Trust, but verify? When trustworthiness is observable only through (costly) monitoring

Timo Goeschl* Johannes Jarke†

November 4, 2014

Abstract

For theoretical and empirical reasons, trust is expected to be lower in economic interactions in which trustors can observe trustworthiness only through (costly) monitoring. We examine this conjecture by investigating the impact of a (costly) monitoring environment on trust using data from 152 subjects participating in a binary version of the finite-horizon modified trust game. The three treatment conditions vary observability and the cost of monitoring. We find that trustors do not trust less, even when trustworthiness can only be observed deliberately and at a cost. When monitoring is costly, the same level of trust is supported by a significantly reduced amount of information on trustworthiness, acquired by trustors mainly in early stages of the repeat interaction. As a result, the efficiency of economic interactions is not lower when trustworthiness is costly to observe. (JEL C92, C72, D03, D80)

1 Introduction

Trust has been acknowledged as an important prerequisite for realizing the gains from cooperating in the many circumstances in which contractability is limited (Arrow, 1972; Greif, 1993; Zak & Knack, 2001). In such circumstances, contracts will be incomplete and when engaging in transactions that expose them to potential exploitation, parties therefore need to have confidence that the other party will not behave opportunistically. Given its significant role in supporting economic transactions (Fukuyama, 1995; Knack & Keefer, 1997), understanding the presence, nature, and scale of trust as well as its determinants has attracted considerable attention over the years. Considerable progress has come from studies of trust under controlled conditions, such as those based on the investment («trust») game in experimental economics (Camerer & Weigelt, 1988; Fehr et al., 1993; Berg et al., 1995).

* 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.
The theoretical literature highlighted early on that among the determinants of trust, both repetition and informational richness stand out as important structural features of the environment in which incompletely contractible interactions take place (Trivers, 1971; Rubinstein, 1979; Kreps et al., 1982; Fudenberg & Maskin, 1986). If agents are sufficiently patient, the «shadow of the future» implicit in repeat interactions is predicted to support trust through a disciplining effect on opportunistic behavior. Likewise, informational richness is predicted to support trust by ensuring that non-opportunistic behavior is observable such that agents are able to condition their actions on information about the past actions of their counterparts, hence justifying investment in a reputation for good behavior.

The core predictions of the theoretical literature on repetition and informational richness have been confirmed by considerable experimental evidence over the last ten years. A series of papers reporting on laboratory experiments with immediate and complete feedback have shown that subjects can accumulate and maintain high levels of trust in the repeated investment game for prolonged periods (Camerer & Weigelt, 1988; Anderhub et al., 2002; Engle-Warnick & Slonim, 2004, 2006a,b; Cochard et al., 2004). These highly encouraging findings have more recently given rise to the question of how robust the observed trust dynamics are to variations in the interaction and information structure. The question of robustness is a natural one given that in most of the initial experiments, players typically remained in fixed matches and fixed roles across rounds, and trustors learned about the history of trustee’s actions automatically, without error, at no cost, and without delay. Many real-life economic interactions take place in settings whose structure is less favorable: Agents, for example, often do not know whether and how often an interaction - in particular in a given constellation - will be repeated. Likewise, knowledge about co-player’s history of actions is often limited. Some of these complexities are reflected in the experimental design of several recent papers (see section 2 for a review) in which structural features of the investment game are exogenously varied. Jointly, their results have demonstrated how surprisingly effective sufficiently rich information structures can be in maintaining high levels of trust even when the repeat interaction takes place not between stable pairs of partners, but between successively different pairs of agents as long as new co-players can observe enough of the history of play. In other words, maintenance of trust imposes minimal requirements on the structure of repeat interactions as long as the information environment is sufficiently rich.

1The general lesson that repetition supports cooperation has also been found in a variety of other specific stage games, such as the Prisoners’ Dilemma game (e.g. Andreoni & Miller, 1993; Cooper et al., 1996; Dal Bó, 2005; Dal Bó & Frêchette, 2011; Duffy & Ochs, 2009), the public good game (Andreoni & Croson, 2008, provide an overview), and the gift exchange game (Kirchler et al., 1996; Fehr et al., 1998; Falk et al., 1999; Gächter & Falk, 2002).

