*, 1989, 47(1), pp. 206-217.

34. Van Huyck, John B.; Battalio, Raymond C. and Beil, Richard O. “Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure.” *The American Economic Review*, 1990, 80(1), pp. 234-248.

35. Xiao, Erte and Houser, Daniel. "Emotion Expression in Human Punishment Behavior." *Proceedings of the National Academy of Sciences*, 2005, 102(20), pp. 7398-7401.

Appendix A: Details of the Econometric Analysis

A.1: Treatment Effects: To confirm the statistical significance of our treatment effects, Table A.1 shows the results of two regressions. For both regressions, the dataset consists of all individual choices from Rounds 11 – 20. The dependent variable is the Period 1 choice. Given that the available choices are naturally ordered categories, we use an ordered probit specification (0 = Low, 1 = Medium, 2 = High). The cut points between categories are not reported since these are not of direct interest. All standard errors are corrected for clustering at the subject level.

$$\begin{aligned} P_{it} = \alpha + \sum_{RdCat=2}^{5}\left(\beta_{RdCat}d_{RdCat}\right) + \varphi Ave\_Per1_i + \gamma\, Aver\_Per2_i + \\ \sum_{RdCat=1}^{5}\left(\gamma_{RdCat}d_{RdCat}\left(d_{P1}+d_{P1C}+d_{PChat}+d_{RChat}\right)\right) + \sum_{RdCat=1}^{5}\left(\psi_{RdCat}d_{RdCat}\left(d_{P1C}+d_{PChat}+d_{RChat}\right)\right) \\ + \sum_{RdCat=1}^{5}\left(\eta_{RdCat}d_{RdCat}\left(d_{PChat}+d_{RChat}\right)\right) + \sum_{RdCat=1}^{5}\left(\nu_{RdCat}d_{RdCat}\left(d_{RChat}\right)\right) + \varepsilon_{it} \end{aligned} \tag{A1}$$

The two models are designed to answer different questions about the data. Model 1 looks for differences between treatments. The equation being estimated for the latent variable is shown above as (A1). The dependent variable is the Period 1 price for subject i in Round t (P$_{it}$). To control for changes over time, rounds have been broken down into five categories: category 1 (RdCat = 1) is Rounds 11 and 12, category 2 (RdCat = 2) is Rounds 13 and 14, etc. Use of a non-linear specification for time is necessary given the non-monotonic time trend in the Pchat treatment.[25] The variables d$_{RdCat}$ are dummies for the five round categories. The variables d$_{P1}$, d$_{P1C}$, d$_{PChat}$, and d$_{RChat}$ are dummies for the P1, P1C, PChat, and RChat treatments respectively. The interactions between dummies are stacked so we get an estimate of the difference between pairs of treatments in each round category. For example, $\bullet_1$ estimates the difference between the P1 and N treatments in round category 1 (Rounds 11 – 12), $\bullet_1$ estimates the difference between the P1 and P1C treatments in round category 1, $\bullet_1$ estimates the difference between the PChat and P1C treatments in round category 1, and $\bullet_1$ estimates the difference between the PChat and RChat treatments in round category 1. The variables Ave_Per1$_i$ and Ave_Per2$_i$ give the average Period 1 and Period 2 prices respectively for subject i from Rounds 1 – 10. These averages are

[25] Using categories containing two rounds rather than round dummies makes reporting results more manageable by reducing the number of parameters and does not affect the conclusions.

calculated setting 0 = Low, 1 = Medium, and 2 = High. These variables are included to better capture the individual effects in the data.

