The Blind Trust Game: Costly Monitoring is Not Detrimental to Cooperation

Timo Goeschl[*] Johannes Jarke[†]

April 10, 2014

Abstract

It is well known that (exogenous) monitoring inaccuracies hamper the opportunism supressing function of repeat interaction but there is no evidence on settings in which they are *endogenous*. We introduce the experimental *blind trust game*: a finite horizon modified trust game in which a cooperating first mover does not automatically learn the stage game outcome at the end of a period, but may demand information about the second mover's action. We manipulate the cost of information acquisition, and find that first movers manage to economize on monitoring expenditures without impinging on cooperation. A succinct dynamical strategy with frequent monitoring at the beginning of a match and only sporadic inspections later on limited cheating sufficiently that payoffs were, even including monitoring costs, not significantly lower in a costly monitoring condition than in a costless monitoring condition. Thus, monitoring imperfections are much less of a problem to cooperation if they can be mitigated endogenously, even at a cost. Furthermore, in a world where monitoring is costly, trusting «blindly» at times can be part of a payoff-maximizing dynamic strategy. (*JEL* C92, C72, D03, D80)

[*] Alfred-Weber-Institute of Economics, Universität Heidelberg. Bergheimer Strasse 20, D-69115 Heidelberg, Germany. Phone: +49 6221 54 8010. E-mail: goeschl@eco.uni-heidelberg.de.

[†] Corresponding author. School of Business, Economics and Social Sciences, Department of Socioeconomics, Universität Hamburg. Welckerstrasse 8, D-20354 Hamburg, Germany. Phone: +49 40 42838 8768. E-mail: johannes.jarke@wiso.uni-hamburg.de.

1 Introduction

The realization of mutual gains from cooperation is jeopardized by opportunism in a variety of social interactions. Well known results from game theory posit that a prospect of repeated interaction in the future can have, under appropriate conditions, a disciplining function in those situations (Trivers, 1971; Rubinstein, 1979; Kreps et al., 1982; Fudenberg & Maskin, 1986).[1] One key condition, that cheats are immediately and effortless detected, arguably lacks empirical accuracy. Previous research shows that cooperation is indeed more difficult to sustain if monitoring is subject to random signal errors (see Gintis, 2009, for a theoretical overview and Sell & Wilson, 1991; Holcomb & Nelson, 1997; Feinberg & Synder, 2002; Aoyagi & Fréchette, 2009; Ambrus & Greiner, 2012 for experiments).

However, in practice information on co-players' actions is routinely not only imperfect, but players are frequently in a position to augment available information through costly effort.[2] In other words, a player's choice whether to cooperate is accompanied by a choice on whether to monitor the co-players' actions, whereas such information acquistion is typically costly. Systematic experimental evidence on how monitoring imperfections impact on cooperation in settings in which they are *endogenous* is lacking.

In this paper, we account for this fact, and extend the literature on experimental games in a new direction by allowing important parts of the information structure of the game to be endogenously determined. Specifically, we introduce the experimental *blind trust game*: a finite horizon modified trust game with endogenous monitoring. In the trust game (see e.g. Camerer & Weigelt, 1988; Anderhub et al., 2002; Engle-Warnick & Slonim, 2006b,a; Slonim & Guillen, 2010), a first mover chooses between an outside option and a cooperative move that renders her or him vulnerable to exploitation by a second mover.[3] The latter may either reciprocate the first mover's cooperation at a personal cost or cheat. The novel feature in the blind trust game is

[1] Experimental findings based on a variety of specific stage games, such as the Prisoners' Dilemma game (e.g. Andreoni & Miller, 1993; Cooper et al., 1996; Dal Bó, 2005; Dal Bó & Fréchette, 2011; Duffy & Ochs, 2009), the public good game (Andreoni & Croson, 2008, provide an overview), the gift exchange game (Kirchler et al., 1996; Fehr et al., 1998; Falk et al., 1999; Gächter & Falk, 2002) or the trust (Camerer & Weigelt, 1988; Anderhub et al., 2002; Engle-Warnick & Slonim, 2004, 2006a) and investment game (Cochard et al., 2004) support this hypothesis. For finite horizon games the predicition hinges on the assumption that players hold a belief that some coplayers are committed to cooperate even in the terminal period, given they have not been cheated previously (Kreps et al., 1982), a belief that is indeed justified as demonstrated by a vast amount of recent evidence on cooperative behavior (e.g. Henrich et al., 2004; Gintis et al., 2005; Fehr & Schmidt, 2006).

[2] Narratively, efforts to overcome imperfect information on co-players' actions have been recognized in a variety of relevant contexts, such as shared resource management (Ostrom, 1990; Seabright, 1993; Rustagi et al., 2010), production teams (Alchian & Demsetz, 1972; Kandel & Lazear, 1992; Dong & Dow, 1993), labor relations (Shapiro & Stiglitz, 1984; Kanemoto & MacLeod, 1991; Lazear, 1993), micro-finance (Armendáriz & Morduch, 2005) or neighborhood watch (Sampson et al., 1997).

[3] The trust game is a simplified version of the sequential Prisoner's Dilemma (e.g. Clark & Sefton, 2001; Dhaene & Bouck, 2010). A trust game with multiple levels of cooperation is the well-known investment game introduced by Berg et al. (1995). See also Cox (2004) and Charness et al. (2011).

that a cooperating first mover does not automatically learn her or his payoff at the end of a period, but may actively acquire information about the second mover's action in that period. This design allows for a notion of trust, recurring in the social science literature (Elster, 2007, p. 345), that is stronger than the typical definition in experimental economics: While in the latter a cooperating first mover is typically said to «trust», according to the former «trust» also requires abstention from monitoring.[4] To avoid confusion, we shall refer to this stronger notion as *blind trust*. The blind trust game allows exactly for this notion, therefore its name.

As a benchmark, we implemented a standard finite horizon trust game with (exogenously) perfect monitoring, that is, first movers were automatically informed about the actions of their co-players. This is a replication of previous research (e.g. Anderhub et al., 2002; Engle-Warnick & Slonim, 2006b,a; Slonim & Guillen, 2010), and we find the typical pattern of frequent cooperation (about two in three cases) until close to the terminal period, and a sharp decline in the final two periods. We contrasted this baseline with two treatment conditions.

In the first main condition monitoring was endogenous but costless. First movers were not automatically informed about their co-player's actions but they could demand this information freely at the end of each period. The monitoring decision was not revealed to the second mover. The second main condition was identical to the first, except that there was a fee on information acquisition. The aim is to gather insights on information acquisition behavior and its consequences for cooperation, efficiency, and distribution. A specific goal is to investigate the dynamic patterns: as a working hypothesis, we draw on a proposition from the management literature which poses that ongoing relationships develop through a process starting with a control-driven stage and converging to a trust-based one over time (Lewicki & Bunker, 1996; Lewicki et al., 1998). In the blind trust game, this proposition predicts that monitoring will occur predominantly in the initial periods, whereas blind trust gets more frequent over time.