2This result is mirrored by similar findings in a PD context. Camera and Casari (2009) show how cooperation can be maintained in an indefinitely repeated prisoner’s dilemma following a strict stranger protocol as long as identities and histories of co-players are public information. Huck, Lünser, and Tyran (2012) show that when trustees can be identified and trustors are allowed to choose their trustees endogenously, the resultant competition between trustees also leads to high levels of trust.
Many economic interactions take place in information environments that are not naturally rich. One familiar deficiency is that the trustee’s response to being trusted is not automatically and freely observable by the trustor, even when trustor and trustee interact repeatedly in stable pairs. Typical reasons are physical distance between the trustor and the trustee (Greif, 1993) or simply pre-existing information asymmetries such as expert knowledge in the health or car repair markets (Wolinsky, 1995; Emons, 1997; Dulleck & Kerschbamer, 2006). In these circumstances, trustors can either stay uninformed or need to make a deliberate decision to remedy their lack of information by monitoring the co-players’ action. The need for as well as the feasibility of monitoring of co-players is a routine feature of the economic world and is particularly salient when monitoring is not casual, but requires costly effort (Ostrom, 1990; Weissing & Ostrom, 1991; Ostrom & Gardner, 1993; Rustagi et al., 2010). The question at the heart of this paper is how such limited observability of trustworthiness and the resulting need for (costly) monitoring impacts on trust in repeat interactions.

Is there a lower level of trust in interactions in which trustors can only ascertain the trustworthiness of their co-player through (costly) monitoring? Theoretical and behavioral considerations (see Gintis, 2009) as well as existing experimental evidence on the impact of exogenously imposed information imperfections in cognate game forms (Sell & Wilson, 1991; Holcomb & Nelson, 1997; Cason & Khan, 1999; Grechenig et al., 2010; Ambrus & Greiner, 2012) give reason to expect lower trust when repeat interactions have to take place in a (costly) monitoring environment. However, there is to our knowledge as yet no experimental evidence that directly addresses this question. Such evidence would not only advance our understanding of the impact of information acquisition costs on trust, but also our understanding of whether and how trustor choose to learn to trust their trustee.

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; 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). As one reader of this paper has pointed out, the basic logic is also enshrined in the Russian proverb «дайте руку, но проверьте» («trust, but verify»).

An early example is Varian (1990, p. 153) who 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 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.

This, in turn, has the potential to help identify plausible drivers of agent’s trust behavior in the investment game. Whether behavior in the investment game captures trust in an adequate fashion is a question of concern both to behavioral economists (Cox, 2004; Karlan, 2005; Ben-Ner & Halldorsson, 2010; McEvily et al., 2012) and social scientists more generally (e.g. Rousseau et al., 1998; Elster, 2007)). As identified in the investment game, trust has been shown to correlate with risk seeking behavior (Karlan, 2005; Schecter, 2006, 2007; Houser et al., 2010) and altruism (Cox, 2004; Ashraf et al., 2006). For Elster (2007), trust is «the result of two successive decisions: to engage in the interaction and to abstain from monitoring the interaction partner» (p. 345, emphasis added). The standard trust game does not include the second step. The monitoring game presented here, however, allows exactly for this succession.
We investigate the impact of a (costly) monitoring environment on trust using data from 152 subjects participating in a binary version of the finite-horizon modified trust game that we term - for short - the monitoring game. In the original trust game (Camerer & Weigelt, 1988; Berg et al., 1995), a first mover chooses between an outside option and a trust move that renders her or him vulnerable to exploitation by a second mover. The second mover may either reward the first mover’s trust at a personal cost or exploit the opportunity. When repeated, the first mover knows at the outset of the next round whether his trust was rewarded or not and can decide accordingly. The essential design variation in the monitoring game is that trusting first movers do not automatically learn their payoff at the end of a round. Instead, the first mover chooses whether to monitor the second mover’s action in that round. If he makes an active monitoring decision, he observes the second mover’s action. In other words, the first mover can choose to «trust, but verify». If the first mover chooses not to monitor, the second mover’s action in that round remains hidden forever. While the second mover is aware that the first mover has the monitoring option, he does not learn whether he is actually being monitored.\(^6\) We implement the stage game in a twelve-round repetition.