The results of Model 1 strongly support the existence of treatment effects on Period 1 choices. With one exception, all of the pairwise comparisons of treatments for a two round block are statistically significant at least at the 5% level and generally at the 1% level.[26] The sole exception is that the initial difference (Rounds 11 – 12) between the PChat and RChat treatments is not statistically significant. The control for average Period 1 price in Rounds 1 – 10 is statistically significant, consistent with the existence of strong individual effects in the data. There is no statistically significant relationship between Period 2 prices in Rounds 1 – 10 and Period 1 prices in Rounds 11 – 20.

$$\text{(A2)}\quad \begin{aligned} P_{it} &= \alpha + \sum_{RdCat=2}^{5}\left(\beta_{RdCat}\delta_{RdCat}d_{N}\right) + \sum_{RdCat=1}^{5}\left(\gamma_{RdCat}\delta_{RdCat}d_{P1}\right) + \sum_{RdCat=1}^{5}\left(\psi_{RdCat}\delta_{RdCat}d_{P1C}\right) \\ &+ \sum_{RdCat=1}^{5}\left(\eta_{RdCat}\delta_{RdCat}d_{PChat}\right) + \sum_{RdCat=1}^{5}\left(\nu_{RdCat}\delta_{RdCat}d_{RChat}\right) + \varphi Ave\_Per1_{i} + \gamma Aver\_Per2_{i} + \varepsilon_{it} \end{aligned}$$

Model 2 looks for changes within treatments over time. The most important question this model addresses is whether the dip and recovery in Period 1 prices for the PChat treatment is statistically significant. The equation being estimated is shown in (A2). The dependent variable has not changed from Model 1, and Ave_Per1$_{\mathrm{i}}$ and Ave_Per2$_{\mathrm{i}}$ are defined as in Model 1. The variables $\mathrm{d_N}$, $\mathrm{d_{P1}}$, $\mathrm{d_{P1C}}$, $\mathrm{d_{PChat}}$, and $\mathrm{d_{RChat}}$ are dummies for the N, P1, P1C, PChat, and RChat treatments respectively. The primary change from Model 1 comes in how the round category dummies are defined. The variable $\bullet_{\mathrm{RdCat}}$ is a dummy for all observation from that round category *and* subsequent rounds. Thus, $\bullet_1$ is a dummy for Rounds 11 – 20, $\bullet_2$ is a dummy for Rounds 13 – 20, $\bullet_3$ is a dummy for Rounds 15 – 20, and so forth. The dummies are set up to estimate the difference in Period 1 play between two consecutive round categories for the same treatment. For example, $\bullet_1$ estimates the difference between round category 1 of P1 treatment and the base (round category 1 of the N treatment), $\bullet_2$ estimates the difference between round category 1 (Rounds 11 – 12) and round category 2 (Rounds 13 – 14) for the P1 treatment, $\bullet_3$ estimates the difference between round category 2 (Rounds 13 – 20) and round category 3

[26] Within a round category, this also holds for the pairwise comparisons that are not explicitly made in Model 1 since the treatments have been stacked from lowest to highest Period 1 prices.

(Rounds 15 – 16) for the P1 treatment, and so on. The • , •, and • parameters measure equivalent differences for the P1C, PChat, and RChat treatments.

Looking at the results for Model 2, the most important issue is whether the u-turn in the PChat treatment is statistically significant. The decline between round categories 1 (Rounds 11 – 12) and 2 (Rounds 13 – 14) is statistically significant at the 5% level and there are smaller (and not statistically significant) declines between round categories 2 and 3 and round categories 3 and 4. The difference between round category 1 (Rounds 11 – 12) and round category 4 (Rounds 17 – 18) is statistically significant at the 1% level.[27] The downward trend then reverses, as the increase between round categories 4 (Rounds 17 – 18) and 5 (Rounds 19 – 20) is statistically significant at the 1% level. Both the initial decline in Period 1 prices for the PChat treatment and the following recovery are statistically significant changes. If we compare Period 1 prices for round categories 1 and 5 in the PChat treatment, the difference is not statistically significant.[28] By the end of the experiment, Period 1 prices in the PChat treatment have returned to the levels reached immediately following the introduction of communication. Turning to the other treatments, decreases in Period 1 prices are statistically significant at the 5% level in all round categories for the P1 and P1C treatments. Thus, even though Model 1 indicates there remain significant differences between the N, P1, and P1C treatments for round category 5 (Rounds 19 – 20), we feel confident in stating that play has not converged in the P1 and P1C treatments and hence these differences would probably not persist if the experiment ran for more rounds. The RChat treatment shows a weak increase in Period 1 prices. If we compare Period 1 prices for the round categories 1 and 5, the difference is statistically significant at the 10% level.[29]