The key results are as follows: First, as expected we find no significant differences between the baseline condition and the first treatment condition. Second, comparing the two main treatment conditions with endogenous monitoring, we find that first movers managed to economize on monitoring expenditures without impinging on cooperation, such that efficiency was, even including monitoring costs, not lower in the costly monitoring condition than in the costless monitoring condition. Specifically, while first movers *always* monitored their coplayers in the costless information condition, they trusted blindly most of the time in the costly information condition. However, they did so sufficiently smartly such that the average first mover did not worse in the costly monitoring condition than in the costless monitoring condition, and the average blind trustor did not worse than the average monitor or defector. The thrust of this adaption is, in support of the mentioned working hypothesis, a marked

[4] According to Elster, trust is «the result of two successive decisions: *to engage in the interaction and to abstain from monitoring the interaction partner*» (p. 345, emphasis by us). The blind trust game allows exactly for this succession.

dynamic shift from frequent monitoring at the beginning of a match towards blind trust with sporadic inspections over time.

In the remainder we proceed as follows. In section 2 we describe the design of the experiment and report the procedures and implementation. The results are presented in section 3. We summarize and conclude in section 4.

2 The Experiment

2.1 Experimental game

As our stage game we used the well-kown (binary) trust game (see e.g. Camerer & Weigelt, 1988; Anderhub et al., 2002; Engle-Warnick & Slonim, 2006b,a; Slonim & Guillen, 2010).[5] The first mover chooses between cooperation (option «pink» in the instructions) and an outside option («yellow»). In case the outside option is chosen, both players get 15 tokens and the period ends. In case the first mover chooses to cooperate, the period continues with the second mover's choice between cooperation (option «brown») or exploitation (option «blue»). If the second mover cooperates, he gets 25 tokens and his co-player 30 tokens. Otherwise, he exploits the first mover by taking 50 tokens for himself while his co-player gets 5 tokens.[6]

As a baseline that directly replicates previous research we implemented a standard 12-fold repetition with perfect monitoring, that is, each player was automatically and perfectly informed about all previous moves. The novel feature in our main conditions is that a cooperating first mover is not automatically informed about the second mover's choice. Specifically, without knowing the second mover's action, a first mover decides whether he wants to monitor the second mover's action or not. In the former case, the first mover was informed about whether their co-player responded with «brown» or «blue», respectively, at the end of the round. In the latter case, (s)he received no information. Second movers were never informed about whether their co-player monitored them or not.[7]

[5] The trust game prototypically captures situations in which efficiency enhancing cooperation is threatened by a possibility of unilateral exploitation. The sequential structure and its simplicity renders it easy for subjects to understand and the interpretation of observed behavior is less difficult than in simultaneous-move games.

[6] This parametrization is rather standard and intended to generate a fair amount of variance in the data. There is an attractive gain from cooperation (25 tokens), but also a quite lucrative incentive for second movers to cheat (25 tokens). Furthermore, the mutual cooperation payoffs are not equal for the two players in order to avoid a «fair focal point».

[7] This is a reasonable approximation of many real-life situations, but there are clearly other situations in which monitoring activities are observed by the targeted player. We opted for this design for three reasons. First, informing the second mover about the monitoring decision may give rise to reciprocal responses («crowding» of intrinsic motivation), for example because monitoring is perceived as an unkind act or a signal of distrust (e.g. Falk & Kosfeld, 2006). We wanted to eliminate this complication in the current study and leave it for future research. Second, we are interested in second movers' pattern of beliefs about being monitored (see below). A third technical reason for this design is explained in note 11.

In order to learn more about the belief dynamics, we supplemented the experimental game by (non-incentivized) elicitations of the participant's first-order beliefs about their co-player's behavior in the current period. In each period, before any decisions were made, first movers were asked to state their belief about whether their co-player will respond with «brown» or «blue» to «pink», and second movers were asked to state their belief whether their co-player will play «pink» or «yellow». Given that «pink» was played in our main conditions, second movers were asked after their decision to state their belief that their decision will be monitored.

2.2 Design

As a benchmark, we implemented a condition with exogenously perfect monitoring, that is, first movers were automatically informed about the actions of their co-players. This is a replication of previous research (Anderhub et al., 2002; Engle-Warnick & Slonim, 2006b,a; Slonim & Guillen, 2010). We contrasted this baseline with two treatment conditions in which monitoring was endogenous, that is, first movers were not automatically informed about their co-player's actions but they could demand this information at the end of each period. Except this monitoring choice, the baseline and the treatment conditions were exactly identical. In the first treatment condition the acquisition of information on the co-player's action in the current period was costless, in the second condition first movers had to incur a fee of five tokens in order to acquire this information. Except for this variation, both main treatment conditions were exactly identical as well. We used a between-subjects design to assign treatments to participants.

2.3 Subjects and procedures

Participants were recruited from the general undergraduate student population of the University of Heidelberg using the online recruitment system ORSEE (Greiner, 2004). In total 152 subjects participated of which 52.6 percent have been female and 85.5 percent German citizen. The mean age was 23.3 years. Subjects were randomly assigned to treatment conditions, 36 in the baseline condition, 56 in the costless monitoring condition, and 60 in the costly monitoring condition. No subject participated more than once or in more than one treatment condition.

All experiments were conducted at the experimental laboratory of the Alfred-Weber-Institute (AWI-Lab) at the University of Heidelberg. Upon entering the laboratory, subjects were randomly assigned to the computer terminals. Besides each terminal, an empty sheet of paper and a pen was prepared which participants were allowed to use for taking notes during the experiment. They were instructed to take this sheet with them after the experiment to ensure that nobody, including the experimenters, could observe their eventual notes. Booths separated the participants visually, ensuring that they made their decisions anonymously and independently. Direct communication among them was strictly forbidden for the duration of the entire session. Furthermore, subjects did not receive any information on the personal identity of any other participant, neither before nor while nor after the experiment.

At the beginning of the experiment, that is, before any decisions were made, subjects received detailed written instructions that explained the exact structure of the game and the procedural rules. All subjects received the same instructions (only the monitoring fee being replaced across conditions) and this was commonly known. The experiment was framed in a sterile way using neutral language and avoiding value laden terms in the instructions (see supplementary material). Post-experimental debriefings attested that no participant had difficulties in comprehending the instructions. The experiment was programmed and conducted with z-Tree (Fischbacher, 2007).