To study the impact of limited observability and a (costly) monitoring option on trust, we used three treatments. The «Baseline» treatment implemented a standard finite horizon trust game with perfect observability, that is, trustors were automatically informed about the trustee’s actions without incurring a cost. This replication of previous research (e.g. Anderhub et al., 2002; Engle-Warnick & Slonim, 2006b,a; Slonim & Guillen, 2010) returned the typical pattern of frequent trust (about two in three cases) until close to the terminal period, and a sharp decline in the final two periods. In the «Costless Monitoring» treatment, trustors were not automatically informed about the trustee’s action, but could remedy this limited observability at no cost by taking an active decision to monitor the trustee. While from a theoretical point of view, «Costless Monitoring» is indistinguishable from perfect observability, the active decision to monitor involves - from a behavioral perspective--an element of distrust towards the trustee (McEvily et al., 2012).\(^7\) Like the Costless Monitoring condition, the «Costly Monitoring» treatment required a deliberate monitoring decision from the first mover in order to observe the trustee’s action, but now monitoring involved a cost equal to one third of the per-period gain from rewarded trust.

Comparing the three treatments of the repeat trust game, we arrive at three key results. The first is that limited observability does not adversely affect trust if a (costly) monitoring option is present. Compared to the Baseline treatment of perfect observability that is observable by the second mover is an important design variation of the blind trust game that we do not explore here. Observability of monitoring mitigates the information imperfections that are the focus of this paper since it provides a channel for the first mover to signal trust and hence induce trustworthiness in the second mover strategically.

\(^6\) There is little guidance on the behavioral implications of a monitoring option. Using monitoring could be emotionally costly to trustors if they have an innate preference for being in a trust relationship (Elster, 2007). Relative to the Baseline treatment, the availability of the option could also prime trustors towards distrust and, hence, reduce trust (Burnham et al., 2000).
ability, trust in both of the Monitoring treatments (Costless and Costly) was not lower. The second result is that the efficiency of economic interactions does not necessarily suffer when trustworthiness can only be ascertained through costly monitoring. Joint payoffs in the Costly Monitoring treatment were as high as those in the Costless Monitoring and the Baseline condition, even after taking monitoring expenses into account. Our third result is that trust can be supported by a limited amount of information on trustworthiness. Similar levels of trust across our three treatments were supported by different information structures, in particular when monitoring was costly: In the Costless Monitoring condition, trustors chose to monitor every single one of the 233 actions by trustees (100 percent). In the Costly Monitoring condition, only 119 out of 254 actions by trustees (47 percent) were monitored. In most rounds, therefore, trustors chose not to monitor when this was costly, with a consistent dynamic pattern of frequent monitoring in early rounds and sporadic inspections in later rounds. Taken together, these results provide first experimental evidence that the maintenance of trust not only imposes minimal requirements on the structure of repeat interactions, but also smaller requirements on the observability of trustworthiness than expected. In addition, the dynamics of information acquisition provide a window onto how trustors choose to learn about the co-player’s trustworthiness, with interesting parallels to neuroeconomic evidence on how trustors build mental models of their trustee as their reputation develops (King-Casas et al., 2005).

In the remainder we proceed as follows. After a review of the related experimental literature in section 2, we describe the design, procedures, and implementation in section 3. The results are presented in section 4. We summarize and conclude in section 5.

2 Related literature

The present study about the level and dynamics of trust when observing co-player’s actions requires (costly) monitoring lies at the intersection of three experimental literatures. The first is a body of research that examines the impact of exogenously manipulating or restricting information within the standard repeated trust game. Burnham et al. (2000) examine the role of priming trustors by introducing the co-player as a «friend» or «foe» and find that despite the priming, learning dynamics lead to behavioral convergence in a repeated setting. Anderhub et al. (2002) study a finite-horizon repeated trust game in which first movers are imperfectly informed about the type of co-player (completely trustworthy or opportunistic) they are matched with. They find that the aggregate dynamics of this modified trust game match predictions of the reputation formation hypothesis in repeated games. Our design also uses a modified finite horizon trust game, but differs from both papers in that the information structure within the fixed interaction evolves endogenously.

The second literature examines the level and evolution of trust when agents take repeat decisions, but not necessarily in fixed matchings. Against this background,