A.2: Effect of Message Types: Equation A3 shows the full specification estimated in Table 3 of the main text, with Cheat_{it} giving the latent variable, where i indexes subjects and t indexes rounds. $\text{Received}_{it}^{\text{Cat}}$ is the variable for subject i receiving a message coded under category "Cat" in Round t and $\text{Sent}_{it}^{\text{Cat}}$ is the variable for subject i sending a message coded under category "Cat" in Round t. The variables d_{Medium} and d_{Low} are dummies for agreements to play Medium or Low, respectively, in Period 1. Agreement to play High is the excluded category. The variables denoted (d_{Round}) are dummies for each round, with the dummy for Round 11 as the excluded

[27] We use a variant on Model 2 to estimate the change between round categories 1 and 4 for the PChat treatment. The parameter estimate for the difference is -.496 with a standard error of .149.

[28] Using a variant on Model 2 to estimate the difference between round categories 1 and 5 for the PChat treatment, the parameter estimate for the difference is .026 with a standard error of .148.

[29] The parameter estimate for this difference is .495 with a standard error of .267.

category. The subject's average Period 1 price in Rounds 1 – 10 is given by Ave_Per1$_i$ and the average Period 1 price in Rounds 1 – 10 of their opponent in Round t is given by Ave_Opponent_Per1$_{it}$.

(A3)
$$Cheat_{it} = \alpha + \sum_{Round=2}^{10} \left(\beta_{Round} d_{Round} \right) + \sum_{\substack{Cat \in \\ included\ categories}} \left(\rho_{Cat} Received_{it}^{Cat} + \sigma_{Cat} Send_{it}^{Cat} \right)$$
$$+ \lambda_H d_{Medium} + \lambda_L d_{Low} + \varphi Ave\_Per1_i + \gamma\, Ave\_Opponent\_Per1_{it} + \varepsilon_{it}$$

Table A.2 displays a number of regressions designed to test the robustness of the results reported in the text. We focus on the strong effect of receiving an explicit threat on reducing cheating in the PChat treatment. One of the central results of our paper is that the effect of explicit threats is far greater than any other type of message in the PChat treatment. While this can be seen from Figures 2 – 3, our confidence in the result relies largely on the strong statistical results reported in Model 1. If these results were not robust, our confidence in the importance of explicit threats would be reduced.

The dataset for the regressions shown in Table A.2, unless otherwise stated, is all observations from the PChat treatment where the subjects reached an agreement. The dependent variable is a dummy for whether a player cheated on their agreement. As in Table 3, controls for the current period, the agreement, and behavior in the first ten rounds are included in the regressions but not reported in the table to save space. The row giving the critical estimate, the effect of receiving an explicit threat, has been highlighted in yellow.

As a point of comparison, Model 1 on Table A.2 repeats Model 1 from Table 3 in the main text. The codes for *sent* messages are included in Model 1 to control for omitted variable bias, but we note in the text that the effect of including these variables on the estimated effects of *received* messages is minimal suggesting there is little omitted variable bias present. This point is confirmed by Model 2. This model is identical to Model 1 except all the categories for *sent* messages have been omitted. The effect of this omission on the log likelihood is large (-325.77 vs. -342.56), but the effect on the parameter estimates for *received* messages is small. The parameter estimate for receiving an explicit threat is slightly reduced, but remains easily significant at the 1% level. The marginal effect is 25% in Model 2 as opposed to 26% in Model 1. Looking at the other message categories, none are significant at even the 10% level. We believe including the sent messages was the right choice especially since the resulting estimates are interesting in their own right. The point is that our conclusions vis-à-vis *received* messages are robust to whether or not controls for *sent* messages are included.