The exact timing of events was as follows. First, the subjects were randomly matched into groups of two. Then twelve rounds of the experimental game described above were played. The binary decisions were made by input boxes to be marked with the computer mouse, beliefs were indicated by a screen slider with a resolution of 100 points. After the twelve rounds, subjects were asked to answer a short questionnaire while the experimenter prepared the payoffs. Subjects were then informed about their payoffs, and then individually called to the experimenter booth, paid out (according to a random number matched to their decisions; no personal identities were used throughout the whole experiment) and dismissed.

In every session subjects received a fixed show-up fee of €3, which was not part of their endowment. The average session had a duration of about 40 minutes and subjects earned €11.37 (€0.03 per token earned) on average, including the fixed show-up fee, with a minimum of €6.75 and a maximum of €15.15. Average earnings exceed the local average hourly wage of a typical student job significantly and can hence be considered meaningful to the participants.

3 Results

In the baseline condition we find the typical pattern of frequent cooperation (about two in three cases) until close to the terminal period, and a sharp decline in the final two periods. This condition therefore constitutes a reasonable benchmark that corresponds to previous research. We expect no significant differences to the baseline condition and the first treatment condition, because there is no reason to forgo information on co-player behavior if it is acquired at no cost. This is what we find. We refer to appendix A for detailed results. Building on this benchmark result, our focus in this section is on a comparison of the two main treatment conditions.

3.1 Key results

The key results of this paper are illustrated in figures 1 and 2. First of all, average joint payoffs were, even including monitoring costs, *not* significantly lower in the

Figure 1: Efficiency: Average payoff per subject and round by treatment condition.

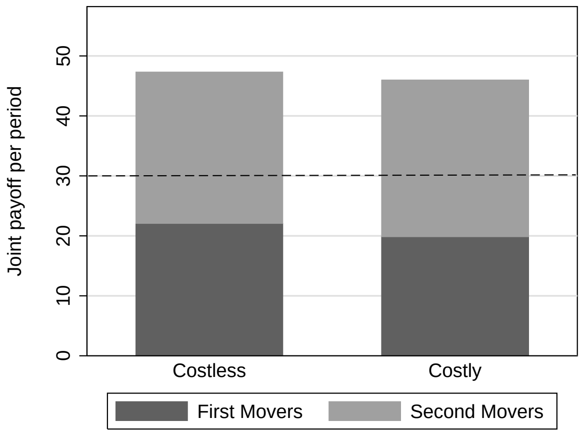


Figure 2: Relative frequency of first mover cooperation pooled over time by treatment condition.

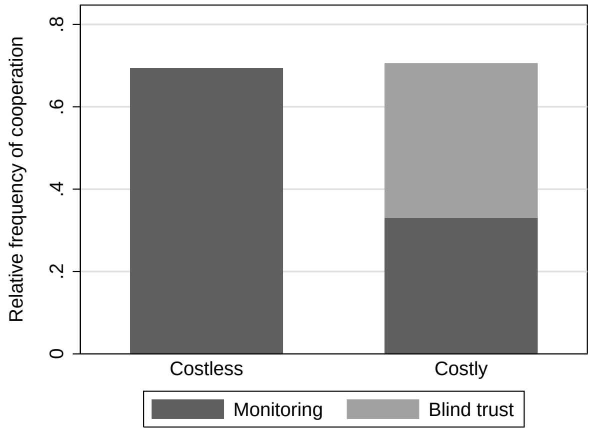
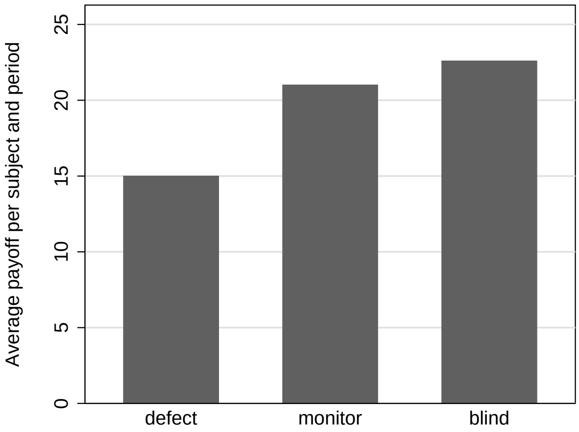


Figure 3: Average first mover payoffs in the costly monitoring condition by behavioral type.



costly monitoring condition than in the costless monitoring condition.[8] Basing this and the following statistical tests on a cross-section in which each observation is a unit-level average taken over all twelve periods,[9] average joint payoffs were not significantly different across conditions (Mann-Whitney rank sum test, $p > .565$). Thus, first movers managed to economize on monitoring expenditures without jeopardizing cooperation. In the costless monitoring condition, first movers cooperated in 69.4 percent of the time (233 out of 336 cases) and *always* monitored their coplayer afterwards. In the costly monitoring condition, they cooperated no less often (254 out of 360 cases, or 70.6 percent) but monitored less than half of the time (119 out of 254 cases, or 46.9 percent).[10] In the remaining cases, first movers trusted their coplayer blindly, that is, cooperating without checking how the second mover responded afterwards. Thus, in the costly monitoring condition blind trust turns out to be the most frequent behavior.

However, it is not obvious why cooperation does not collapse under those circumstances. Blind trustors are easily exploited, perhaps over multiple periods without noticing it. But figure 3, that compares the payoff across a first mover's three possible courses of action,[11] illustrates that the average first mover played out her

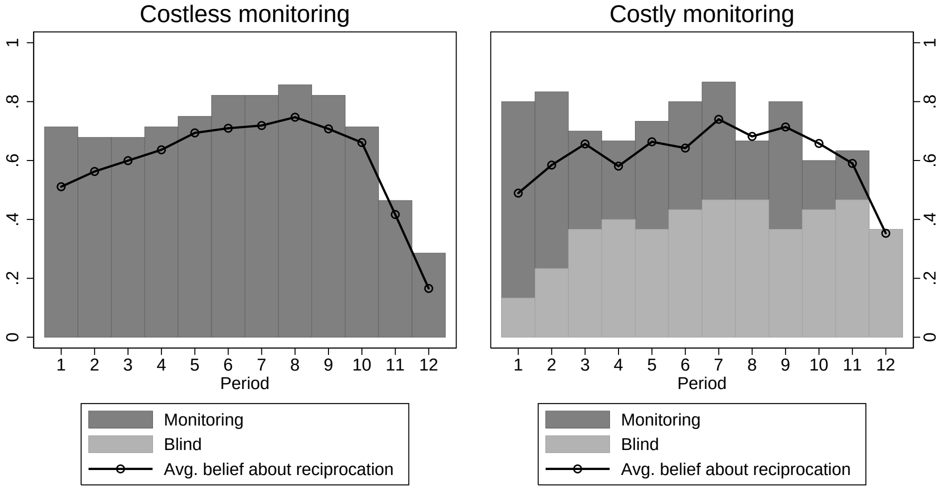
[8] Recall that, excluding monitoring costs, the minimum joint payoff per period was 30 tokens (i.e. 15 per period and subject) and the maximum 55 tokens (27.5 per period and subject). While expenditures on monitoring were, of course, zero in the costless monitoring condition, first movers spent per subject and period 1.65 tokens on monitoring in the costly monitoring condition (i.e. 19.8 tokens per match). Averaging only over those first movers who actually cooperated, they spent 2.34 tokens per period (i.e. 28.1 tokens per match).

