On the Economics of

TRANSPARENCY AND COOPERATION

JOHANNES S. JARKE

On the Economics of Transparency and Cooperation

$\star \star \star$

Dissertation
zur Erlangung des akademischen Grades *Doctor rerum politicarum* (Dr. rer. pol.)
eingereicht an der Fakultät für Wirtschafts- und Sozialwissenschaften an der
Universiät Heidelberg

vorgelegt von
Diplom-Volkswirt
Johannes Stephan Jarke
geboren in Viernheim

Gutachter:
Professor Timo Goeschl, Ph.D.
Professor Dr. Jean-Robert Tyran

Tag der Einreichung: 27. November 2012

Tag der Verteidigung: 5. Februar 2013

The picture on the title page is an image of «Suspicion» (oil on 40x30 inch canvas)
by Ann Karin Glass, and is reproduced with permission.

Für meine Eltern *Für Nelli*

Acknowledgments

I would like to express my gratitude to my mentor and coauthor of chapters 3 through 5, Timo Goeschl, for always giving me the support, advice and freedom I needed for the work embodied in this thesis. I am also deeply grateful for the hospitality of Jean-Robert Tyran and his team at the VCEE in Vienna where parts of this work has been composed. Both examiners provided thoughtful and valuable comments and asked the right questions to improve my research and my professional development.

Many other people have a footprint in this work and I am afraid I cannot name them all personally here. A number of participants of seminars, workshops, and conferences in Berlin, Chicago, Cologne, East Anglia, Exeter, Graz, Heidelberg, Kreuzlingen, Mannheim, Oslo, Paris, and Stanford have provided valuable comments. Ernst Fehr and Urs Fischbacher deserve credit for providing valuable material. More distantly, I want to name just a few people that sparked and nourished by interest in economics with outstanding teaching: Klaus Nielebock, Jürgen Eichberger, Oskar Gans, Jörg Oechssler, Lars Feld, again Timo Goeschl, Tri-Vi Dang, Inge Kaul, Melanie Schienle, Enno Mammen, Ernst-Ludwig von Thadden, Holger Bonin, Melanie Arntz, Catherine Eckel, and Rick Wilson (in more or less this sequence). I want also thank members of the 2008 freshmen class at the GESS in Mannheim, the 2010 class of the ESNIE in Cargése, and the 2011 class of the EITM Summer Institute in Mannheim. Carmen Riecherzhagen deserves special thanks for her kind mentoring at the German Development Institute in Bonn. Last but not least I want to thank my local colleagues, in particular Ole Jürgens, Daniel Römer, Johannes Diederich, Andrea Leuermann, Johannes Lohse, Daniel Heyen, and Sara Kettner.

Contents

Introduction

«A lack of transparency results in distrust and a deep sense of insecurity.»

THE DALAI LAMA

This thesis is about the relationship between transparency and cooperation. Here is how I define the terms for the present purposes. Cooperation is the performance of an act that is beneficial to others. Transparency is the availability of information about others' actions. Fixing a particular group of individuals, the basic question can be posed as follows: What does the propensity of those individuals to cooperate with one another has to do with the availability of information they have about each others' actions?

In the remainder of this introduction I argue (i) why this question is relevant, (ii) why the answer to this question is not obvious, (iii) that drawing on the established paradigm to study cooperation theoretically and experimentally is a useful approach, (iv) that the respective literature provides valuable insights, (v) but that one critical aspect has generally received little attention so far. Based on the latter point I am going to motivate and outline the original contributions contained in this thesis.

Why this question is relevant

Cooperation is fundamental to economic, political, and social life. By working together as a team, people can achieve things that they cannot achieve by working alone. However, whenever cooperating is individually costly, shirking pays. In the marketplace, for example, exchanging products and services honestly is mutually beneficial but cheating on delivery, quality or payment is tempting. Likewise, all parties benefiting from a shared resource are better off by using it responsibly and sustainably, but for each of them there is a temptation for extracting as much as possible for instant personal gain. Entrusting ones' democratic decision rights to representatives or delegates allows for mutual gains from specialization in political decision-making but abuse of office is tempting. But also conspiring in cartels or criminal organizations holds mutual benefits for the conspirators, but is vulnerable

to cheating. Generally, such tensions arise whenever individuals may engage in actions that result in social benefits that exceed social costs while for the performing individual private costs exceed private benefits.

How can people work together anyway, and which factors hamper cooperation? This is one of the most fundamental and important questions science, and in the social sciences in particular (Kennedy & Norman, 2005; Pennisi, 2005). A long-standing and influencial thrust in the literature highlights spontaneous social sanctions, reaching from rather explicit to more tacit forms within continuing relationships and communities, as potentially powerful suppressors of opportunism, not only in distant human history but also in many domains in the contemporary world to which law and formal contracts do not reach. Documented examples include the management of shared natural resources (e.g. Ostrom, 1990; Seabright, 1993; Bardhan, 2000; Kikuchi et al., 2000; Otsuka & Tachibana, 2000), long-distance trade (e.g. Milgrom et al., 1990; Greif, 1989, 1993, 2006), production teams, cooperatives, and business networks (e.g. Okun, 1981; Dong & Dow, 1993; Craig & Pencavel, 1995; Pencavel, 2002; Kawagoe, 2000; Ohno, 2000; Fafchamps, 2000; Platteau & Seki, 2000; Baker et al., 2004), labor relations (e.g. Gintis, 1976; Akerlof, 1982, 1984; Akerlof & Yellen, 1990; Shapiro & Stiglitz, 1984; Bowles, 1985; MacLeod & Malcomson, 1989, 1998; Kanemoto & MacLeod, 1991), customer relationships (e.g. Okun, 1981; Blinder et al., 1998; Ryssel et al., 2004; Apel et al., 2005) credit unions and associations (e.g. Hossain, 1988; Banerjee et al., 1994; Besley & Coate, 1995; Fessler, 2002; Armendáriz & Morduch, 2005; Eroglu, 2010), risk-sharing and insurance associations (e.g. Morduch & Sicular, 2000), open-source software production (e.g. Benkler, 2002; Osterloh et al., 2008), politics (e.g. Calvert, 1987; Goldstein & Freeman, 1990), and even residential neighborhoods (e.g. Tilly, 1981; Sampson et al., 1997; Browning et al., 2005).

Information the involved parties receive about each others behavior is frequently emphasized as a critical factor moderating the efficacy of such informal governance. «How can we trust them if we don't know what they are doing?» *Transparency International* puts it, the world's major anti-corruption organization (Beddow & Sidwell, 2012, p. 16). It advocates thorough monitoring of representatives, delegates, and officials by the citizenry as the most forceful instrument to fight corruption, and defines its function as a facilitator and amplifier. Indeed, this view has gained considerable momentum in recent years. While there is long tradition of public sector transparency in the Nordic countries (Manninen, 2006), the 1966 United States *Freedom of Information Act* can be viewed as an ignition of a worldwide spread of similar provisions.[1] Even Germany with its long standing tradition of administrative secrecy passed a federal freedom of information law in September 2005. Supra-nationally, a regulation (EC 1049/2001) of the European Parliament and the Council grants a right of access to documents of the three institutions to any Union citizen since 2001. Moreover, article 10 of the European Convention on Human Rights of 1953 provides a right to freedom of expression, including freedom to receive and impart information. On 11

[1] See `www.right2info.org` for a constantly updated list.

October 2006, 240 years after the first freedom of information legislation was enacted in Sweden, the *Inter-American Court of Human Rights* became the first international tribunal to hold explicitly that access to government information is a fundamental human right, indeed (Blanton, 2006).

Since around two decades, environmental regulation is undergoing similar transformation. Again beginning in the United States in the 1980s (Sunstein, 1999; Percival et al., 2009), it is culminating in the «third wave» (after command-and-control and market-based instruments, respectively) of environmental regulation (Tietenberg, 1998; Tietenberg & Wheeler, 2001; Dasgupta et al., 2006b). In the United States, the setup of the *Toxic Release Inventory* (TRI), that publishes toxic emissions from polluting facilities nationwide, was a more practical response, as governmental regulators recognized that conventional regulatory instruments were ill-equipped to handle the hundreds, and quickly growing number, of pollutants that may have harmful effects. But the TRI was applauded as being so successful in controlling pollution that it quickly spread around the world. Most euphorically the idea was taken up in the «developing world». Lacking sufficient capacities to handle pressing health, safety and environmental problems through legal liability or governmental regulation, it is often viewed as the only promising avenue to manage those problems. In collaboration with the World Bank, many environmental agencies in those countries adapted public information programs to the local needs (Dasgupta et al., 2006b). In international law, this development culminated in the adoption of the *Aarhus Convention*,[2] one of its three «pillars» devoted to safeguarding access to environmental information, such that members of the public can get easily informed on what is happening in the environment around them, in particular to identify potential polluters and monitor what they are doing. Similar programs to aid community monitoring have also been set up in the health sector (Dranove, 2002; Björkman & Svensson, 2009, 2010) or the public education sector (Reinikka & Svensson, 2005). Facilitating trust by aiding monitoring and pooling of information about particular traders' honesty has also a very long tradition in the marketplace (Milgrom et al., 1990; Greif, 1989, 1993, 2006), and underlays much of consumer protection law (Prosser, 1971; Whitford, 1973), labeling (Beales, 1994; Golan et al., 2001; Giannakas, 2002; Jin & Leslie, 2003), or feedback systems in online marketplaces (Resnick & Zeckhauser, 2002; Resnick et al., 2006; Jøsang et al., 2007). What all those approaches share is the conviction that mutual monitoring and sharing of information about particular actors' conduct is an essential precondition for effective informal enforcement of cooperation.

Why the answer to this question is not obvious

Two rather different views, which are intimately related to different answers to the question why people cooperate in the first place, illustrate that the answer to the posed

[2] It clumsy full name being *UN/ECE Convention on Access to Information, Public Participation in Decision-making and Access to Justice in Environmental Matters.*

question is not obvious anyway. On the one had, one might argue that «without monitoring there can be no credible commitment» (Ostrom, 1990, p. 45), without credible commitment there can be no trust, without trust there can be no cooperation. This line of reasoning has intuitive appeal, indeed, and it is common. Relatedly, there is an influential stream in the social sciences, and economics in particular, to explain social cooperation by extrinsic incentives, rewards and punishments, that induce inherently selfish individuals to cooperate. In order to apply such rewards and punishments, the argument can straightforwardly be continued, a minimum requirement is that the behaviors to be sanctioned are observable. If they are not, incentives that inhibit cheating are diluted and hence there is no rational reason to trust.

On the other hand, a different explanation of social cooperation emphasizes the role of moral values and the internalization of social norms, leaving for extrinsic incentives only a bit part or even a detrimental role. According to this view, cooperativeness is a cultural primitive, by and large short-term invariant and determined by rather distant social history, not by the incentives current situations create. This view is more prevalent in sociology (see Durkheim, 1893, for a seminal account), but it has also been articulated by a number of influential economists, such as Kenneth Arrow (1971, p. 20), who argued that as an alternative to government enforced contracts

> «society may proceed by internalization of these norms [of social behavior, including ethical and moral codes] to the achievement of the desired agreement on an unconscious level. There is a whole set of customs and norms which might be similarly interpreted as agreements to improve the efficiency of the economic system … by providing commodities to which the price system is inapplicable.»

Following seminal studies of Putnam (1993b,a), Fukuyama (1995), La Porta et al. (1997), and Knack & Keefer (1997), a body of empirical studies has accumulated in the past decade which use some measure of trust, cooperativeness, or «social capital», within a population as an *independent* variable to predict various economic outcomes (e.g. Zak & Knack, 2001; Guiso et al., 2004, 2008).

A direct consequence of viewing cooperative behavior as a short-term constant which is not adjusted to specific incentive structures in given situations and domains is that it should be rather *independent* from information about others' behavior. Dore (1987) and Putnam (1993a), for example, argue that trust and cooperativeness are culturally transmitted habits formed during *centuries* of social interaction. Along the same lines, Fukuyama (1995) asserts that they are the result of shared values that induce people to generally suppress their private interests to the benefit of the group. Motivated by the problem of credible commitment, a number of economists even suggested that humans might have evolved bio-psychological adaptions that may suppress self-interest (e.g. Schelling, 1978; Hirshlifer, 1978, 1985, 1987; Akerlof, 1983; Frank, 1988).[3] One might even *reverse* the relationship, as monitoring

[3]Frank (1988), however, does consider that such adaptions might also be conditional. I come back to this below.

and other forms of control may create a milieu of suspicion, undermine civic or moral duties, or may be perceived as an signal of mistrust that is responded with *less* not more cooperation (Titmuss, 1970; Frey & Jegen, 2001).

Relatedly, a body of evidence has accumulated in recent years (see section 2.2) that shows that people frequently act against their own material self-interest, and the evidence may be interpreted in a way that tactical considerations may actually of minor importance. In fact, based on these results economists argued that cooperation within relationships and communities is to a considerable extent non-strategic (e.g. Bowles & Gintis, 2002, 2005b). But what, then, does this imply for the relationship between transparency and cooperation? Do people also cooperate if they have very little or no information about interaction partners' past conduct? Do they actually respond to more of such information with more cooperation? The aim of this thesis is *not* to argue that either of these views is incorrect but to take them as a good reason to investigate the *fine-structure* of the relationship between the propensity of a given group of individuals to cooperate with one another and the availability of information they have about each others' actions in more detail.

Approach

The real world is extremely complex. It is very difficult to grasp, let alone understand, the rich situations of our everyday life by mere observation because a myriad of things is going on simultaneously, and they usually interact in complex ways. For this reason economists build models of the world. The art of modeling is to consider one phenomenon, problem, or supposed relationship at a time, distill its generic essence and abstract from the richness of the real world in a way that its becomes comprehensible. Any model is necessarily «wrong» in that it does not replicate the real world, but this is the very reason to build models in the first place. The same applies for empirical research, and economic experimentation in particular. Economic experiments can be viewed as simple models populated by real people. They *need* to be simple to be useful. Like models they help us to create and refine hypotheses, and strengthen or weaken the priors about them, but hypotheses always remain hypotheses that ultimately must be evaluated in the world we actually live in.[4] Therefore, modeling and experimentation are utilities that help to understand the world better, but are always to be understood as *complements* not substitutes to other approaches.[5]

There exists a generic paradigm in the behavioral and social sciences to represent the problem of cooperation in such a stylized way that it becomes easy to study it

[4] To view not only the results of theoretical models («thought experiments») but also the results of experiments as new predictions perhaps help to avoid the misunderstanding that behavioral experiments provide for readily generalizable «truths».

[5] There is a sizable economic literature considering the role of transparency for corruption and fiscal discipline (e.g. Hameed, 2005; Kolstad & Wiig, 2009; Peisakhin, 2012), environmental management (e.g. Stephan, 2002; Hamilton, 2005; Dasgupta et al., 2006a), investment decisions (e.g. Drabek & Payne, 2002; Gelos & Wei, 2005), or monetary policy (see Geraats, 2002, for an overview).

theoretically and experimentally, as a *game*. I will introduce this game in section 1.1.1. Within this game, I conceptualize transparency just as the accuracy with which information about a player's actions is conferred to the other players. This is, of course, not the only way to conceptualize the concept, but I hope to convince the reader that it is an illuminating one. It corresponds closely to the physical definition (an action is easily observed behind a wall of glass, but not behind a wall of stone) and also corresponds closely to the common usage of the term in a social context.

The advantage of studying the relationship between transparency thus defined and cooperation in the established paradigm of cooperation games is that the latter have been studied extensively, both theoretically and experimentally, and it is therefore possible to screen the literature with the specific focus of the guiding question: What does the propensity of individuals in a given group to cooperate with one another has to do with the availability of information they have about each others' actions? I will do this chapters 1 and 2. Although the reviews contain no fundamentally novel material, I hope its organization around the specific focus outlined above is nonetheless a useful contribution to the literature.

Received literature

Theory

In the first chapter I begin with outlining the generic problem of cooperation within a simple model. Mutual cooperation is beneficial to all, but cheating (assuming that cooperation is costly) always pays. In a selfish world cooperation therefore always needs to be enforced. Cooperating must provide for a return benefit (reward or avoided punishment) that renders cooperation profitable. I will highlight that if such enforcement needs to be performed by the individuals in the group themselves, then information about the others' actions becomes generally critical, since rewards and punishments are by definition conditioned on past play.

In the remainder of the first chapter, I investigate the qualifications and fine-structure of this tentative conclusion. Beginning with a setting in which players interact repeatedly and every action is perfectly observed by all other players, I illustrate that strategies that condition behavior on this information can enforce cooperation if one of two conditions are met: (i) The players believe that their coplayers might be precommitted to play the conditionally cooperative strategy as well, or (ii) the horizon of the interaction is indefinite. If neither condition is satisfied, unconditional defection is the unique stable outcome, which is another way to say that players do not care about what the others did before. If one of the conditions apply, then significant cooperation is possible by the application of conditional strategies that reward cooperative behavior and punish non-cooperative behavior. Since all cooperation in these settings hinges on the fact that all actions are perfectly transparent and hence each player is perfectly informed about her or his coplayers' history, one might expect that applying rewards and punishments is more difficult if such information becomes

limited, and as a result cooperation might be much more difficult to sustain.

I investigate to what extent this intuition is correct in sections 1.3 and 1.4. In section 1.3 I consider a setting in which the information about the other players' actions is limited in one special way: it may contain errors but all individuals in the group receive the *same* signals. In this case, monitoring is said to be *imperfect* but *public*. The literature has found that the beforementioned result extends to this case, that is, (almost) any degree of mutual cooperation is possible by means of strategies that condition rewards and punishments on the information on the coplayer's behavior *available*. For empirical purposes, it is important to observe that the result is also consistent with actual cooperation being infrequent even when noise is small. A related result provides a reason to hypothesize that the propensity of individuals in a group to cooperate with another will at least not decease as information about one anothers' actions becomes more accurate.

In section 1.4 I turn to a setting in which the information about the other players' actions is limited in a more general way: it may not only be inaccurate but different players may also receive different information about particular players' actions. In this case, monitoring is said to be *private.* If information about past actions is dispersed among the players in the group in the form of many small packages of private information, it becomes generally very difficult to coordinate on stable cooperation. On the other hand, the settings in which cooperation can be sustained under public (perfect) monitoring turn out to be much more encompassing, including not only situations in which the *same* individuals interact for an indefinite period of time but even situations in which the same individuals meet hardly more than once. Furthermore, mechanisms that honestly transmit information about the players' behavior to other players, to bring monitoring closer to public, render cooperation more likely. However, how such information processing should happen in a group of selfish individuals remains unexplained.

I conclude the first chapter by remarking, that all cooperation in the models reviewed is *tactical* in the sense that individuals only cooperate with one another if it is in their (long-term) self-interest. A number of questions follow immediately: Do real people actually perform such tactics? Do real people *only* cooperate for tactical reasons? What strategies do they actually play in repeated games? What information about the coplayers' actions to they use? How do they respond if such information becomes limited?

Evidence

In the second chapter I take up those questions and selectively screen the empirical literature in order to find answers. I will proceed in roughly two steps. As a first step, I will focus on games with perfect monitoring and investigate when and to what extent subjects play strategies that are conditioned on information about the coplayers' past actions. This will be a useful first step in order to identify situations in which the degree of information players have about the history of play likely has an impact.

In section 2.1 I review evidence on behavior in repeated games with perfect monitoring. Results show that a majority of subjects actually cooperates tactically. They use the available information about the coplayers' past behavior to condition their current behavior. However, there is a persistent residuum of cooperation even in (approximated) one-shot interactions which cannot be accounted for by tactical considerations. In section 2.2 I consider this residuum in more detail, and put special emphasis on evidence that is informative with respect to the present guiding question. A large body of evidence shows clearly that people do *not* always cooperate for tactical reasons, but there is also evidence that such cooperation is conditioned on coplayers' behavior anyway.

If some subjects are willing to incur a cost to confer a benefit to someone else, are there also subjects who are willing to incur a cost to impose a cost on someone else? I review the evidence on this question with a focus on the degree of conditioning on coplayers' past behavior, in which case such behavior can serve as a punishment, in section 2.3. The evidence shows that such spiteful behavior is very common in various subject pools worldwide. Furthermore, it is frequently used as a punishment of non-cooperative behavior, even by unaffected third-parties, with strongly disciplining effects. It follows predictable patterns, despite the fact that pecuniary incentives seem to play a significantly smaller role than for cooperative behavior. But there are exceptions and evidence that the degree of conditioning varies.

In sum, the evidence from experiments under perfect monitoring shows that if information about the other players' past behavior is readily available, this information is generally used in order to play a discriminatory strategy, even in situations in which material returns from doing so is ruled out. As a result, there are material incentives to cooperate when the same players meet more than once. This is reflected in a notably higher frequency of cooperation in repeated games compared to (approximated) one-shot games. Nonetheless, there is also significant cooperation in the latter, in particular if costly punishment is available. In one sentence: if information about the other players' past behavior is readily available, cooperation generally occurs and can be very high if the interaction is repeated.

As a second step, I consider conditions in which this information is *actually* limited. While imperfect monitoring is an issue in the theoretical literature for quite some time, there is only a very small (but growing) experimental literature, which I review in the final part of the second chapter. Section 2.4 deals with cooperation experiments in which the signals the players receive about their coplayers behavior are subject to some random disturbance. In cooperation experiments with an indefinite horizon and imperfect public monitoring, the available evidence shows that there is cooperation at any level of noise, but always less than maximal. Furthermore, the frequency of cooperation decreases as noise increases, and cooperation in conditions with very inaccurate information is not more frequent than under one-shot play. Finally, subjects seem to resort to more lenient and forgiving variants of conditional strategies, that wait for a coplayer to defect two or three times before reverting to punishment mode, and return to cooperation after a punishment phase has occurred.

These results suggest that one cannot just interpolate results obtained in a perfect monitoring setting to ones with informational constraints. This conclusion is reinforced by the available evidence on imperfect monitoring in cooperation games with an option to reduce coplayers' payoffs. It is shown that payoff reductions become *more* frequent and cooperation less frequent as monitoring gets more noisy.

Finally, in section 2.5 I consider studies that implement different conceptualizations of *private* monitoring. The received evidence is affirmative with respect to the hypothesis that information transmission, or the publication or disclosure of behavioral records may facilitate cooperation relative to conditions in which monitoring is private. The frequency of cooperation is generally lowest in private monitoring conditions, and information transmission in some form or the other increases cooperation. If matching is non-random, disclosure also reduces the clientelization effects observed under private monitoring. The evidence is more ambiguous with respect to the question whether random matching games with public monitoring provide for equally efficient results as repeated interaction among the *same* individuals. To summarize with respect to our guiding question, the available evidence suggest that the level of cooperation (and efficiency) is generally increasing in the accuracy and symmetry of information players have about each others' actions.

In sum the insights obtained feed back on the two divergent views about why people cooperate and what role information about coplayers' behavior might play. The evidence shows that both views have merit. People cooperate to a significant extent tactically, use reciprocity strategically to realize mutual benefits as envisaged by game theory, and thereby draw strongly on information about coplayers' past behavior. But people also forgo own material gains to cooperate or reduce incomes of others. A significant fraction of those latter behaviors are conditioned on information about coplayers' past play anyway. In a one-sentence summary: there is cooperation even in situations in which people have minimal or no information about others' behavior, but increasing such information generally facilitates cooperation significantly.

One aspect that is seldomly considered

After having reviewed the existing literature I point to one aspect that is shared by both theory and experimentation: information structures are generally considered as an exogenous parameter. In reality, information about others' behavior is not just available, non-available, or something in between, but is must be produced, acquired and transmitted. This implies that transparency, as defined here, is not just there or absent, it is *created* by the players themselves, information structures are *endogenous*.

This fact is vividly illustrated by Milgrom et al. (1990) in their account on the revival of trade in Europe during the early middle ages. During this time, merchants evolved their own private code, known as the *Lex Mercatoria*, that was mutually enforced, and enforcement relied on shared information about individual members' conduct. But given that getting informed (monitoring) and information transmission is generally costly, what incentives did the merchants have to engage in such activi-

ties? The the vast majority of received theoretical and experimental literature is silent on this issue because it considers information structures generally as exogenous. It thereby abstracts from the possibility that players may alter them themselves. In other words, the standard approach in the theoretical experimental literature is to confront the players with a given information structure, and investigate how they behave *without adjusting the strategy space relative to the case of perfect monitoring*. However, when studying settings in which the players' information is imperfect it seems intuitive that at least one extension of the strategy set is natural: to acquire better or additional information!

Doing so might have a number of interesting implications. First, if we allow players to acquire more accurate information (at a cost), then they might adopt strategies that produce entirely different cooperative and sanctioning behaviors than the strategies that they adopt under an exogenously given information structure. Second, if we allow players to acquire more accurate information at a cost, then the player also has the option to forgo the opportunity. This implies, for example, that there must be some (potentially intrinsic) incentives «to become adequately informed about how others had behaved» even though this «might be personally costly» Milgrom et al. (1990, p. 1). Investigating the nature of those incentives might be an interesting avenue for research. Milgrom et al. (1990), for instance, argued the there were particular institutional features in the merchant community that provided incentives to disclose evidence against violators of the code to the community and to become informed in the first place. They documented that the system was indeed very successful in enforcing honest behavior. This view puts the institutional and technological environment as a determinant of information acquisition and transmission, and therefore the degree of «equilibrium transparency», center stage. This conception is the motivation for all original contributions in chapters 3 through 7.

Outline of the original contributions

The first four original contributions reported in chapters 3 through 6 are empirical studies that augment established experimental paradigms to study cooperation and sanctioning by options to acquire more accurate information (at a cost) about their coplayers' previous behavior before acting themselves. In the final chapter 7 I draw on the above line of reasoning that highlights the technological environment as a determinant of information acquisition and transmission, and therefore the degree of «equilibrium transparency», and investigate its relation to measures of generalized trust using a large sample of micro-level survey data. I now provide a brief outline.

In chapter 3, which is co-authored with Timo Goeschl, we study mutual monitoring and punishment behavior in a simple one-shot cooperation experiment. We extend upon earlier research that introduced informational imperfections *exogenously* by allowing for their *endogenous* mitigation through costly acquisition of information. Specifically, players can decide to resolve imperfect information about coplayer behavior at the cooperation stage through a perfect, but costly monitoring technol-

ogy. We study how monitoring and punishment respond to changes in information costs. The key findings are a follows. First, positive information costs have the predicted impact of decreasing the supply of punishment. Second, while all sanctioning is discriminate at zero information costs, a distinct share of subjects switches to «blind» punishment rather than refraining from sanctioning when information is costly. Third, positive information costs lead to defectors making up a smaller share of the punished and receiving smaller punishment compared to zero information costs. Fourth, I find that beliefs and risk aversion contribute to explaining individuals' monitoring and punishment behavior. Fifth, turning to cooperation, single-round cooperation under positive monitoring costs is maintained because subjects overestimate monitoring and punishment, in contrast to the baseline of zero information costs.

In chapter 4 (also co-authored with Timo Goeschl) we draw on the setup of the previous chapter and study monitoring and punishment behavior by materially unaffected *third parties*, and compare it to their second party counterparts. We find the following. First, positive information costs have the predicted impact of decreasing the supply of punishment. Second, while virtually all sanctioning is discriminate at zero information costs, a distinct share of subjects switches to «blind» punishment rather than refraining from sanctioning when information is costly. Third, conditional on the decision to remain uninformed, it is risk averse subjects that tend to punish blindly, and more risk tolerant subjects that tend to refrain from punishment. Fourth, positive information costs lead to defectors making up a smaller share of the punished and receiving smaller punishment compared to zero information costs. Fifth, while third party punishment provides for weaker incentives to cooperate than second party punishment already when monitoring is costless, positive information costs render this difference even larger.

In chapter 5 (again co-authored with Timo Goeschl) we study finite horizon modified trust games with endogenous information structures. At the end of each period, the first mover may acquire information about the second mover's action in that period. We exogenously vary the cost of monitoring, and find that the introduction of information costs results in less monitoring and an emergence of blind trust as a new behavior type, where (i) blind trust is the dominant behavior under costly monitoring, (ii) first mover cooperation is more frequent, and (iii) payoffs are higher if information is costly. Furthermore, the average first mover in the costly monitoring condition did not worse than in the costless monitoring condition, and the average blind trustor did not worse than the average monitor or defector. We find that this static differences between conditions stem from differences in the respective dynamic patterns. The preferred interpretation is a «second-order reputation building» hypothesis, according to which some second movers try to strategically exploit the costliness of monitoring by investing in a sufficiently favorable reputation in the initial periods in which they are likely to be monitored in order induce blind trust and reap larger gains from exploitation in later periods. We provide further results that support this interpretation.

The setup of the previous chapter is extended upon in chapter 6, in which I investigate the question whether and how the length of the horizon influences the patterns of monitoring and blind trust by exogenously varying the length of the horizon of a given match. I find the following: First, the key results of the previous chapter are qualitatively replicated for matches with a considerably shorter horizon, but blind trust remains clearly the exception here. Second, the average first mover's behavior is independent from the length of the horizon when monitoring is costless, while blind trust and overall first mover cooperation is significantly more frequent in long horizon than short horizon matches when monitoring was costly. Third, while in long horizon matches first mover cooperation is significantly more frequent under costly than under costless monitoring, this difference vanishes in short horizon matches. In consequence, efficiency is *ceteris paribus* higher in long horizon than in short horizon matches, and *ceteris paribus* not lower under costly than under costless monitoring. We underpin these results by an investigation of the behavioral dynamics and a possible interpretation.