---

8 Recall that in the Baseline treatment, trustors were automatically informed about trustees’ actions.
different designs examine the impact of exogenous variations in the information structure on trust. Bohnet & Huck (2004) compare fixed matching and random re-matching with and without providing to trustors the history of the trustee’s actions in previous interactions with third parites. They find that a repeat interaction environment with stranger matching and information about their co-players’ actions is essentially as efficient as a fixed matching environment. Bracht & Feltovich (2009) add a pre-stage to the standard trust game in which the first mover in a mutually one-shot interaction receives either information about the co-player’s last decision with another player or cheap talk from the co-player or both. They find that the observability of the recent history of actions leads to high levels of cooperation. Charness et al. (2011) examine the role of indirect reciprocity in a repeated trust game and compare the effect of providing information on the co-player’s history of actions in different roles, as a first and a second mover. They find strong evidence that information on past behavior as a trustor is as effective for reputation building as past behavior as a trustee, affirming the role of indirect reciprocity as a mechanism supporting trust. The general conclusion from this literature is that information-rich environments are highly conducive towards trust and trustworthiness.9 Huck et al. (2012) confirm this basic finding in a repeated binary trust game, but also find that allowing for endogenous matching (competition) can perform even better even when the information environment is exogenously restricted to trustee’s identity rather than their history of play. Our paper differs in two important aspects from this literature: Our subjects interact repeatedly in stable pairs across all rounds (no rematching) and, more importantly, the specific information structures are endogenously determined by the trustor rather than being exogenously imposed.10 These differences reflect our specific interest in how trust responds to poor, but remediable information environments.

The third literature to which this research relates examines the role of costly monitoring in principal-agent relationships. Nagin et al. (2002) conduct a field experiment on how variations in the probability that their sales figures will be audited impacts on the trustworthiness of online call center employees. They find that rational cheating is the dominant behavior, but also that there is considerable heterogeneity: A significant share of employees do not decrease their trustworthiness in response to a

---

9Interestingly, in contrast to the cooperation literature (Holcomb & Nelson, 1997; Sainty, 1999; Aoyagi & Fréchette, 2009; Bornstein & Weisel, 2010; Grechenig et al., 2010; Ambrus & Greiner, 2012), we are not aware of experimental results in the repeated trust game in which co-player’s actions are observed with a certain error probability.

10Within the trust game paradigm, there is also an interesting parallel to a recent paper by McEvily et al. (2012) whose experimental design introduces a costly option of insuring against vulnerability in the investment game. In a trust measurement experiment, they offer second movers in series of five one-shot exchange games with strangers the possibility to change the structure of the interaction through a costly option to avoid being exposed to the first mover’s decision. They find that subjects apply the option selectively based on expected trust. Some readers of the present paper have commented that the decision to monitor can also be interpreted as an option that reduces vulnerability relative to a situation in which the trustor does not learn about the trustee’s actions. However, if vulnerability is the main concern of trustors, simply choosing not to trust provides full insurance against vulnerability in our design.
decrease in the monitoring probability. Dickinson & Villeval (2008) study the choice of a costly monitoring intensity by a first-moving principal and the effort response by second-moving agents with two treatments, a stranger or partner matching protocol and a payoff function for the principal that was either increasing in the agent’s output or fixed. They find that effort is increasing in monitoring intensity, but that this disciplining effect on opportunism is tempered by a crowding out effect in the partner matching protocol. We differ from this line of research in that the first mover decides on a round-by-round basis whether to engage in monitoring, and that monitoring is costly. As a result, our design disentangles the effect of we introduce a cost. Trustor can determine WHEN he wants to seek information. Monitoring BEHAVIOR across rounds! THIs is because our interest is different: Impact of the fact that monitoring is costly on trust, including the trustor.

3 The Experiment

3.1 Experimental game and design

The stage game of the experiment is the well-known (binary) trust game (see e.g. Camerer & Weigelt, 1988; Anderhub et al., 2002; Engle-Warnick & Slonim, 2006b,a; Slonim & Guillen, 2010). The first mover chooses between investing (option «pink» in the instructions) and an outside option («yellow»). If the outside option is chosen, both players get 15 tokens and the period ends. If the first mover chooses to invest, the period continues with the second mover’s choice between splitting (option «brown») or keeping (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. This payoff structure is standard and intentionally asymmetric. As a ‘Baseline’ treatment that directly replicates previous research, we implemented a standard 12-fold repetition of the stage game with perfect information: Each player was informed about the co-player’s action automatically, at no cost, without error, and without delay. This treatment constituted the benchmark of the experiment.

The two treatment conditions of the experiment introduced imperfect information into the trust game. In both treatments, a cooperating first mover was no longer automatically informed about the second mover’s action. Specifically, without knowing the second mover’s action, a first mover decided whether he wanted to monitor the second mover’s action or not. If so, the first mover was informed about whether their
co-player responded with «brown» or «blue», respectively, at the end of the round. Otherwise, (s)he received no information. Second movers were never informed about whether their co-player monitored them or not. In the first treatment, monitoring the second mover required a deliberate decision to do so, but involved no cost. In the second treatment, first movers had to incur a fee of five tokens in order to acquire this information. Except for these variation, both treatment conditions and the benchmark were exactly identical. We used a between-subjects design to assign treatments to participants.