[9] Note that the individual observations in our data set are not independent in a rigorous statistical sense, that is, strictly speaking each of the 76 matches constitute one independent observation. The procedure used here is a common response to this fact (e.g. Vanberg, 2009).

[10] The frequency of cooperation is not significantly different across conditions (Mann-Whitney, p = .530), but the frequency of monitoring is (Mann-Whitney, p = .000).

[11] Note that we can do this because second movers were not informed about the monitoring deci-

Figure 4: Aggregate dynamics of first mover cooperation and beliefs about reciprocation. The left column depicts the costless monitoring condition, the right column the costly monitoring condition. The dark shaded area represents the relative frequency of cooperation accompanied by monitoring, the light shaded area represents the relative frequency of blind trust, where both areas are stacked such that the joint area depicts first mover cooperation rates. The connected line depicts the average first mover's belief about the second mover reciprocating.



or his blind trust in a sufficiently wise way to keep exploitation in check: In the costly monitoring condition blind trustors did in fact slightly better than both monitors or defectors. As a result first movers did not earn significantly less in the costly monitoring condition (21.5 tokens per period excluding, and 19.8 tokens including monitoring costs) than in the costless monitoring (22.1 tokens).[12].

It order to understand better how this works out, we have to analyze the full dynamics of the game.

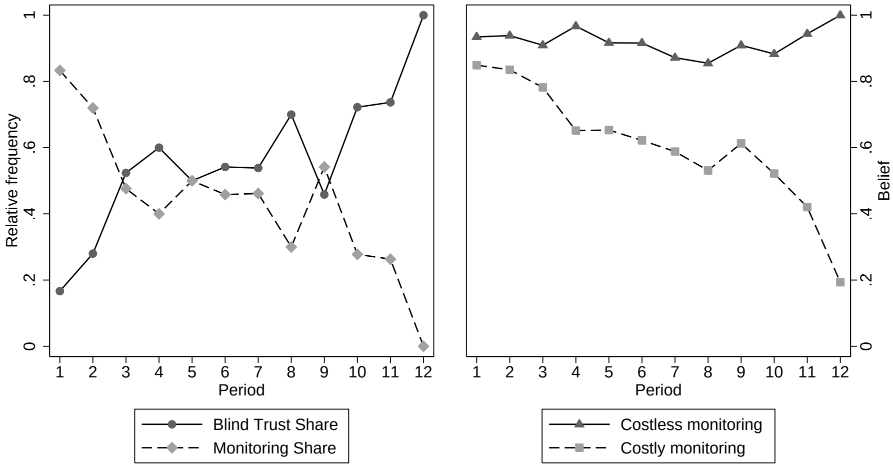
3.2 Dynamics in the costless monitoring condition

As reported above, first movers always monitored their co-player in the costless monitoring condition. This is what one might expect based on elementary economic reasoning, because any non-negative valuation of the information suffices for rendering its acquisition a best response. Not a single one among the 233 instances of coop-

sion. If they would have been, they might have adjusted their behavior and the comparison becomes problematic. This methodical point was the second major reason for this design choice.

[12] If monitoring costs are excluded, all differences are clearly insignificant (Mann-Whitney, $p = .686$). When monitoring costs are included, the difference to the costless monitoring condition is marginally significant (Mann-Whitney, $p = .082$). The small difference is offset by the average second mover who earned slightly more in the costly monitoring condition (26.2 tokens) than in the costless monitoring condition (25.3 tokens), but this difference is statistically insignificant (Mann-Whitney, $p = .307$).

Figure 5: Relative dynamics of monitoring and blind trust in the costly monitoring condition (left-hand column) and the average second mover's belief about being monitored (right-hand column).



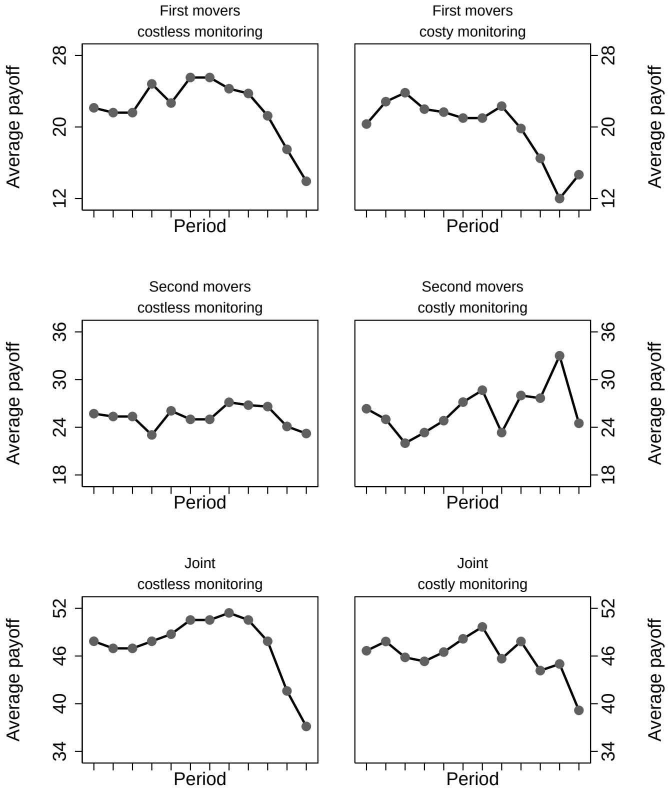
eration (out of 336 opportunities) was blind. As a result, in matches of the costless monitoring condition first movers were *de facto* perfectly informed about the entire history of the game at any time, just as in the baseline condition. The pattern of cooperation exhibited the typical form: The left panel of figure 4 depicts the full aggregate dynamics of first movers' behavior. First movers cooperated in 73.0 percent (225 out of 308) of the time in non-terminal and in 28.6 percent (8 out of 28) of the time in terminal periods.