Finally, in chapter 7 I carry the conceptualization to the field. A previous study by Fisman & Khanna (1999) found a positive relationship between landline connectivity and survey-measured generalized trust using highly aggregated data and a limited set of control variables. In chapter 7 I replicate this relationship using (i) individual level micro-data from the German Socio-Economic Panel (SOEP), (ii) an improved, more fine-git, and experimentally validated measure of generalized trust, (iii) and using alternative estimation methods. I extend on the study by (i) controlling for a much larger set of variables, (ii) drawing on an experimentally validated measure of beliefs into others trustworthiness, and (iii) extending the set of considered communication technologies to include cellular phones, internet (both narrow and broad bandwidth), and television. I find evidence suggesting that the effect is, as predicted by economic theory, mediated by beliefs, that it becomes significantly weaker as variables on level of education and local network inclusion are controlled for, and that there is a significantly positive effect of broad bandwidth internet connectivity, and a significantly negative effect of TV set possession.

Chapter 1

Transparency & Cooperation: Conceptions & Theory

What does the propensity of individuals in a given group to cooperate with one another has to do with the availability of information they have about each others' actions? The purpose of this chapter is to briefly screen the major theoretical approaches to cooperation in economics with this guiding question as a specific focus.

In the first section I begin with outlining the generic problem of cooperation within a simple model. Mutual cooperation is beneficial to all, but cheating (assuming that cooperation is costly) always pays. In a selfish world cooperation therefore always needs to be enforced. Cooperating must provide for a return benefit (reward or avoided punishment) that renders cooperation profitable. I will highlight that if such enforcement needs to be performed by the individuals in the group themselves, then information about the others' actions becomes generally critical, since rewards and punishments are by definition conditioned on past play.

In the remainder of this chapter, I investigate the qualifications and fine-structure of this tentative conclusion. Beginning with a setting in which players interact repeatedly and every action is perfectly observed by all other players, I illustrate that strategies that condition behavior on this information can enforce cooperation if one of two conditions are met: (i) The players believe that their coplayers might be precommitted to play a conditionally cooperative strategy as well, or (ii) the horizon of the interaction is indefinite. If neither condition is satisfied, unconditional defection is the unique (sequential) equilibrium outcome of any repeated PD, which is another way to say that players do not care about what the others did before. If one of them is satisfied, then significant cooperation is possible by the application of conditional strategies that reward cooperative behavior and punish non-cooperative behavior. Since all cooperation in these settings hinges on the fact that all actions are perfectly transparent and hence each player is perfectly informed about her or his coplayers' history, one might expect that applying rewards and punishments is more difficult if information about the coplayer's behavior becomes limited, and as a result cooperation might be much more difficult to sustain.

I investigate to what extent this intuition is correct in sections 1.3 and 1.4. In section 1.3 I consider a setting in which the information about the other players' actions is limited in one specific way: it may contain errors but all individuals in the group receive the *same* signals. In this case, monitoring is said to be *imperfect* but *public*. The literature has found that the folk theorem extends to this case, that is, (almost) any degree of mutual cooperation is possible by means of strategies that condition rewards and punishments on the information about the coplayer's behavior *available*. For empirical purposes, it is important to observe that under imperfect public monitoring punishment can occur at times *on* the equilibrium path such that actual cooperation may be quite low even when noise is small. A related result provides a reason to hypothesize that the propensity of individuals in a group to cooperate with another will at least not decease as information about one anothers' actions becomes more accurate.

In section 1.4 I turn to a setting in which the information about the other players' actions is limited in a more general way: it may not only be inaccurate but different players may also receive different information about particular players' actions. In this case, monitoring is said to be *private.* If information about past actions is dispersed among the players in the group in the form of many small packages of private information, it becomes generally very difficult to coordinate on a cooperative equilibrium. On the other hand, the settings in which cooperation can be sustained under public (perfect) monitoring turn out to be much more encompassing, including not only situations in which the *same* individuals interact for an indefinite period of time but even situations in which the same individuals meet hardly more than once. Furthermore, mechanisms that honestly transmit information about the players' behavior to other players, to bring monitoring closer to public, render cooperation more likely. However, how such information processing should happen in a group of selfish individuals remains unexplained.

In section 1.5 I conclude with two remarks that concern the need for empirical facts and the exogeneity of information structures.

1.1 The generic problem of cooperation

Etymologically, the term «cooperation» comes from Latin *cooperationem*, which is put together from *com-* («with») and *operari* («to work»), thus meaning literally «a working together» (Onions, 1966; Klein, 1971). In this way, cooperation is commonly said to occur when two or more individuals engage in joint actions that result in mutual benefits (see for example the *Collins English Dictionary* or *The American Heritage Dictionary of the English Language*). Examples include the mutually beneficial exchange of goods and services, contributions to finance or maintain shared facilities, working together in a team, managing shared natural resources, collusion among firms, or participating in collective actions such as demonstrations, strikes, boycotts, coups, or warfare.

1.1.1 Cooperation

A model that is able to represent cooperation should have three ingredients. First, cooperation is a *social* phenomenon, such that at the very minimum two individuals are needed to capture it. Second, cooperation involves *activity*. Finally, cooperation is defined relative to the *consequences* of this activity.

Strictly speaking, only *individuals* can «act», such that the primitive of *joint* or *collective* action must be *individual* action (Dugatkin, 1997). In this sense, the primitive of mutually beneficial cooperation is in the behavioral sciences typically defined as *individual action that is beneficial to other individuals* (Hamilton, 1964a,b). To capture this formally in the simplest way, consider an individual, which I call *One* (female articles), that lives in a population N of size $n \in \{\mathbb{Z} : n \geq 2\}$, and may choose from a set $A_1 = \{0, \alpha\}$ of mutually exclusive courses of action a_1, where $a_1 = \alpha > 0$ represents the performance of some action, and $a_1 = 0$ represents the logical negation (maintaining the *status quo*). The performance of the action has consequences for *One* (private consequences) and other individuals in the population (social consequences). Specifically, assume that $a_1 = \alpha$ implies a cost $\kappa \in \mathbb{R}$ (which may be negative in which case it is a benefit) for *One*, and an external benefit $\eta \in \mathbb{R}$ (which may be negative in which case it is a cost) that is somehow shared among the remaining $n-1$ individuals. Costs and benefits are normalized such that maintaining the *status quo*, $a_1 = 0$, entails neither private nor social consequences, $\kappa = \eta = 0$. Costs and benefits are to be understood in a *ceteris paribus* fashion, that is, the performance of the act is the *only* change in an otherwise constant world.[1] For the time being, I adopt the assumption $n = 2$ for sake of parsimony, a generalization to $n > 2$ is straightforward (I come to this below). I call the second individual *Two* and use male articles.

I now apply (more or less) standard terminology to classify actions with respect to their private and social consequences (e.g. West et al., 2007b). Action $a_1 = \alpha$ is called *cooperative* if and only if $\eta \geq 0$, otherwise the action is called non-cooperative. One may call α *weakly cooperative* or *socially neutral* in case $\eta = 0$, and *strictly cooperative* if $\eta > 0$. A cooperative action is called *mutualistic* if and only if $\kappa \leq 0$, otherwise ($\kappa > 0$) it is called *altruistic*.[2]

[1] Generally, the appropriate metric on the set of consequences depends on the problem at hand. In biology benefits and costs are measured in terms of direct genetic or cultural fitness consequences, that is, relative reproductive success in a constant environment (see Lehmann & Keller, 2006; West et al., 2007b,a, for recent overviews). Such consequences capture usually very long time frames, which is natural for evolutionary biologists' interest in «ultimate» explanations (Tinbergen, 1963). Ecologists, anthropologists and social scientists, including economists, have a more proximate perspective, that is, they are interested in the more immediate consequences, such that different metrics are useful (Tinbergen, 1963; Brosnan & de Waal, 2002; Nowak, 2006; Jensen et al., 2006; Jensen, 2010). Specifically, it is standard practice in economic applications to operationalize costs and benefits in terms of pecuniary consequences or effects on consumption possibilities (broadly construed). See Becker (1976a) for an account on the relationship between the operationalizations in biology and economics.

[2] One may call actions for which $\eta < 0$ and $\kappa \leq 0$ hold *selfish* or *egoistic*, and actions with consequences $\eta < 0$ and $\kappa > 0$ *spiteful*. However, there is usually no loss of generality in restricting attention to α being cooperative, since non-cooperative acts are just the converse of cooperative acts, that is, the omission of a non-cooperative act can be represented as a cooperative act by just switching labels: we

The specification can be readily extended to a larger (possibly continuous) action set A_1, that captures the notion of «levels» of a cooperative behavior (e.g. helping a little, helping a lot) or different behaviors that differ in their cooperativeness in the sense of having differential external effects.[3] Again, this complication does not produce important additional insights, so for the time being I stick to the more parsimonious dichotomous case.

With this simple generic representation of what it means to «cooperate», it is easy to conceptualize the notion of engaging in «joint action for mutual benefit»: If both individuals can perform a cooperative action for which $\eta > \kappa$ holds, then working together, *exchanging* cooperative acts reciprocally, is mutually beneficial. Both players realize a net benefit of $\eta - \kappa > 0$ such that each of them is better off, even if $\kappa > 0$, by working together compared to each of them working alone. In other words, mutual cooperation yields a strict *Pareto improvement* over mutual defect.

This is the kind of self-interested cooperation Adam Smith (1776, p. 19, emphasis added) described when he famously wrote that

> «man has almost constant occasion for the help of his brethren, and it is in vain for him to expect it from their benevolence only. He will be more likely to prevail if he can interest their self-love in his favour, and show them that it is for their own advantage to do for him what he requires of them. Whoever offers to another a bargain of any kind, proposes to do this. *Give me that which I want, and you shall have this which you want*, is the meaning of every such offer; and it is in this manner that we obtain from one another the far greater part of those good offices which we stand in need of. It is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity but to their self-love, and never talk to them of our own necessities but of their advantages.»

It is apparent that such generic exchange of cooperation models not only exchanges of services in the narrow sense, but virtually all situations in which *One* and *Two* can work together to achieve something that is not possible to achieve by working alone. In this case η is an individual's contribution to the common project and κ is the personal opportunity cost of contributing. It models also the avoidance or resolution of costly conflict or the sustainable conservation of shared resources: in this case κ is the personal opportunity cost and η the other's benefit of restraint.

---

can label the performance of the non-cooperative act by $a_1 = 0$, and the abstention from it by $a_1 = \alpha$. It is easy to see that this just reverses the inequalities in the above definitions. There can be cases, however, in which the distinction is useful (see for example section 2.3). It should also be noted that those definitions neither refer to motivations nor have any normative connotations (see Sober & Wilson, 1998, for an comprehensive discussion).

[3] Our simple dichotomous case is nested in this general model with $\eta(a_1) : A_1 \to \mathbb{R}$ and $\kappa(a_1) : A_1 \to \mathbb{R}$ representing the social and private consequences, respectively, as a function of a_1, requiring $\eta(0) = 0$ and $\kappa(0) = 0$, and defining with some abuse of notation $\eta(\alpha) := \eta$ and $\kappa(\alpha) := \kappa$.

Henceforth, one might be led to conclude that if there are mutual gains feasible as above, a «win-win-situation», then exploiting them is a «rational» thing to do, just like grabbing cash on the table. Thus, «rational selfishness», in the sense of cooperating when it personally pays off, appears to be enough for widespread cooperation to occur. Indeed, this is what Adam Smith (1776, p. 488f, emphasis added) concluded later in his seminal work:

> «[E]very individual necessarily labours to render the annual revenue of the society as great as he can. He generally, indeed, neither intends to promote the public interest, nor knows how much he is promoting it. By preferring the support of domestic to that of foreign industry, he intends only his own security; and by directing that industry in such a manner as its produce may be of the greatest value, he intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention … *By pursuing his own interest he frequently promotes that of the society* more effectually than when he really intends to promote it.»

This idea became the basis for the Walrasian model, the dominant model in economic theory until quite recently, which culminated in the welfare theorems of Kenneth Arrow and Gerard Debreu (Arrow & Debreu, 1954; Debreu, 1959; Arrow & Hahn, 1971). In essence, they are an almost trivial consequence of the assumption that if there are mutual gains, as in the above stylized situation involving *One* and *Two*, they are realized.

1.1.2 The problem of opportunism

While Smith's view highlights one aspect of self-interest, it neglects another critical part: whenever $\kappa > 0$, then independently from *Two's* behavior *One* is *always* better off by refusing to cooperate. To see this, consider the payoff consequences of *One*'s behavioral alternatives in the two possible contingencies. In case *Two* cooperates, *One* gets $\eta - \kappa$ by cooperating and η by the non-cooperative move, whereas the latter is always greater by assumption $\kappa > 0$. If *Two* does not cooperate, *One* gets $-\kappa$ by cooperating and zero by the non-cooperative move, again the latter being always greater by assumption $\kappa > 0$. Thus, *independently* from *Two*'s behavior, cooperation does never pay.

Under the traditional behavioral assumptions employed in economics (see Mas-Colell et al., 1995; Bowles, 2004; Kirchgässner, 2008, for detailed expositions), which imply that a course of action is performed if and only if it pays off in material terms relative to the other alternatives, it follows that whenever $\kappa > 0$ *One* will refuse to cooperate *independently* from *Two*'s behavior, that is, non-cooperation is said to be a *strictly dominant strategy*.[4] The same is true for *Two*, such that both players will

[4] The behavioral assumptions imply that the case $\kappa \leq 0$ is not very interesting, since such cooperative acts will be performed by immediate self-interest already; the external benefits are windfall gains.

refuse to cooperate and the mutual gains will not be exploited. This generic problem is well known as the *Prisoner's Dilemma* (PD).[5] It readily extends to any arbitrary number of players: suppose there is a finite number n of players, each having the two behavioral alternatives as before, but now the external benefit η generated by one player's cooperation is equally shared among the remaining $n-1$ players. This is well known as the *public good game*. It is easy to show that under the mentioned behavioral assumptions defection still is a strictly dominant strategy for all players. This was highlighted by Samuelson (1954). In all these situations, everyone would be better off if everyone could cooperate, however, everyone defects, «ruefully but rationally, confirming one another's melancholy expectations» (Putnam, 1993b).

Thus, the «invisible hand» and the Prisoner's Dilemma predict outcomes that could hardly be any farther apart: While the former predicts that aggregate wealth of a population will be maximized if each individual pursues her or his own profit, the latter reverses this statement. This puzzled economists (and other social scientists alike) since the PD was devised by Merrill Flood and Melvin Dresher in the early 1950s (Flood, 1952). The stark difference stems from differing assumptions about the possibilities for the players to *commit* to, or bind themselves to a particular course of action (see Bowles, 2004, for an overview). The central assumption underlying the Walrasian model is that perfect commitment is possible at no cost. This is often termed «complete contracting», and it eliminates any possibility for cheating on agreements by assumption.

The PD turns the perfect commitment assumption on its head in the sense that no binding commitments are possible at all.[6] Whenever complete contracting is infeasible, contributing to some potentially mutually beneficial endeavor exposes individuals to possible exploitation by others acting opportunistically. The temptation for opportunistic behavior is the personal gain accruing over and above the mutually cooperative payoff by «taking a free ride» on the others' cooperation, which in our simple model is equal to $\eta - (\eta - \kappa) = \kappa$.

1.1.3 Reward and punishment

The problem could obviously be «solved» by just «taxing» or «subsidizing» away the gain from opportunism, that is, provide appropriate counter-incentives by a respective reward and/or punishment scheme that renders cheating unattractive and cooperation a selfish best response. That is, if we assume that there is some central enforcement automaton that monitors the players' behavior and imposes a punishment $p \geq \kappa$ on

In particular, such acts are performed *independently* from anyone else's behavior. Keeping with the literature on cooperation, I will restrict attention to the interesting case $\eta > \kappa > 0$. Even if we want to assume that the players are not entirely selfish, there is still always a *temptation* to defect, which renders the problem still (or even more) interesting. I come to this below.

[5] In economics this problem has many names, among others social dilemma, problem of cooperation, free-riding problem, problem of opportunism, tragedy of the commons, problem of collective action, or moral hazard.

[6] A corollary is that in a selfish world pre-play communication, promises or agreements are completely irrelevant because they are not credible. I come to this below.

any defection, or a reward of $r \geq \kappa$ on any cooperative act, or some combination of both, then the players are deterred from cheating. To see this, observe that *One*'s payoff from cooperating, given that *Two* cooperates as well, is $\eta - \kappa + r$, and the payoff from defecting is $\eta - p$, such that cooperation is a best response if and only if

$$\eta - \kappa + r \geq \eta - p \Leftrightarrow r + p \geq \kappa$$

The same condition results contingent on *Two* defecting. Thus, if all actions by all players are subject to an incentive scheme for which this condition holds, then mutual cooperation instead of mutual defect becomes the unique equilibrium in dominant strategies (the idea goes back to Pigou, 1932).

1.1.4 Conclusion

Recall that I defined transparency as the availability of information about the other individuals' actions. Does it matter in the settings reviewed above? That is, does the propensity of individuals in a given group (here *One* and *Two*) to cooperate with one another has to do with the availability of information they have about each others' actions? The answer is «no». In the Walrasian world with perfect and costless commitment, the players can costlessly bind themselves to the mutual cooperation outcome, so once committed there is no reason to monitor the coplayers behavior. Likewise, if no precommitment is feasible but there is an enforcement automaton which is perfect in the sense that it can monitor all players actions and imposes sufficient rewards and punishments diligently, both players have a strictly dominant strategy which is by definition independent from the coplayer's behavior. The same is true if the enforcement automaton is imperfect or absent, this time defection being the strictly dominant strategy. Generally, information about the other individuals' action is apparently only relevant if a player wants to *condition* own behavior on the behavior of the other players. We might expect this to be the case if the rewards and punishments mentioned above need to come at least partially (this includes their contribution to centralized enforcement, for example by reporting, complaining, verifying or litigating, which all is costly) from the players themselves. I turn to this case in the next section.

1.2 Mutual enforcement

Assume that *One* and *Two* play the PD, and *Two* is informed about *One*'s action before choosing his own.[7] Observe that this setting involves an opportunity for *Two* to reward *One* for cooperative and punish her for non-cooperative behavior. Reward comes just in the form of reciprocation and punishment in the form of withholding it,

[7] That the sequence of moves in a game models the *information* structure and not the actual temporal structure of actions is easily illustrated: Suppose that *One* and *Two* play the PD, and that *Two* chooses temporally before *One*. Suppose further, however, that *One* is not informed about *Two*'s action at the point of her own choice. It is easy to see, then, that *One*'s situation is no different from one in which she chooses temporally before *Two*.

that is, to *cooperate conditionally* on *One*'s cooperation. If we assume that *Two* plays this strategy, and this is known to *One*, then *One* faces the incentive scheme $r + p = \eta$. Thus, since defecting (relative to cooperating) still generates a gain κ but now also carries an opportunity cost of η, then by assumption $\eta > \kappa$ cooperation becomes a strict best response to her. It is easy to check that these strategies are mutual best responses, that is, the strategy profile constitutes a Nash equilibrium point. Thus, one might argue that conditional cooperation, or (direct) *reciprocity*, enforces mutual cooperation even if there is no external enforcer at all.

However, this equilibrium violates a basic notion of dynamic consistency. To see this, observe that under the mentioned behavioral assumption of selfishness the conditionally cooperative strategy «cooperate if *One* cooperates, defect if *One* defects» involves a promise that is not *sequentially consistent*, that is, once *One* actually cooperated, it is no more in *Two*'s interest to adhere to the promise of reciprocating. This is just because $\kappa > 0$. If *One* perfectly knows about *Two*'s interests, it is not plausible to assume that she will believe such non-credible promises (and threats), and for this reason game theorists generally reject equilibria that rely on non-sequentially consistent strategies and call possibly remaining ones *sequential equilibria* (or *subgame perfect equilibria* in games with complete information, see van Damme, 1983). The notion of sequential equilibrium requires that *Two*'s strategy must specify an action contingent on all of *One*'s possible actions that are still in his interest to perform once he is informed about *One*'s actual action.[8] In other words, it does not allow for threats and promises off the equilibrium path that are not credible. It is easy to check that mutual defecting is the unique sequential equilibrium of the sequential PD. Thus, in equilibrium, *Two* still defects unconditionally, which is another way to say that he does not care about information about what *One* did before.

1.2.1 The shadow of the future

Intuitively, one might expect that the problem outlined above might disappear if the interaction does not end after *Two*'s action but there are many future opportunities to respond on one anothers' actions, that is, there is a long *shadow of the future*. The basic idea is probably best captured by the common German idiom «*man trifft sich immer zweimal im Leben*» («you always meet twice», in English commonly known as «be nice to people on your way up because you will meet them on your way down»). The essence is that cooperation might be turned into a long-term best response if cooperation is exchanged conditionally («I'll scratch your back if you'll scratch mine») and there is a long enough horizon of future interaction. The prospect for repeat business might work this way just as the prospect of being shunned can be a strong motivator to be honest and generous to friends, neighbors, and colleagues.

[8] Generally, it requires that the action taken by any player at any point in the game must be part of an optimal strategy from that point onward, given her or his beliefs about previous play to this point (which must, to the extent possible, be consistent with Bayesian updating on the hypothesis that the equilibrium strategies have been used to date) and given that future play will unfold according to the equilibrium strategies.

To begin with, the dynamic consistency problem explained above does *not* disappear if there is a commonly known definite end point of the interaction. To see this, suppose that there is just another stage after *One* and *Two* played the (simultaneous or sequential) PD in which *One* may reward *Two*'s cooperation with a benefit η at a personal cost κ. It is easy to show that *One*'s strategy «cooperate and reward *Two* only for cooperation» together with *Two*'s strategy «cooperate if *One* cooperates» constitutes a Nash equilibrium. But again, *One*'s promise is not credible, and hence this is not a sequential equilibrium. The unique sequential equilibrium is mutual defect with no reward given. What if *One* may punish *Two*'s non-cooperation with a benefit $-\eta$ (still at cost κ) instead of rewarding cooperation with benefit η? Apparently, nothing changes to the previous setting, since the threat of punishment is not credible.

One may continue to add another stage in which *Two* has the option to reward or punish *One*, but the above reasoning actually extends to any arbitrary finite duration of the interaction. Thus, a shadow of the future *per se* is not enough for changing anything: by backwards induction universal defect remains the unique sequential equilibrium if the common future has a (commonly known) definite end point (this stunning result has been vividly illustrated by Selten, 1978). Thus, we can extend the above conclusion in that after any history the player at move does not care about information what the other player did before.

1.2.2 Finite horizons, belief perturbations and imitation

Selten (1978) argued that the backwards induction prediction is in a way descriptively implausible, as one should expect the players to cooperate at least for some time when there is still a long time remaining in a finitely repeated game. He advocated a bounded rationality approach, but far more influential was the approach of Kreps & Wilson (1982); Milgrom & Roberts (1982); Kreps et al. (1982) that retained the standard behavioral assumptions and showed that cooperation over a large fraction of a finitely repeated game's duration is indeed possible if there is a (potentially only small) degree of incomplete information, that is, uncertainty about how the coplayers might behave.

Kreps et al. (1982) perturbed a *finite* horizon PD supergame with a little incomplete information. Specifically, they assumed that players hold a prior $\varepsilon \in]0,1[$, close to zero, that their matched coplayer is precommitted to a strategy called *tit-for-tat* (TFT) : «cooperate in the initial period and then copy what the coplayer did in the previous period» (Axelrod & Hamilton, 1981). That is, TFT is an in-kind reciprocal strategy that rewards cooperation with cooperation, and retaliates against defection with *one* period of defection. Precommitted means that this type of player sticks to the strategy no matter how the game unfolds, and particularly also in the terminal period. Kreps et al. (1982) show that in any sequential equilibrium of this perturbed supergame, each conventional (i.e. rational and selfish) player *imitates* the committed type throughout most of the time, except some last few periods before termination. That is, despite ε being arbitrarily small, rational and selfish players can end up in a

situation of self-confirming beliefs that the coplayer will play TFT: each player plays TFT in response to the expectation that the coplayer plays TFT, and this expectation is always confirmed in equilibrium. Generally, Kreps et al. (1982) show that in *any* sequential equilibrium of a finitely repeated PD there exists an upper bound on the number of stages where one player or the other defects, and this bound depends only on ε but is independent of the duration of the game (i.e. the number of repetitions).

The intuition is that a standard rational-selfish player may have an incentive to cooperate in the first period in the hope of «passing by» as a committed type and reap the gains from cooperation in subsequent periods, until it pays to cheat when the terminal period approaches. Thus, cooperation in initial periods can be thought of as an investment in maintaining a cooperative *reputation*, its return coming in the form of the coplayer's cooperation in later periods. «In common usage, reputation is a characteristic or attribute ascribed to one person (firm, industry, etc.) by another ... Operationally this is usually represented as a prediction about likely future behaviour» (Wilson, 1985, p. 27), which is «inferred from ... past actions» (Camerer & Weigelt, 1988, p. 1). This is exactly what happens along the equilibrium path: *One*, being uncertain about *Two*'s type, updates her prediction about *Two*'s likely behavior based on *Two*'s observed actions as the game unfolds. The prediction is *Two*'s reputation in the eyes of *One*. As long as *Two* cooperates, *One*'s belief in *Two* cooperating in the next period as well, i.e. *Two*'s «good» reputation, is maintained; as soon as *Two* defects, *One*'s belief degenerates, that is, *Two*'s reputation turns immediately «bad». Thus cooperating in the initial periods along the equilibrium path is often called «reputation building» (Camerer & Weigelt, 1988), but I have put this term into quotation marks because no reputation is actually *built*, but merely *maintained*. If the players share a belief that their coplayer could be a committed type, all players prefer to imitate this type (at least until the terminal period approaches), such that the «equilibrium reputation» they are playing is constructively maintained along the equilibrium path. But cooperation in the initial periods of the supergame can certainly be interpreted as an *investment* in reputation maintenance, because it entails immediate opportunity costs (κ) in return for a compensating payoff stream in the future. Apparently, for such an investment, and therefore sticking to the imitation equilibrium strategy, to be worthwhile, the (remaining) duration of the game needs to be long enough for the prospective return to cover the cost of the investment asset (a good reputation). Furthermore, since the value of a good reputation is decreasing in the (remaining) duration of a finitely repeated game, so is the incentive to cooperate. In any case, since reputations are inferred from the coplayer's past actions, adequate information about those actions are essential for this process to work.

A striking feature of the model is that the committed types need not actually exist, they only need exist in the heads of the conventional players. Even if all players are rational and selfish, the mere belief in the existence of committed or «crazy» types is sufficient to sustain the imitation equilibria. Kreps et al. (1982) assume that all players *are* rational and selfish, but this is not common knowledge. Thus, one interpretation of this prior belief is that it is just a result of «fuzzy minds» or super-

stition. However, for being empirically useful the model must pose some restrictions on observable behavior, and if any behavior can be «explained» by assuming an appropriate belief in a particular type, then the theory is vacuous. In fact, Fudenberg & Maskin (1986) show that the set of permissible types is virtually unrestricted. Thus, a second interpretation is that the assumption of committed types may be substantive, that is, the prior is justified by the *actual existence* of precommitted types. In any case the reason for the players' belief is an empirical question (I will come back to this in the following chapter). Theorists preferred the «save» route by rejecting any precommitment power or beliefs in such power, and it turns out that the only way in which cooperation can be shown to be part of an equilibrium strategy in this setting is to eliminate the definite termination point, that is, to assume that there is uncertainty about the exact time after which the players disband. I will illustrate how this works by means of a very simple example here.

1.2.3 Indefinite horizons and trigger strategies

Assume that *One* and *Two* play the (simultaneous-move) PD repeatedly for an indefinite duration, that is, after each round of play the interaction will continue for one more period with probability $\sigma \in]0,1[\subset \mathbb{R}$. We denote the action profile (a_{1t}, a_{2t}) in period t by $\mathbf{a}_t$, and call the sequence $(\mathbf{a}_0, \ldots, \mathbf{a}_t)$ the repeated game's *history* through time t. Player i's payoff in period t, after actions $\mathbf{a}_t$ have been performed, is represented by $w_i(\mathbf{a}_t)$. At the end of each period, all players are perfectly informed about the other players action performed in that period. That is, we consider a world in which at each point in time *every player perfectly knows exactly what the other player has done so far*. Assume that player i's preferences are represented by the expected payoff from the entire interaction, which is given by

$$\mathfrak{w}_i = w_i(\mathbf{a}_0) + \sigma w_i(\mathbf{a}_1) + \sigma^2 w_i(\mathbf{a}_1) \ldots = \sum_{t=0}^{\infty} \sigma^t w_i(\mathbf{a}_t)$$

Denoting the expected duration of the interaction by $\mathfrak{d}$, it holds that

$$\mathfrak{d} = 1 + \sigma + \sigma^2 + \ldots$$

and multiplying both sides with σ we obtain $\sigma\mathfrak{d} = \sigma + \sigma^2 + \sigma^3 \ldots$, such that

$$\mathfrak{d} = 1 + \underbrace{\sigma + \sigma^2 + \ldots}_{\sigma\mathfrak{d}} \Leftrightarrow \mathfrak{d} = 1 + \sigma\mathfrak{d} \Leftrightarrow \mathfrak{d} = \frac{1}{1-\sigma}$$

We can use this to write the expected per-period average payoff as

$$\bar{\mathfrak{w}}_i = (1-\sigma) \sum_{t=0}^{\infty} \sigma^t w_i(\mathbf{a}_t)$$

which can be meaningfully compared to the payoffs of the stage game.

As a first step, assume the players have only the two completely unconditional strategies available: always cooperate (ALC) or always defect (ALD). In case the coplayer plays ALD, it is a strict best response for player i to play ALD as well. To see this, observe that i realizes an expected average payoff of zero by playing ALD, and an expected average payoff of $-\kappa < 0$ by playing ALC. Likewise, in case the coplayer plays ALC, i realizes an expected average payoff of $\eta - \kappa$ by playing ALC, and an expected average payoff of $\eta < \eta - \kappa$ by playing ALD, such that it is strict best response for player i to play ALD. In other words, ALD is strictly dominant over ALC.