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.

3.2 Subjects and procedures

Participants were recruited from the general undergraduate student population of the University of Heidelberg using the online recruitment system ORSEE (?). In total 152 subjects participated of which 52.6 percent were female and 85.5 percent German. 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

---

13 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 ??.
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.

3.3 Predictions

For finite horizon games, the theoretical prediction that trust arises in repeat interactions hinges on the assumption that players hold a belief that some coplayers are committed to reward trust 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 substantive amount of recent evidence on cooperative behavior (e.g. Henrich et al., 2004; Gintis et al., 2005; Fehr & Schmidt, 2006; Thöni et al., 2012). Acquiring a reputation for trustworthiness requires that trustors initiate the trust relationship and observe the trustee’s response. The cycle of trust-reward-observation-reputation-trust typically maintains trust in finite-horizon repeat trust games over extended periods for roughly two thirds of trustors before trust declines in the final two rounds (Anderhub et al., 2002; Engle-Warnick & Slonim, 2004, 2006a). We therefore expect to replicate this result in the «Baseline» treatment in which observability of trustworthiness is perfect.

In the presence of limited observability and costless monitoring, the only difference to the ’Baseline’ treatment consists in the trustor having to take an active decision to monitor in order to observe trustworthiness. In a laboratory environment,
this involves minimal effort. From a theoretical perspective, there is therefore no reason to expect a difference in behavior between the «Costless Monitoring» and the 'Baseline' treatment. Behaviorally, costless monitoring could be seen as an admission of distrust towards the trustee and an attempt to reduce vulnerability to possible exploitation (McEvily et al., 2012). Using monitoring could therefore be emotionally costly to trustors if they have an innate preference for being in a trusting relationship: Monitoring «might be incompatible with the agent’s emotional attitude toward the other person» (Elster, 2007, p. 346). Relative to the Baseline treatment, the availability of the option could also prime trustors towards distrust and, hence, reduce trust (Burnham et al., 2000). In light of the limited evidence on the presence and scale of these putative mechanism, however, we predict no measurable impact of costless monitoring on trust.

In the presence of limited observability and costly monitoring, there are several reasons for expecting that trust will be lower in the monitoring game. One is that every instance of monitoring reduces the gains from trusting: The maximum gain from trust in the current round is 15 tokens while the cost of monitoring is 5 tokens. The second is that trustors who trust and then save on monitoring costs are easily exploited, possibly over multiple periods, without being able to condition future trust on observed trustworthiness. Thirdly, trustees that correctly anticipate non-monitoring in the current round have a higher probability of getting away with cheating, which reduces the relative gains from reputation building. Fourth, repeated reciprocation by the trustee does not accumulate into a reputation unless the first mover incurs the monitoring cost. If the trustee anticipates less observation of trustworthiness due to monitoring costs, the rewards from investing in trustworthiness are reduced relative to perfect observability. We therefore predict that both first and second movers will respond to the costly monitoring environment with lower trust and less trustworthiness.

4 Results

We proceed with the presentation of results as follows. In a preliminary step, we establish that the «Baseline» condition can serve as a benchmark. We then show as result 1 that the «Baseline» treatment and the «Costless Monitoring» treatment generate statistically indistinguishable behavior. The subsequent results 2 to 4 are then based on a comparison of the «Costless Monitoring» treatment and the «Costly Monitoring» treatment. These results answer to the prediction that trust is lower when second movers’ actions need to be monitored at a cost in order to be observed by the first mover. We conclude the results section by examining the endogenous information structures that underpin cooperation in a costly monitoring environment.

4.1 Replication check

We first show that the «Baseline» treatment, i.e. a trust game with perfect observability, passes a basic replication check. In the baseline, first movers trusted 63.9
percent (138 out of 216) of the time. Second movers reciprocated 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. 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.

Both in terms of levels and in terms of dynamics, the baseline closely corresponds to previously published evidence on first and second mover behavior in the finite horizon trust game (Anderhub et al., 2002; Engle-Warnick & Slonim, 2004, 2006a). On this basis, we progress to the key results of the experiment in which we investigate the treatment effects relative to a baseline that is in line with the literature.