3.3 Dynamics in the costly monitoring condition

We now consider the costly monitoring condition in more detail. The right-hand panel of figure 4 illustrates that the average first mover's strategy exhibits a succinct dynamical pattern: a shift from monitoring towards blind trust over time. In the first period, blind trust was rare and the vast majority of first movers invested the fee to get information about their co-player's response. Over time, monitoring gets successively less and blind trust more frequent. The left-hand panel of figure 5 illustrates this shift more clearly, depicting the frequency of blind trust and monitoring, respectively, as a fraction of all instances of first mover cooperation. Thus, first movers predominantly resort to monitoring at the beginning of a match while tending to trust blindly towards the end. This lends support to the control-towards-trust hypothesis (Lewicki & Bunker, 1996; Lewicki et al., 1998) mentioned above. However, our results also adds a qualification to this hypothesis, as it is only supported under costly monitoring.

Interestingly, the average second mover anticipates this dynamic pattern. The

Figure 6: Average payoffs over time.



right-hand panel of Figure 5 shows the average belief of second movers of being monitored over time. It is constantly high in the costless monitoring condition, but matches the decreasing trend of actual monitoring decisions in the costly monitoring condition quite closely. Thus, there are stronger incentives to cheat towards the end of a match in the costly monitoring condition because there is a prospect that this remains undetected and unsanctioned.[13]

In fact, this incentive leaves traces in the distribution of payoffs between first and second movers. Figure 6 illustrates the dynamics of realized payoffs. As evident from the left-hand column, the distribution of payoffs between first and second movers was

[13] Conventional economic theory suggests that the temptation to cheat is decreasing in the perceived likelihood of being detected, and *vice versa*. On average, second movers respond consistently with this predicition: In both conditions, the average cooperating second mover had a stronger belief of being monitored (.920 in the costless monitoring condition, .680 in the costly monitoring condition) than the average defecting second mover (.887 in the costless monitoring condition, .467 in the costly monitoring condition), where the difference is significant in the costly (Mann-Whitney, $p < .001$) but not in the costless monitoring condition ($p = .485$). Rank correlation between beliefs of being monitored and reciprocation is positive and significant when monitoring is costly (Kendall's $\tau_b = 0.206$, $p < .001$).

quite stable in the costless monitoring condition, with first movers reaping on average 46.6 percent of the joint payoff, minimally 37.5 percent (round 12) and maximally 51.9 percent (round 4). In the costly monitoring condition, depicted in the right-rand column, the average second mover reaps notably larger payoffs in the second half of the match, both compared to the first half and the costless monitoring condition. In the first six periods, the average first mover got on average a share of 47.0 percent of the joint payoff (47.9 percent in the costless monitoring condition), in the final six periods 39.2 percent (45.2 percent in the costless monitoring condition) with a minimum in the penultimate period (26.7 percent). Thus, trusting blindly is risky.

An individual level analysis of the data suggests that (i) first movers seek to minimize this risk as far as possible by monitoring heavily at the beginning, and slacking off somewhat only if the co-player proves cooperative, and (ii) that there is individual heterogeneity in which they deal with this risk. Illustrations of the individual dynamics are available in appendix B. Summarizing the data, the following can be shown: First, using data on elicited first movers' beliefs, it turns out that the tendency to trust blindly is positively related to the first movers' confidence in reciprocation.[14] Second, using post-experimental survey data on individual preferences we find that the tendency to trust blindly is negatively related to the degree of risk aversion and betrayal aversion (see appendix C).

4 Conclusion

We have shown that monitoring imperfections are much less of a problem to cooperation if they can be mitigated endogenously. Even if monitoring was costly, first movers managed to economize on monitoring expenditures without jeopardizing cooperation. A succinct dynamical strategy with frequent monitoring at the beginning of a match and only sporadic inspections later on limited cheating sufficiently that payoffs were, even including monitoring costs, not significantly lower in a costly monitoring condition than in a costless monitoring condition. Thus, in a world where monitoring is costly, trusting «blindly» at times can be part of a payoff-maximizing dynamic strategy.

Our hope is to stimulate further game theoretic research on endogenous monitoring. Some two decades ago, Hal Varian (1990, p. 153) commented that the agency literature

> «typically assumes that principals are unable to observe the characteristics or the actions of the agents ... However, in reality, it is often not the

[14] This is also illustrated in figure 4. Correlation between the first movers' belief about reciprocation and their own cooperation is strongly positive and significant both the costless monitoring condition (Kendall's τ = .627, p = .000) and the costly monitoring condition (τ = .558, p = .000). However, separating first mover cooperation by cooperation coupled with monitoring and blind trust, it turns out that the latter is correlated more strongly with first movers' beliefs about reciprocation (τ = .378, p = .000) than the former (τ = .152, p < .001). To put it differently, the average belief that the second mover will reciprocate was .820 for blind trustors, .736 for monitors, and .211 for defectors.

> case that agents' characteristics or effort levels are really unobservable; rather, they simply may be very costly to observe. One may choose to model high-costs actions as being infeasible actions, but in doing so, one may miss some interesting phenomena.»

Indeed, the current paper reveals such phenomena that remain hidden if monitoring imperfections are modelled (or designed) as exogenous.

An interesting avenue for further research is a detailed investigation of the strategies the players play in the blind trust game and whether they are in some kind of equilibrium. Specifically, the expectation that there is a higher probability of being monitored during early stages may give second movers an incentive to a «higher order» of reputation building: In standard repeated games with perfect monitoring (but incomplete information), strategically acting second movers can build a favorable reputation in order to induce the first mover to cooperate until close to termination (Kreps et al., 1982). In the blind trust game, second movers can induce the first movers not only to cooperate, but also to refrain from monitoring. We suspect that the incentive for the latter («second-order reputation building») is much stronger than the former («first-order reputation building»), because under perfect monitoring the maximum number of periods in which a strategically acting second mover can exploit the first mover is equal to one (assuming that the first mover will not cooperate again once cheated), while in the blind trust game there is the possibility of cheating over *multiple* periods once the first mover starts to trust blindly. Intuitively, (some) second movers may deliberately try to «earn» a reputation in the initial periods in which they are likely to be monitored, favorable enough to be trusted blindly later on. But this strategy can only be part of some kind of *mixed* strategy equilibrium, because trusting blindly with certainty is not a best response to it. The sporadic inspections many first movers perform in later periods is a hint in this direction.

References

Alchian, A. A. & Demsetz, H. (1972). Production, information costs, and economic organization. *American Economic Review*, 62(5), 777–795.

Ambrus, A. & Greiner, B. (2012). Imperfect public monitoring with costly punishment–an experimental study. *American Economic Review*, 102(7), 3317–3332.

Anderhub, V., Engelmann, D., & Güth, W. (2002). An experimental study of the repeated trust game with incomplete information. *Journal of Economic Behavior & Organization*, 48(2), 197–216.

Andreoni, J. & Croson, R. (2008). Partners versus strangers: Random rematching in public goods experiments. In C. R. Plott & V. L. Smith (Eds.), *Handbook of Experimental Economics Results* (pp. 776–783). Amsterdam, The Netherlands: North-Holland.