But now suppose that instead of the unconditional ALC, players have the following conditionally cooperative strategy available called *grim trigger* (GRIM): begin with cooperating initially, keep cooperating if the coplayer reciprocated in the previous period, and punish defection with perpetual defection. Thus, GRIM rewards reciprocation and threatens to punish cheating with permanent shunning. If the coplayer plays GRIM, ALD yields a payoff of η in the initial period and zero thereafter for player i, whereas GRIM yields $\eta - \kappa$ forever. Thus, GRIM is a strict best response to GRIM if and only if

$$\frac{\eta - \kappa}{1 - \sigma} > \eta \Leftrightarrow \sigma > \frac{\kappa}{\eta} \tag{1.1}$$

Thus, if this condition holds, GRIM is a best response to itself and both players playing GRIM therefore a Nash equilibrium. In this equilibrium both players permanently cooperate, deterred from cheating by the coplayer's threat to retaliate in the future. Note, however, that both players playing ALD remains an equilibrium as well: if the coplayer plays ALD, playing GRIM yields player i a payoff of $-\kappa$, whereas ALD yields a zero payoff, such that ALD is still a best response to itself. Without further assumptions we cannot say which equilibrium will be played.[9]

It is straightforward to include impatience or myopia into the model by assuming that players discount future payoffs by a common discount factor $\delta \in]0, 1[\subset \mathbb{R}$, such that player i's expected payoff becomes

$$\mathfrak{w}_i = \sum_{t=0}^{\infty} \sigma^t \delta^t w_i(\mathbf{a}_t)$$

and condition 1.1 adjusts to

$$\sigma\delta > \frac{\kappa}{\eta} \tag{1.2}$$

Note that the players' patience and the probability of continuation are substitutes in fulfilling the above condition.

[9] Axelrod & Hamilton (1981) provided an evolutionary argument for equilibrium selection: kin altruism might have helped reciprocity to get started. In closely related and stable family bands, reciprocators can invade an all defector population, and once reciprocators became prevalent, their advantage no longer requires relatedness.

GRIM is the most unforgiving punishment strategy one can imagine, but much more forgiving strategies work as well. Consider, for example, the most forgiving retaliatory strategy possible, the beforementioned strategy *tit-for-tat* (TFT). The exercise to show that TFT is a best response to itself is identical to the one above: if the coplayer plays TFT, responding with TFT yields payoff $\eta - \kappa$ permanently for player i, while ALD yields η in the initial period and zero afterwards, such that universal TFT is a Nash equilibrium if and only if condition 1.2 holds. Again, this equilibrium is fully cooperative, but permanent mutual defect remains an equilibrium as well.

A lot of other trigger strategies have been explicitly proposed and shown to be able to enforce cooperation.[10] In fact, the number of possible trigger strategies is unlimited, but all lead to the same simple rule: condition 1.2. Generally, Fudenberg & Maskin (1986) showed that trigger strategies (of whatever specific form) can sustain any average payoff vector $(\bar{\text{w}}_1, \bar{\text{w}}_2)$ that Pareto dominates $(0,0)$, up to the full cooperation outcome $(\eta - \kappa, \eta - \kappa)$, in a subgame perfect equilibrium of the supergame, provided $\sigma\delta$ is sufficiently close to unity.[11] In other words, under those conditions *any* degree of cooperation can be obtained as an equilibrium outcome, including enduring full cooperation. This result is known as the *folk theorem*, because the authors noted that they have written down what already has been folklore for years among game theorists.

Such trigger strategy equilibria can also be interpreted as embodying a notion of reputation (Vega-Redondo, 2003): after any history in the game, each player forms her or his expectations about the coplayers' future behavior (their reputations from the perspective of the player considered, see above) based on the latters' behavioral record. For example, in the grim trigger strategy equilibria outlined above, a player's reputation can be interpreted as being «good» as long as (s)he sticks to the cooperative equilibrium path, but immediately shifts to «bad» after just one deviation. In this sense, a trigger strategy can be interpreted as a behavioral rule that prescribes to cooperate with coplayers with a «good» reputation, and to defect on players with a «bad» reputation. This interpretation helps to draw the following two connections. First, it provides for an alternative view on the stark differences between finitely and

[10] In a famous agent-based computer simulation tournament (Axelrod, 1984), TFT turned out to be the most successful in the set of competing strategies. For the present purposes it should be noted that this tournament was implemented with perfect monitoring, that is, arbitrary conditioning on past play was feasible.

[11] Note that I have only illustrated a Nash equilibrium argument above. In principle, such an equilibrium may involve non-credible threats and promises, that is, it may not be in the interest of a player to carry out the punishment phase once an instance of cheating occurred. However, it is easy to extend the demonstration to subgame perfectness by verifying that there is no single period where one of the players can make a profitable deviation from the equilibrium strategies (the «one-stage deviation principle»). Fudenberg & Maskin (1986) did this. The original result is also much more general in other respects: it holds for any finite number of players and any kind of stage game payoff function. The idea is that once one player deviates, *all* the other players impose the punishment payoff on the defector. It asserts that any average payoff vector that is an element of the feasible and individually rational payoff set, in the PD the convex hull of the four stage game payoff vectors bounded by zero, can be sustained by an equilibrium in the repeated cooperation game, provided $\sigma\delta$ is sufficiently close to unity.

indefinitely repeated games and the imitation equilibria considered above. At any point in time in a repeated game, and given that *Two* plays grim trigger or TFT, the value to *One* of maintaining a «good reputation» by cooperating (maintaining a «clean» record) is equal to the expected present value of a stream of mutual cooperation payoffs $\eta - \kappa$ accruing over the (expected) remaining duration of the game. As shown above, in an indefinitely repeated game with continuation probability σ, this expected duration is equal to $\frac{1}{1-\sigma}$ in *any given period*, that is, it is *constant* over time. It follows that the value of maintaining a good reputation is also constant over time, namely equal to $\frac{\eta-\kappa}{1-\sigma}$, and if this value is no smaller than the investment required to maintain a good reputation, which is equal to the immediate gain from cheating η, then investing in the maintenance of a good reputation by cooperating is a best response *in any period*. This is just another way to describe the conditions 1.1 or 1.2.

In principle, the same applies to the finite horizon case. However, under complete information it is not credible for *Two* to actually stick to the trigger strategy but the only credible strategy is universal defect (by backwards induction), such that maintaining a good reputation by cooperating is worthless to *One* in any period. Thus, since the required investment is by assumption greater than zero, but maintaining a good reputation pays off with zero returns, *One* will never invest. This is just another way to describe the unique sequential equilibrium in a finitely repeated PD. If there is, however, some small probability that *Two* might be committed to a trigger strategy, then we know that the value of a good reputation can be sufficiently valuable such that cooperation is profitable for some fraction of the finite duration of the game (see section 1.2.2).

Second, since the players condition their behavior on the other's reputation, which is in turn inferred from the other's past actions, or more precisely from the available *information* about the other's past actions, a sufficient amount of such information seems essential for reputation building. This interpretation will be useful in the remaining sections.

1.2.4 Conclusion

Summing up, in a perfectly transparent world in which every action is perfectly observed by all other players, strategies that condition behavior on this information can enforce cooperation if one of two conditions are met: (i) The players believe that their coplayers might be precommitted to play a conditionally cooperative strategy, or (ii) the horizon of the interaction is indefinite. If neither condition is satisfied, unconditional defection is the unique (sequential) equilibrium outcome of any repeated PD, which is another way to say that players do not care about what the others did before. If (i) is satisfied, then cooperation over a large fraction of a finitely repeated PD is possible (even among entirely selfish players) by an equilibrium process of «reputation maintenance» in which players base their predictions of their coplayers' future play on information about their past play, and therefore have an incentive to imitate a behavioral record of a «precommitted type».

If (ii) is satisfied, the folk theorem shows that (even entirely selfish) players may

sustain any degree of mutual cooperation by means of strategies that reward cooperative behavior with reciprocity and punish non-cooperative behavior by (some periods of) shunning.[12] It formalizes the intuition of tactical cooperation, «I scratch you back if you scratch mine», in personal relationships. One may expect such tactical cooperation to be more likely the higher the gains from cooperation for a given $\sigma\delta$, and for given κ and η, more likely the «larger» the shadow of the future, that is, the closer $\sigma\delta$ is to unity.

With respect to our guiding question (What does the propensity of individuals in a given group to cooperate with one another has to do with the availability of information they have about each others' actions?), it can be concluded that all cooperation in these setting hinges on the fact that all actions are perfectly transparent and hence each player is perfectly informed about her or his coplayers' history. Players use this information to condition their own behavior.[13] This conditioning provides for incentives to cooperate along the equilibrium path.

Those results do *not* readily extend to setting in which information about the coplayer's behavior is limited. Intuitively one might expect that applying rewards and punishments is more difficult if information about the coplayer's behavior becomes limited. As a result, there might be phases of retaliation based on false accusations, incentives to cooperate might be diluted, and hence cooperation much more difficult to sustain. Whether this is correct is the subject of the following section.[14]

1.3 Imperfect public monitoring

Monitoring is called *perfect* if each action taken by each player is observed by all other players automatically, immediately and without error, which implies that at any point in time t all players perfectly know the entire history of the interaction $(\mathbf{a}_0, \ldots, \mathbf{a}_t)$. This is the case I considered in the previous section. It is stunning that a simple strategy such as TFT is able to enforce cooperation among purely selfish individuals in this setting. However, the sensitivity of the strategy to precise monitoring has soon be highlighted after its «invention» (Selten & Hammerstein, 1984; Fudenberg & Maskin, 1990):[15] TFT cannot correct errors, such that one wrong signal (or accidental defection) leads to a long (possibly infinite) sequence of retaliation. A number of modified strategies that tolerate some «defect» signals have been proposed (Fudenberg & Maskin, 1990; Nowak & Sigmund, 1992, 1993). In particular, more forgiving strategies may perform better than harsh ones under noisy monitoring, because the latter are too quick in punishment of apparent defections such that cooperative equilibria resting on such strategies become extremely fragile (Bendor et al.,

[12] This result can be viewed as both a success and a failure. I come back to this in the concluding section 1.5.

[13] To be precise, this is the case in all equilibria but one: the universal defect equilibrium.

[14] Following the literature, I will focus on the indefinite horizon case in what follows.

[15] The argument works not only with monitoring imperfections but with any kind of error in perception or performance.

1991). Thus, since more tolerant strategies can avoid such escalations, but necessarily invite exploitation, there seems to be a trade-off under imperfect monitoring (Axelrod & Dion, 1988). However, Bendor (1993) found counterexamples in which imperfect monitoring can actually facilitate cooperation in a population of harsh strategies, and concludes that «the idea that inferential uncertainty always harms nice strategies and always impairs the evolution of cooperation must be sharply qualified.» Indeed, later work in game theory has shown that monitoring imperfections *per se* need not to be a serious problem, as long as all players get the *same* noisy signal.

Like many other ideas that were later generalized in game theory, this case was motivated by an application in industrial organization. In a seminal contribution, Stigler (1964) suggested that tacit industrial collusion may be difficult to sustain when price cuts can be made in secret, because there may be erroneous punishment phases («price wars») if there is uncertainty whether cheating has actually occurred. Porter (1983a) and Green & Porter (1984) took up this idea and considered a Cournot duopoly in which quantity choices are private information and demand is uncertain. Specifically, in each period, firms first choose output levels secretly, and then a random demand shock is realized and the market price determined. The shock is unobserved by all players (they share a common belief, though) but the market price is commonly observed. Monitoring is public in this case because all players observe the same signal, the market price, but it is imperfect because (due to demand shocks) the firms' actions can only be deduced with error.[16] In sum, the market price provides a public but imperfect signal about the players' actions. Green & Porter (1984) showed that in such situations firms still can maintain some collusion using statistical inferences on unobserved actions and punish potential deviations by aggressive output choices.[17] Key is the publicly observed signal which is correlated with the unobserved actions, in this case the market price (see also Green & Porter, 1984; Abreu et al., 1986).[18]

Abreu et al. (1990) provided an early generalization of this idea of imperfect public monitoring in a standard generic repeated game framework.[19] They allowed monitoring to be imperfect, in the sense of an exogenous probability of error, but retained the assumption that (possibly erroneous) signals are observed *publicly by all players*. Specifically, a player in the game cannot observe other players' actions

[16] Without demand stochasticity monitoring would be quasi-perfect because the rival's action could be perfectly inferred from the price. Through a random demand shock, such inference becomes noisy because a low price may be caused either by low demand or by a rival's cheat.

[17] Specifically, the equilibrium trigger strategies put forth are as follows: Firms produce at the jointly monopolistic level as long as the realized price is above a certain threshold, but revert to the one-shot Cournot quantity for a fixed number of periods when it falls below the threshold. Due to the random component in demand, false accusations and periodic price wars, i.e. punishment, occurs *on* the equilibrium path. The latter is an efficiency loss compared to trigger strategy equilibria under perfect monitoring where punishment never occurs on the equilibrium path.

[18] There is a body of empirical work following the Green-Porter model, most of it based on data from the 1880's Joint Executive Committee (JEC) railroad cartel (Porter, 1983b, 1985; Lee & Porter, 1984; Cosslett & Lee, 1985; Ellison, 1994b; Vasconcelos, 2004).

[19] I follow Kandori (2002).

directly but after performing her or his action each player i observes a signal $s_{it} \in S_i$. Monitoring is called *public* if the signal generated by one player's action is identical for all players, that is, $s_{1t} = s_{2t}$ (generally $s_{1t} = \ldots = s_{it} = \ldots = s_{nt}$). Thus, denoting the signal vector $(s_{1t}, s_{2t}) := \mathbf{s}_t$, the sequence $\{\mathbf{s}_0, \ldots, \mathbf{s}_t\}$ is the *public history* of the supergame through time t. In this setup, the perfect monitoring case considered above is obtained by assuming $s_{it} = \mathbf{a}_t$ for all i and t. Observe that perfect monitoring ($\mathbf{s}_t = \mathbf{a}_t$ for all t) implies public monitoring, but not the reverse, since monitoring may be public but imperfect ($\mathbf{s}_t \neq \mathbf{a}_t$ for some, possibly all, t). Assuming that $\mathbf{s}_t$ is realized conditional on the current action profile $\mathbf{a}_t$, the probability of a particular signal vector $\mathbf{s}_t$ being $\pi(\mathbf{s}_t | \mathbf{a}_t)$, and letting $w_i(a_{it}, s_{it})$ represent player i's realized payoff in case (s)he performs action a_{it} and receives signal s_{it}, each player i's expected stage game payoff is given by[20]

$$\hat{w}_i(\mathbf{a}_t) = \sum_{\mathbf{s} \in S} w_i(a_{it}, s_{it}) \, \pi(\mathbf{s}_t | \mathbf{a}_t)$$

Defining the expected payoff over the entire interaction as before using this new stage game payoff function, Fudenberg et al. (1994) extended their earlier folk theorem to this case (see Fudenberg & Yamamoto, 2011, for generalizations), by showing that essentially the same region of payoffs (i.e. in the repeated PD any strictly positive average payoff vector up to the full cooperation payoff) can be sustained in a supergame equilibrium with the additional qualification that the signal space S has sufficient cardinality, that is, there are sufficiently many signals possible.[21] In addition, the qualifier «essentially» means that the full cooperation payoff can be approximated as closely as desired (but never reached entirely). In other words, monitoring imperfections indeed result *always* in an efficiency loss, since punishment can occur on the equilibrium path, which can be made arbitrarily small, however.

With respect to our guiding question (What does the propensity of individuals in a given group to cooperate with one another has to do with the availability of information they have about each others' actions?) it follows that the basic conclusion from the previous section carries over to a world in which the players cannot observe directly what the other player has done so far but receive some noisy signal from a commonly known conditional distribution which is the same for all players. Then again entirely selfish players may sustain (almost) any degree of mutual cooperation by means of strategies that condition rewards and punishments on the information about the coplayer's behavior *available*, provided condition 1.2 holds. However, as the original folk theorem for the case of perfect monitoring, it does only claim that (almost) all degrees of mutual cooperation are *possible*, it does not provide a sharp prediction on the level of cooperation one might expect to actually observe under a given level of monitoring inaccuracy. Specifically, the approach taken is that one fixes

[20] This specification assures that the realized payoff does not convey any information on top of what is already contained in a_{it} and s_{it}. Note that one can, without loss of generality, redefine the signal as $\hat{s}_{it} = (s_{it}, w_i)$. See Kandori (2002).

[21] Specifically, for the PD the folk theorem of Fudenberg et al. (1994) implies that for a generic signal distribution $\pi(\mathbf{s}_t | \mathbf{a}_t)$ the same set of payoffs can be approximately sustained under the usual conditions and if the S has a cardinality of at least three elements.

a desired level of cooperation and then shows that for any group size and for any error rate, there *exists* a $\sigma\delta$ sufficiently close to unity that this level of cooperation can be realized. Conversely, starting with a given $\sigma\delta$, the attainable payoffs may be quite low even when the group size and the signal error is small, let alone the general open question of how the players might coordinate on a particular supergame equilibrium.

But for empirical purposes it would be nice to have at least a hypothesis about what might actually happen if monitoring gets more or less noisy. A result by Kandori (1992b) is useful here. Drawing on the framework of Abreu et al. (1990), he formalized the idea that «improved» monitoring helps the players to enforce cooperation. As an illustration in the above example, suppose the demand becomes less stochastic so that low prices are a better signal of cheats. Can the firms achieve a more collusive outcome, perhaps with less punishment (price wars) along the equilibrium path? Indeed, Kandori (1992b) shows that the pure-strategy sequential equilibrium payoff set never becomes smaller and often expands (in the sense of set inclusion) if the quality of the signal improves (in Blackwell's sense). The intuition is simple: If cheating is detected more accurately, punishment becomes more directed, that is, there are less losses due to erroneous punishments and stronger incentives to cooperate *ex ante*. Therefore, we can summarize as follows: In repeated games with an indefinite horizon and imperfect public monitoring, the *available* information about the other player's action is used to play conditional strategies, and it can be hypothesized that cooperation will be more frequent if *more* information (in the sense of more accurate signals) becomes available.

1.4 Private monitoring

In the previous section, signals were allowed to be noisy, but every player still received the same (noisy) signal. Therefore, at each point in the game all players have the same information about the history of play. One may imagine this as a scenario in which the players are located in a class room and the public signal is written on the blackboard after each round. In this scenario, each player (i) knows what signals the coplayers received, and that (ii) signals were drawn always from the same commonly known distribution. Using statistical inferences, the players can, in principle, mitigate efficiency losses due to omitted and misdirected punishments to a large extent.

But what if different players receive different signals, that is, if $s_{it} \neq s_{jt}$ for $i \neq j$ and for some, possibly all, t (this includes the case where some players receive no signal at all, i.e. their signal is «infinitely noisy» and hence uninformative)? In this case monitoring is called *private*. To continue along the above collusion example, and following Kandori (2002), imagine a situation that is identical to the one above, except there is no single market price but secret price cuts can be offered to customers individually. The firms still cannot observe the coplayers secret offers, but this time there is also no publicly observed signal from which every player can form a identical belief; all a firm can do is to extract information from her *own* sales only.

It is apparent that in this case each player is neither sure about the actions per-

formed by the other players, *nor about the signals the other players received*. In other words, each player has different information about the history of play, and no player knows what information her or his coplayers have. For instance, suppose *One* and *Two* play the PD and *One* receives information that *Two* defected, whereas this signal might be wrong. If *One* believes this information and chooses to punish *Two*, then *Two* does not know whether *One* defected «on the cooperative equilibrium path» because she received a defect signal or whether she left the «cooperative equilibrium path». It is apparent that reasoning of this kind becomes extremely complex as the game unfolds both for the players and the analyst.[22] The literature in this field is consequently very limited.

I will illustrate this case by means of a setting that might intuitively come into mind when thinking of situations in which «individuals can hardly observe what the others are doing»: A large group (community or population) of individuals that meet in smaller groups in each period, and everything that happens within this subgroups is private information to them, unobserved by the other individuals in the larger group. This scenario is apparently very common in practice.[23] I will come back briefly to the (small) general literature on private monitoring at the end of this section.

1.4.1 Public and private monitoring in matching games

The scenario I just outlined has been formalized by means of *matching games* (Rosenthal, 1979), repeated games in which in each period the players in a larger group are matched together into subgroups to play a game such as the PD. It turns out that the number of players and the matching scheme is completely unimportant, that is, it does not matter whether the *same* individuals interact frequently or just once, provided that monitoring is *not* private (see Milgrom et al., 1990, and below).

But for the moment, observe that depending on the perspective, matching games may be viewed as both a generalization and a special case of the setting considered in the previous sections. When considering the matching game as a whole, it is a special case. When considering the game the subgroups are playing, it is a generalization, as the setting considered above, in which the *same* players play a game repeatedly, can be captured in the matching game setting. One way is to assume that in any period the *same* subgroups are matched. Then in each period and each subgroup the same players meet repeatedly as in the setting above. Another way is to assume that any pair (I assume the subgroups are pairs for sake of continuity) of players is matched in one out of τ periods, where τ depends on the size of the larger group. We may choose τ very large, such that the setting represents the practically important case in which any individual may interact frequently with a large number of other individuals, but

[22] The technical challenges are that (i) such games lack the convenient recursive structure that games with public monitoring have (see Abreu et al., 1990), such the method of recursive dynamic programming cannot be applied, and that (ii) the players, and therefore the analyst, must conduct very complex statistical inference about the coplayers' private histories (Kandori, 2002).

[23] A further reason why I consider this case in more detail is that it is very useful to interpret particular experimental designs with respect to their information structure. I will come back to this in chapter 2.

infrequently with any single one.

If we assume that players can recognize each other, and keep account of the outcomes of past interactions with particular coplayers, then the folk theorem results from the above setting can be straightforwardly extended to this case.[24] To illustrate, assume that monitoring is perfect within each match (but private in the larger group, as players only observe what happens in their own matches), and keep the fixed continuation probability σ. Then after any match the probability that the same two players will meet once again is equal to σ^τ, since the same pair of players only meets every τ periods. Furthermore, the present value at time t of a payoff realized in the next meeting, $w_i(\mathbf{a}_{t+\tau})$, is now $\delta^\tau w_i(\mathbf{a}_{t+\tau})$. Then, an argument parallel to that of above shows that cooperating is a strict best response if and only if

$$\frac{\eta - \kappa}{1 - (\sigma\delta)^\tau} > \eta$$

or equivalently

$$(\sigma\delta)^\tau > \frac{\kappa}{\eta} \tag{1.3}$$

Note that this condition is identical to condition 1.2 except that $\sigma\delta$ on the left hand side is now multiplied with itself τ times, such that condition 1.3 is more restrictive than 1.2 for any $\tau > 1$. Thus, the larger τ, that is, the less frequent the interaction between the same individuals, the more restrictive the above condition, and for sufficiently large τ, it does not pay to cooperate (note that one-shot games can be approximated by letting τ converge to infinity), such that direct retaliation fails to enforce cooperation.

However, it can be shown that if one drops the assumption that players are not informed about what happens outside their own matches, that is, if monitoring is public (in fact, perfect) then full cooperation can be sustained by *indirect enforcement*, that is, by a trigger strategy sequential equilibrium in which not the cheated player her- or himself but third parties gang up upon the defector to wipe away any gain from cheating (Milgrom et al., 1990; Kandori, 1992a; Ellison, 1994a; Okuno-Fujiwara & Postlewaite, 1995). In such equilibria, everybody cooperates by the justified expectation «give and you shall receive, cheat and you will be expelled».[25] But even more: the restrictions on $\sigma\delta$ are no more stringent than under direct retaliation, that is, *perfect monitoring can perfectly substitute for repeated interaction among the same individuals*. Thus, the restriction of the traditional folk theorem that the *same* players need to interact repeatedly turns out to be not critical but the symmetry of information about the players' behavior takes center stage.

[24] If the game is finite, it is apparent that the backwards induction argument also holds in this case: independent from the information a player has about the history of play (in particular the current coplayer's behavioral record), it is a dominant strategy to defect in the terminal period, and by backwards induction also in all other periods.

[25] This is closely related to the biological literature on *indirect reciprocity* (Alexander, 1987, see Nowak & Sigmund, 2005 for an overview).

The result was proved by Milgrom et al. (1990) and Kandori (1992a), showing that under perfect (public) monitoring *any* payoff vector that is for some $\sigma\delta$ supported by a sequential equilibrium in a repeated game with the same two players is also supported by a sequential equilibrium in the matching game with arbitrary matching group size and matching rule for *the same* $\sigma\delta$. The proof is so short that I can quote it here in Kandori's own words:[26]

> «The players in the matching game start by playing the equilibrium path of the two-player game. If the type-i player deviates ($i = 1, 2$), all type-i's are punished by all of the other type, using the same punishment strategies as in the two-player game. The same principle applies to any further deviations, and simultaneous deviations are ignored. Then, it is clear that the incentives of each player are identical with those in the two-player game because each player encounters the same sequence of action profiles as in the two-player repeated game, only the opponents being changed over time. Hence the prescribed strategies are in fact a perfect equilibrium.»

I find this result remarkable because it implies that it is the publicness of monitoring not the size of the population or the matching protocol that matters. Under public monitoring it is entirely unimportant who punishes a defector, that is, indirect enforcement in matching games works just as well as direct personal retaliation in a repeated game with a fixed set of players. In this sense, «even if no pair of traders come together frequently … then transferable reputations for honesty can serve as an adequate bond for honest behavior *if members of the trading community can be kept informed about each other's past behavior*» (Milgrom et al., 1990, p. 3, emphasis in the original), that is, «observability in the community is a substitute for having a long-term frequent relationship with fixed partners» (Kandori, 1992a, p. 68).

From this result one might be led to conclude that we should always expect more (or at least no less) cooperation in settings where monitoring is public compared to settings where it is private. There is a caveat to this intuition as Kandori (1992a) and Ellison (1994a) showed. I will briefly outline the idea.

1.4.2 Contagious equilibria

Consider a matching game with private monitoring. Specifically, to rule out the possibility for direct retaliation, assume that players cannot recognize each other, which approximates a situation in which each pair meets at most once (we might say that the individuals in each match are *perfect strangers*). Furthermore, each player observes only the history of actions in her or his own matches and information transmission or pooling among players is ruled out. Thus, the matching game is a game with *private monitoring*, since in each stage, each player's action is (perfectly or imperfectly) observed only by a *subset* of all other players, namely those with which the player is

[26] Kandori (1992a) proved also a version in which only defectors are punished.

matched in the current period. In other words, each player is only informed about a tiny part of the overall history: her or his personal experience.

Under these adverse conditions it appears that repetition of the stage game Nash equilibrium (i.e. zero cooperation) is the only possible outcome. But Kandori (1992a) and Ellison (1994a) showed that this need not to be the case. There is a possibility, for $\sigma\delta$ sufficiently close to unity, to sustain full cooperation in the population supported by a sequential equilibrium that may be called «contagious equilibrium». Here is the intuition how it works. Suppose players adopt a generalized grim trigger strategy that prescribes to cooperate as long as one defection is observed in the population, after which the strategy prescribes to defect perpetually. Since every player observes only the outcome of his or her own matches, a defection can only be detected by the coplayer that is currently matched to the defector. Thus, the defector cannot be immediately punished because the subsequent coplayers are not informed about the defection and hence cooperate. However, assuming that all players stick to the trigger strategy, the coplayer of the defector will defect in the subsequent periods, and all affected coplayers will defect in the periods after that as well, and so on. Thus, any defection «rebounds» on the defector through a contagious process of defection spreading through the population. Thus, although delayed, defections will be punished in much the same way as above, and if $\sigma\delta$ is sufficiently close to unity, is does not pay to defect. However, this does not establish a sequential equilibrium yet, because it must also be in the interest of every player to actually start off the contagious punishment phase (which was simply assumed above). That is, after observing a defection, it must be a best response to defect as well, that is, continuing to cooperate should not be beneficial. This turns out to be the case, because the original defector will continue to defect (the trigger strategy requires to defect in response to detected defections, including one's own) and hence for the defector's coplayer continuing to cooperate the next period delays the contagion but cannot stop it. Thus, a player whose coplayer defected will continue to cooperate only if sufficiently impatient or the probability of termination is sufficiently high. It again follows that this is ruled out if $\sigma\delta$ is sufficiently close to unity.

It should be noted that the deterrence of cheats in a contagious equilibrium is stronger than under direct retaliation (embedded in the matching game), since in the latter a defector is punished only by the player (s)he cheated upon (which occurs only every τ periods in the above random matching example), while under the contagious equilibrium a defector forgoes also some of the future payoffs with other players (Kandori, 1992a). That is, if for a player condition 1.3 holds with equality under direct retaliation, such that the player is indifferent between cooperate and defect, (s)he strictly prefers to cooperate under the contagious equilibrium. Thus, there are situations in which cooperation is sustained by the contagious equilibrium, but not by direct retaliation.

However, the deterrence of cheats in a contagious equilibrium is weaker than in the indirect (community) enforcement case under perfect (public) monitoring, because punishment comes immediately in the latter but delayed in the former. The

delay is the greater the larger the population, such that the conditions for a cooperative contagious equilibrium are quite restrictive. In fact, Kandori (1992a) shows that for a large enough population there is no way of sustaining cooperation by a contagious equilibrium, because it takes a long time for the contagion to rebound on the defector. Thus, there exist cases in which cooperation can be sustained under perfect (public) monitoring but not under the private monitoring.