Another reason the results of Model 1 might not be robust is uncontrolled individual effects. Model 1 includes multiple features designed to control for individual effects: standard errors are corrected for clustering at the individual level and controls are included for the player's behavior in Rounds 1 – 10 and their opponent's behavior in Rounds 1 – 10. Going even further, we can include fixed effects in the regression to control for individual effects. Doing this within the framework of a probit is problematic because many subjects either never or always cheat (21 of 64 subjects). We therefore move to a linear probability model to avoid dropping large numbers of observations. Model 3 includes fixed effects for the subject making choices. Model 4 also contains fixed effects for their opponents. In both cases the standard errors are corrected for clustering at the subject level, correcting for the heteroskedasticity associated with use of a linear probability model. In Model 3, the control for the player's behavior in Rounds 1 – 10 is dropped (since it is subsumed by the fixed effect) and in Model 4 the control for the opponent's behavior in Rounds 1 – 10 is also dropped.

For both Model 3 and Model 4, receiving an explicit threat significantly reduces the likelihood of cheating. In both cases the estimated marginal effect of receiving an explicit threat is similar to the 26% estimated for Model 1. There are no significant effects from receiving other types of messages. Turning to sent messages, the strongest effect, both in terms of magnitude and statistical significance, is from sending an explicit threat. The marginal effect of explicit threats is smaller for Models 3 and 4 than in Model 1, in line with the reduced marginal effect when Model 1 is modified to control for whether a subject *ever* sent an explicit threat (this reduces the estimated marginal effect from 40% to 35%). Sending a promise of trustworthy behavior continues to have a significant negative effect on cheating, and sending an appeal for trustworthy now also has a significant negative effect. Using a fixed effects model has a mild effect on our results but does not affect our primary conclusion that explicit threats are by far the most important type of message in determining whether players will cheat on a collusive agreement. We view Model 1 as preferable to Models 3 and 4 since it does not require moving to a linear probability model, but this is primarily a matter of taste since the results are very similar.

A final concern for Model 1 is that the received messages are endogenous – to be precise, the received messages may be correlated with the error term. The most obvious sources of such correlation are failure to control for individual effects or the effect of sent messages. We hope that at this point it is apparent that these are unlikely to cause the estimated effect of receiving an explicit threat. However, it never hurts to be extra careful. The only channel through which a subject can affect the messages he receives is the messages he sends, since no other interaction between subjects occurs prior to Period 1. A cautious reader could argue that messages contain nuances that our coding cannot capture, so that controlling for commonly coded sent messages is not sufficient. It can also be argued that because sent messages may well be endogenous, the estimates for received messages will be biased as well. Given that the results for Models 1 and 2 are quite similar, this is unlikely, but only an instrumental variables approach can adequately address concerns about endogeneity. Model 5 therefore presents an "all of the above" approach to establishing whether there is a causal relationship between cheating and receiving explicit threats. This is a two-stage least squares model. As instruments for the received messages we use the messages sent by the opponent in the previous round, the messages received by the opponent in the previous round, and the opponent's outcome in the previous round. We drop any observations where the same subjects were matched for the previous round, so a subject's opponent's actions and outcomes from the previous round should be uncorrelated with the subject's current error term. All Round 11 observations are dropped since there are no previous round messages to use as instruments. To control for individual effects, fixed effects are included for a subject and his opponent. Once again this is a linear probability model. No controls are included for sent messages since these are endogenous.

Model 5 once again yields the result that receiving an explicit threat has a strong, statistically significant effect on reducing cheating. The estimated effect of receiving an expression of distrust is also significant, although the marginal effect is lower than the effect of receiving an explicit threat and the statistical significance is marginal. We do not think that the instrumental variables approach taken by Model 5 is the best way to study the data. A large percentage of the data is discarded and a great deal of power is lost through the use of instruments. To be frank, the magnitude of the parameter for explicit threats is implausible even accounting for this being a linear probability model and it is difficult to see why receiving an expression of distrust should lead to *more* trustworthy behavior. Nonetheless, we feel that it is worthwhile showing that going the extra step of instrumenting for received messages does not affect our main conclusion: there is a causal relationship between receiving an explicit threat and cheating on collusive agreements. This effect is stronger than the effect of any other category of message.