4.2 Treatment effects

A visual comparison of the «Baseline» treatment and the «Costless Monitoring» treatment is provided in figure 1. As can be seen, the aggregate results on trust and trustworthiness match closely, and none of the differences are statistically significant (Mann-Whitney, \( p > .301 \)). Furthermore, the dynamic patterns are quite similar. As mentioned earlier, not a single one among the 233 instances of trust in the costless monitoring condition went unmonitored, such that trustors were de facto perfectly informed about the entire history of the game at any time, just as in the Baseline condition. We conclude on the basis of this comparison that the introduction of an active monitoring step alone does not impact on a behavioral trust measure, a result in line with the theoretical prediction. The possible behavioral mechanisms must therefore be quantitatively small, if at all present. Due to the greater sample size and similarity in design, we will frequently use the «Costless Monitoring» condition as a statistically more demanding basis for establishing treatment effects associated with the «Costly Monitoring» condition.

Figure 1a illustrates the three key findings of this paper. The first of these findings is that trust is not lower when trustworthiness can only be observed through (costly) monitoring option is present. In the «Baseline» treatment, first movers trusted 63.9 percent of the time (see 4.1) and in the «Costless Monitoring» condition, 69.4 percent of the time (233 out of 336 cases). In the «Costly Monitoring» condition, they trusted no less often (254 out of 360 cases, or 70.6 percent). The frequency of trust is not significantly different across conditions (Mann-Whitney, \( p = .530 \)), based on a cross-sectional analysis in which each observation is a unit-level average taken over all twelve periods.\(^{14}\) The evidence on the basis of aggregate measurements therefore leads us to conclude that despite a less favorable information environment, trust behavior in the investment game under limited, but remediable observability was not lower than under perfect observability.

The result on trust extends to efficiency and gives rise to our second key result: The efficiency of economic interactions does not necessarily suffer when trustwor-\(^{14}\)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 follows Vanberg (2009), and takes account of this fact.
Figure 1: Behavior and payoffs by treatment condition.

(a) Pooled over all rounds

(b) Dynamics in the baseline condition

(c) Dynamics in the costless monitoring condition

(d) Dynamics in the costly monitoring condition
thiness can only be ascertained through costly monitoring. Average joint payoffs, including monitoring costs, were not significantly lower in the «Costly Monitoring» treatment than in the «Costless Monitoring» condition. Again using the cross-section of unit-level averages taken over all twelve periods, the average joint payoffs were not significantly different across conditions (Mann-Whitney rank sum test, $p > .565$). This result implies that trust cannot have been maintained through permanent costly monitoring. Indeed, trusting first movers monitored less than half of the time (119 out of 254 cases, or 46.9 percent) and trusted without monitoring in more than half of the interactions. This monitoring frequency differs significantly from the zero monitoring cost environment (Mann-Whitney, $p=.000$). In other words, while quantitatively similar, trust behavior in the «Costly Monitoring» treatment differed qualitatively because it relied on a very different endogenous information structure than in the «Baseline» trust game with perfect observability and the «Costless Monitoring» condition in which first movers always monitored. Our third key result is therefore that trustors do not trust less when trustworthiness can only be observed at a cost, but instead acquire only a fraction of the available information on trustworthiness. Note that trustees broadly anticipated that trustors would monitor judiciously when observing trustworthiness is costly: In the «Costly Monitoring» treatment, trustees expected that 63 percent of their actions will be monitored (92 percent in the zero monitoring cost environment).

The treatment effects results raise a number of interesting questions about the information structures that support trust in the «Costly Monitoring» condition. We analyze their dynamics in the following section.

4.3 Dynamics of Monitoring and Trust

We reported above that in 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: Figure 1c 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.