To summarize, the idea of a «contagious equilibrium» may be viewed in a way that each player does not trust specific coplayers but the «community as a whole», whereas a single defection lets this trust collapse in the whole population. Each individual that is cheated starts cheating all of her or his future coplayers. Apparently, cooperation sustained by such an equilibrium is extremely fragile. With a small degree of monitoring imperfection *within* matches (or «trembling hands»), a single error in one match lets cooperation collapse within the *entire* population. Thus, even Kandori (1992a) considers them as not very realistic. It can therefore be concluded that we should expect more (or at least no less) cooperation in settings where monitoring is public compared to settings where it is private, with the qualification that (unlikely) instances exist in which this is not the case.

1.4.3 Information transmission and publication

We may also take the contagious equilibrium as a benchmark to investigate cases between the boundary cases of entirely public and entirely private monitoring. To begin with, note that the weaker deterrence of cheats in a contagious equilibrium under private monitoring compared to the public monitoring case stems from the fact that punishment comes immediately in the latter but delayed in the former. This points toward the conclusion that mitigating the «privateness» (or increasing the «publicness») of monitoring increases deterrence in a contagious equilibrium. To see this, suppose that players still cannot identify each other but that more than one (the current coplayer), possibly all other players observe a defection in a given match. Then the contagion is accelerated, the faster the more players are informed, such that the critical $\sigma\delta$ above which full cooperation can be sustained as a sequential equilibrium can only decrease relative to the fully private monitoring case. In extreme, if all players get to know that a defection has occurred somewhere in the population, then the punishment phase begins in the very next period after the defection, such that contagious punishment has equal force than in the perfect monitoring case considered above. Thus, information transmission increases deterrence.

However, the fragility of the equilibrium remains, and in a sense information transmission may be viewed to even worsen things along these lines, since a single signaling error lets cooperation collapse *immediately*, while it takes some time under private monitoring. Thus, to get rid of those implausibly fragile contagion equilibria one must necessarily relax the anonymity assumption, that is, players need not only to detect *whether* a defection has occurred in the group but also *who* defected. In this case, the set of possible trigger strategies is significantly expanded because punishments can then be targeted on defectors while keeping to cooperate with other

cooperators.

Indeed, if one assumes the existence of some information transmission mechanism, then one can show that an appropriate adaption of a trigger strategy that prescribes to cooperate with a coplayer having a «good» reputation and defect towards with a coplayer having a «bad» reputation is a best response to itself. Generally, suppose that a «label» is attached to each player and the label carries all «necessary» information on a players past behavior (the setup follows Okuno-Fujiwara & Postlewaite, 1995). In each match, the two players observe each other's label and then play the stage game. After the game, the labels are updated as a function of the original labels and the actions performed. In principle, such labels can carry arbitrary amounts of information, including each players complete history. If such information transmission mechanism exists, then there can be an additional (it adds to the incentives arising through direct reciprocity) incentive to cooperate in order to cultivate a cooperative *reputation*, to «keep one's record clean».

The aim of much of the literature is to identify a class of very simple labels, that is, labels that transmit a minimal amount of information about the behavioral record, perhaps only a one dimensional statistic, and still suffice to sustain a cooperative equilibrium. There are two prominent specifications of such labels (see Nowak & Sigmund, 2005, for an overview), elaborated in evolutionary models, that have been shown to be able to support indirect reciprocity as an evolutionary stable strategy (and hence a perfect or even proper equilibrium, see Weibull, 1996 for the correspondence). The first is due to Sugden (1986), and is called *standing*. According to this mechanism, an individual who cooperated in the previous period is in good standing, and the only way an individual can fall into bad standing is by defecting on a coplayer who is in good standing. Note that an individual can always defect when her or his coplayer is in bad standing without losing her or his good standing status. In this more general setting the TFT strategy is replaced by the following standing strategy: cooperate if and only if your current coplayer is in good standing.[27] Despite it seeming simplicity, the standing model is still informationally demanding. It requires that each individual knows the current standing of each other individual in the group, the identity of each other individual's current coplayer, and whether each individual cooperated or defected against her or his current coplayer. Since reputation (standing in this case) hinges on monitoring, errors in determining the standing of individuals may be frequent if monitoring is imperfect. Particularly serious are errors in second-order information, that is, about the standing or behavior of the current coplayer's previous period coplayer. Then it becomes difficult to judge whether the current coplayer's defection in the previous period was warranted or not. This will occur with high frequency if information is partially private rather than public (not everyone has the same information).

The informational requirements of the second commonly studied mechanism, *image scoring* proposed by Nowak & Sigmund (1998), are somewhat milder in that

[27] If errors are possible, the strategy can be extended such that after accidental defections, it prescribes to cooperate unconditionally in the next period in order to restore the status as a player in good standing.

players need not know the reputation of cooperators' coplayers. Nowak and Sigmund show that the strategy of cooperating with others who have cooperated in the past, independent of the reputation of the cooperator's partner, is stable against invasion by defectors, and weakly stable against invasion by unconditional cooperators once defectors are eliminated from the population. Takahashi (2010) provides a general result by showing that cooperation can be sustained in a matching game if players are informed about their current coplayers' past play only.

What virtually all approaches in this stream of literature share is the assumption that there exists a mechanism or institution that processes the respective information honestly, without explaining its specific workings or where it comes from. This, of course, assumes parts of central problem away. In reality, information about others' behavior is not just available, non-available, or something in between, but is must be produced, acquired and transmitted. «Perhaps the most important question which is unanswered», concluded (Kandori, 1992a, p. 77) «concerns the way in which the information transmission ... is implemented.» But how this should occur in a population of selfish individuals has not been (and probably cannot be) shown (Bowles & Gintis, 2011). What does incentivize players to share information they have and deter them from spreading lies if it is in their interest? What does incentivize players to get informed in the first place? Those questions underlie all original contributions in chapters 3 through 7 and I will come back to this repeatedly below.

1.4.4 Conclusion

As mentioned above, obtaining general results for the case of private monitoring is technically challenging. There is a small but growing literature that tries to extend the folk theorem to this case (Fudenberg & Levine, 1991; Sekiguchi, 1997; Bhaskar & Obara, 2002; Piccione, 2002; Ely & Välimäki, 2002; Matsushima, 2004; Hörner & Olszewski, 2006, 2009; Yamamoto, 2007; Kandori, 2011). Fortunately, for the present purposes they provide not much further insights than the above illustrations. Three related facts from this stream of literature can be summarized. First, the constructed equilibria are extremely fragile, since they generally rely on strictly mixed strategies, which appears to render coordination on a particular equilibrium extremely difficult. Second, the equilibria are engineered in a way that players are indifferent between acting on the information they receive about the other players' behavior and ignoring it, such that the slightest perturbation destroys this indifference and players behave the same way whatever signal they receive (Gintis, 2009). Third, the above results require restrictive conditions on the monitoring technology. In particular, almost all of them are based on «close-to-public» monitoring, that is arbitrarily small disturbances from public monitoring, whose actual order of magnitude is typically not specified or discussed (Gintis, 2009). But this just reinforces our earlier hypothesis: if a folk theorem can only be proved under approximately public monitoring, then anything that brings monitoring closer to public, to reduce the *asymmetry* in the distribution of information about the history of play, can be expected to be conductive to cooperation.

With respect to our guiding question (What does the propensity of individuals in a given group to cooperate with one another has to do with the availability of information they have about each others' actions?) we can henceforth conclude the following. The propensity of individuals in a given group to cooperate with one another is constrained, besides mere inaccuracy, in particular by the privateness of monitoring. If information about past actions is dispersed among the players in the group in the form of many small packages of private information, it becomes generally very difficult to coordinate on a cooperative equilibrium. On the other hand, the settings in which cooperation can be sustained under public (perfect) monitoring turned out to be much more encompassing, including not only situations in which the *same* individuals interact for an indefinite period of time but even situations in which the same individuals meet hardly more than once. Furthermore, mechanisms that honestly transmit information about the players' behavior to other players, to bring monitoring closer to public, render cooperation more likely. However, how such information processing should happen in a group of selfish individuals remains unexplained.

1.5 Conclusion

What does the propensity of individuals in a given group to cooperate with one another has to do with the availability of information they have about each others' actions? After having briefly reviewed the major theoretical approaches to cooperation in economics, the following can be wrapped up with respect to this guiding question.

In section 1.1 have started with outlining the generic problem of cooperation in a simple model. Mutual cooperation is beneficial to all, but cheating (assuming that cooperation is costly) always pays. In a selfish world cooperation therefore always needs to be enforced. Cooperating must provide for a return benefit (reward or avoided punishment) that renders cooperation profitable. I highlighted that if such enforcement needs to be performed by the individuals themselves, then information about the others' actions becomes generally critical, since rewards and punishments are by definition conditioned on past play.

In the remainder of this chapter, I investigated the qualifications and fine-structure of this tentative conclusion. Beginning with a setting in which players interact repeatedly and every action is perfectly observed by all other players, it was shown that strategies that condition behavior on this information can enforce cooperation if one of two conditions are met: (i) The players believe that their coplayers might be precommitted to play a conditionally cooperative strategy, or (ii) the horizon of the interaction is indefinite. If neither condition is satisfied, unconditional defection is the unique (sequential) equilibrium outcome of any repeated PD, which is another way to say that players do not care about what the others did before. If one of them is satisfied, then significant cooperation is possible by the application of conditional strategies that reward cooperative behavior and punish non-cooperative behavior. Since all cooperation in these settings hinges on the fact that all actions

are perfectly transparent and hence each player is perfectly informed about her or his coplayers' history, one might expect that applying rewards and punishments is more difficult if information about the coplayer's behavior becomes limited, and as a result cooperation might be much more difficult to sustain.

I investigated to what extent this intuition is correct in sections 1.3 and 1.4. In section 1.3 I considered a setting in which the information about the other players' actions is limited in one specific way: it may contain errors but all individuals in the group receive the *same* signals. In this case, monitoring is said to be *imperfect* but *public*. The literature has found that the folk theorem extends to this case, that is, (almost) any degree of mutual cooperation is possible by means of strategies that condition rewards and punishments on the information on the coplayer's behavior *available*. The result does not allow for predictions about which cooperation level can be actually expected for a given degree of information accuracy, or how the degree of cooperation may change if information becomes more (or less) accurate. For empirical purposes, it is important to observe that under imperfect public monitoring punishment can occur at times *on* the equilibrium path such that actual cooperation may be quite low even when noise is small. A related result provides a reason to hypothesize that the propensity of individuals in a group to cooperate with another will at least not decease as information about one anothers' actions becomes more accurate.

In section 1.4 I turned to a setting in which the information about the other players' actions is limited in a more general way: it may not only be inaccurate but different players may also receive different information about particular players' actions. In this case, monitoring is said to be *private.* If information about past actions is dispersed among the players in the group in the form of many small packages of private information, it becomes generally very difficult to coordinate on a cooperative equilibrium. On the other hand, the settings in which cooperation can be sustained under public (perfect) monitoring turned out to be much more encompassing, including not only situations in which the *same* individuals interact for an indefinite period of time but even situations in which the same individuals meet hardly more than once. Furthermore, mechanisms that honestly transmit information about the players' behavior to other players, to bring monitoring closer to public, render cooperation more likely. However, how such information processing should happen in a group of selfish individuals remains unexplained (I come back to this in the second remark below).

I conclude with two remarks. First, all cooperation in the models reviewed in this chapter is tactical in the sense that individuals only cooperate with one another if it is in their (long-term) self-interest. Two related questions follow immediately: (i) Do real people actually perform such tactics? (ii) Do real people *only* cooperate for tactical reasons? Concerning (i), the repeated game models provide for little restrictions on the set of attainable outcomes. The folk theorem shows that virtually «anything goes», which can be interpreted as both a success and a failure. On the one hand, it tells us that (provided the conditions are satisfied) equilibria with significant cooperation along the equilibrium path *exist*, that is, it is *not impossible* for entirely selfish

players to coordinate on a fully cooperative outcome. In principle, this is good news. However, in terms of explanatory/predictive power, the fact that «anything goes» can also be interpreted as the theorem's central weakness: since the model is compatible with a myriad of very diverse outcomes, it does not pose much testable restrictions on empirical data, such that its empirical usefulness can be questioned (Vega-Redondo, 2003). The theorem does not suggest how the equilibria whose existence it demonstrates can actually be accessed and sustained, and in particular it does not predict which one(s) from the continuum of equilibria is more likely to be observed than another. The question of how people can ever coordinate on one of the equilibria is one of the central open questions in game theory. In addition, and concerning (ii), the traditional behavioral assumptions, in particular universal selfishness, have become under criticism. In any case, what is needed is rigorous evidence about how real people actually behave in the generic circumstances outlined in this chapter. Do they really cooperate tactically only, that is, only if there is a prospect for a material return? What strategies do they actually play in repeated games? What information about the coplayers' actions to they use? How do they respond if such information becomes limited? In chapter 2 I selectively review the body of received experimental evidence in order to finds answers on these questions.

The second remark concerns how information structures are typically modeled: as an exogenous parameter. In the previous section I already mentioned that this is exemplified, *inter alia*, by assumptions about honest information acquisition and transmission. In reality, information generally, and information about others' behavior in particular, is not just available, non-available, or something in between, but is must be produced, acquired and transmitted. This implies that transparency, as defined here, is not just there or absent, it is *created* by the players themselves. But how this should occur in a population of selfish individuals has not been (and probably cannot be) shown (Bowles & Gintis, 2011). What does incentivize players to share information they have and deter them from spreading lies if it is in their interest? What does incentivize players to get informed in the first place?

This problem is vividly illustrated by Milgrom et al. (1990) in their account on the revival of trade in Europe during the early middle ages. During this time, merchants evolved their own private code, known as the *Lex Mercatoria*, that was decentrally enforced by mechanisms similar to the ones outlined in section 1.4.3. Importantly, enforcement in the merchant community relied on shared information about individual members' conduct. But given that getting informed (monitoring) and information transmission is generally costly, what incentives did the merchants have to engage in such activities? The received theoretical literature is largely silent on this issue because it considers information structures generally as exogenous (in turns out that the situation is not much different in the experimental literature, which I will review in the next chapter). It thereby abstracts from the possibility that players may alter it themselves. In other words, the standard approach in the theoretical literature reviewed above is to confront the players with a given information structure, and investigate how they behave *without adjusting the strategy space relative to the*

case of perfect monitoring. However, when studying settings in which the players information is imperfect it seems intuitive that at least one extension of the strategy set is natural: to acquire better or additional information! Doing so might have a number of interesting implications. For example, there must be some incentives «to become adequately informed about how others had behaved» even though this «might be personally costly» Milgrom et al. (1990, p. 1). Investigating the nature of those incentives might be an interesting avenue for research. Milgrom et al. (1990), for instance, argued the system of judges used to enforce commercial law before the rise of the state provided incentives to disclose evidence against violators of the code to the community and to become informed in the first place. They documented that the system was indeed very successful in enforcing honest behavior. This view puts the institutional and technological environment as a determinant of information acquisition and transmission, and therefore the degree of «equilibrium transparency», center stage. This conception is the motivation for all original contributions in chapters 3 through 7 and I will come back to it at the end of the following chapter.

Chapter 2

Transparency & Cooperation: Empirical Insights

«Logically, the conclusions follow from the assumptions. But empirically, scientifically, the assumptions follow from the conclusions!»

ROBERT AUMANN (in van Damme, 1998, p. 206)

In this chapter, I take up the questions raised at the end of the previous chapter and selectively screen the empirical literature in order to find answers: Do real people actually perform such tactics as envisaged in the theoretical models? Do real people *only* cooperate for tactical reasons? What strategies do they actually play in repeated games? What information about the coplayers' actions to they use? How do they respond if such information becomes limited? In other words, I review the empirical literature in order to investigate what strategies real people *actually* play in the generic situations described in the previous chapter. I proceed in roughly two steps.

As a first step, I focus on games with perfect monitoring and investigate when and to what extent subjects play strategies that are conditioned on information about the coplayers' past actions. This is a useful first step in order to identify situations in which the amount of information players have about the history of play likely have decisive consequences. In section 2.1 I review evidence of behavior in repeated games with perfect monitoring. Results show that a majority of subjects actually cooperates tactically. They use the available information about the coplayers' past behavior to condition their current behavior. However, there is a persistent residuum of cooperation even in (approximated) one-shot interactions which cannot be accounted for by tactical considerations. In section 2.2 I consider this residuum in more detail, and put special emphasis on evidence that is informative with respect to our guiding question. A large body of evidence shows clearly that people do *not always* cooperate for tactical reasons, but there is also evidence that such cooperation is conditioned on coplayers' behavior anyway.

If some subjects are willing to incur a cost to confer a benefit to someone else, are there also subjects who are willing to incur a cost to impose a cost on someone else? I review the evidence on this question with a focus on the degree of conditioning on coplayers' past behavior, in which case it can serve as a punishment, in section 2.3. The evidence shows that such spiteful behavior is very common in various subject pools worldwide. Furthermore, it is frequently used as a punishment of non-cooperative behavior, even by unaffected third-parties, with strongly disciplining effects. It follows predictable patterns, despite the fact that pecuniary incentives seem to play a significantly smaller role than for cooperative behavior. But there are exceptions and evidence that the degree of conditioning varies, individually and spacially.

In sum, the evidence from experiments under perfect monitoring shows that if information about the other players' past behavior is readily available, this information is used in order to play a discriminatory strategy, even in situations in which material returns from doing so are ruled out. As a result, there are pecuniary incentives to cooperate, which is reflected in the notably higher frequency of cooperation in repeated games compared to (approximated) one-shot games. But there is also significant cooperation in the latter, in particular if costly punishment is available. In one sentence: if information about the other players' past behavior is readily available, cooperation generally occurs and can be very high if the interaction is repeated.

Although the study of perfect monitoring settings is a useful first step, it is not sufficient to draw definite conclusions about conditions in which information about coplayers' actions is *actually* limited. This is because players might adapt to the situation by adopting entirely different strategies, such that we cannot just interpolate behavioral tendencies from a perfect monitoring to an imperfect monitoring setting. A hint in this direction comes from the evolutionary literature touched in the previous chapter, which suggests that more forgiving strategies may perform better than harsh trigger strategies in environments in which monitoring is noisy. Computer simulations with evolving finite automata playing PD population games confirm that informational frictions can lead to significant differences among the evolving strategies (e.g. Bendor et al., 1991; Bendor, 1993; Miller, 1996). Hypotheses in the similar direction has been advanced in cultural anthropology (Boyd & Richerson, 1988a, 1995): in situations in which monitoring is noisy or costly, rules-of-thumb or beliefs about the «right» action (eventually with moral imperative) may be adaptive. As a second step, I will therefore review experiments that *actually* limit the players information about past play in different ways. Research in this area is quite limited to date, and all four experimental studies reported upon in the second part of the thesis contribute to this literature. So what happens if such information is limited?

While imperfect monitoring is an issue in the theoretical literature for quite some time (sections 1.3 and 1.4 in the previous chapter), there is only a very small (but growing) experimental literature, which I review in sections 2.4 and 2.5. Section 2.4 is devoted to cooperation experiments in which the signals the players receive about their coplayers' behavior are subject to some random disturbance (corresponding to

the setting outlined in section 1.3). In PD experiments with an indefinite horizon and imperfect public monitoring, the available evidence shows that there is cooperation at any level of noise, but always less than maximal. Furthermore, that the frequency of cooperation decreases as noise increases, and cooperation in conditions with very inaccurate information are not higher than under one-shot play. This supports the hypothesis derived in section 1.3, and poses empirical restrictions on the predictions of the folk theorem under imperfect public monitoring. Furthermore, subjects seem to resort to more lenient and forgiving variants of the classical TFT and grim trigger strategies, that wait for a coplayer to defect two or three times before reverting to punishment mode, and return to cooperation after a punishment phase has occurred. These results support the above claim that merely interpolating results obtained in a perfect monitoring setting to ones with informational constraints is problematic. This conclusion is reinforced by the available evidence on imperfect monitoring in cooperation games with an option to reduce coplayers' payoffs. It is shown that payoff reductions become *more* frequent and cooperation less frequent as monitoring gets more noisy.

Finally, in section 2.5 I consider studies that implement conceptualizations of *private* monitoring. The received evidence is affirmative with respect to the hypothesis that information transmission, or the publication or disclosure of behavioral records, may facilitate cooperation relative to conditions in which monitoring is private. The frequency of cooperation is generally lowest in private monitoring conditions and information transmission in some form or the other increases cooperation. If matching is non-random, disclosure also reduces clientelization effects observed under private monitoring. The evidence is more ambiguous with respect to the question whether random matching games with public monitoring provide for equally efficient results as repeated interaction among the *same* individuals, as suggested by the theorem of Kandori (1992a) reviewed in section 1.4.1. To summarize with respect to our guiding question, the available evidence suggests that the level of cooperation (and efficiency) is generally increasing in the accuracy and symmetry of information players have about each others' actions.

Wrapping up, the insights obtained feed back on the two divergent views (expressed in the general introduction) about *why* people cooperate, and what role information about coplayers' behavior might play accordingly. The evidence shows that both views have merit. People cooperate to a significant extent tactically, use reciprocity strategically to realize mutual benefits as envisaged by game theory, and thereby draw strongly on information about coplayers' past behavior. But people also forgo own material gains to cooperate or reduce incomes of others. A significant fraction of those latter behaviors are conditioned on information about coplayers' past play anyway. Thus, there is cooperation even in situations in which people have minimal or no information about others' behavior, but increasing such information generally facilitates cooperation significantly. However, returning to an issue raised at the end of the previous chapter: what if players have the opportunity to later the information structure they are confronted with themselves?

Before I begin, one remark is to be made. «Science is the knowledge of consequences, and dependence of one fact upon another», puts it Thomas Hobbes over four centuries ago (Hobbes, 1651, ch. 5). Thus, the challenge of any empirical science is (i) precise and valid *measurement*, and (ii) *identification* of causality. Both are a peculiar strength of the *experimental method*. Experiments are strong in precision and validity of measurement because the researcher controls (or even completely designs) the data generating process, and observes and records the data by her- or himself. The researcher therefore knows exactly how the data set to be analyzed has been constructed.[1] Well designed experiments also allow for causal inferences with a minimal set of identifying assumptions, as focal factors can be truly exogenously varied, and the impact of confounding variables can be minimized by means of direct control (observables) and randomization (unobservables). In sum, it approximates the ideal causal inference model in which one parameter is varied at a time while holding all other influences constant as far as possible (Rubin, 1974, 2008). I therefore focus strongly on experimental research, because a reasonable degree of measurement precision as well as control of incentives and information conditions is critical for obtaining empirically sound results in the domain under consideration here. The standard approach in experimental economics combines the general methods of behavioral experiments with pecuniary incentivation, anonymity (except if its relaxation is a focus of the study), and strict control of information flows. However, the experimental approach has also been subject to criticism in economics, that has, however, less to do with the experimental method itself but more with generalizations drawn from experimental results. I address the concerns by considering evidence from field experiments as far as available. With respect to the settings considered in this chapter, the key behavioral patterns are generally observed in both the lab and the field alike. Nevertheless, generalizations are always to be made with due care (see e.g. Schram, 2005; Guala & Mittone, 2005, for discussions of the issue).

2.1 Mutual enforcement under perfect monitoring

I begin by taking up two fundamental questions with which I concluded chapter 1: (i) Do real people actually perform such tactics as envisaged in the conventional approaches to cooperation in economics? (ii) Do real people *only* cooperate for tactical reasons? In this section I will focus on (i) and only touch (ii). In the following section 2.2 I consider (ii) in greater detail.

Experimental research on cooperation has been conducted for quite some time. The (experimental) simultaneous-move PD is presumably one of the most extensively

[1] This is not the case for existing data sets, such as surveys or official records. The researcher is typically not present when surveys are completed, but this task is delegated to interviewers or the respondents themselves (however, in section 2.2 and chapter 7 I will touch on how survey and experiments can be fruitfully combined). This problem may also apply to experiments whenever the subjects are not directly observed, such as in experiments conducted over the internet or some natural field experiments. Likewise, official records are typically created by many different people, sometimes under changing definitions and guidelines, and usually for different purposes than the researcher's question at hand.

studied experimental games, and there is a huge literature stemming predominantly from social psychologists in the 1960s and early 1970s (see Rapoport & Chammah, 1965; Rapoport, 1974; Colman, 1982, 1999; Roth, 1995, for overviews).[2] There is also a vast early literature on public good games from the 1980s (see Dawes, 1980; Dawes & Thaler, 1988; Ledyard, 1995, for overviews). I will focus on a selection of more recent experiments, because they are more tailored to speak to the theories reviewed in the previous chapter, and include appropriate control conditions (see below).

2.1.1 Tactical cooperation under indefinite horizons is prevalent

The basic prediction taken from the theoretical approaches in the previous chapter is that we should expect (conditional) cooperation especially in indefinitely repeated games with perfect monitoring, because there are tactical reasons to do so and information about the coplayers behavior is readily available. However, it was also highlighted that the multiplicity of equilibria is a central weakness of the folk theorem, since virtually any outcome between zero and full cooperation can be sustained in an equilibrium. So a useful first step would be to see whether the trigger strategy equilibria indeed exert some attraction to human players, and if «yes», which ones.

Perhaps somewhat surprisingly, there is only a very small experimental literature on this issue. An early body of studies provided rather mixed results. While Roth & Murnighan (1978) found in a repeated PD experiment with random termination that the frequency of cooperation decreases with the termination probability (and more cooperation under parameter constellations in which cooperation is an equilibrium in pecuniary payoffs), those results were not replicated in a similar experiment by Murnighan & Roth (1983).[3] In addition, Roth (1995) pointed to methodological problems in both studies, most importantly that subjects played with the experimenter and not with each other, and that a questionable payoff procedure was used. Similar methodological problems render it difficult to interpret the results of Feinberg & Husted (1993), who find a small increase in cooperation as the discount factor (implemented by a reduction of payoffs after each round) was decreased. Palfrey & Rosenthal (1994) also find only a very small difference between a one-shot condition and a condition with termination probability 0.1 in a public good experiment, which was, however, unnecessarily complicated.

A few recent papers with strongly improved designs have taken up the issue again. Dal Bó & Fréchette (2011) conducted a repeated PD experiment (266 undergraduates from New York University) in which they varied the termination probability (0.25 and 0.5) and the gains from mutual cooperation (low, intermediate, high). Subjects played

[2] The first PD experiments were (probably) conducted by Flood (1952, 1958) at the RAND Corporation. Some economists were quick in proposing a behavioral approach to the PD, that is, to refine the behavioral assumptions based on experimental research (e.g. Lave, 1962), but this approach gained momentum in economics only about three decades later.

[3] In both studies, cooperation rates were still far from the maximum, even under a relatively low termination probability.

a large number of matches (between 22 and 77) with randomly changing coplayers. The main predictions can be directly derived from condition 1.2 in the previous chapter: the frequency of cooperation should be decreasing in the termination probability and increasing in the gains from cooperation.

Specifically, the parameter values were chosen such that, assuming subjects maximize their payoffs, there were some conditions in which cooperation is a subgame-perfect and risk dominant equilibrium (low probability, high gains), some conditions in which cooperation is a subgame-perfect but not risk dominant equilibrium (low to high probability, intermediate gains), and some conditions in which cooperation was no equilibrium at all (high probability, low gains).[4] The results are threefold: First, in the conditions in which cooperation is not supported by an equilibrium, actual frequencies of cooperation decreased with experience from modest to very low levels (around 5 to 10 percent). Second, in the conditions in which cooperation is supported by an equilibrium, cooperation sometimes increases, sometimes decreases, and sometimes remains rather stable over the long haul, but the frequency of cooperation is on average clearly higher than in the previous conditions (around 40 to 50 percent). Third, the frequency of cooperation is typically (not always though) markedly increasing with experience in conditions in which it is supported by a risk dominant equilibrium; in the conditions with favorable parameters (low probability, high gains) cooperation rates converge to very high (sometimes full) levels. In fact, average cooperation rates converge to around 60 to 70 percent in the conditions in which cooperation is supported by a risk dominant SPE, and around 20 percent in the conditions in which cooperation is supported by a SPE that is not risk dominant. Thus, at the minimum, one may conclude that subjects respond markedly to the incentives of the situation, cooperating more the lower the termination probability and the higher the gains from mutual cooperation. These effects a present with little experience already, often in the first match, but become much stronger with increasing experience. However, the results also show that even experienced subjects can have difficulties in sustaining mutual cooperation in situations in which full cooperation is an equilibrium. In sum, the evidence obtained provides for empirical restrictions on the set of outcomes that the folk theorem predicts as attainable; overall, being an equilibrium action seems to be a necessary but not sufficient condition for sustained cooperation over the long haul.

But the above findings do not necessarily support the hypothesis that human players actually play the kind of trigger strategies as assumed in the theory of indefinitely repeated games. In particular, the fact that people respond to the termination probability and the size of the gains from mutual cooperation is also consistent with the kind of strategic imitation that also predicts cooperation in *finite* horizon games (section 1.2.2), or at least partially with some kind of intrinsic motivation to cooperate, such as maximizing the joint surplus (see section 2.2 below). Systematic errors seem rather unlikely for cooperation that remains after many trials (I come back to this

[4] An equilibrium is considered risk dominant if it has the largest basin of attraction of all equilibria (Harsanyi & Selten, 1988).

below). To differentiate between those possibilities, it is necessary to include appropriate control conditions into the experiment.

What would be a reasonable control condition? First, a finite horizon condition would clearly rule out the any folk-theorem-type of cooperation. Second, a condition in which subjects play (at least approximately) one-shot matches rules out *all* repeated game incentives, that is, any remaining cooperation cannot be tactical. The priors subjects bring to the lab may induce some cooperation in initial trials, but it appears unlikely that beliefs alone can account for any residual cooperation after subjects have gained sufficient experience.