Appendix B: Full List of Codes

This appendix shows the original list of codes that was given to the coders. Notes in square brackets discuss interpretations of the codes and changes that were made after the coding process had started.

Period 1 Codes

1. Proposal of Action
    a. Proposed Action period 1
        i. Both A
        ii. Both B
        iii. Both C
    b. Proposed Action period 2
        i. D
        ii. E
        iii. F
2. Response to Proposal
    a. Disagreement
    b. Weak Agreement
    c. Clear Agreement

[We initially hoped to distinguish the intensity of agreement with proposals. We abandoned this when it became clear that there was no valid way to make this distinction. The final version of the coding combined 2b and 2c into a single category for agreement.]

3. Proposed Threats
    a. Nonspecific Threat
    b. Concrete Threat with Medium
    c. Concrete Threat with Low
    d. Mutual Threat
    e. Explicitly non-contingent
4. Response to Proposed Threats

    a. Disagreement
    b. Weak Agreement
    c. Strong Agreement
    d. Extension to Mutual Threat
    e. Request for explanation

[Categories 4b, 4c, and 4d were combined into a single category as it proved too difficult to distinguish between the varying degrees of agreement.]

5. Request for Proposals
6. Explanation
    a. In reference to own proposal
    b. In reference to other's proposal
    c. In reference to own proposed threat
    d. In reference to other's proposed threat
    e. Appeal to mutual benefits
    f. Appeal to "fairness"
    g. Discussion of incentive to cheat
    h. Safety or risk
    i. Specific reference to payoff table
    j. Explanation of contingencies
7. Cheating
    a. Weak Cheating
    b. Clear Cheating
    c. Strong Cheating

[This was not a coding category per se. To help us identify interesting dialogues, we asked the coders to keep track of cases where they thought somebody had cheated on an agreement, with subcategories for the intensity of cheating. This is *not* the variable used to measure cheating in the analysis contained in the main text. See the main text for a description of how cheating was measured.]

8. Boredom
9. Trust and Fairness
    a. Indicating that you should be trusted

b. Indicating that you trust the other person

c. Indicating that you *do not* trust the other person

d. Appeal for mutual trust

e. Appeal for trustworthy behavior

f. Appeal to fairness

10. Past Play

a. Reporting about having been cheated

b. Self-reporting about past own behavior

c. Judgmental comments about others' behavior

d. Agreement about judgmental comments

e. Sympathy

f. Inaccurate reporting

Period 2 Codes

11. Comments on Previous Period

a. Positive feedback after first period cooperation

b. Positive feedback after both deviate first period

c. Apology for cheating

d. Suggesting to cheat in future rounds to make up for loss

e. Rationalizing cheating

f. Clarifying whether deviation was deliberate or accident

g. Admonition for cheating

h. Admonition for lying

[Categories 11g and 11h are not well distinguished, so we have combined them into a single category for purposes of analysis.]

12. Proposal of Action (period 2)

a. D

b. E

c. F

13. Response to Proposal

a. Disagreement

b. Weak Agreement

c. Clear Agreement

d. Mutual Statement of Same Action

[Categories 13b, 13c, and 13d were combined into a single category as it proved too difficult to distinguish between the varying degrees of agreement.]

14. Promise not to lie in period 2

15. Request for Proposals

16. Explanation

a. In reference to own proposal

b. In reference to other's proposal

c. Appeal to mutual benefits

d. Pointing out that there are no cheating incentives in period 2

e. Appeal to "fairness"