Figure 1d illustrates the average first mover’s actions in the «Costly Monitoring» condition. These actions exhibit a clear dynamic pattern: a shift from monitoring towards non-monitoring over time. In the first period, the vast majority of first movers incurred the monitoring cost in order to observe the trustworthiness of their co-player. Over time, monitoring becomes successively less frequent. The left-hand panel of figure 3 illustrates this shift more clearly, depicting the frequency of trust with and

---

15Recall 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).
**Figure 2:** Dynamics of beliefs by treatment

(a) Trustors’ belief in their trustee acting trustworthy
(b) Trustees’ belief in being monitored

**Figure 3:** Frequency if verification as a fraction of all instances of trust in the costly monitoring condition.
without verification, respectively, as a fraction of all instances of trust. Thus, first movers predominantly resort to monitoring at the beginning of the repeat interaction and then maintain trust without verification towards the end. This lends support to the control-towards-trust hypothesis (Lewicki & Bunker, 1996; Lewicki et al., 1998). Note, however, that it requires a positive monitoring cost to reveal the intertemporal structure of the demand for information on trustworthiness; requiring a deliberate monitoring decision is not sufficient. The intertemporal pattern of monitoring seen in figure 3 also has an interesting parallel with the neuroeconomic literature that studies how neural responses to each other’s decisions evolve in trustor’s and trustee’s dorsal striatum over the course of the repeated interaction in a trust game (e.g. King-Casas et al., 2005; Delgado et al., 2005; Krueger et al., 2007; Fareri et al., 2012). King-Casas et al. (2005), in particular, find evidence that trustors engage in «model building of the partner», a process that they find to be complete after five to six rounds of interaction. This pattern broadly coincides with the number of rounds over which trustors monitor the trustor with a high degree of consistency.

The average trustee anticipates the monitoring dynamics with reasonable accuracy. The right-hand panel of Figure 3 shows the average belief of second movers of being monitored over time. It is constantly high in the «Costless Monitoring» condition, but broadly 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 therefore without impact on the reputation. This incentive leaves traces in the distribution of payoffs between trustors and trustees. Figure 4 illustrates the dynamics of realized payoffs. As evident from the left-hand column, the distribution of payoffs between first and second movers was 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-hand 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). This underlines the risk to trustors of trusting without verification in the costly monitoring game.

An examination of the individual-level dynamics of trust substantiates the pres-

---

16 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 prediction: 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 $\tau_b = 0.206, p < .001$).
**Figure 4:** Average payoffs over time.
ence of significant heterogeneities behind the average results and add some additional subtlety, in particular with respect to the beliefs underpinning individual behavior (see appendix A). Individual-level data in the «Costly Monitoring» condition show that (i) most trustors consistently gathered information in early rounds of the repeat interaction even when their beliefs about the trustee’s trustworthiness were initially pessimistic; (ii) monitoring was reduced only if the trustee was observed to be trustworthy early in the interaction; and (iii) among a significant subset of trustors, the trust and monitoring dynamics exhibit large heterogeneity that is not fully explained by beliefs. Examining the elicited beliefs more closely, we find that trust without monitoring is positively related to the trustor’s confidence in reciprocation.

Using post-experimental survey data on individual preferences, we also find that the incidence of trust without monitoring is negatively related to the degree of risk aversion and betrayal aversion (see appendix B).

5 Conclusion

Recent experimental evidence based on the investment game has revealed how surprisingly robust trust is to structural imperfections in the environment in which the economic interactions take place. Adequately rich information environments are sufficient to ensure that trust survives in repeat interactions whose coupling is as loose as random rematching. This raises the question of whether trust is perhaps much more vulnerable to imperfections that compromise the richness of the information structure. One common and much discussed imperfection is the one that we studied in this paper: trustors frequently need to engage in (costly) monitoring in order to ascertain whether the trustee rewarded their trust or not.

Compared to a setting with perfect observability, a setting with limited, but remediable observability of trustworthiness should be give rise to less efficient interactions. The reason is that the effort and cost of monitoring negatively affects both the relative gains from trust versus opting out for the trustor and the relative gains from reputation-building trustworthiness versus immediate exploitation for the trustee. This means that when the limited observability of trustworthiness can only be remedied at a cost, trust should either be lower or the aggregate costs of adequately monitoring trustworthiness should reduce the joint surplus. Despite these factors, neither the level of trust nor efficiency was negatively affected in the repeated interactions in which the observability of trustworthiness was compromised. The explanation is that when trustees are mostly trustworthy, trustors require relatively little information on trustee’s trustworthiness in order to build up and maintain trust.

17This is also illustrated in figure ???. Correlation between the first movers’ belief about reciprocation and their own trust is strongly positive and significant both in the ‘Costless Monitoring’ condition (Kendall’s τ = .627, p = .000) and the ‘Costly Monitoring’ condition (τ = .558, p = .000). However, separating trust with monitoring and trust without, the latter turns out to be 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 non-monitoring trustors, .736 for monitoring trustors, and .211 for defectors.
Trustors ostensibly believe in behavioral ‘types’, i.e. some degree of behavioral persistence in trustee’s actions (rather than sophisticated Bayesian strategizing), and rely on later spot checks to confirm that their identification strategy in early rounds was correct. These results therefore point to the possibility that the maintenance of trust not only requires minimal structure in the repeat interactions, but also considerably less observability of trustworthiness than previously thought. We commented on the parallels to neuroeconomic studies on the evolution of trust above.