Andreoni, J. & Miller, J. H. (1993). Rational cooperation in the finitely repeated Prisoner's Dilemma: Experimental evidence. *Economic Journal*, 103(418), 570–585.

Aoyagi, M. & Fréchette, G. (2009). Collusion as public monitoring becomes noisy: Experimental evidence. *Journal of Economic Theory*, 144(3), 1135–1165.

Armendáriz, B. & Morduch, J. (2005). *The Economics of Microfinance*. Cambridge, MA, USA: MIT Press.

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

Camerer, C. & Weigelt, K. (1988). Experimental tests of a sequential equilibrium reputation model. *Econometrica*, 56(1), 1–36.

Charness, G., Du, N., & Yang, C.-L. (2011). Trust and trustworthiness reputations in an investment game. *Games and Economic Behavior*, 72(2), 361–375.

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

Cochard, F., Van, P. N., & Willinger, M. (2004). Trusting behavior in a repeated investment game. *Journal of Economic Behavior & Organization*, 55(1), 31–44.

Cooper, R., DeJong, D. V., Forsythe, R., & Ross, T. W. (1996). Cooperation without reputation: Experimental evidence from Prisoner's Dilemma games. *Games and Economic Behavior*, 12(2), 187–218.

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, 46(2), 260–281.

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

Dal Bó, P. & Fréchette, G. R. (2011). The evolution of cooperation in infinitely repeated games: Experimental evidence. *American Economic Review*, 101(1), 411–429.

Dhaene, G. & Bouck, J. (2010). Sequential reciprocity in two-player, two-stage games: An experimental analysis. *Games and Economic Behavior*, 70(2), 289–303.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3), 522–550.

Dong, X.-Y. & Dow, G. K. (1993). Monitoring costs in Chinese agricultural teams. *Journal of Political Economy*, 101(3), 539–553.

Duffy, J. & Ochs, J. (2009). Cooperative behavior and the frequency of interaction. *Games and Economic Behavior*, 66(2), 785–812.

Elster, J. (2007). *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. Cambridge, England: Cambridge University Press.

Engle-Warnick, J. & Slonim, R. L. (2004). The evolution of strategies in a repeated trust game. *Journal of Economic Behavior & Organization*, 55(4), 553–573.

Engle-Warnick, J. & Slonim, R. L. (2006a). Inferring repeated-game strategies from actions: evidence from trust game experiments. *Economic Theory*, 28(3), 603–632.

Engle-Warnick, J. & Slonim, R. L. (2006b). Learning to trust in indefinitely repeated games. *Games and Economic Behavior*, 54(1), 95–114.

Falk, A., Gächter, S., & Kovács, J. (1999). Intrinsic motivation and extrinsic incentives in a repeated game with incomplete contracts. *Journal of Economic Psychology*, 20(3), 251–284.

Falk, A. & Kosfeld, M. (2006). The hiddencosts of control. *American Economic Review*, 96(5), 1611–1630.

Fehr, E. (2009). On the economics and biology of trust. *Journal of the European Economic Association*, 7(2-3), 235–266.

Fehr, E., Kirchler, E., Weichbold, A., & Gächter, S. (1998). When social norms overpower competition: Gift exchange in experimental labor markets. *Journal of Labor Economics*, 16(2), 324–351.

Fehr, E. & Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism - experimental evidence and new theories. In S.-C. Kolm & J. M. Ythier (Eds.), *Handbook of the Economics of Giving, Altruism and Reciprocity*, volume 1 of *Handbooks in Economics* (pp. 615–691). Amsterdam, The Netherlands: North-Holland.

Feinberg, R. & Synder, C. (2002). Collusion with secret price cuts: An experimental investigation. *Economics Bulletin*, 3(6), 1–11.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2), 171–178.

Fudenberg, D. & Maskin, E. (1986). The Folk Theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3), 533–554.

Gächter, S. & Falk, A. (2002). Reputation and reciprocity: Consequences for the labour relation. *Scandinavian Journal of Economics*, 104(1), 1–26.

Gintis, H. (2009). *The Bounds to Reason*. Princeton: Princeton University Press.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E., Eds. (2005). *Moral Sentiments and Material Interests. The Foundations of Cooperation in Economic Life*. Cambridge, Massachusetts, USA: MIT Press.

Greiner, B. (2004). *The Online Recruitment System ORSEE 2.0–A Guide for the Organization of Experiments in Economics*. Working Paper Series in Economics 10, University of Cologne, Cologne.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., & Gintis, H., Eds. (2004). *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Oxford, UK: Oxford University Press.

Holcomb, J. H. & Nelson, P. S. (1997). The role of monitoring in duopoly market outcomes. *Journal of Socio-Economics*, 26(1), 79–93.

Kandel, E. & Lazear, E. P. (1992). Peer pressure and partnerships. *Journal of Political Economy*, 100(4), 801–817.

Kanemoto, Y. & MacLeod, W. B. (1991). The theory of contracts and labor practices in Japan and the United States. *Managerial and Decision Economics*, 12(2), 159–170.

Kirchler, E., Fehr, E., & Evans, R. (1996). Social exchange in the labor market: Reciprocity and trust versus egoistic money maximization. *Journal of Economic Psychology*, 17(3), 313–341.

Kreps, D. M., Milgrom, P., Roberts, J., & Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, 27(2), 245–252.

Lazear, E. P. (1993). Labor economics and the psychology of organizations. *Journal of Economic Perspectives*, 5(2), 89–110.

Lewicki, R. J. & Bunker, B. B. (1996). Developing and maintaining trust in work relationships. In R. M. Kramer & T. R. Tyler (Eds.), *Trust in Organizations: Frontiers of Theory and Research* (pp. 114–139). Thousand Oaks, California, USA: Sage Publications.

Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. *Academy of Management Review*, 23(3), 438–458.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge, UK: Cambridge University Press.

Rubinstein, A. (1979). Equilibrium in supergames with the overtaking criterion. *Journal of Economic Theory*, 21(1), 1–9.

Rustagi, D., Engel, S., & Kosfeld, M. (2010). Conditional cooperation and costly monitoring explain success in forest commons management. *Science*, 330, 961–965.

Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277(5328), 918–924.

Seabright, P. (1993). Coping with asymmetries in the commons: Self-governing irrigation systems can work. *Journal of Economic Perspectives*, 7(4), 93–112.

Sell, J. & Wilson, R. K. (1991). Levels of information and contributions to public goods. *Social Forces*, 70(1), 107–124.

Shapiro, C. & Stiglitz, J. E. (1984). Equilibrium unemployment as a worker discipline device. *American Economic Review*, 74(3), 433–444.