Dal Bó (2005) did this in a comprehensive repeated PD experiment (390 undergraduate students from the University of California) that included both indefinite and finite horizon experiments with respective one-shot control conditions. The indefinite horizon experiment (198 subjects) was similar to the one reviewed above, involving two treatment conditions with termination probabilities 0.25 and 0.5, respectively, and a control condition with termination probability 1 (i.e. one-shot). Note that these termination probabilities imply *expected* horizons of one, two, and four periods, respectively (the actually realized lengths were with averages of 1.91 and 3.73 periods, respectively, slightly below the expectations). Dal Bó (2005) used this fact to implement a finite horizon experiment (192 subjects) that could be suitably compared to the indefinite horizon experiment: Here the two treatment conditions had *definite* durations of two and four periods, respectively, and the control condition was just one-shot. In both experiments, every subject played all three conditions, whereas the researcher counterbalanced partly by reversing the sequence. In each of the three conditions, each subject played between 5 and 10 matches, depending on the session, thus between 15 and 30 matches in total.

Again, the main prediction for the indefinite horizon case can be directly derived from condition 1.2 in the previous chapter: the frequency of cooperation should be decreasing in the termination probability.[5] Consistent with the experiment reviewed above, this turned out to be the case: Averaging over all matches (excluding the initial three to allow for learning) and rounds, cooperation occurred in around 9 percent in the one-shot condition, 27 percent in the condition with termination probability 0.5, and 37 percent in the condition with termination probability 0.25 (pairwise significant differences).[6]

[5] In order to resemble the scenario usually assumed in theory, Dal Bó (2005) also introduced a public randomization device in the form of a screen that displayed a random number between 1 and 1,000 every ten seconds. It turned out that subjects did not pay any attention to the device, even after being told that they may use it to choose one of their actions.

[6] Dal Bó (2005) also replicated the effect of a change in payoffs by implementing two payoff function conditions in a way that under termination probability 0.5 the unique equilibrium outcome in pecuniary payoffs was mutual cooperation in one and mutual defection in the other. Again, it turned out that behavior was quite sensitive to this subtle change in payoffs: cooperation occurred significantly more often under the payoff function under which it was predicted (18.8 versus 3.2 percent), and defection occurred more often, albeit not significantly so, under the payoff function under which it was predicted (28.6 versus 25.5 percent).

A significantly higher frequency of cooperation (55 percent) in a repeated PD with termination prob-

2.1.2 Tactical cooperation under finite horizons is prevalent as well

The main prediction for the finite horizon condition can be directly derived from the reasoning in section 1.2.2: Under the assumption of some incomplete information about coplayers' types, the length of the horizon is a key determinant of the value of maintaining a cooperative reputation, such that the imitation equilibrium model predicts a non-increasing rate of cooperation in the initial periods of a match, and a sharp decline as the terminal period approaches. Furthermore, the frequency of cooperation in the first period should be increasing in the length of the (remaining) horizon. Again, dropping all observations from a player's first three matches to allow for learning (but see below), Dal Bó (2005) finds within-match dynamics that are quite consistent with the prediction: Cooperation in the first round is increasing in the remaining match horizon, and decreasing over time in the two treatment conditions. The average cooperation rate in the first period was 10.3 percent in the one-shot treatment, 13.3 percent in the two-period match treatment, and 34.6 in the four-period match treatment. In the second period, cooperation diminished to 6.9 percent in the two-period, and 21.6 percent in the four-period treatment. In the latter, cooperation decreased further in the third and fourth period, respectively to 19 and 10.6 percent. Thus, the dynamic patterns are quite consistent with the imitation model, but there is also non-negligible cooperation in terminal periods (including the one-shot control condition).

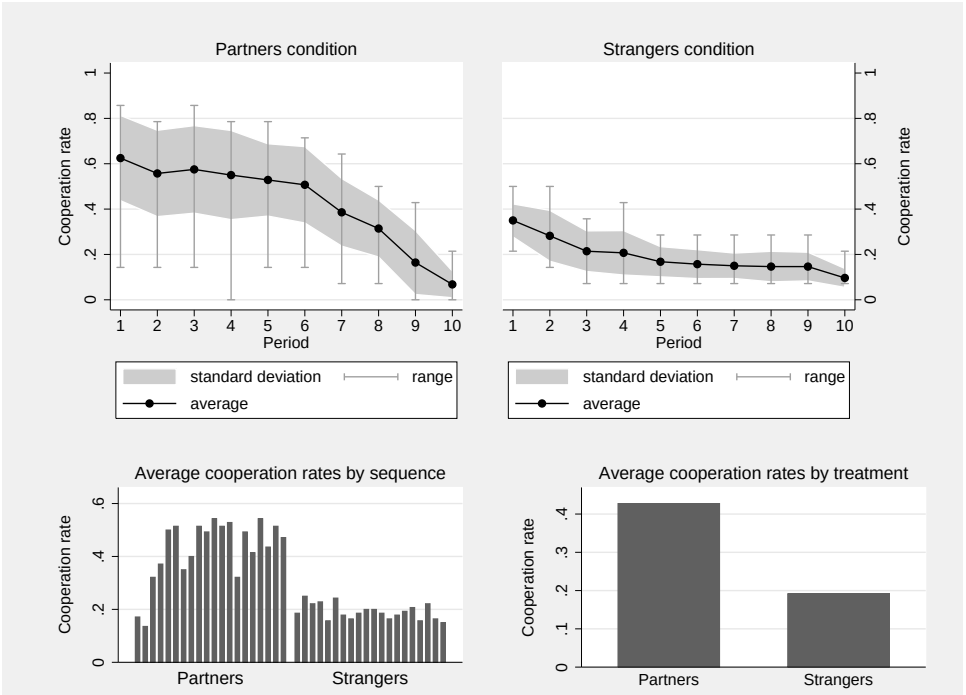
There is, however, some conflicting evidence on long-term learning in finite horizon games. In the experiment by Dal Bó (2005), cooperation rates tended to deteriorate somewhat over the long haul.[7] In the one-shot treatment, average cooperation rates started at 26.6 percent in the first match, deteriorated successively to 12 percent in the fifth match, and settled between 6 and 10 percent afterwards. Likewise, in the four-period match treatment, average cooperation rates started at somewhat higher 31.6 percent in the first match, and declined successively, albeit somewhat less than in the one-shot treatment (25.1 percent in the fifth match, settling between 15 and 23.8 percent afterwards). Interesting is the two-period match treatment: Here, cooperation rates were most stable over the long haul, but are generally *lower* than in the one-shot treatment, starting at 19.8 percent in the first match and settling between around 6 and 12 percent.

In a classic repeated PD experiment (28 students from the University of Wisconsin) conducted by Andreoni & Miller (1993), in which subjects also played a rela-

ability 0.1 compared to a random rematching condition (6 percent) has also been found by Duffy & Ochs (2009).

[7] The same trend was found in a trust game experiment by Engle-Warnick & Slonim (2004), to be reviewed in more detail below. In the first supergame of their experiment, first movers cooperated in 82 percent of the time, and 81 percent of first mover cooperation was reciprocated. In the final supergame, first mover cooperation declined to 53 percent of the time, while 90 percent of their cooperation was reciprocated. Interestingly, the average payoff per period decreased steadily for both players as they gained experience over the long haul, for first movers from about 62 cents average over the first ten matches to about 52 cents in the last ten, for second movers from about 47 to 42 cents. See also the experiment of Anderhub et al. (2002) to be reviewed below.

Figure 2.1: Cooperation in a finite horizon PD experiment.



Data source: Original data of James Andreoni and John H. Miller, Economic Journal Vol. 103, No. 418 (May 1993).
Graph composition: Johannes Jarke.

tively large number of matches, the long-term trend points in the other direction (see lower left panel of figure 2.1). In the treatment condition (the «partners condition») each of the 14 subjects played 20 ten-period matches with a fixed coplayer in each match, respectively, and a random rematch occurred between matches. The control condition (the «stranger condition») was identical except that players were randomly re-matched in each of the 200 periods.[8] Interestingly (see the lower-left panel of figure 2.1), the full effect in the partner condition did require a few matches to unfold. In fact, in the initial two matches cooperation rates are actually lower in the partner than in the stranger condition, but strongly *increase* afterwards and remain high for most of the time in the former but not in the latter. More evidence is needed to resolve the exact reasons for the conflicting evidence, because the experiments differ along too many dimension to compare them in a meaningful way. Likely candidates are the number of players in a session (very few in Andreoni's and Miller's experiment) or the length of the matches (very short in the other experiments). But also the sampling population might play a role. The evidence reviewed in the next subsection is informative on this issue as well.

Anyway, that (i) the frequency of cooperation is generally higher in «partners» than in «strangers» conditions, and (ii) the distinct within-match dynamics in the former are very robust results (see also below). They have also been documented in public good games. While there is some conflicting evidence in early experiments of this kind (Andreoni, 1988, 1990; Croson, 1996, see Andreoni & Croson, 2008 for an overview),[9] the literature following one of the most rigorous studies (Keser & van Winden, 2000) converged to the following two stylized facts that mirror the above evidence for dyadic games: (i) cooperation in a «partners» condition is higher than in a «strangers» or «perfect strangers» condition, (ii) there is a drop of cooperation if the terminal period approaches in «partners» conditions, and (iii) cooperation in «strangers» conditions (and the terminal period of «partners» conditions) is significantly above zero.

2.1.3 What strategies are actually played?

Returning to the experiment by Dal Bó (2005), it remains to compare the finite and indefinite horizon outcomes. Dal Bó (2005) expects cooperation in the first round to be no less in the indefinite than in the finite horizon condition for any given (expected) remaining duration of a match.[10] This turned out to be clearly the case. While the

[8] The design has one weakness, though: in the control condition the probability of meeting the same subject again in the future was not negligible given the moderate number of subjects in a session and the large number of matches. However, Cooper et al. (1996) replicated the experiment with a much more sophisticated matching scheme in the control condition, which assured that no pair interacted more than once, but obtained virtually identical results.

[9] See also Weimann, 1994; Palfrey & Prisbrey, 1997, Burlando & Hey, 1997; Brandts et al., 2004 for indications that the sampling population plays a role.

[10] Note that this comparison is meaningful because in the first period the expected remaining horizon in the indefinite horizon condition is exactly equal to the certain remaining horizon in the finite horizon condition. There is no general theoretical result on such a comparison, so the hypothesis is intuitively

first-round cooperation rates between the two control (one-shot) conditions do not differ significantly (9.2 percent in the condition with zero continuation probability, 10.3 percent in the one-period match condition), they are clearly higher in the indefinite horizon condition, 30.9 versus 13.3 percent and 46.2 versus 34.6 percent in the (expected) two- and four period horizon conditions, respectively. In the subsequent periods, cooperation remained significantly higher in the indefinite horizon conditions; the differences actually increased, as cooperation rates deteriorated much less in the indefinite than in the finite horizon treatments.[11]

The results suggest that all three of the possible forces mentioned above must be operative. The fact that there is a modest but clearly positive and stable frequency of cooperation even in one-shot games suggest that not all observed cooperation in repeated games can be attributed to the strategic incentives reviewed in the previous chapter.[12] Putting this aside, the patterns in finitely repeated games are by and large consistent with strategic imitation (but see the final paragraph below). Finally, the differences between the finitely and indefinitely repeated games of the same (expected) duration suggests that there are indeed different things going on in the latter than in the former.

But what exactly? We still do not know what kind of strategies the players are actually playing. There is indeed very little evidence on this issue, but two recent studies by Engle-Warnick & Slonim (2004, 2006b) make an honest effort to infer repeated game strategies from observed actions in finitely and indefinitely repeated trust games. Engle-Warnick & Slonim (2004) conducted a trust game experiment (146 students from the University of Pittsburgh) with two conditions: an indefinite horizon condition with termination probability 0.2, i.e. an expected duration of five periods (the actually realized duration was 5.2 on average), and a finite horizon condition with definite duration of five periods. Each subjects played 50 matches (which was not told to them, however, only that they will play «many» matches), pairs were randomly re-matched between matches, and subjects kept their roles as first or second movers, respectively, for the entire experiment.

Similar as in the experiments reviewed above, the frequency of cooperation and pecuniary efficiency did not differ significantly between conditions in early matches,

---

derived.

[11] Normann & Wallace (2012) found also more frequent cooperation before a repeated PD was terminated with low probability compared to high probability and certain termination, although the differences are small and insignificant. This might be explained by the fact that subjects played only a single match, whereas the above experiments suggest that learning is important. Furthermore, a somewhat peculiar feature of the design is that subjects played 22 rounds for sure in any condition, before the treatments kicked in. An interesting novel feature, however, is a fourth condition in which the horizon was completely unknown (ambiguous) to the subjects. No significant difference to the random termination treatments was found.

[12] Note, however, that the control conditions in the experiments by Dal Bó & Fréchette (2011) and Dal Bó (2005) do not rule out any strategic incentives beyond doubt, because there is a non-negligible probability that the same pair of subjects meet in more than one match due to the large number of repetitions relative to a modest number of players in a session. Though, a large number of experiments that rigorously rule out this possibility do not find any different results (see section 2.2).

but diverged as subjects gained experience, with first movers cooperating significantly more often in the indefinite than in the finite horizon condition (contingent second mover cooperation rates did not differ significantly). The researchers attempted to infer strategies from observed behavior by fitting the data to finite automata (Engle-Warnick & Slonim, 2006a, go into much greater detail). In the indefinite horizon condition, only one first mover strategy was inferred, and this strategy was always a best response for one of the inferred second mover strategies: grim trigger. For second movers, between two and four strategies were inferred for each interval of ten matches. «Always cooperate» and the four-period counter («cooperate three rounds, then defect once») were inferred across all matches, the other strategies were inferred only sporadically. Apparently, one second mover strategy («always cooperate») is always a best response to the inferred first mover strategy (grim trigger). In the finite horizon condition, clearly more strategies were inferred, and dynamically, strategies disappear and emerge, at least for first movers, significantly more often than in the indefinite horizon condition. Between four and two first mover strategies were inferred, including various counter strategies («cooperate x rounds, then defect once»), «always defect», «always cooperate», or grim trigger. The set of inferred strategies becomes smaller over time, with only «always defect» and grim trigger surviving, only the former being a best response to the inferred second mover strategies.[13] For second movers, various counter strategies come and go (four and three period counters survive), «always defect» emerges towards the end, and again «always cooperate» was inferred across all matches.

The interesting novel insights that Engle-Warnick & Slonim (2004) provide are twofold. First, the strategy inference provides evidence that subjects use distinct strategies in the two conditions from the beginning already, an insight that would not have been obtained by comparing aggregate performance alone (because there were no significant differences in the beginning). Second, the fact that subjects seem to settle on a very small set of strategies in the indefinite horizon but not the finite horizon condition, in which heterogeneity persists, suggest a reason why the frequency of cooperation begins to diverge over time. Furthermore, the fact that the only second mover strategy that was inferred across all matches, in both conditions, and even after many trials was «always cooperate» suggests that there is a hard core of subjects which are somehow intrinsically motivated to cooperate. Interesting would be an individual level inference in order to get an idea about how frequent (in relative terms) the different strategies actually are.

Dal Bó & Fréchette (2012) did this with a different approach to infer strategies involving the «strategy method» (Selten, 1967). The basic design was similar to the the studies of Dal Bó (2005) and Dal Bó & Fréchette (2011) reviewed above, built around a sequence of indefinitely repeated PDs with different termination probabilies and gains from mutual cooperation conditions (they had 246 undergraduates from New York University). After playing the games in the usual fashion for around 20

[13] Note that behavior predicted by the strategic imitation model may be observationally identical to grim trigger.

minutes, subjects were asked to design complete strategies that will play in their place. They find that the most popular strategies were «always defect», TFT and grim trigger. However, the prevalence of the strategies depend markedly on the termination probability and the gains from mutual cooperation: in treatments less conductive to cooperation (likely termination, low gains) «always defect» is the by far most popular strategy (between 46 and 58 percent). As those conditions become more conductive, «always defect» becomes significantly less frequent (only 9 to 14 percent under most conductive conditions), and trigger strategies become dominant (between 48 and 69 percent). Interestingly, grim trigger is more prevalent in treatments with high gains, while TFT is more prevalent in treatments with a low termination probability. Thus, the strategies players adopt appear to depend significantly on the parameters of the game. This result may help to resolve the conflicting evidence on the use of trigger strategies in earlier studies (Sell & Wilson, 1991; Feinberg & Husted, 1993). I will come back to this important insight in section 2.4.

Summing up so far, what goes on in indefinitely repeated games appears relatively simple: trigger strategies are the dominant way to play the game. On the other hand, what happens in finitely repeated games remains more mysterious. The overall patterns appear to be quite consistent with the strategic imitation model, but at a closer look there are also some oddities. This conclusion is reinforced by an experiment that was particularly well tailored at testing the strategic imitation model. Anderhub et al. (2002) conducted a trust game experiment (36 not further specified subjects) in which they also varied the length of the horizon, but added as a special feature that players could explicitly randomize over their actions, that is, a subject had to specify the probabilities (in percent) for her or his two actions to be carried out. Each participant played a sequence of six matches with a different coplayer each, where the horizon of the matches comprised, always in this order, three, six, two, ten, three, and six periods, respectively. Roles were randomly assigned and remained fixed for the entire sequence. Consistent with the experiment by Dal Bó (2005), they found that changing the length of the horizon had little effect on the standard qualitative within-match dynamics of frequent cooperation in the non-terminal periods and a strong termination effect, but that cooperation in the first period of a longer match tends to be more frequent than in the first period of a shorter match. The explicit randomization option was hardly used. The most interesting analysis performed by Anderhub et al. (2002) was the investigation of the more fine-grit predictions of the strategic imitation model. Two clear predictions are that (i) second movers should never cooperate in the terminal round, and that (ii) once a second mover defected in a non-terminal round, the coplayer should immediately turn to defection for the rest of the match. It was found that the majority of second movers indeed defected in at least some of the six terminal rounds, but there was also a small minority that did never exploit a cooperating first mover in the terminal period of a match. There were also some second movers who sometimes rewarded and sometimes exploited in the terminal period of a match. Concerning the second prediction, roughly a third of those first movers that were exploited in a non-terminal round never cooperated

again, whereas the remaining two thirds forgave. This confirms the conclusion that strategic cooperation plays a role, but cannot account for all observed cooperation.[14] There is, however, a qualification. Since the first and fifth, and the second and sixth supergame, respectively, had the same length Anderhub et al. (2002) could also check for the role of experience. Indeed, learning shifted behavior clearly towards the theoretical benchmark. From first to the fifth and the second to the sixth supergames, respectively, the share of monotone strategies increased significantly with the result that first and second mover cooperation *increased* in the non-terminal periods and *decreased* in the terminal period.

2.1.4 Conclusion

In this section I have investigated two fundamental questions with which I concluded chapter 1: (i) Do real people actually perform such tactics as envisaged in the conventional approaches to cooperation in economics? (ii) Do real people *only* cooperate for tactical reasons? In this section I focused on (i) and touched (ii). The following conclusions with respect to those questions can be drawn from the evidence reviewed in this section. First, a majority of subjects actually cooperates tactically. In indefinitely repeated cooperation games, the evidence shows that cooperation is more likely to prevail if it is a supergame equilibrium (in particular if it is risk dominant) than when it is not. As predicted by the folk theorem, subjects respond to changes in the termination probability and the size of the gains from mutual cooperation. Most importantly with respect our guiding question (What does the propensity of individuals in a given group to cooperate with one another has to do with the availability of information they have about each others' actions?), subjects use the available information about the coplayers' past behavior to condition their current behavior. The evidence suggests that the most commonly employed strategies are the two trigger strategies considered in section 1.2.3: TFT and GRIM.

The evidence shows that there is generally also significant cooperation in finitely repeated cooperation games. The average within-match dynamics are consistent with the strategic imitation model considered in section 1.2.2: significant cooperation over some fraction of the duration, with cooperation in the first period being increasing in the remaining duration, and a sharp decline as the terminal period approaches. However, while most people seem clearly to condition their behavior on past play, the individual level fine-structure of behavior is often inconsistent with the strategic imitation model, and subjects seem to have more difficulties to settle on a particular

[14] The limited consistency with this theory is confirmed by individual level analysis. A key aspect of the equilibrium strategies is that the probability of cooperation should be, for both players, at least non-increasing over time. While the fraction of such monotone strategies is relatively high in matches with shorter horizons, it is less so with longer horizons, and only 28 percent of the first movers and 33 percent of the second movers played monotonically in all six supergames. Thus, while the theory of reputation equilibria appears to have some merit, it is clearly only part of the story. This conclusion is further reinforced by the regularly *increasing* patterns of cooperation in non-terminal periods found in other experiments (see below), which are also inconsistent with the theory.

strategy.

In sum, tailored at our guiding question this means that if information about the other players' past behavior is readily available, this information is used in order to play a discriminatory strategy. As a result, there are material incentives to cooperate which is reflected in the notably higher frequency of cooperation in repeated games that in (approximated) one-shot games. However, and this brings me to question (ii), there is some persistent residual of cooperation even in (approximated) one-shot interactions which cannot be accounted for by tactical considerations. Andreoni & Miller (1993) show that behavior in their stranger condition is consistent with an incomplete information Nash equilibrium in which individuals share a common prior on the probability of experiencing cooperation of around 0.2, which is similar to the «homegrown» subjective prior of 0.17 estimated by Camerer & Weigelt (1988) in a similar experiment. But where should such a prior come from, and more importantly, why should it persist over many trials if not justified? I will now turn to experiments that address those peculiarities more directly.

2.2 Non-tactical cooperation

One key conclusion from the experiments reviewed above is that tactical (conditional) cooperation plays an important role, but that there is also a residuum of observed cooperation that must have different reasons: either the respective subjects make errors or do not understand the strategic incentives of the experimental game, or they are motivated not only by their pecuniary payoffs.

To start with, the «residuum» is not a peculiarity. The findings that (some) people (sometimes) cooperate even if pecuniary returns from doing so are ruled out have now been replicated in a vast number of studies using a variety of different experimental games, protocols or sampling populations, and in the laboratory as well as in the field (see Fehr & Schmidt, 2003, 2006; Henrich et al., 2001, 2004, 2005; Gintis et al., 2005; Meier, 2007; Kagel & Roth, 2011, for overviews). There is even substantial cooperation in *strict* one-shot cooperation games without any kind of repetition (e.g. Marwell & Ames, 1979; Gächter et al., 2004; Walker & Halloran, 2004; Dufwenberg et al., 2011; Cubitt et al., 2011). In this section I will put special emphasis on evidence that is informative with respect to our guiding question. Namely, to what extent is even such non-tactical cooperation conditioned on coplayers' actions?

2.2.1 Non-tactical cooperation is predictable

The notion that humans are not always motivated by their immediate self-interest alone is of course an old hat in sociology and social psychology (see general introduction). But economists have traditionally been wary to overhasty adjust assumptions on preferences, for good reason, since a theory quickly becomes vacuous if *any* behavior can be «explained» by positing respective preferences. Furthermore, the common observation that many people *want* to be helpful, generous, philanthropic

and generally a good person does not refute the selfishness assumption, because those desires may just be a «mental program» that actually implements a long-term selfish strategy.

But also economists have pointed to many apparent forms of cooperation under conditions that make it hard to believe that they rely on (enlightened) self-interest alone (Bohnet & Frey, 1997). A large fraction of people sacrifice time and resources to vote or participate otherwise in political or other collective actions, despite their own contribution having a negligible impact on the overall outcome (Olson, 1965; Mueller, 2003; Shabman & Stephenson, 1994). Likewise, often people take care of shared resources and comply with environmental standards (Russell et al., 1986; Harrington, 1988) or pay their taxes honestly (Andreoni et al., 1998) even if the likelihood of sanctions is extremely low. Apparently, people also voluntarily contribute to finance local amenities, donate to charities or tolerate large scale income redistribution (Roberts, 1984). This is backed by more general claims that throughout human history, relationships and communities were frequently disbanded by warfare, famine and other catastrophes, such that folk theorem arguments (recall the condition 1.2 in section 1.2) run into severe difficulties in plausibly accounting for the degree of cooperation that has frequently been observed precisely in the midth of such adverse conditions (e.g. Knauft, 1991; Gintis, 2000; Fehr & Henrich, 2003). «Physicists also do not argue that we can explain sunset and sundown by assuming that the Sun orbits around the Earth although this incorrect assumption can provide a superficially plausible explanation for these phenomena» argues Fehr et al. (2009, p. 363). But rigorous empirical research is needed before adjusting behavioral assumptions. And the controlled experiments is again indispensable here, because in the field virtually «all human interaction involves a seamless transition from past involvements to future anticipations» (Seabright, 2010, p. 66), such that too many uncontrolled factors, on which adequate data is often lacking, simultaneously affect the outcomes which renders it extremely difficult, if not impossible to rigorously distinguish between extrinsically motivated («enforced») and intrinsically motivated cooperation.

The major competing hypothesis that have been put forward to explain apparently non-tactical cooperation is that not preferences diverge from the traditional behavioral model but decision-making and performance, that is, subjects are just not as rational in forming beliefs and performing their actions as assumed by game theory, that they are more guided by intuition and rules of thumb and do not backward induct. However, in principle game theory does not require that players select strategies consciously (e.g. Gintis, 2009), that is, even intuition may be sufficiently adapted to a situation at hand to produce behavior that is «rational» in the usually defined sense of consistency. The argument therefore boils down to the hypothesis that subjects apply intuitions from the real world «maladaptively» in the laboratory experiment initially, i.e. that they make errors, and that they need some time to learn the incentives of the game (e.g. Selten & Stoecker, 1986; Binmore, 1998; Anderson et al., 1998).

Clearly, people differ in their powers of comprehension and reasoning, and some people might require a longer period of practice than others. But doubt on the «error

hypothesis» as a *major* explanation is cast already by experiments in which subjects play a significant number of matches (reviewed above). In those experiments a small but robust residual of cooperation persists even with significant experience. Engle-Warnick & Slonim (2004), for example, inferred the «always cooperate» strategy even in the last ten trials, after 40 trials have already been played. Furthermore, cooperation typically starts high in new trials after deteriorating in the previous one, even with significant experience. For example, in the experiment by Andreoni (1990) reviewed above, the frequency of cooperation rebounds, after dropping to very low levels at the end of a ten period match, back to about equally high levels in the first period of a new match. This «restart effect» has also been found by Andreoni (1988) in a public good game in which there was a previously unannounced second sequence of stage games (see also Croson, 1996; Cookson, 2000).

Another fact that supports the hypothesis that there is something systematic behind non-tactical cooperation is the observation that (most) subjects respond clearly and quite consistently to changes in the inventive structure created by the experimental design. For example, the comparisons in the previous section show that when incentives for strategic reputation building are removed there are behavioral adjustments in the very first period already. Likewise, the end-game-effect in finitely repeated games is pronounced even in the first match or if only a single match is played. This suggests that, at least a majority of subjects are able and do analyze the simple situations created in the laboratory *before* they start playing, and plan their strategy accordingly.

By means of an extensive giving game experiment in which they vary the cost (κ) and the external benefit (η) of cooperation, Andreoni & Miller (2002) show that subjects systematically respond to changes in those parameters. A majority of subjects even behave consistent with the generalized axiom of revealed preference (GARP), a basic consistency requirement that assures that behavior can be represented by a well-behaved utility function. Note that the giving game is a one-shot unidirectional game in which only one of the two players (the donor) has the option of performing a cooperative action from which the other player (the recipient) benefits, just as described in section 1.1. Thus, cooperation *never* pays off, such that this design rules out any pecuniary motivation to cooperate. Relatedly, a number of one-shot (or controlled rematching) public good experiments find that cooperation is generally increasing in the external benefit η, which is equally shared among the other group members in this case (Isaac & Walker, 1988; Brandts & Schram, 2001; Goeree et al., 2002; Zelmer, 2003; Carpenter, 2007b).[15] Relatedly, if the marginal per capita return of the public good is held constant, that is, η is increased with the number of coplayers such that $\frac{\eta}{n-1}$ remains constant, then cooperation is generally independent from the size of the group (Marwell & Ames, 1979; Isaac & Walker, 1988; Isaac et al., 1994; Zelmer, 2003; Carpenter, 2007b; Cárdenas & Jaramillo, 2010).

There are also studies that systematically test the «error hypothesis» against the «non-pecuniary motivation» hypothesis. In a public good experiment, Andreoni

[15] See also Anderson et al. (1998) for a theoretical approach.

(1995) found confusion to be a dominant reason for cooperation in the very first round, but it declined strongly after just a few rounds to about 10 to 20 percent of cooperative moves. The fraction of cooperative moves coming from subjects which understand the free-riding incentive quite well but choose to cooperate nevertheless follows the opposite pattern. Several experiments with different designs strengthen the evidence that errors play some role, it particular in the very first period, but relatively swiftly become a minor issue (e.g. Palfrey & Prisbrey, 1997; Houser & Kurzban, 2002; Brandts & Schram, 2001; Willinger & Ziegelmeyer, 2001; Charness & Haruvy, 2002). The evidence in favor the «non-pecuniary motivation» hypothesis, on the other hand, has become very strong (see Fehr & Schmidt, 2003, 2006; Henrich et al., 2001, 2004, 2005; Gintis et al., 2005; Meier, 2007; Kagel & Roth, 2011, for overviews), whereas much research on the fine-structure of those motivations is still to be done.[16]

2.2.2 Non-tactical cooperation is partly conditional

Even if one rejects the «error hypothesis» as a major explanation, there are alternative hypotheses concerning individual motivations.[17] A classic hypothesis is that people might have altruistic preferences in the sense that they experience subjective utility from knowing about someone else receiving a benefit (Becker, 1974). Another hypothesis is that people experience subjective utility, a «warm inner glow», just from the *act* of contributing to the common good (Andreoni, 1990). Or people may just be principled by their moral values or social norms that generally suppresses selfish behavior (see the introduction). In sum, there might be intrinsic motivations to cooperate that are *independent* from the behavior of others, and therefore independent from information about those behaviors.