f. Appeal that past play does not matter

g. Statement that punishment results from first period behavior

h. Absence of reasons for punishments

Appendix C: Derivation of the TPBG from an Infinite Horizon Game

The game played in the first period is based on a standard model from oligopoly theory, a symmetric Bertrand duopoly with homogeneous goods. The game is simplified by only allowing three prices: Low (L), Medium (M), and High (H). Let $\bullet^{i}$ be industry profits if demand is served at price i, and assume $\bullet^{H} > \bullet^{M} > \bullet^{L} > 0$ and $\bullet^{L} > \bullet^{M}/2 > \bullet^{H}/4$. The following matrix (with player 1's strategies being the rows and player 2's strategies the columns) shows the payoffs for the Period 1 game:

$$(C1)\qquad \begin{array}{cccc} \textit{Player } 1 \textit{ payoffs} & L & M & H \\ L & \frac{\pi^{L}}{2} & \pi^{L} & \pi^{L} \\ M & 0 & \frac{\pi^{M}}{2} & \pi^{M} \\ H & 0 & 0 & \frac{\pi^{H}}{2} \end{array}$$

The unique Nash equilibrium of the game shown in (C1) is (L,L). In a typical collusion game we would model the competition between firms as an infinite repetition of the stage game shown in (1), with future payoffs discounted by the discount factor •. Such an infinite horizon game would yield a continuation game with an infinite number of strategies. To reduce the strategy space while still capturing the essential features of the infinitely repeated game, we instead use the pay-off matrix shown in (C2) for the continuation game, where $\bullet^{i} = (\bullet/(1 - \bullet))*(\bullet^{i}/2)$ and • is the discount factor. The rows are player 1's strategies and the columns are player 2's strategies.

$$(C2)\qquad \begin{array}{cccc} \textit{player } 1 \textit{ payoffs} & L & M & H \\ L & \Pi^{L} & \delta[\pi^{L} + \Pi^{L}] & \delta[\pi^{L} + \Pi^{L}] \\ M & \delta\Pi^{L} & \Pi^{M} & \delta[\pi^{M} + \Pi^{L}] \\ H & \delta\Pi^{L} & \delta\Pi^{L} & \Pi^{H} \end{array}$$

Given the definition of $\bullet^{i}$, the payoff matrix in (C2) has three equilibria, in each of which the players choose the same strategy. These equilibria are Pareto ranked with (H,H) being the Pareto dominant equilibrium.

The second period game is derived from the matrix of continuation profits of the infinitely repeated version of (C1) when players are restricted to symmetric stationary equilibrium strategies in which players play the same pair of symmetric actions forever. In the infinite horizon version of (C1) the optimal punishment is to revert to play of (L,L) forever. Hence, the worst equilibrium in (C2) corresponds to the optimal punishment of the infinite horizon game. The payoffs on the diagonal of (C2) then correspond to the discounted payoffs from the three strongly symmetric stationary equilibria that can be sustained with a threat to revert to the optimal punishment equilibrium. The off-diagonal payoffs give the discounted payoffs following a deviation in the second period (i.e. the first period of the continuation game) followed by the most severe punishment equilibrium: If a player is cheated and therefore has a higher price than the other player, he earns zero payoffs in the first period of the continuation game and $\bullet^{L}/2$ thereafter. If a player deviates and undercuts a symmetric equilibrium at price H (M), he receives the industry profit $\bullet^{H}$ ($\bullet^{M}$) in the first period of the continuation game and then $\bullet^{L}/2$ forever. The payoff matrix in (C2) can therefore be interpreted as a reduced form of the infinite horizon game when attention is restricted to the symmetric optimal punishment equilibria. This is the set of equilibria that is often analyzed in applications of collusion theory in industrial organization.

We assume that the incentive conditions are satisfied so that {(L,L),(L,L)}, {(M,M),(M,M)}, and {(H,H),(H,H)} are subgame perfect equilibrium outcomes of the TPBG if players play (L,L) in the second period after any deviation in the first. Colluding at either M or H in Period 1 is therefore feasible, allowing us to detect whether communication leads to full collusion or not.

The two stage game used in the experiments is obtained by setting $\bullet^{L} = 78$, $\bullet^{M} = 138$, $\bullet^{H} = 168$, $\bullet = 2/3$, and subtracting a fixed cost of 24 from all payoffs in every period.