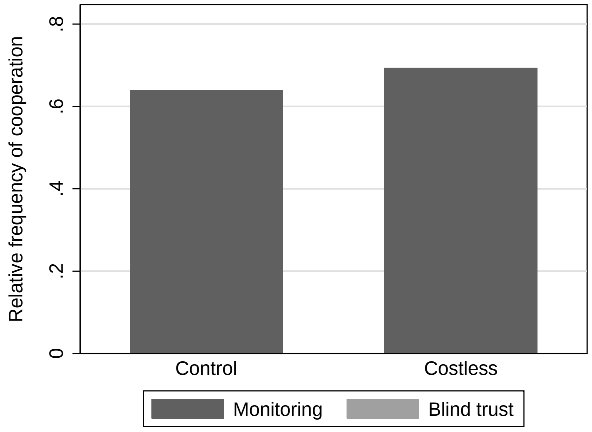
Slonim, R. & Guillen, P. (2010). Gender selection discrimination: Evidence from a trust game. *Journal of Economic Behavior & Organization*, 76(2), 385–405.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46(1), 35–57.

Vanberg, C. (2009). Why do people keep their promises? an experimental test of two explanations. *Econometrica*, 76(6), 1467–1480.

Varian, H. R. (1990). Monitoring agents with other agents. *Journal of Institutional and Theoretical Economics*, 146(1), 153–174.

Figure 7: Relative frequency of first mover cooperation pooled over time in the baseline condition and the costless monitoring condition.



A Dynamics in the baseline condition

In the baseline condition, first movers cooperated in 63.9 percent (138 out of 216) of the time. Second movers reciprocated in 82.6 percent (114 out of 138) of the time. The average (per period) joint payoff was 46.0 tokens, 21.8 for first movers and 24.2 for second movers. A comparison of those figures to the costless monitoring condition is depicted in figures 7, 8 and 9. None of the differences are statistically significant (Mann-Whitney, $p > .301$). Furthermore, the dynamic patterns (see figure 10) are quite similar. As mentioned earlier, not a single one among the 233 instances of cooperation in the costless monitoring condition was blind, such that first movers were *de facto* perfectly informed about the entire history of the game at any time, just as in the baseline condition.

B Individual-level dynamics

In this section we underpin the above aggregate results by briefly taking a look a individual dynamics, and show *en passant* that there is some individual heterogeneity hiding behind the averages. Figure 11 depicts the individual first mover dynamics for the costless monitoring condition. The bars indicate whether the player cooperated in a given period, where a bar is shaded in dark gray if accompanied by monitoring and light gray if blind, whereas the latter apparently never occurred in the costless monitoring condition. The markers at the top and the bottom of the bars represent the actual second mover's responses, where a marker at the top means cooperation and a marker at the bottom means defection. Finally, the black lines depict the first movers' beliefs about their coplayer's response.

It is evident that dynamics in individual matches differ. Particularly interesting are the individual belief patterns. Almost half of the first movers start with a rather

Figure 8: Relative frequency of reciprocation pooled over time in the baseline condition and the costless monitoring condition.

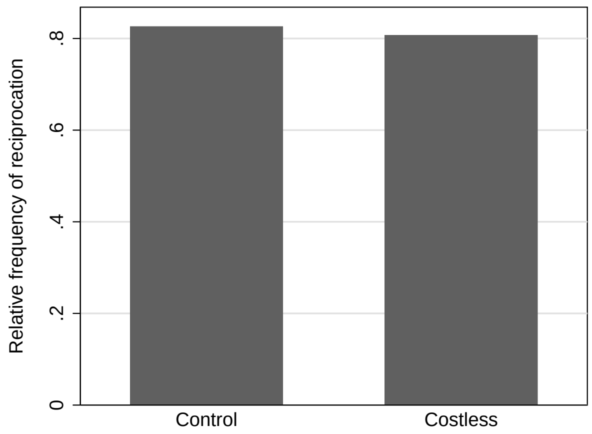


Figure 9: Efficiency: Average payoff per subject and round in the baseline condition and the costless monitoring condition.

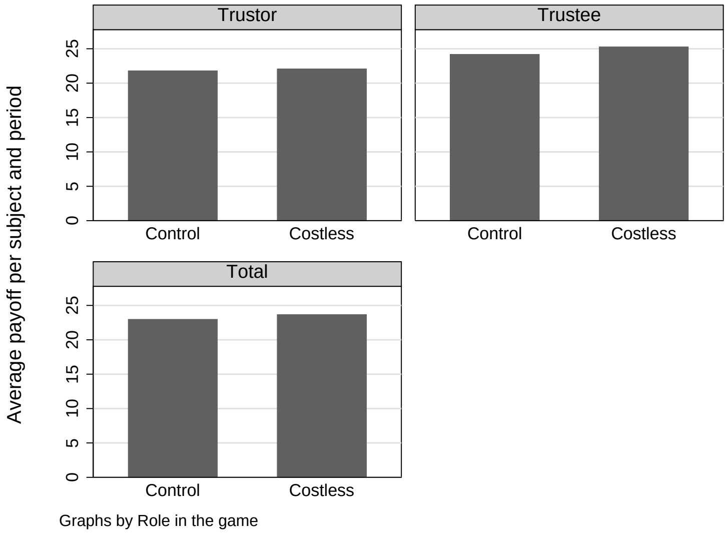


Figure 10: Aggregate dynamics of first mover cooperation and beliefs about reciprocation. The left column depicts the baseline condition, the right column the costless monitoring condition.