But the idea that intrinsically motivated behavior may be conditional on others' behavior has long been hypothesized in biology and anthropology (e.g. Darwin, 1871; Williams, 1966; Trivers, 1971; Sahlins, 1972), sociology (e.g. Gouldner, 1960; Blau, 1962), social psychology (e.g. Kelley & Stahelski, 1970), and even by a number of influential economists (e.g. Friedman, 1962; Buchanan, 1967; Cornes & Sandler, 1984; Frank, 1988). Indeed, Milton Friedman (1962, p. 191) mentioned vividly half

[16] The question of how humans might have evolved pro-social dispositions that suppress selfish behavior in particular circumstances to the benefit of others is a vibrant area of research in biology and anthropology (e.g. Sober & Wilson, 1998; Boehm, 1999; Hammerstein, 2003; Boyd & Richerson, 2005; Kappeler & van Schaik, 2006; Henrich & Henrich, 2007). However, such ultimate questions are only of peripheral importance to economists with their more proximate perspective.

[17] Such non-pecuniary motivations are commonly called *social preferences* (or *other-regarding preferences*), and are typically conceptualized by a preference order over payoff distributions in the relevant reference group in the sense that individuals who exhibit social preferences behave *as if* they value the payoff of relevant reference players positively or negatively (Fehr & Fischbacher, 2005). Generally, the relevant reference agents are always relative to the situation under consideration, that is, a person may have different reference agents in different domains. In the laboratory, however, the researcher is able to have some control along these lines: the common assumption in experimental economics is that the relevant reference players are those matched into the same group (Fehr & Schmidt, 2006).

a century ago that

> «we might all of us be willing to contribute to the relief of poverty, provided everyone else did. We might not be willing to contribute the same amount without such assurance.»

With respect to our guiding question, this can be put alternatively as follows: people respond to information about other people's (non-)cooperative behavior. The evidence reviewed in the previous section shows clearly that this is true when doing so carries a prospect of a material return. But does this extend to settings in which such material returns are very unlikely?

A way to investigate this question is to observe second mover behavior in *one-shot* sequential games (Camerer & Fehr, 2004). The advantages are two-fold: First, because such games end after the second mover's action, pecuniary motivation can be ruled out as a reason for her or his behavior. Second, since the second mover follows the first mover's action (i.e. is informed about this action before her or his own move), one can study how the former's behavior depends on the latter's. First mover behavior can still be tactical, but is interesting insofar as it may reflect an anticipation of the second mover's non-tactical cooperation (which may be viewed as even stronger evidence on the existence of instrinsic cooperative motivations).

In a seminal study, Berg et al. (1995) conducted a strictly one-shot investment game experiment (64 undergraduates from the University of Minnesota).[18] They used a sophisticated double-blind protocol in order to mitigate subjects' concerns about being observed as far as possible. Out of the 32 first movers, five sent their complete endowment, nine sent between 6 and 8 Dollars, six sent half of their endowment. Only two out of 32 sent nothing, two sent one Dollar. Thus, first movers cooperated to a considerable extent: almost 94 percent cooperated (sent positive amounts), 63 percent sent at least half of their endowment. Among the second movers, the majority (24 out of 30, or 80 percent) cooperated in return. Most first movers were at least compensated, such that the exchange was mutually beneficial to both players. But some first movers were exploited, their coplayer returning little or nothing at all.

Although the behavior of the second movers in the trust or investment game is frequently termed «reciprocity» from the outset, it is important to observe that their cooperation may be, at least partially *unconditional*.[19] Thus, identifying whether and to what extent second mover cooperation in the trust or investment game is conditional is actually not that easy, because a (within-subject) counter-factual is missing

[18] The investment game is an extension of the trust game in which subjects can send also intermediate amounts between the two extremes. The first mover's transfer is usually doubled or tripled, the second mover's transfer is not modified.

[19] Berg et al. (1995) anticipated this caveat and tried to avoid transfers based on distributional considerations by endowing *both* players with 10 Dollars each. There are, however, some plausible motives inducing unconditional cooperation that were not ruled out by this feature. For example, if a first mover cares about the joint surplus of both players, then the fact that the amount sent is tripled provides a strong incentive to cooperate, independently from the second mover's return. The same is true for the second mover in the gift-exchange game since there the return transfers are multiplied.

by design: one can only observe the actual path of play, but to identify conditioning, one needs also to know what the second mover *would* have done in case the first mover had behaved differently.

One way to address this problem is to just ask second movers about the reasons of their behavior. Bolle (1998) conducted a slightly modified trust game (first mover binary decision, second mover multiple) experiment (95 first year economics students in Frankfurt/Oder, Germany). Each first mover had to indicate both what the second mover *will* return, and what he or she *should* return, and each second mover could justify her or his behavior free of form. Bolle (1998) obtained similar results as Berg et al. (1995). Not surprisingly, the distribution of normatively expected return transfers is concentrated at much higher levels than actual ones. The majority expected (normatively) an equal split of the surplus. The vast majority of second movers felt obliged to at least compensate the first mover, whereas a minority of those perceived it as just to maximize their income under this restriction, while the rest expressed it just to reward the first mover's willingness to take the risk of being exploited by returning somewhat more than the compensating amount. However, all stayed well below the equal split of surplus with their judgment. Some of the second movers provided an equity justification, namely an equal share of the joint payoff to be fair. The most common justification among the second movers who returned very little was that they did not felt bound by any rule, commitment or promise to return anything. In sum, those more-than-compensating return transfers that were justified as a «reward» point toward a conditional motivation, whereas the statements suggest that distributional concerns are important as well.

Cox (2004) attempted to decompose behavior in the one-shot investment game more rigorously (188 students from the University of Arizona). There was one condition with the standard investment game, and two conditions in which subjects played a giving game, respectively. The giving games were structurally identical to the situations of the first and second mover in the investment game, respectively, but with the actual move of the coplayer eliminated.[20] Cox (2004) argues that the differences of the average amounts sent and returned between the actual investment game and the two giving games should identify tactical cooperation, conditional on the belief of a compensating return («trust») in case of first movers, and cooperation conditional on the first mover's action in case of second movers («reciprocity»). Note that this argument is not unproblematic, because the comparison is (i) between-subjects and (ii) across quite different games which may be, despite being structurally identical to the investment game, not perceived as the same kind of situation. Anyway, assuming that this design is successful in disentangling unconditional and conditional cooperation, Cox (2004) finds that both partially account for cooperation in the investment game. On average, the first movers sent 5.97 Dollars in the investment game and 3.63

[20]The first was identical to the first mover's situation in the investment game except that the second mover did not have an opportunity to return anything. The second was identical the second mover's situation, whereas the endowments were adjusted by the average amounts set in the investment game in order to render the two games comparable.

Dollars in the respective giving game, suggesting that about 40 percent of the first movers' cooperation is tactical and the remaining 60 percent is unconditional. Likewise, the second movers returned on average 4.94 Dollars in the investment game and sent 2.06 Dollars in the respective giving game, suggesting that roughly 60 percent of the second movers' cooperation is conditional on first movers' behavior and the remaining 40 percent unconditional. The larger fraction of unconditional cooperation among the first movers is consistent with a surplus maximization motive (Charness & Rabin, 2002; Engelmann & Strobel, 2004), because the amounts sent are tripled.

Ashraf et al. (2006) address one shortcoming of the previous approach by employing a within-subject design. In addition, they used the strategy method where the second movers had to decide on a contingent action for every possible amount sent by the first movers. The three-games design was very similar to Cox's, but they did not adjust endowments in the giving game that resembled the second mover's situation in the investment game. They are somewhat more cautious in relating the results from the three games by forgoing to interpret the differences quantitatively but instead focusing on correlations. Ashraf et al. (2006) argue that if tactical cooperation is the dominant component (relative to unconditional cooperation) of first mover cooperation in the investment game, then the latter should be stronger correlated with the expected return than with their own transfer in the giving game. Likewise, if conditional cooperation is the dominant component (relative to unconditional cooperation) of second mover cooperation in the investment game, then the latter should be stronger correlated with the amount received than with their own transfer in the giving game. Their results are affirmative in the first case, and somewhat less in the second, but conditional second mover cooperation, according to their measure, is clearly present. Furthermore, they conducted the same experiment in three different countries, the USA (112 students from universities in Boston), Russia (118 students from universities in Moscow), and South Africa (129 students from universities in Cape Town), and found little differences between the locations. Using also a within-subjects design, Chaudhuri & Gangadharan (2007) finds similar results and more clear indications of conditional cooperation in second movers.

Further evidence comes from experiments using the gift-exchange game in which researchers aim to measure conditional cooperation in second movers by within-period correlations between first and second mover cooperation (e.g. Fehr & Falk, 1999; Falk et al., 1999; Fehr & Gächter, 2000b; Gächter & Falk, 2002). The following patterns that have been replicated many times: First, a notable fraction of second movers cooperate above the minimum in response to above-the-minimum cooperation by their respective first movers. In addition, the two's cooperation levels are positively correlated. Second, there is also a notable fraction of second movers that cooperate above the minimum consistently, *independent* from the level of their first mover's cooperation. Third, on average first movers cooperate above the minimum, that is, they are «rent leavers» (Bohnet & Frey, 1997). Fourth, while the average level of second mover cooperation (and total surpluses in turn) is typically significantly above the minimum, it is still far from the maximum.

The gift exchange game is probably the most studied experimental game so far, and there is a vast number of studies that vary design elements in one-shot or random matching laboratory gift-exchange game experiments to check the robustness of non-tactical conditional cooperation. Such elements comprise *inter alia* stake sizes (Rigdon, 2002; Fehr et al., 2002b),[21] the extent of experimenter-subject interaction (Rigdon, 2002), instructions, framing and presentation (Charness et al., 2004; Engelmann & Ortmann, 2009), whether the equilibrium in material payoffs is a corner or interior point (Pereira et al., 2006; Engelmann & Ortmann, 2009), the matching mechanism (Engelmann & Ortmann, 2009), whether second mover's get information about transfers received by other second movers (Maximiano et al., 2007; Charness & Kuhn, 2007; Abeler et al., 2010; Clark et al., 2010; Siang et al., 2011), or whether second movers' cooperation is a pecuniary transfer or real effort (Brüggen & Strobel, 2007). Some manipulations have a notable effect on the strength of conditional behavior, in some cases it gets quite weak (i.e. information about the coplayer's previous behavior is close to being ignored), but the qualitative conclusion that it is common turned out to be quite robust with respect to variations in the laboratory implementation. There are also a number of studies that embedded the gift-exchange game in a competitive experimental market (Fehr et al., 1993; Kirchler et al., 1996; Fehr et al., 1997, 1998a,b; Fehr & Falk, 1999; Hannan et al., 2002; Brandts & Charness, 2004; Brown et al., 2004, 2008). They generally find little differences to the standard experiments, that is, there is still a positive correlation between first and second mover cooperation even under strong competition, both on the demand and the supply side.

However, while all of those approaches are rather indirect (and not entirely unproblematic, in particular the experiments using multiple games) in measuring the extent of conditional cooperation, a very direct and simple way is the familiar sequential PD. Clark & Sefton (2001) conducted an experiment using this game (120 students from the University of Manchester, and 120 students from Penn State University), in which any subject played ten periods under a round robin rematching scheme, approximating one-shot play. They found that first movers cooperated between 32 and 58 percent, and second movers between 15 and 23 percent. The interesting question, however, are the frequencies of second mover cooperation conditional on the first mover actions: if the first mover defected, only between 3 and 6 percent of the second movers responded with cooperation, while between 35 and 39 percent of them

[21] Similar manipulations have been done with the trust game. Johansson-Stenman et al. (2005) find no differences in second mover behavior for different, even considerable (transfers equivalent to 1.683 USD) stake sizes. Most other studies that examined the effects of stake size on behavior in experiments have analyzed the ultimatum game (see the following section). In all studies, first movers' behavior is independent of the stake size (with the exception of Fu et al., 2007, who found that transfer decrease with stakes), but the respondents tend to accept smaller (relative to the endowment) offers with high stakes compared to conditions with low stakes (Hoffman et al., 1996; Slonim & Roth, 1998; Cameron, 1999; Munier & Zaharia, 2002; Fu et al., 2007). No significant effects are found in the dictator game (Forsythe et al., 1994; Cherry et al., 2002; Carpenter et al., 2005; List & Cherry, 2008). Cooperation in the centipede game is reduced significantly with lower stakes (Parco et al., 2002).

cooperated in response to the first mover cooperating. This experiment illustrates that simplest designs can produce the clearest insights.

Several additional facts underpin their robustness. First, the average degree of conditioning is virtually constant over the ten periods, even slightly increasing. Second, the researchers also implemented a condition with all payoffs doubled, and a condition in which only the temptation payoff (defect in response to cooperation) was doubled, and found that in both conditioning was weaker but not eliminated. Finally, the results are approximately the same in both locations. A weakness of the experiment is that Clark & Sefton (2001) only observed actual paths of play, such that their comparison is between-subjects. It would be interesting to have second movers' complete strategies elicited in a sequential PD by means of the strategy method. I am not aware of such an experiment.

2.2.3 Non-tactical and tactical cooperation interact

The reviewed evidence in section 2.1 shows that the *average patterns* of behavior in finitely repeated games fit the imitation model outlined in section 1.2.2 quite well, but the *fine-structure* of behavior not exactly. In particular, many subjects appear to adopt less harsh strategies. For instance, in the experiment by Anderhub et al. (2002) reviewed above a large fraction of first movers granted coplayers that defected early in the game «another chance». Might this have to do with the non-tactical, backward-looking conditional cooperation discussed above?

Cochard et al. (2004) conducted an experiment (72 students sampled from different not further specified universities in France) using the investment game. In the main condition, the game was repeated seven-fold with random rematching in between.[22] Again, the researchers found the standard dynamic patters: After significant cooperation in the initial periods, it collapsed in the terminal period. While there was no second mover who returned zero in the first five periods, 31 and 56 percent did so in the penultimate and ultimate period, respectively. However, 56 and 11 percent of the second movers, respectively, still returned more than two thirds of their bud-

[22] They also implemented a control condition which was a strict one-shot version of the investment game. Comparing average cooperation rates across treatment conditions, cooperation was notably higher in the partners condition than in the one-shot condition. The average first mover transfer was 7.5 tokens in the partners condition and 5.0 in the one-shot condition. Second movers who received a positive transfer returned on average 56.1 percent in the partners condition and 38.2 in the one-shot condition. In sum, the average joint payoff was about 17 percent higher in the partners condition (35.0) than in the stranger condition (30.0). Interestingly, the additional surplus accrued almost exclusively to the first movers: second movers earned on average the same in both conditions (19.7 tokens in the partners and 19.3 in the one-shot condition), while first movers earned about 43 percent more in the partners condition (15.2 tokens in the partners and 10.7 in the one-shot condition). It should be noted that their comparison is not a particularly clean one since more than one factor is changed across treatments. Specifically, the repeated game offers the opportunity to learn while the one-shot game does not. Relatedly, they had to change the conversion rate between Francs and tokens in order to keep overall earnings approximately equal, but this may change the incentives in the individual stage game. A more fortunate design would have been, in may opinion, the inclusion of a seven period game with stranger or total stranger rematching protocol.

get in the last two periods. The latter is clearly inconsistent with tactical reputation building. Using panel data estimation methods, Cochard et al. (2004) tried to isolate the extent of non-tactical conditioning in both first and second movers. Concerning the second movers, they defined (non-tactical) conditional cooperation as positive within-period correlation between first and second mover cooperation. Likewise, (non-tactical) conditional cooperation in first movers was defined as positive across-period correlation between first and second mover cooperation. They find both types of correlation.

Gächter & Falk (2002) did a similar exercise using data from a gift-exchange game experiment (116 students sampled from various educational facilities in Vienna). They compared a «partners» condition to a «(perfect) strangers» condition.[23] They found that first mover transfers were quite similar in both conditions, but second mover cooperation was, on average, significantly higher in the partners condition, but with the familiar drop to approximately the same level as in the perfect stranger condition in the ultimate period. Pooling the data, the correlation between first mover and second mover cooperation is stronger in the partners compared to the perfect strangers condition, that is, reciprocal behavior is more pronounced in the former, but also not entirely absent in the latter. This kind of reciprocity cannot be accounted by strategic reputation building, because a second mover in the stranger condition will certainly never play with the same coplayer again.

Gächter & Falk (2002) use their data to classify the second movers in the perfect stranger condition into two types: A subject is classified as a «strong reciprocator» if her or his return cooperation is highly significantly correlated with the first mover's cooperation. A subject is classified as selfish if (s)he defects (chooses the minimum cooperation level) in at least six of the ten periods. According to these criteria, 53.3 percent of the second movers in the perfect stranger condition were strong reciprocators, and 20 percent were selfish.

In the partners condition, the share of second movers with a highly significant correlation between own and their coplayer's cooperation was notably higher at 67.8 percent. However, since in the partners condition there are strategic reasons to behave reciprocally, this share includes both strong reciprocators and the imitators. To separate them, Gächter & Falk (2002) supplement the previous «strong reciprocity» criterion by requiring a return cooperation of strictly above the minimum in the terminal period. Analogously, a subject is an «imitator» if the latter qualification is violated, that is, if (s)he defects in the final period. Finally, a subject is classified as «myopically selfish» if her or his return cooperation fails to be significantly correlated with the first mover's cooperation and if (s)he defects in the terminal period. According to these criteria, 48 percent turn out to be «strong reciprocators», 20 percent «imitators», and 21.4 percent «myopically selfish» types. Thus, putting the former two types together, 68 percent of second movers in the partners condition conditioned

[23] Kirchler et al. (1996) conducted the same experiment before, but (1) their presentation of results was somewhat sparsely, and (2) they changed the exchange rate between tokens and money among conditions (lower in the partners condition), which may confound the results.

their cooperation on the information about their coplayers' behavior. This likely accounts for the higher average cooperation rates compared to the perfect stranger condition.[24]

Gächter & Falk (2002) also point to an important difference in the patterns of finitely repeated games with a simultaneous-move stage game and those with a sequential stage game. In their experiment less first movers can be classified as reciprocal («strong» or «imitator»), in the sense of responding to their coplayer's previous period action, than second movers. Under the assumption of successful randomization, it is implausible to expect there to be more conditional cooperators among the second movers than among the first movers, such that there must be a different explanation. Gächter & Falk (2002) speculate that conditioning across periods may be less obvious to subjects in supergames with a sequential stage game than in supergames with a simultaneous stage game; in the former within-period reciprocity may be focal. I am not aware of a study that investigates this conjecture rigorously, but it may be useful to consider this point in interpreting results from existing studies and perhaps in designing new experiments.

In sum, the results reviewed in this section displace a significant part of the mystery of what goes on in finitely repeated cooperation games. A significant fraction of subjects seems to resort to a non-tactical form of conditional cooperation which rewards cooperation with cooperation in return even in the terminal period. The existence of those types, in turn, provides for an empirical justification of the imitators' belief in «committed types», that entered the theoretical model (section 1.2.2) as an unexplained assumption.

2.2.4 Conditional cooperation in public good games

I take a specific look at public good games because it is not at all clear how conditioning of cooperation on coplayers' behavior should extend to this game. To see this, suppose a player in a three-player public good game is confronted with one coplayer cooperating and the other coplayer defecting. How should a conditional cooperator behave in this situation? On the one hand, a conditionally cooperative strategy prescribes to return the cooperation of the other cooperator, on the other hand it prescribes to defect on the defector.

Furthermore, simultaneous-move games are generally not ideally suited to study conditional cooperation, and non-tactical conditioning in particular, because (i) subjects need to respond to a belief about their coplayers' action, such that they have a pecuniary reason to cooperate with players whom they believe are also cooper-

[24] Gächter & Falk (2002) conducted also some trials in which another sequence of ten periods was conducted with the same subjects to check for learning effects. All results turn out to be robust with respect to experience. In fact, the reciprocal patterns become even more pronounced in the second sequence.

In a smart modification of a finite horizon public good experiment, Sonnemans et al. (1999) replaced one group member by another after a prespecified number of periods. They also found evidence of both tactical (forward looking) and non-tactical (backward looking) behavior.

ating, and (ii) unconditional cooperation may «pollute» behavior in a difficult-to-disentangle way.

Although it was not their primary aim to test for conditional cooperation, Keser & van Winden (2000) and Brandts & Schram (2001) both identified behavioral patterns in their public good experiments that strongly pointed towards its existence. In the experiment by Brandts & Schram (2001), subjects reported a complete contribution function contingent on various marginal rates of transformation between a public and a private good. Their results show that subjects' behavior cannot be explained exclusively as the result of errors, but that they exhibited quite consistently one of two behavioral types: (i) selfish payoff maximizers and (ii) intrinsically cooperative types, where some features of the data indicated the latter's cooperation may be conditional on the coplayers behavior. They suggested that the *interaction* between these two types may be important in accounting for the behavioral dynamics, namely initially high and successively declining contributions.

Extending this conjecture, Fehr et al. (2002a) suggested that the pattern results from non-tactical conditional cooperation («strong reciprocity»), together with some selfish bias or presence of selfish coplayers in the group (Fehr & Fischbacher, 2003, 2004a).[25] To illustrate, assume that there are free riders and conditional cooperators in a population and players are randomly drawn in each period to play a sequence of public good games. Free riders never contribute anything and conditional cooperators contribute conditional on their belief about the other group members' contributions. Monitoring is private in the sense that subjects are informed only about the contributions in their current group, but perfectly so. For free riders this information is irrelevant because they will not adjust their behavior upon receiving it. Conditional cooperators, however, may respond to this this information in the next period if it conflicts with their prior belief about the cooperativeness of the other players in the population. This effects may be exaggerated by the fact that the average conditional cooperator does not fully match the others' average contribution (Fischbacher et al., 2001).

Croson (2007), Neugebauer et al. (2009), Fischbacher & Gächter (2010), and Dufwenberg et al. (2011) elicited each subject's beliefs about other group members' contributions, and found a marked and statistically significant correlation between those beliefs and own contributions. Fischbacher & Gächter (2010) add an individual level analysis and find some heterogeneity in the sense that some subjects exhibit a positive correlation between beliefs and contributions, whereas other subjects do not cooperate even if rather optimistic about the coplayers' contributions. Clearly, there are at least three problems with this interpretation. First, beliefs are endogenous and are therefore difficult to control by the experimenter. Second, a free rider who believes that others contribute zero and actually contributes nothing him- or herself is observationally equivalent to a pessimistic conditional cooperator who only contributes a little because he or she believes others will free ride. Third, beliefs may reflect a «false consensus effect», that is, people may project their own motivation

[25] See Ambrus & Pathak (2011) for a formalization.

unto others (e.g. Kelley & Stahelski, 1970; Orbell & Dawes, 1993).

Fischbacher et al. (2001) and Fischbacher & Gächter (2010) circumvent these problems by using an adaption of the «strategy method» (Selten, 1967) in order to elicit a conditional contribution function for each subject: for each (integer) average contribution of the other group members, a subject had to specify a contribution, where the contingency with the true average was actually implemented.[26] Furthermore, the game was played strictly one-shot. The researchers classified their subjects according to their contribution function into two main classes: they labeled a subject as a *free rider* if and only if (s)he contributed nothing in all contingencies, and a *conditional cooperator* if the contribution function is (weakly) increasing in the others' average contribution. Note that this involves a specification of the above-mentioned problem of how to extend the concept of conditional cooperation to the public good game. Fischbacher et al. (2001) found 50 percent of the subjects to be conditional cooperators and 30 percent to be fee riders. Likewise, Fischbacher & Gächter (2010) found 55 percent conditional cooperators and 23 percent fee riders. The rest shows either more complicated patters (about one out of ten subjects had a «hump shaped» contribution function) or could not be classified. Many other studies with related designs find similar heterogeneity, and all find that the vast majority can be classified as either free riders or conditional cooperators, with the latter constituting the majority (e.g. Burlando & Guala, 2005; Page et al., 2005; Bardsley & Moffatt, 2007; Ones & Putterman, 2007; Muller et al., 2008).

Interestingly, Fischbacher & Gächter (2010) also explicitly tested the theory outlined above. In the experiment, any subject participated in both a strict one-shot public good game with contribution function elicitation as in Fischbacher et al. (2001), and a ten period repeated public good game with random rematching in which subjects' beliefs about the coplayers' average contribution were elicited. As predicted, contributions in the repeated game declined over time from around 40 percent to around 10 percent on average. The researchers then combined each subject's contingent contribution function and elicited beliefs to predict her or his contributions in the repeated game, and compared those predictions to the actual contributions. They find that the predicted contributions match the actual contributions quite closely, although the former are generally slightly below the latter (whereas this gap vanishes towards the end). Thus, this experiment supports the above theory, and in particular highlights the important role of beliefs (and therefore information) about other's cooperation.[27]

[26] One might object to this method that it involves some experimenter demand effect in the sense that if somebody is given the opportunity to condition one's contribution on others' contributions, (s)he will tend to actually do so. This would overstate the prevalence of conditional cooperation. However, building on the design of Fischbacher et al. (2001), Kocher (2008) tested for such demand effects by running one experiment that just reversed the order of how the contingencies were presented, and another experiment that introduced a deliberate choice between an unconditional contribution and a conditional contribution. The results show that the existing findings on conditional cooperation are very robust to the implemented changes even on a quantitative basis of comparison.

[27] A corollary is that any factor that alters these beliefs will influence cooperation. Gächter et al. (2012) conducted a public good experiment in which one group member was designated as the «leader»,

2.2.5 Conditional cooperation extends to the field

At least in the laboratory, conditional cooperation seems to be a qualitatively robust phenomenon. But does this extend to the field, in particular with different subjects pools than students? To start with, the phenomenon is replicable and appears to be even stronger in the field. A number of experiments retain the standard laboratory design but recruit non-standard subjects, such as soldiers (Fehr et al., 1998a), sports-card traders (List, 2006), MBA's (Hannan et al., 2002), or CEOs (Fehr & List, 2004). They confirm the results obtained with student subjects, with non-students tending to behave even more pro-socially and conditionally so. The same results have been obtained by Falk et al. (2010) who did find no evidence that a representative sample from the Zurich citizen behave different in the investment game than a student sample from the University of Zurich (rather the opposite). They also find no evidence for a self-selection bias (more pro-social students are not overrepresented due to self-selection into the experiments) in student samples. Further one-shot trust game experiments performed in various locations worldwide (e.g. Willinger et al., 2003; Ashraf et al., 2006; Greig & Bohnet, 2008; Bohnet et al., 2008) produced results which at times differ quantitatively but not qualitatively.

Here I focus on particularly strong evidence that is provided by implementations of the experimental games used in the laboratory in large representative surveys. This approach allows to replicate the lab experiments with a large, heterogeneous sample of people and for linking large amounts of socio-demographic data to experimentally elicited behavior. Fehr et al. (2002c, 2003) integrated the investment game into the German Socio-Economic Panel (SOEP).[28] They had a sub-sample of 429 subjects that did not differ significantly in observables from the general 2003 wave of the SOEP comprising about 25 thousand individuals. The second movers' responses were elicited with the strategy method and first movers' beliefs were elicited by non-incentivized direct question. On average, the results replicate familiar patterns from the laboratory: Out the first movers, only 17 percent transferred nothing, 60 percent transferred between 2 and 5 Euros, and 11 percent transferred the full endowment of 10 Euros. Furthermore, first-movers' transfers were strongly positively correlated with elicited beliefs about return transfers.

But again, our primary focus is on second mover behavior. Most importantly, the transfer of the average second mover is increasing in the amount they received, that is, conditioned on information about their coplayer's behavior. Relating behavior to socio-demographics, young second movers (below the age of 35) give significantly

that had to make a contribution decision first. The other players were informed about this decision and then made their decisions simultaneously. The researchers also elicited the followers' beliefs about the other followers' contributions, both before and after being informed about the leader's action, which allowed them to determine how the leader's contribution influences the beliefs about other followers' contributions. They find that the leader's contribution positively influences the followers' beliefs about other followers' contributions significantly. Furthermore, the followers' actual contributions match their beliefs quite closely.

[28] The SOEP is a large representative household panel. The experiment had the same protocol as the investment game in the laboratory and subjects were paid by cheque sent a few days later.

less whereas old second over (above 65) give significantly more than the intermediate age group (35 to 55). Second movers without German citizenship, who are unemployed or recently separated from their (marriage) partner return less than the respective reference group, those with good health tend to return more. Such variables as gender, income situation, expressed worries about the own economic situation, or education status were unrelated to behavior in both first and second movers.

The study was significantly extended to a much larger sample and various accompanying experiments aimed at investigating the sensitivity of the design (the strategy method vs. direct response method, removal of the equal split option, high stakes), and is reported in Naef & Schupp (2009a). Most interestingly, Naef & Schupp (2009a) used the full sample in order to compare behavior of students and non-students. They found no significant difference between them as second movers.[29] Bellemare & Kröger (2007) conducted a similar trust game experiment with a sample that was representative for the Netherlands, and also found that students transferred less than a representative population sample.[30] In a trust game experiment conducted with 18 to 84 year old participants recruited from an online panel, Garbarino & Slonim (2009) find socio-economic and demographic information to explain little of first and second mover behavior. Gächter et al. (2004) also found no relationship between socio-economic background and behavior in a one-shot public good experiment. Thöni et al. (2012) find the usual patterns in a public good experiment with a sample representative for the Danish population. More research is certainly warranted, but the current evidence points towards the conclusion that cooperative behavior and its conditioning on coplayers' actions does not differ overwhelmingly along socio-demographic differences.