We see a number of promising avenues of further research on the basis of variations of the monitoring game presented here. Real-world settings differ in monitoring costs, and even while we consider the costs in the present experiment (at one-third of the per-round gain from rewarded trust) as large and meaningful, there is no natural upper limit to monitoring costs. This raises both the question of the shape of trustors’ demand curve for information about trustworthiness and the question of the existence and level of a possible threshold of monitoring costs for trust to break down irreparably. The monitoring game also excluded deliberately the question of the observability of trustor’s monitoring actions by the trustee. While the experimental literature on monitoring reviewed in section 2 goes some way towards understanding the higher-level game that can arise through trustors sending signals of high trust by observably not monitoring, the stochastic nature of the monitoring rate in a multi-agent setting explored there fails to exhaust the full ramifications of observable non-monitoring in repeat interactions, in particular in fixed matches (Elster, 2007).

An interesting avenue for further research is a detailed investigation of the strategies the players play in the monitoring game and their equilibrium properties. 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 monitoring game, second movers can induce the first movers not only to trust, 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 trust again once cheated), while in the monitoring game there is the possibility of cheating over multiple periods once the first mover starts to trust without monitoring. 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 without verification later on. But this strategy can only be part of some kind of mixed strategy equilibrium, because trust without verification with certainty is not a best response to it. The sporadic inspections many trustors perform in later periods is a hint in this direction.
References


Dickinson, D. & Villeval, M.-C. (2008). Does monitoring decrease work effort? the complementarity


Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel...


## A Individual-level dynamics

An examination of the individual dynamics of trust reveals interesting heterogeneities behind the averages.

Figure 5 depicts the individual first mover dynamics for the ‘Costless Monitoring’ condition. The bars indicate whether the first mover trusted in a given period, where a bar is shaded in dark gray if accompanied by monitoring and light gray if not. The latter never occurred in the costless monitoring condition. The markers at the top and the bottom of the bars represent the trustee’s responses, where a marker at the top means trustworthy behavior 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 pessimistic prior regarding reciprocation (see in particular pairs 1–6, 11, 12, 17, 19, 20, 24, and 27). For these subject, learning requires significant risk-taking. Only one trustor refused to do so (pair 4), and hence forewent all feasible gains from trust, the rest tested the coplayer at least once over the course of interaction. Subjects whose trust was not exploited appear to become rapidly more confident. Trustors who ‘gave it a try’ (see pairs 7–10, 13–16, 18, 21–23, 25–26, and 28) and 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
**Figure 5:** Individual first mover dynamics in the costless monitoring condition.

![Graphs showing individual first mover dynamics in the costless monitoring condition.](image)

**Figure 6:** Individual first mover dynamics in the costly monitoring condition.

![Graphs showing individual first mover dynamics in the costly monitoring condition.](image)
of opting out. 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 broadly consistent with their beliefs about reciprocation even at the individual level, and that their beliefs respond to monitored second mover behavior in expected fashion.

In the ‘Costly Monitoring’ condition, the same general conclusion can be drawn, but with the important qualification that the same conclusion arises despite frequent non-monitoring. The majority of first movers started the match with monitoring. If second movers reciprocated once or twice, most trustors rapidly moved to trust without monitoring, moderated by occasional monitoring and responding with opting out if monitoring revealed a cheat (see pairs 1, 4–8, 11, 12, 15, 19, 22 and 28–29 for those patterns). Almost half of the first movers behaved in line with a heuristic strategy that can be summarized as shifting from trust with monitoring to trust with only sporadic monitoring as long as no cheating is detected.

Among the other half of trustors, we find considerable individual heterogeneity. There is one first mover (pair 3) who did not trust even once. One first mover (pair 12) started with very pessimistic beliefs, yet trusted the trustee’s trustworthiness, but despite becoming very optimistic over time in line with observed high trustworthiness, monitored in every round (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 without monitoring for all twelve periods, despite 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 and monitor that is not accounted for by beliefs alone.

B 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 $p = .307$. 