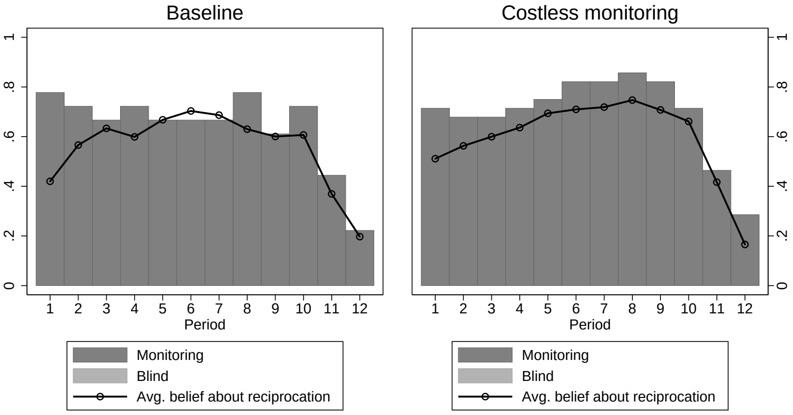
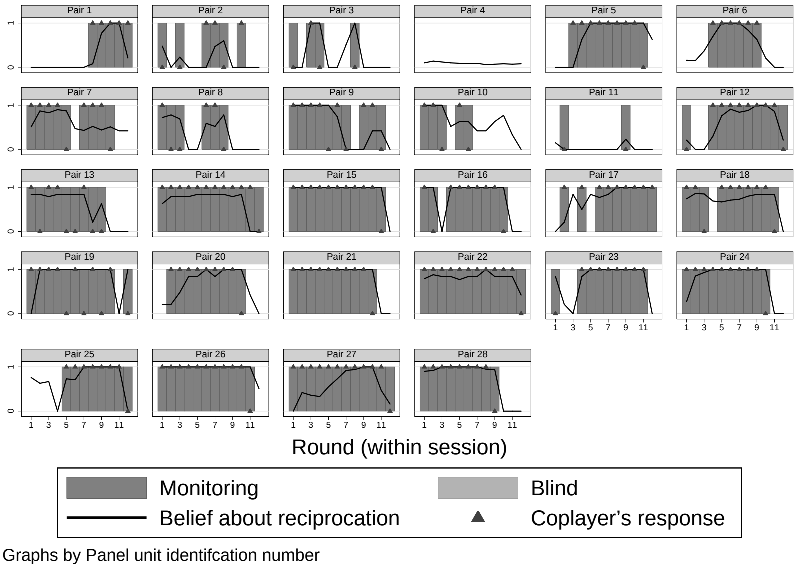
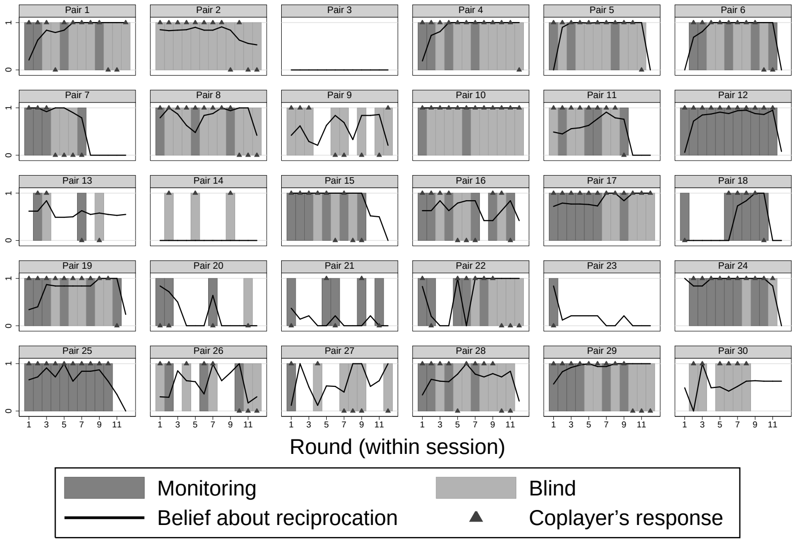


Figure 11: Individual first mover dynamics in the costless monitoring condition.



Graphs by Panel unit identifcation number

Figure 12: Individual first mover dynamics in the costly monitoring condition.



Graphs by Panel unit identifcation number

pessimistic prior regarding reciprocation (see in particular pairs 1–6, 11, 12, 17, 19, 20, 24, and 27). For them, it takes to go for some risk in order to learn. Only one first mover refused to do so (pair 4), and hence forewent all feasible gains from cooperation, the rest tested the coplayer at least once over the course of interaction. Apparently, those whose cooperation is not exploited swiftly become more confident. The pessimistic testers and initially rather optimistic first movers (see pairs 7–10, 13–16, 18, 21–23, 25–26, and 28) who were disappointed usually became more, or remained, pessimistic and sometimes punished detected cheats (all cheats were detected) in non-terminal periods with at least one period of non-cooperation. Note that almost all first movers anticipated the coplayer's strategic incentive to defect in the final period. In sum, this individual investigation reveals that first mover's behavior is quite consistent with their beliefs about reciprocation even at the individual level, and that their beliefs are quite responsive to monitored second mover behavior.

The same general conclusion can be drawn from investigating individual patterns in the costly monitoring condition. However, there is apparently an important difference due to frequent blind trust. Clearly, first movers cannot learn anything without monitoring. Thus, the majority of first movers start the match with monitoring. If second movers reciprocate one or a few times, then they often start going for the risk of trusting blindly, at times performing some random inspections in turn and responding with defection if the inspection revealed a cheat (see pairs 1, 4–8, 11, 12, 15, 19, 22 and 28–29 for those patterns). Thus, alomost half of the first movers behaved consistent to the strategy outlined above, namely shifting from initial testing to blind trust over time as long as no cheating is detected.

But there is apparently some individual heterogeneity. There is one first mover (pair 3) who did not cooperate a single time. One first mover (pair 12) started very pessimistically, tested her or his coplayer, became very optimistic over time as the latter turned out to be cooperative, but nevertheless never trusted blindly (spending 55 tokens on monitoring alone). A similar pattern resulted in pair 25. Another first mover (pair 2), starting with a very optimistic prior, trusted blindly for all twelve periods, despite of foreseeing the second mover's strategic incentive to cheat towards the terminal period. Thus, there clearly appears to be some individual heterogeneity in the propensity to trust blindly that is not accounted for by beliefs alone.

C Supplementary evidence from post-experimental debriefings

Blind trust in our experiment is positively correlated with an experimentally validated survey measure of individual risk preference ($\tau = .219$, $p = .000$). The item contains the question «Are you generally willing to take risks, or do you try to avoid risks?», and respondents answer the question on a 11-point Likert scale ranging from 0 (very risk averse) to 10 (very risk seeking). The item is used in the German Socio-Economic Panel (SOEP) and has been shown to be good predictor of behavior in experiments with decisions under risk (Dohmen et al., 2011). Blind trust is negatively correlated with a measure of negatively reciprocal inclination ($\tau = -.149$, $p = .001$), that has been argued to be a good proxy for betrayal aversion (Fehr, 2009). The items have also been implemented in the SOEP and read «If I suffer a serious wrong, I will take revenge as soon as possible, no matter what the cost» and «If somebody offends me, I will offend him/her back», and respondents can answer on a 7-point Likert scale ranging from 1 («does not apply to me at all») and 7 («applies to me perfectly»). I take the sum of both responses a measure of negatively reciprocal inclination. This suggests also a possible explanation for the existence of blind trust in the first period of the costly monitoring condition. Those four subjects who trusted blindly in the first period are on average more risk tolerant (risk tolerance item score 6.00 vs. 4.83) and less betrayal averse (negative reciprocity item score 4.50 vs. 7.17) than all other subjects in the sample; for them saving five tokens of information fee may already be enough compensation for bearing the risk of being exploited. Those differences are of course not significant since there are only four observations in one group, but the preference-based determinants are close to: a Mann-Whitney test on the difference in betrayal aversion is marginally significant ($p = .076$), a test on the difference in risk aversion yields a $p = .307$.

D Additional experiment with a two-period horizon