Still, the above studies keep the highly structured, critics might say «artificial», experimental games from the laboratory and people generally know that they participate in an experiment. Can conditional cooperation also be demonstrated in natural environments, and even if subjects are not aware of their participation in an experiment? Existing evidence points towards the answer «yes, but...». Gneezy & List (2006) conducted a field experiment in a labor relation setting involving two distinct tasks: data entry for a university library and door-to-door fund-raising for a research

[29] As a first mover, students transfer on average significantly *more* (61 percent of the endowment) than non-students (50 percent of the endowment). This difference upholds even when controlling (statistically) for age, income, and the level of education, which are key characteristics along which students and non-students typically differ. Furthermore, since Naef & Schupp (2009a) found students to be less risk averse than non-students, risk preferences may be conjectured to be the primary cause of the differences in transfers, but the differences remain highly significant after controlling for a measure of risk tolerance.

[30] In related large-scale experiments, Carpenter et al. (2008) found with a sample representative for a community in Vermont (USA) that students transferred significantly less than non-students in a dictator game. In a representative ultimatum game experiment in Taiwan, no difference was found between students and non-students (Fu et al., 2007). Harrison et al. (2002) found that in Denmark, students have a six percentage point higher discount rate than non-students. In a similar study in Denmark, students were found to be more risk-averse than non-students (Harrison et al., 2007). Concerning the latter result, note the conflict to Naef & Schupp (2009a).

center. Consistent with laboratory gift-exchange games worker performance in the first few hours on the job is considerably higher in a «gift treatment» than in a «non-gift treatment». After the initial few hours, however, no difference in outcomes is observed, and overall the gift treatment yielded inferior aggregate outcomes for the employer: with the same budget the employer would have logged more data for the library and raised more money for the research center by using the market-clearing wage rather than by trying to induce greater performance with a gift of higher wages. Kube et al. (2006, 2011b) extended the design to allow also for negative reciprocity, and found that the gift (wage increase) does not work well in the long run (if at all) as well, whereas a wage reduction (the negative reciprocity treatment) had a significant and lasting negative impact on performance. Thus, there appears to be a marked asymmetry of «positive» (reciprocating cooperative behavior in-kind) and «negative» reciprocity (reciprocating non-cooperative behavior in-kind).

In a similar field experiment designed to measure worker responses in a tree-planting firm to a monetary gift from their employer (Bellemare & Shearer, 2009), firm managers told a crew of tree planters they would receive a pay raise for one day as a result of a surplus not attributable to past planting productivity. A comparison of planter productivity on the day the gift was handed out with productivity on previous and subsequent days of planting on the same block showed that the gift had a significant and positive effect on daily planter productivity, controlling for planter-fixed effects, weather conditions and other random daily shocks. Kube et al. (2011a) finds in a similar labor context that non-monetary gifts have a much stronger impact than monetary gifts of equivalent value. They also observe that when workers are offered the choice, they prefer receiving money but reciprocate as if they received a non-monetary gift. Furthermore, monetary gifts can effectively trigger reciprocity if the employer invests more time and effort into the gift's presentation. This suggests that the employer's *intention* and not only the consequences are important in triggering reciprocity (see also section 2.3.6).

Falk (2007) reports evidence from a field experiment in collaboration with a charitable organization, that sent roughly 10,000 solicitation letters to potential donors. One-third of the letters contained no gift, one-third contained a small gift, and one-third contained a large gift. In support of the hypothesis, the relative frequency of donations increased by 17 percent if a small gift was included and by 75 percent for a large gift compared to the no gift condition. Also in a charity context, Frey & Meier (2004) exploited the fact at the University of Zurich each students has, upon registering for the new semester, the opportunity to donate, in addition to the tuition fee, to one of two charity funds, one helping needy students with subsidized loans (CHF 7 donation), and the other generally supporting foreign students (CHF 5 donation). Students in the treatment conditions received information about about the donation behavior of the other students, in one condition that 64 percent of the other students donated, in the other condition that 46 percent made a donation in the past (those frequencies were actual frequencies from past but different semesters). Students in the control condition received no information. Frey & Meier (2004) found (i) positive

and significant correlation between students' expectations (which were also elicited) and own donations, and (ii) that students who were informed that 64 percent donated in the past are more likely to donate than those who received the information that only 46 percent donated.[31]

In a similar natural field experiment, Heldt (2008) asked tourists using a cross-country skiing slope for a donation to help for keeping the slope well-prepared. He also manipulated the information people got, and found that those who were told that 70 percent of the other tourists donated contributed significantly more than those who received no further information. In a natural field experiment by Martin & Randal (2008), visitors to an art gallery in New Zealand were given the opportunity to leave a donation to the museum in a transparent box. In one condition there was already some money in the box, in another condition the box was empty. They found that people donate significantly more in the former than in the latter condition.

Shang & Croson (2009) conducted a natural field experiment around a fund-raising tour for a public radio station. In the three treatment conditions subjects were told that they had just another member who contributed either USD 75, 180, or 300. In the control condition no such information was given. The researchers found that callers who were confronted with a previous pledge of USD 300 donated significantly more than people in the control condition. Callers in the other two information conditions also contributed more than the control group, but those differences were not statistically significant.

Alpizar et al. (2008) conducted a natural field experiment at a national park in Costa Rica in which visitors were asked for donations for the park's maintenance. They found that (i) contributions made in public in front of the solicitor were a quarter higher than contributions made in private, (ii) giving subjects a small gift before requesting a contribution or (iii) telling subjects that the typical contribution of others is USD 2 (a small contribution, instead of telling nothing) both increases the *likelihood* of a positive contribution but decreases the *size* of the average contribution compared with providing no reference information, whereas (iv) providing a high reference level (USD 10) increases both the likelihood and the average size of the donation.

There are also studies that connect laboratory and field experiments. Benz & Meier (2008) replicated the natural field experiment described above (Frey & Meier, 2004) in a laboratory, which involved exactly the same donation decision to the same funds as in the field experiment. As a control condition, they also conducted a second experiment, which involved the same donation decision but to another charity unrelated to the university. They find that behavior in the two lab experiments are significantly correlated with behavior in the field experiment, approximately to the

[31] Within the same experiment, Meier (2005) explores the role of framing effects and found that their influence is limited. People behave in a conditionally cooperative way if informed either about the number of contributors or about the equivalent number of non-contributors. The positive correlation between group behavior and individual behavior is, however, weaker when the focus is on the defectors. He also finds gender differences in social comparison.

same degree. This holds up in more refined statistical analysis. Along the same lines, Carpenter & Seki (2011) conducted a finitely repeated public goods experiment (with and without opportunities to express disapproval) with Japanese fishermen in the laboratory, and related their behavior in the experiment to collected data on their daily fishing activities. The researchers used the data from the lab to statistically derive five measures of social preferences for each fisherman: his level of unconditional cooperativeness; his conditional cooperativeness; the propensity to disapprove; the fisherman's response to received social disapproval; and finally, the level of the unconditional response to disapproval. The results show that fishing productivity is significantly related to the experimentally derived measures of social preferences.

Finally, there are also some informative studies that use survey data or other existing data sets. Using survey data from 30 West and East European countries, Frey & Torgler (2007) found marked negative correlation between perceived tax evasion and tax morale. Using survey data from the European Values Survey (EVS), Torgler et al. (2009) found a similar (small but significant) positive relation between perceived environmental cooperation (reduced public littering) and voluntary environmental morale. Using retrospective survey data on charitable giving (the Giving the Netherlands Panel Study 2005), Wiepking & Heijnen (2011) find that in the case of door-to-door donations, social information affects perceived social norms for giving and, through this perception, influences the level of actual donations. They also found that people in different income categories donate roughly the same amounts in separate instances (they use the same social information), and as a result people in lower income households donate a higher percentage of their income to charitable organizations. Ferrary (2003) finds gift exchange being the principle rule of exchange in the social networks that underlie the industrial networks of *Silicon Valley*.

In sum, the evidence from field experiments support the evidence from the laboratory qualitatively in that conditioning on other's cooperation is common, while the order of magnitude may differ.

2.2.6 Conclusion

In this section I considered one fundamental question with which I concluded chapter 1 in greater detail: Do real people *only* cooperate for tactical reasons? A large body of evidence shows clearly «no». But in this section I have put special emphasis on evidence that is informative with respect to our guiding question: What does the propensity of individuals in a given group to cooperate with one another has to do with the availability of information they have about each others' actions?

The received body of evidence shows that part of observed non-tactical cooperative behavior is not, but a significant part *is* conditioned on the coplayer's behavior in settings in which this information is readily available. This holds even for one-shot interactions for which the traditional behavioral model based on rational selfishness predicts no cooperation and therefore no conditioning at all (see section 1.1). However, the evidence is not yet conclusive on the quantitative importance. While measurement of the degree of conditioning is in principle not subject to the same prob-

lems as measurement of *absolute* degrees of non-tactical cooperation (see e.g. Zizzo, 2004; Bardsley, 2008), because it is conceptually a *difference* between the level of cooperation between contingencies, there is still much unexplained variation in the degree of measured non-tactical conditional cooperation. Nonetheless, around half of the subjects typically exhibits a tendency to return cooperation with cooperation even if a material return from doing so is ruled out.

But if some subjects are willing to incur a cost to confer a benefit to someone else, are there also subjects who are willing to incur a cost to impose a cost on someone else (such behavior was defined as spite in section 1.1.1)? I turn to this question in the next section, again with a focus on the degree of conditioning on coplayers' past behavior.

2.3 Spite and punishment

Aggression and spite, that is, behaviors that are immediately costly to both the aggressor and the target are widespread in the animal kingdom including, of course, in humans (Clutton-Brock & Parker, 1995; Knauft, 1991; Pinker, 2011). Dominance, relative standing, and envy are important proximate motivators of such behavior (Clutton-Brock & Parker, 1995; Boehm, 1993; Kirchsteiger, 1994; Mui, 1995).[32] There are a number of experimental studies that document human subjects' willingness to «burn others money» at a cost to themselves, even in simple unidirectional matches in which they neither respond to the target's behavior nor can enforce any behavior in the future, and a high fraction seems to do this in order to be *relatively* better off (e.g. Zizzo & Oswald, 2001; Zizzo, 2003, 2004; Abbink & Sadrieh, 2009; Abbink & Herrmann, 2011).

However, if applied conditionally, spite can also be used as a means of punishment. Behavioral psychologists speak of «positive punishment», the presentation of an aversive stimulus, in contrast to «negative punishment», the omission or withdrawal of an appetitive stimulus (e.g. Powell et al., 2002; Mazur, 2002; Schachtman & Reilly, 2011). In the context of the enforcement of cooperation, the latter form of punishment corresponds to the withdrawal of cooperation, which is the by far dominant form of punishment considered in economic theory. Is conditional spite in fact used an additional means to enforce cooperation? If yes, then the scope of situations in which transparency might matter is even more expanded, punishment is by definition conditioned on past play.

2.3.1 Spite and punishment in finite horizon games with perfect monitoring

The classic demonstration of spiteful behavior as an enforcer of cooperation is observed second mover behavior in the ultimatum game (Güth et al., 1982). In the ultimatum game, the first mover can offer to transfer an amount of money from her or

[32] Arguments in bargaining theory also suggest that surplus destruction can increase bargaining power (Avery & Zemsky, 1994).

his endowment to the second mover, whereas the latter can either accept, which case the transfer is implemented, or burn the entire endowment such that both get nothing. On average, first movers offer on average about 40 percent of their endowment, about 16 percent of the offers are rejected, and rejections decrease in the amount offered, whereas offers below 20 percent are typically rejected half of the time (Camerer, 2003; Oosterbeek et al., 2004). Thus, the evidence suggests that the average second mover conditions her or his option to burn the money on the first mover's behavior, such that it serves to enforce more cooperative behavior in the latter. These results have been replicated a large number of times (see Güth & Tietz, 1990; Güth, 1995; Camerer & Thaler, 1995; Camerer, 2003, for overviews), and do not differ significantly even with very large stakes (e.g. Hoffman et al., 1996; Slonim & Roth, 1998; Cameron, 1999; Munier & Zaharia, 2002; List & Cherry, 2008). There is, however, some cultural heterogeneity, to which I will come back below (see Henrich et al., 2001, 2004, 2005, 2006; Oosterbeek et al., 2004).

In keeping with much of the recent literature, I focus on the public good game here for the following reason. In dyadic games, it is in principle easier to punish the coplayer by withholding cooperation («negative punishment») because such punishment is automatically targeted on a specific player, while in games with more than two players this becomes difficult. This is not to say that the option for «positive» punishment is unimportant in dyadic games, since it may avoid long periods (or many domains, if the players interact in multiple domains simultaneously) of defection. But on top of that, it is easy to see that in public good games it becomes difficult to punish defection of one or a few group members by withdrawing cooperation, because all other (cooperating) group members are punished as well (Boyd & Richerson, 1988b). The collateral damage of punishing defection with defection is therefore sizable.

The most comprehensive and influential study that allowed for costly reductions of coplayers' payoffs was conducted by Fehr & Gächter (2000a).[33] In their experiment (40 students from the University of Zurich), they used a finite horizon public good game with ten periods played under two conditions: The control condition was just an ordinary public good game, whereas in the treatment condition there was a second stage in each period in which subjects where informed about each coplayers' contribution, and given the opportunity to reduce each coplayer's payoff at a cost to themselves, whereas higher reductions were also more costly to the actor. Conditions were varied within-subject by having each subject play one match in one condition, and then under the other condition (the restart was not announced and the researchers also reversed the sequence across sessions).

The results are remarkable. First, contribution levels were moderate and decreasing over time to low levels in the control condition, and strongly increasing to almost maximal cooperation in the treatment condition. As a result, average contributions

[33] Seminal studies that implemented an option of explicit punishment into public good (or closely related) games were Yamagishi (1986) and Ostrom et al. (1992). Yamagishi (1986) allowed players to contribute to a sanctioning system that automatically imposed punishments in the main PGG, which is apparently itself a public good. Ostrom et al. (1992) studied punishment in a shared resource game.

were much higher in the latter (85 percent) than in the former (38 percent). Why does the option to reduce coplayers' payoffs have this effect? First, the discriminate use of this option as a punishment of free riding is to some extent *anticipated* since differences in contributions between conditions are already present in the very first period. Second, punishment of free riding is actually carried out, as on average a subject's payoff is more strongly reduced the more her or his contribution falls short of the coplayers' average contribution in the same period. However, not all reductions are used in this manner, since there is also a small but positive minority of high contributors whose payoff was reduced. However, free riders clearly faced the option to escape significant payoff reductions by increasing their contributions to the public good. In fact, almost 90 percent of the subjects who actually experienced payoff reductions increased their contribution immediately in the subsequent period. These responses induced the increase in contributions over time and and a decrease in actual payoff reductions, such that, after falling short initially, aggregate efficiency exceeded that in the control condition. Numerous experiments have replicated these results (see Balliet et al., 2011, for a meta study), some with a considerably longer horizon of 50 periods, after which aggregate efficiency more clearly exceeded that of a control condition without payoff reduction option (Gächter et al., 2008; Ambrus & Greiner, 2012).

2.3.2 Pecuniary returns are of minor importance

In the experiments using finite horizon games reviewed above, payoff reductions can in principle be motivated by prospective pecuniary returns, since targeted punishment may (and usually does, as noted above) modify the target player's behavior towards more cooperation in subsequent periods. As conditional cooperation in finite horizon games, such behavior may be purely tactical insofar as subjects invest into costly punishment up to the point where costs are equal to the expected future return. The evidence reviewed in the previous section shows that parts of conditionally cooperative behavior cannot be tactical in this sense, and it turns out that costly punishment is even less.

A first hint comes already from the experiment by Fehr & Gächter (2000a) reviewed above. With the punishment option there was still almost full cooperation in the terminal period, which strongly contrasts with finitely repeated games without costly punishment option (see section 2.1). This suggesting that the threat of punishment is credible even in the terminal period.

Clear evidence comes from two additional conditions in which subjects were rematched into new groups after any period, in one randomly («stranger condition», 72 subjects), and in another in a round-robin fashion such that no pair of subjects met more than once («perfect stranger condition», 24 subjects). In fact, the results from both conditions differed not much to the finite horizon condition. There were the same dynamic differences between the conditions with and without option to decrease payoffs, resulting in much higher average contributions in the former (58 percent in the stranger condition) than in the latter (19 percent). Contributions in

the perfect stranger condition were slightly below the stranger condition, but with approximately the same differential impact of a punishment option. Fehr & Gächter (2002) replicated the perfect stranger condition with a slightly improved design (a constant fine-to-fee ratio, see Casari, 2005 on this issue) and a larger sample (240 students from the University and ETH of Zurich), and found the same results. Thus, while the overall level of cooperation is higher in the finite horizon condition than in the other two conditions,[34] the impact of introducing an opportunity for direct costly punishment is by and large the same. More interestingly, the actual payoff reduction patterns are approximately the same in the (perfect) stranger treatment than in the partner treatment, and still around 80 percent of the subjects who actually experienced payoff reductions increased their contribution immediately in the subsequent period.

The result have been replicated many times (see Balliet et al., 2011, for a meta-study), and is consistent with the large amount of evidence on the ultimatum game. Experiments by Masclet et al. (2003) and Noussair & Tucker (2005) show that pecuniary consequences are not necessary for punishment to have an effect on cooperation, even entirely symbolic expressions of disapproval work, but pecuniary punishment has a stronger effect, and a combination of both is particularly powerful (see also Balliet et al., 2011).

There are, however, documented exceptions with respect to the payoff enhancing effect of mutual sanctioning. One set of exceptions has to do with the structure of interaction. For instance, the effect is weak if group structure is asymmetric in the sense that some group members have negative costs of contributing (Reuben & Riedl, 2009), or if payoff reductions are inconsistently applied (van Prooijen et al., 2008; Drouvelis, 2010). The second set of exceptions, likely related to the latter point, has to do with cross-cultural variation with respect to the cooperation enhancing effect of introducing the option to reduce others' payoffs. I come back to this below.

But even in cases in which cooperation is enhanced, the payoff reductions and their costs to the actor may eat up all efficiency gains such that aggregate payoffs are frequently below a condition without option to decrease payoffs (e.g. Page et al., 2005; Bochet et al., 2006; Sefton et al., 2007; Dreber et al., 2008; Egas & Riedl, 2008; Herrmann et al., 2008; Masclet & Villeval, 2008; Nikiforakis, 2010). This may be particularly severe if subjects respond to received payoff reductions not by increasing their contribution but by decreasing payoffs in return, that is, if «feuds» emerge (Denant-Boemont et al., 2007; Nikiforakis, 2008). However, it should be noted that the typical of the cited studies use a relatively short duration of usually ten periods. If the horizon is longer such that subjects have the opportunity to learn, aggregate efficiency is usually clearly higher than in the control treatments without payoff reduction option (Gächter et al., 2008).

[34] The fact that cooperation is higher in the partners treatment (see also Masclet et al., 2003) suggests that conditional cooperation and the opportunity for payoff reductions reinforce each other (Shinada & Yamagishi, 2007).

2.3.3 Even unaffected third parties take action

A set of seminal experiments by Fehr & Fischbacher (2004b) showed that even third parties who are completely unaffected (in pecuniary terms) from the actions of the target player are willing to spend money to reduce the latter's payoff. More importantly, the majority conditions those reductions on the target player's behavior towards her or his coplayers. Specifically, in one experiment (72 students from the University of Zurich) two players played an experimental PD strictly one-shot, and a third player was informed about those players actions and had the option to reduce either player's payoff at a cost, similar to the experiment reviewed above. The experiment allowed for powerful conclusion with respect to conditioning because the strategy method was employed. Almost half of the third parties reduced the payoff of cheaters (defectors on cooperators), and they forwent on average around 35 percent of their payoff to do so. Around one fifth reduced also the payoff of defectors whose coplayer defected as well. However, consistent with the evidence reviewed above a minority (around 8 percent) also reduced the payoff of cooperators.

In a second experiment (48 subjects) in which the behavior of second parties (the affected coplayer in the PD) and third parties was compared showed that slightly more second parties were willing to reduce a defecting coplayer's payoff (67 percent) than third parties were willing to reduce the payoff of cheaters (59 percent), and the the former spent on average clearly more than the latter. Furthermore, more third parties reduced the payoff of cooperators (15 percent) than second parties (8 percent). Those results suggest that there may be subtle differences in the motivations underlying third party and second party payoff reductions, but overall the patterns do not differ very much.

This experiment has been replicated several times, at times with slightly differing designs (e.g. Charness et al., 2008; Ottone et al., 2008; Leibbrandt & López-Pérez, 2009; Kusakawa et al., 2012, see also chapter 4), and even in a number of very diverse non-standard subjects pools (Henrich et al., 2001, 2004, 2005, 2006; Marlowe et al., 2008). However, while the latter experiments find that payoff reductions are performed in all locations, the *magnitude* varies substantially across populations. I will come back to this below.

2.3.4 Non-tactical but systematic

Since, in contrast to cooperative behavior, payoff reductions turned out such unresponsive to the matching protocol, and exhibited even the distinctive pattern in the perfect stranger treatment in which any possibility from return benefits was ruled out, Fehr & Gächter (2000a, 2002) speculated that conditionally applied spite has a functional role as punishment of uncooperative behavior, which is, however, to a very large degree non-tactical, that is, intrinsically motivated. Observed patterns of intrinsically motivated sanctioning behavior, and in particular the patterns and degree of targeting and conditioning, can be plausibly underpinned by psychological mechanisms. In particular, it has been argued that affect might play an important role

(Fessler & Haley, 2003). In a post-experimental debriefing, Fehr & Gächter (2002) confronted each participant with a number of hypothetical scenarios and asked them to indicate their intensity of particular emotions towards a specific coplayer on a Likert scale in each scenario. Subjects reported that they were angry if the free rider contributes less than they did, and clearly more angry when they contributed a high amount than when they contributed a rather low amount themselves. Subjects reported to be happy when their coplayer contributed more than themselves, whereas their reported happiness is independent from whether they contributed high (but still less than the other subject) or low themselves. In other words, the evaluation of a coplayers contribution appears to be independent from the own contribution as long as the former is above the latter, whereas anger on free riders seems to depend on the own contributions. This is quite consistent with the results on betrayal aversion by Bohnet & Zeckhauser (2004); Hong & Bohnet (2007); Bohnet et al. (2008): people dislike being exploited, and get angry on someone who does.

Similar evidence comes from a number of related games (Pillutla & Murnighan, 1997; Bosman & van Winden, 2002; Bosman et al., 2005; Hopfensitz & Reuben, 2009; Reuben & van Winden, 2010; Cubitt et al., 2011). Informative are also experiments combined with neuroscientific methods. For instance, Sanfey et al. (2003) had their subjects play the ultimatum game, while the subjects' brains were scanned by functional magnetic resonance imaging (fMRI). Scan contrasts showed that low offers elicited activity in brain areas related to both emotion (anterior insula) and cognition (dorsolateral prefrontal cortex). Furthermore, significantly heightened activity in the anterior insula during rejections of unfair offers suggests an important role for emotions in punitive behavior. Interestingly, arousal is significantly mitigated when low offers are generated by a computer instead of a human coplayer.

de Quervain et al. (2004) used positron emission tomography (PET) to to scan the subjects' brains while they learned about a coplayer's defection and determined the punishment. They found elevated activity in the dorsal striatum (which has been implicated in the processing of rewards that accrue as a result of goal-directed actions) associated with an act of punishment with payoff consequences relative to symbolic punishment without payoff consequence. Moreover, subjects with stronger activations in the dorsal striatum were willing to incur greater costs in order to punish. The researchers interpret these findings as support for the hypothesis that people derive satisfaction from punishing free riders and that the activation in the dorsal striatum reflects the anticipated satisfaction from punishing defectors.[35]

Knoch et al. (2006) find that a disruption of the right dorsolateral prefrontal cortex (DLPFC) reduces unfair ultimatum game offers while subjects still judge such offers as very unfair. Knoch et al. (2010) find that baseline cortical activity in the right prefrontal cortex to be an individual marker for the tendency to reduce others'

[35] Buckholtz et al. (2008) found similar patterns in «third parties» while they determined the appropriate punishment for presented crimes, «raising the possibility that the cognitive processes supporting third-party legal decision-making and second-party economic norm enforcement may be supported by a common neural mechanism in human prefrontal cortex» (p. 930).

payoffs. Ben-Shakhar et al. (2007) report on significant correlation between self-reported anger and physical indicators of arousal. See also Fehr & Camerer (2007); Seymour et al. (2007); Fehr (2009b) for more general overviews.

However, intrinsically motivated does not mean chaotic. Quite the contrary, it has been shown that payoff reductions follow quite systematic patterns. A number of studies have shown that subjects respond predictably to changes in the cost and impact of payoff reductions (Anderson & Putterman, 2006; Carpenter, 2007a; Egas & Riedl, 2008; Nikiforakis & Normann, 2008; Ambrus & Greiner, 2012): The are less payoff reductions the more expensive, and stronger reductions of defectors' payoffs (and more cooperation) with a stronger impact for a given cost.

2.3.5 The degree of conditioning is culturally specific

That people get angry on coplayers which exploit their cooperation is of course consistent with a variety of motivations. They might get angry because the defector earns more than themselves, or because the defector makes them look like the «sucker», and the like. The experiments on third party intervention suggest that there may also be a strong normative component, that is, that defecting is viewed as unjust in the respective situation. The issue of what specific motivations underlie payoff reductions is far from settled and I will not go intro the details here.[36] But these considerations suggest a conjecture that is particularly important for the present purposes: both payoff reductions and the responses to them may differ culturally because different moral norms may apply. The above-mentioned results that show that the mere expression of disapproval works as a punisher suggests that part of the responses to pecuniary payoff reductions are not due to the material consequences alone. It is suspected that those expressions trigger feelings of guilt or shame which may be mitigated by responding with redress (Barr, 2001; Fessler & Haley, 2003; Bowles & Gintis, 2005a). Hopfensitz & Reuben (2009) provide confirmatory evidence. However, since those moral emotions may be triggered only if there is indeed a known moral norm that has been transgressed (Tangney, 1999; Tangney & Dearing, 2003; Tangney et al., 2007), this effect likely differs between populations in which different sets of norms apply. Recent cross-cultural evidence (see below) if affirmative. Along similar lines, the motivations underlying payoff reductions may also differ culturally because different moral norms and notions of status and dominance may apply. While it has been shown that most people from virtually anywhere in the world reduce others' payoffs at a cost, those cultural differences might especially have an impact on the degree of *conditioning*, that is, to what extent payoff reductions are used as a punishment of uncooperative behavior and to what extent they are meted out independently from the target player's previous actions.

[36] The evidence is conflicting in a number of respects. A series of recent experiments suggest that distributional concerns are important (see Leibbrandt & López-Pérez, 2009, 2011), but those experiments were designed in a way that renders those concerns particularly salient. Other experiments suggest that consequential motives are of minor importance but intentionality and responsibility is critical (e.g. Sutter, 2007; Falk et al., 2008; Stanca, 2010). I will come back to this below.

A number of recent cross-cultural studies is affirmative. Interestingly, in public good games with the option to reduce payoffs (see above), there seem to be very little differences in the degree in free-rider's payoff reductions across cultures, but marked differences in the extent to which cooperator's payoffs are reduced as well (Herrmann et al., 2008; Gächter & Herrmann, 2009; Gächter et al., 2010). For example, Gächter et al. (2010) found that very little costs are imposed on cooperators in Protestant Western (Potestant Europe, UK, USA, Australia) and Confucian countries (China, Korea), but high costs are inflicted on cooperators in Southern European and Arabic countries. In Athens, payoffs of cooperators were even slightly stronger reduced than those of defectors. In sum, there are non-trivial differences in the degree of conditioning on coplayers' actions in the public good game. Perhaps not surprisingly, the differences in conditioning translate proximately to the strength of eventual cooperation enhancing effects. Similar differences have been discovered with respect to payoff reductions by third parties (Henrich et al., 2001, 2004, 2005, 2006; Marlowe et al., 2008). In ultimatum game experiments, there is also little cross-cultural variation in first mover behavior, but second mover behavior differs (Oosterbeek et al., 2004). In particular, in some societies both offers and rejection rates were very low (in fact among the Machiguenga framers in Peru all but one offer was accepted, and offers were very low), suggesting that a norm of fair sharing seems not to exist everywhere (Henrich et al., 2001, 2004, 2005, 2006).

Various possible reasons for punishment of cooperation have been suggested. For instance, Herrmann et al. (2008) found that it is particularly prevalent in countries with widespread corruption and correspondingly weak rules of law and democratic institutions. Causality may, of course, run in either direction. At the micro-level, the most obvious reason might be relative payoff status as discussed above. Revenge may also be a motive (Denant-Boemont et al., 2007; Nikiforakis, 2008; Herrmann et al., 2008). There may also be differences in distributional fairness preferences or norms (Fliessbach, 2007; Thöni, 2011). Different norms of honor or conformism may also play a role (Carpenter, 2004; Carpenter et al., 2004; Gächter et al., 2010). There might also people who dislike «do-gooders» or «moral show-offers» (Monin, 2007). To sort out between those explanations is an important area for future empirical research.

2.3.6 An intermediate conclusion

In this section I have considered spite, behavior that is costly to the performing individual and the target individual(s), again with a focus on the degree of conditioning on coplayers' past behavior. If spite is conditioned, it can serve as a punishment. The evidence shows that spiteful behavior is very common in various subject pools worldwide. Furthermore, it is frequently used as a punishment of un-cooperative behavior, even by unaffected third-parties, with strongly disciplining effects. It follows predictable patterns, despite the fact that pecuniary incentives seem to play a significantly smaller role than for cooperative behavior. But there are exceptions and strong evidence that the degree of conditioning differs significantly across cultures.

This has repercussions for the different motivational theories that attempt to explain both conditional cooperation and conditional spite within a single behavioral model (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000; Charness & Haruvy, 2002; Dufwenberg & Kirchsteiger, 2004; Sobel, 2005; Falk & Fischbacher, 2006; Cox et al., 2007; López-Pérez, 2007, 2008, 2010). Together with conflicting evidence in experiments that designed to discriminate between them as well as sensitivity to experimental parameters and designs suggest that it is still to early for alternative behavioral models that exhibit a certain generality in the sense that they produce sufficiently precise predictions that hold «everywhere and everytime». At lot more evidence is needed for such an endeavor. However, I want to highlight one aspect for the present purposes. One empirically supported motivational theory is that there might be people who value equitable payoff distributions, that is, dislike distributions which a sufficiently unequal (Boehm, 1993, 1999; Loewenstein et al., 1989; Dawes et al., 2007; Fehr et al., 2008; Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000). Individuals with such preferences can be willing to sacrifice own payoffs to increase other's payoffs that are below an equitable benchmark, and decrease other's payoffs that are above the benchmark. Emotions of compassion and sympathy on the one hand, and envy on the other might be important psychological processes underlying such preferences (Krebs, 1970; Trivers, 1971; Becker, 1974, 1976b; Kirchsteiger, 1994; Mui, 1995). It is easy to see that such preferences can induce unconditionally cooperative or unconditionally spiteful behavior. However, they can also induce behavior in certain games that is *observationally identical* to conditional behavior. To see this, note that actions typically alter payoff distributions. If *One*, for example, performs an action that alters the payoff distribution in a way that, according to *Two*'s preferences, *One* has gotten unjustly rich (poor), then *Two* might be willing to incur a personal cost to decrease (increase) *One*'s payoff. The respective action looks like a punishment (reward) of *One*'s previous action, but *Two*'s behavior is not really conditioned on *One*'s *action* but only on its consequences. To recognize the distinction, *Two*'s «response» would have been the same whether *One* had performed the action intentionally or accidentally, or if not *One* would have implemented the unjust payoff distribution but anybody else or even a some natural (random) event.

It is a matter of definition whether one wants to differentiate such behavior from or include it in the category of conditional cooperation or punishment. The literature is not always explicit and consistent on this issue. I take the former route because the distinction is important for the present purposes, since for a player with equity preferences, the actual path of play (other players' actions) is no relevant information over and above the payoff consequences. In other words, the player is indifferent between any history of play that leads to the same payoff distribution, that is, it does not matter to her or him how a particular payoff distribution was reached. This implies that this individual, if informed about the payoff consequences of a particular coplayer's action(s), does not alter her or his behavior in response to information about *how* this payoff distribution has come about.

Intuition suggests that it matters for most people how an outcome has been brought

about. In fact, the above considerations suggest an obvious test to differentiate between «pseudo-reciprocal» (induced by equity preferences) and actually reciprocal behavior: let the payoff distributions in fact be implemented exogenously, such as a random mechanism, a computer player, or the experimenter. Such experiments have been conducted and they generally support the intuition (Blount, 1995; Charness, 2004; Falk et al., 2005; Sutter, 2007; Falk et al., 2008; Stanca, 2010; Dhaene & Bouck, 2010). Although the fine-structure of social preferences and their distribution in the population is far from settled and an active area of research (e.g. Charness & Rabin, 2002; Charness & Haruvy, 2002; Engelmann & Strobel, 2004; Fowler et al., 2005; Fehr & Gächter, 2005; Johnson et al., 2009; Bartling et al., 2009), one can safely conclude that a significant fraction of non-materially motivated actions are actually conditioned on coplayers' actions and not just their consequences (Gintis et al., 2005; Bowles & Gintis, 2011).

2.4 Imperfect monitoring

While imperfect monitoring is an issue in the theoretical literature for quite some time (sections 1.3 and 1.4 in the previous chapter), there is only a very small (but growing) experimental literature. Some papers have rather little connection to the theoretical approaches,[37] some recent papers draw closely on them. In this section I focus on studies with some random disturbance in the signals players receive, in most of them the signals are public. Thus the studies correspond mostly to the setting outlined in section 1.3. In the following section I will focus on studies that implement different conceptualizations of private monitoring, corresponding therefore mostly to the setting outlined in section 1.4.

2.4.1 Cooperation increases in the quality of a public signal

In an experiment (240 students from Ohio State University) drawing directly on the imperfect public monitoring setting considered in section 1.3, Aoyagi & Fréchette (2009) used an indefinitely repeated PD with termination probability 0.1 in which subjects received a public signal after each round, which was perfect in the control condition and imperfect in the treatment conditions. Imperfection was implemented by perturbing the payoffs with a random shock, just as in the Green-Porter model outlined in section 1.3. They had three treatment conditions, low noise, medium noise, and high noise. In addition, they implemented two further comparison conditions, one condition in which there was no public signal at all (i.e. «infinite noise») and one quasi-one-shot condition with perfect monitoring in which players were randomly re-matched after any period.

[37] Cason & Khan (1999) conceptualize «imperfect monitoring» as a delay in which players learn the coplayers' actions. They implemented this idea in an experiment using a finitely repeated public good game in which subjects are informed about the coplayers' contributions every six periods. They did not find any significant differences to a condition in which subjects learn contributions in every period.

Recall the testable prediction for the case of imperfect public monitoring outlined in section 1.3: The frequency of cooperation can be expected to increase (or at least not decrease) as the accuracy of the public signal increases. The results of Aoyagi & Fréchette (2009) support this hypothesis. Specifically, they find the following. First, there is cooperation at any level of noise. Second, the frequency of cooperation is lower than the theoretical maximum (approximately full cooperation, see section 1.3) for the lowest level of noise. Third, the frequency of cooperation decreases as noise increases. Fourth, for high but finite noise, the frequency of cooperation is approximately equal to that in the quasi-one-shot condition. Fifth, for infinite noise (i.e., no public signal at all), the frequency of cooperation is lower than that in the quasi-one-shot condition. There are also distinctive patterns in the dynamics of play: While in the low noise treatments subjects tend to cooperate more often as they accumulate experience, the opposite is the case for the high noise treatments.

2.4.2 Are different strategies used under imperfect monitoring?

Aoyagi & Fréchette (2009) also attempted to infer strategies from observed behavior. They use a different approach than the ones mentioned in 2.1.3, namely assuming that subjects play threshold strategies with regime shifts between the cooperation and punishment states. These strategies start out in the cooperation state, switch to the punishment state when the public signal falls below a certain threshold, which may depend on the current state and the own action choice, and return to the cooperation state when the signal exceeds another threshold. In all but one noise treatment, they find that the data is best described by the simplest threshold strategies which only have a single threshold: those strategies simply check the most recent public signal and cooperate if is above some threshold, and defect otherwise.

In a similar indefinitely repeated PD experiment (278 students from Harvard University), Fudenberg et al. (2010) used another technique to infer strategies. Imperfect monitoring in their experiment was implemented by an $\frac{1}{8}$ chance that the respective other action was signaled, whereas each player was informed about whether the signal of her or his own action to the other player was erroneous. They find that the most common and successful cooperative strategies are somewhat more lenient variants of the classical TFT and grim trigger strategies: they wait for a coplayer to defect two or three times before reverting to punishment mode. They also find a high level of forgiveness: many subjects were willing to return to cooperation after a punishment phase has occurred when the gains from mutual cooperation were high. Thus, if monitoring is imperfect it can pay to «cool down» and try another time. This result supports the claim I have made in the introduction to this chapter that one cannot just interpolate results obtained in a perfect monitoring setting to ones with informational constraints.

2.4.3 Imperfect monitoring and costly punishment

What if experimental games with an option to reduce coplayers' payoffs at a cost, such as those reviewed in section 2.3, are augmented with imperfect monitoring? Are subjects more reserved or cautious with meting out payoff reductions if information about the coplayers' behavior might be wrong?

Carpenter (2007b) conduced a quasi-one-shot (random rematching) public good game experiment (735 students from Middlebury College) with an option to decease the payoff, just as considered in section 2.3. He varied the marginal per capita return (low and high), the group size (five and ten subjects), and the fraction of coplayers whose actions a subject could «monitor».[38] I put the term in parentheses because subjects were actually informed about all coplayers' contributions, but restrictions were introduced on how many of them a subject could punish afterwards (I comment on this below). Four such conditions were implemented: In the no-monitoring condition, no payoff reductions were possible at all. In the full-monitoring condition, players could decrease the payoffs of all coplayers. In the half-monitoring condition, this was possible only for half of the coplayers. Finally, in the single-monitoring condition each player could decrease the payoff of only one coplayer.

The results show that a *ceteris paribus* variation of the group size does not lead to an decline of contributions (actually a slight increase) and an *ceteris paribus* increase in the MPCR boosts cooperation significantly. Those results are consistent with previous experiments reviewed in section 2.2. The focus was to separate group size effects and monitoring constraints, which are often intermingled in discussions of public good provision settings. Specifically, he hypothesized that there is no group size effect independent from monitoring constraints. The results are affirmative, as average contributions are significantly higher in the full- and half- monitoring conditions than in the no-monitoring conditions. However, they are slightly lower in the single-monitoring condition. The problem with the design of this experiment is that it not actually impose limits on *monitoring* but merely on which coplayers could be *sanctioned*. This rules out the potentially important option of reducing payoff despite being imperfectly informed about the target players' behavior.

There are two recent studies that address this shortcoming.[39] Ambrus & Greiner (2012) conducted an experiment (339 students from the University of New South Wales) using a finite horizon public good game (with fixed matching) in which they introduce a specification of imperfect *public* monitoring. After the contribution stage, players received a signal on the contribution vector of the coplayers. In the perfect monitoring condition, the signal was accurate. In the imperfect public monitoring condition, there was a 10 percent chance that for each player that contributed a posi-

[38] The MPCRs and group sizes were chosen such that the marginal return for the group was equal in the large groups with low MPCR and the small groups with high MPCR.

[39] There are also studies who introduce other kinds of asymmetric information, such as Bornstein & Weisel (2010) who vary whether subjects know about each others endowment. Such experiments can be viewed as related to monitoring imperfections, as low contributions are not necessarily signals of cheats.

tive amount the signal contained a «zero contribution». Since all players got always the same contribution vector signals, the game was one of imperfect public monitoring (see section 1.3).[40] The researchers crossed these conditions with conditions differing in the payoff reduction technologies: there was one condition without option to decrease coplayers payoffs, one with a reduction option with intermediate fine-to-cost ratio, and one with a reduction option with high fine-to-cost ratio. They found the following. First, contributions were generally lower under imperfect (public) monitoring than under perfect monitoring, whereas the difference was only statistically significant in the high-fine condition. The same was true for aggregate efficiency, whereas here only statistically significant in the intermediate-fine condition. Second, in the two conditions in which payoff reductions were feasible, their averages were also generally higher under noisy monitoring than under perfect monitoring. Thus, subjects seem to be not more cautious under noisy monitoring. In sum, while the introduction of noise into public signals had no measurable effects in the ordinary public good game, there were more payoff reductions, less contributions, and lower payoffs under imperfect public monitoring than under perfect monitoring, with the worst outcomes in the intermediate-fine condition. The likely reason is that contributors face more payoff reductions under noisy monitoring and reduce their cooperation in response.

Similar evidence has been found by Grechenig et al. (2010), who used a finitely repeated public good game (with fixed matching) for their experiment (192 students from the University of Bonn) in which they introduced a specification of imperfect *private* monitoring. After contributions were made in a given round, all players received signals on the contributions of the coplayers. Each signal received by a player i on j's contribution had a particular probability of being wrong, whereas errors were drawn independently from a stationary distribution. Thus, the game involved imperfect *private* monitoring (see section 1.4), because each player received her or his own signal that was private information. One control condition and three treatment conditions were implemented: In the control condition, there was a 50 percent chance that a signal is wrong and players had no option to decrease the coplayers contributions. In the three treatment conditions there was an option to decease the coplayers payoff after receiving signals, and the chances of a wrong signal were 50, 10, and zero percent, respectively. The experiment revealed the following. First, in the treatment conditions the average amount of «punishment» was *in*creasing in the chance of error. Specifically, more noise increased the frequency of punishment acts, but decreased the intensity of punishment for a specific punishment act, whereas the former effect was stronger. Behind this is the fact that subjects conditioned their payoff reductions less on the signal received if it was noisy, that is, tended to ignore noisy signals and reduce payoffs more unsystematically. Cooperation in the public good game (and overall efficiency) turned out to be not significantly different between the conditions without and with a little noise (10 percent error chance), respectively, but

[40]Note that this is a very specific kind to introduce noise. The theoretical literature (section 1.3) generally assumes that the actions of all players may be signaled as any other action from the support.

with more noise (50 percent error chance) it was significantly lower. This is because high contributors tended to decrease their contributions upon payoff reductions received, which occurred relatively frequently. Comparing the control condition to the treatment condition with the same error probability showed that the option to decrease coplayers payoffs did not have a significant impact on cooperation.

2.4.4 Conclusion

In this section I reviewed cooperation experiments in which the signals players receive about their coplayers' behavior are subject to some random disturbance. In PD experiments with an indefinite horizon and imperfect public monitoring, the available evidence shows that there is cooperation at any level of noise, but always less than maximal. Furthermore, the frequency of cooperation decreases as noise increases, and cooperation in conditions with very inaccurate information is not not more frequent than under one-shot play. This supports the hypothesis derived in section 1.3, and poses empirical restrictions on the predictions of the folk theorem under imperfect public monitoring. Furthermore, subjects seem to resort to more lenient and forgiving variants of the classical TFT and grim trigger strategies, that wait for a coplayer to defect two or three times before reverting to punishment mode, and return to cooperation after a punishment phase has occurred. These result supports the claim I have made in the introduction to this chapter: interpolating results obtained in a perfect monitoring setting to ones with informational constraints is indeed problematic.

This conclusion is reinforced by the available evidence on imperfect monitoring in cooperation games with an option to reduce coplayers' payoffs. Payoff reductions become *more* frequent and cooperation less frequent as monitoring gets more noisy. As Grechenig et al. (2010) note, those results are «surprising, since people could simply choose not to make use of punishment» if monitoring is imperfect. I find these results surprising as well, because they are clearly counter-intuitive. When experiments produce counter-intuitive results, it is always a good idea to (i) reason for possible explanations, and (ii) investigate whether the results may be a peculiar artifact of the experimental design. To be specific, one might ask what might be different if subjects would have had the opportunity to do something about the noise before reducing payoffs? In this case, the informational constraints would have been not *imposed* but *chosen* by the subject, and it seems likely that this has some effect on her or his payoff reduction meted out afterwards, because it becomes more difficult to «externalize the blame» for harming others. I take up this point in the new experimental studies reported in chapters 3 and 4.

To summarize with respect to our guiding question, the available evidence shows that the level of cooperation (and efficiency) is generally increasing in the accuracy of information players have about each others' actions.

2.5 Private monitoring and publication

In the this section I focus on studies that implement different conceptualizations of *private* monitoring, corresponding to the setting outlined in section 1.4, and provide evidence on the effects of increasing its «publicness». In section 1.4 I concluded that the theoretical literature points towards the hypothesis that information transmission, or the publication or disclosure of behavioral records, may facilitate cooperation relative to conditions in which monitoring is private. A few recent studies are informative with respect to this hypothesis.

2.5.1 People care about their reputation, even affectively

In section 1.4 I briefly reviewed the theoretical literature showing that, by assuming the existence of some honest information processing mechanism from the outset, it can be proven that a conditional strategy that bases the decision to cooperate on the information about the coplayer's past behavior can occur in equilibrium. A number of recent experiments confirm that people condition their cooperation on such information (e.g. Wedekind & Milinski, 2000; Milinski et al., 2002b,a; Semmann et al., 2005; Seinen & Schram, 2006; Engelmann & Fischbacher, 2009; Stanca, 2009). As a result, individuals with a cooperative record receive more help than people with a less favorable reputation, which creates incentives to care for one's reputation, self-reinforcing the mechanism. Interestingly, people respond to (cues of) observation even if reputation management does certainly not provide any pecuniary returns, and even to mere cues of being observed (such as a picture of staring eyes) without any actual observation (Haley & Fessler, 2005; Bateson et al., 2006; Milinski & Rockenbach, 2007; Burnham & Hare, 2007; Boone et al., 2008; Rigdon et al., 2009; Boero et al., 2009; Fehr & Schneider, 2010). This suggests that part of people's attention and concern for one's reputation or social image is processed affectively without ever reaching consciousness. This is, of course, also a methodologically relevant point to consider in conducting and evaluating experimental research.

The hypothesis with which I introduced this section immediately follows: Starting from a situation with private monitoring, does «more publicness» has an effect on the frequency of cooperation?

2.5.2 Publication of behavioral records facilitates cooperation

There are two public good experiments relevant to the question posed above. In an experiment by Rege & Telle (2004) that was built around a simple one-shot public good game, there was a control condition in which subjects simply made, as usual, their contribution decision anonymously, and a treatment condition in which a subject's contribution was publicly but silently recorded on a blackboard. All other participants could see the decision. Since the game is one-shot, there is no pecuniary incentive to behave differently in the two conditions. However, the researchers found

that contributions were substantially higher in the treatment condition than in the control condition.

Andreoni & Petrie (2004) varied confidentiality in a public good experiment along two dimensions: First, in one condition subjects were unmasked by presenting their digital photos to their coplayers, in the control condition the usual anonymity applied. Second, in one condition the subjects contributions are revealed to the coplayers, in the control condition individual contributions remained private. They found that, relative to a baseline in which neither information on contributions nor individual identification was available, both adding just information on contributions or adding just the identity of the giver had no significant effect on cooperation. However, the researchers find a substantial impact of 59 percent more contributions if both is used in combination. In a supplementary experiment in which subjects had a choice to remain anonymous, virtually no subject did so and cooperation was even higher. Thus, disclosure appears to have a significantly positive effect on cooperation in public good games.

There are also some recent studies that speak more directly to the theoretical framework outlined in section 1.4. Recall that the hypothesis was derived that cooperation should be more frequent in a condition with public monitoring than in a condition with private monitoring. In a seminal study, Schwartz et al. (2000) conducted a random matching PD experiment (100 undergraduates from the University of Arizona) with indefinite horizon (random termination) involving three conditions. In the «no disclosure» condition subjects received no information about the currently matched coplayer's past action. In the «delayed disclosure» condition they received the coplayers behavioral record, except the last three periods. Finally, in the «immediate disclosure» condition the full behavioral record was revealed. They found that cooperation was more frequent in the «immediate disclosure» (42 percent) than in the other two conditions, only slightly less frequent in the «delayed disclosure» (36 percent), and much less frequent in the «no disclosure» condition (19 percent). In addition, subjects cooperated more frequently when encountering a player who has tended to cooperate in the past, and less frequently when encountering a player who has tended to defect in the past.

Duffy & Ochs (2009) also compared a private monitoring condition with two conditions in which players received different amounts of information about their current coplayer's past behavior. In addition, they also included a fixed matching condition. In their PD experiment (344 undergraduates from the University of Pittsburgh), all subjects played approximately 100 rounds of a standard indefinitely repeated PD with termination probability 0.1. In a control condition, each subject played a sequence of approximately ten supergames with a fixed coplayer within each supergame, and a new coplayer in each new supergame (rematching between matches ensured that each pair met at most once). Thus, the control condition resembled the standard repeated game setup with the same players interacting repeatedly in a given match. The treatment conditions implemented a random matching population game as described in section 1.4 of the previous chapter. Specifically, for each supergame, subjects were

matched into subgroups of size $n = 14$, and then also played the same stage game in pairs, whereas in each round pairs were randomly re-matched from this population.[41] The termination probability was, of course, identical to the control condition. After a supergame was terminated, another supergame was started with the same population until the around 100 rounds were reached.[42]

The main treatment variable was the information subjects received about the current coplayer's last action. In the private monitoring condition, subjects received no information about current coplayer's previous action at all. In the noisy-information condition, subjects were informed about the average payoff realized in the current coplayer's previous match.[43] In the full-information condition, subjects were perfectly informed about the current coplayer's action in the previous round.

The population size was sufficiently small that, given the parametrization of the experiment, a fully cooperative contagious equilibrium exists even in the private monitoring condition. Intuitively, however, one might expect that the more accurate the information players have about their current coplayers' previous action, the more swiftly a contagious punishment phase may unfold, and therefore the stronger the incentives to cooperate. Duffy & Ochs (2009) indeed find modest differences in the frequency of cooperation between the conditions. Taking session level averages, cooperation is more frequent in the information transmission conditions (17.6 and 17.3 percent in the noisy and full information conditions, respectively) than in the private monitoring condition (7.5 percent), whereas only the differences between the noisy and the private monitoring condition is significant. Furthermore, the frequency of cooperation turned out to be much higher in the control condition (54.9 percent). From those facts the researchers concluded that «it is the *matching protocol rather then the information about the opponent's history* that plays the more important role in the achievement of high frequencies of cooperation» (p. 806, emphasis in the original). I find this conclusion somewhat premature for two reasons. First, the amount of information about the coplayer's history was changed *with* the matching protocol as well. In the control condition, a player had automatically information about the entire history of her or his current coplayer, whereas in the treatment conditions this information is always limited. Thus, an appropriate treatment condition that tests for the effect of the matching protocol in isolation would be one with *perfect public* monitoring. Second, the *specific* kinds of limitations the researchers implemented may be the reason for the low cooperation rates in the treatment conditions rather than limitations *per se*. Specifically, it may be a lack of second-order information, information about the current coplayer's previous coplayer's action, that rendered the first-order information rather useless (see section 1.4.3).

[41] The researchers also implemented a condition with population size $n = 6$, but found no differences in the main results.

[42] The researchers did not tell subjects the target duration of 100 periods. The experiment just continued until the experimenters ended the session unexpectedly.

[43] Since in the PD there are only three possible realizations of average payoffs ($\eta - \kappa$, κ, $\frac{\eta-\kappa}{2}$), of which two are perfectly informative and one leaves uncertainty about which player defected, this may be called as a kind of imperfect monitoring.

To see this, consider an experiment (192 students from Penn State University) by Bolton et al. (2005), which was designed to investigate the amount of information about coplayers' past behavior necessary to support cooperation in a random matching game. In particular, they compared a condition in which players received only information about the currently matched coplayer's immediate past action to a condition in which they also received one step of *recursive* information: information about the currently matched coplayer's previous round coplayer's action. This allows for a richer set of conditional strategies; in particular a player can then evaluate whether the current coplayer defections were «justified». The key results are that introducing (relative to a private monitoring condition) the availability of first-order information had only a moderate effect, whereas the availability of second-order information unambiguously boosted cooperation. Thus, if enough information is available, subjects clearly play strategies that are more complex than just conditioning on the current partner's past action(s).

An experiment of Camera & Casari (2009) addresses the shortcomings of the study of Duffy & Ochs (2009). In their experiment (160 undergraduates from Purdue University) a population has size $n = 4$ and in each period they are randomly matched in pairs to play a PD, so each subject had a one-third probability of meeting any other subject in the population in each period. After each round, interaction terminates with probability 0.05, that is, the expected duration of a supergame was 20 periods. Each subject played five (random matching) supergames, where the rematching between supergames ensured that any pair of subjects were never matched into the same population. They implemented three main treatment conditions that differed in the amount of information available to subjects. In the private monitoring condition, subjects could observe actions and outcomes in their pair but not the identity of their coplayer. The remaining two conditions were public monitoring conditions, that is, each action was observable to all other players in the community. However, in the non-anonymous public monitoring condition, histories were associated with identities of subjects. In the anonymous public monitoring condition, subjects observed histories but no associated identities (actions were listed in random order without identifiers). Thus, the private monitoring condition resembles two informational frictions: Players (i) cannot observe identities of coplayers, and (ii) cannot observe the behavioral record of others, only the actions taken in their matches, respectively. The two public monitoring conditions relax these frictions in turn.

Parameters in the experiment were set to ensure that the efficient outcome can be sustained as one of the possible equilibria, when agents adopt the simple grim trigger extension discussed in section 1.4: Every player cooperates unless being informed about a defection in which case (s)he defects perpetually. The results are threefold. First, cooperation was frequent even under private monitoring (59.5 percent averaged over all supergames).[44] Second, through introducing public monitoring without

[44] In comparison to the experiment by Duffy & Ochs (2009) this frequency appears high but is likely the result of a relatively large gain from mutual cooperation in the stage game, a very low termination probability, and a quite small population size. In fact, Camera & Casari (2009) also conducted a con-

identity markers nothing changed at all (58.6 percent). Third, introducing public monitoring with identity markers resulted in a significantly higher frequency of cooperation (81.5 percent). Some attempts to infer strategies suggest a reason for this. First, the researchers find evidence that under private monitoring, the average subject switched from cooperation to defection after being cheated; in particular, the behavioral patterns for many subjects are consistent with a trigger, most often with a grim trigger strategy. Second, while in the anonymous public monitoring condition subjects could, in principle, switch to punishment mode once they have discovered a defection anywhere in the population (which extremely accelerates a contagious punishment phase), they appear to avoid such global punishment. Namely, while the average subject displayed a strong and persistent decrease of cooperation after experiencing a defection her- or himself, the response to a defection outside her or his matches was much weaker. Third, this was also the case in the non-anonymous public monitoring case, i.e. the average subject still largely ignored defections against other population members, but here the average subject tended to target her or his punishment to future encounters with the *same* subject, and not indiscriminately in all subsequent matches. In other words, the average subject cooperated with coplayers that have not cheated (only) on him- or herself yet, and defects discriminately against coplayers that have. Furthermore, Camera & Casari (2009) find in all conditions no evidence of forgiving, that is, grim trigger strategies seem to be the most frequent strategy employed, whereas it is *personalized* whenever possible. It follows that individual-specific information is likely much more effective that aggregate information.[45]

Note that the above conclusion were based on play of the average subject, that is, at the aggregate level the data look consistent with the notion of grim trigger play. However, an individual level analysis of the data reported in Camera et al. (2012) relativized this conclusion. They find that only one out of four subjects behaves in a manner consistent with the use of the grim trigger strategy, and subjects tried out a variety of strategies. Most importantly, Moreover, most subjects appear to adopted strategies that are *not* conditional on coplayers' actions. In comparison to the evidence reviewed in section 2.1.3, subjects appear to behave differently in population games than in repeated games with the same coplayers. The literature is, however, still in its infancy such that the is much left to be discovered.

dition with a population size of $N = 14$ and found a markedly lower frequency of cooperation (23.6 percent).

[45]The results of another private monitoring condition, in which subjects had the opportunity to reduce the payoff of coplayers at a cost speak along the same lines. Camera & Casari (2009) find the frequency of cooperation close to the non-anonymous public monitoring case (74.2 percent), which is why subjects frequently resort to the direct punishment option instead of reverting to non-cooperative mode in subsequent periods.

2.5.3 Information islands and pools

In matching games in which monitoring is private, Ghosh & Ray (1996) suggested an additional way than information transmission how players might mitigate the informational frictions: non-random matching.[46] Indeed, in reality there are many domains in which one can by and large freely decide with whom to interact and how long to continue a relationship. There are many economically significant examples of such situations. Perhaps most importantly, in non-monopolistic markets consumers can freely choose among the products and services of different suppliers, and whether they want to become, or cease to be, a *customer* of one or a few suppliers. Relatedly, in labor markets employers may choose whom to hire, and employees may choose for whom to work. In financial capital markets, investors are careful whom to trust their savings.[47]

Renner & Tyran (2004) conducted a framed experiment around an opaque market (market outcomes cannot be publicly observed) for an experience good with two quality levels and a short supply side, which is essentially a matching game with private monitoring. They implemented two conditions that differed with respect to whether buyers and sellers can develop long-term relationships. In the customer market condition, buyers can trade repeatedly with the same seller and can thus develop long-term trading relationships. Specifically, at the opening of the customer market one buyer and one seller are matched randomly. The seller then submits a price to his buyer and chooses privately the quality he wishes to provide contingent on the buyer's acceptance. If the buyer accepts his seller's offer, he learns the quality of the good and is re-matched with the same seller in the next period (there were five trading phases with ten periods each in total). Thus, the customer relationship is continued as long as the buyer accepts her or his seller's offers. If, however, the buyer rejects an offer, the customer relationship is terminated. As a consequence, both the buyer and the seller move to the anonymous market condition from this period onwards.[48] In this condition buyers and sellers do not know with whom they trade. Specifically, in any period all sellers simultaneously post prices to all buyers. Buyers receive a list of anonymous price offers among which they can choose in random order. No identifying information is revealed to the players upon acceptance, such that trading is completely anonymous. Renner & Tyran (2004) found that

[46] Methodically, there is an advantage of studying random association and selective association separately, since reputation has (at least) two distinct functions when selective association is feasible: On the one hand, it may serve as a conditioner of behavior *within* a given relationship. On the other hand, the decision whether to initiate or continue a relationship in the first place may also be conditioned on reputation.

[47] What these examples suggest is that selective association is intimately related to competition. If the demand side is short, that is, there are many suppliers in market but demand is limited, market shares become a scarce resource. This means nothing but that suppliers will compete for associations with a limited number of potential customers. Likewise, if the supply side is short, demanders will compete for associations with a limited number of suppliers.

[48] Note that returning to the customer market condition was impossible. Thus, punishment took the form of *grim trigger*.

long-term customer relationships are frequently upheld, causing a clientelization of the market. The average product quality was about three times higher in the customer compared to the anonymous market. Prices were also higher such that upholding the relationship was profitable for both sellers and buyers on average. In addition, comparing the average payoff differences between sellers and buyers by treatment shows that surpluses are shared almost equally in the customer market, while in the anonymous market the distribution of incomes are quite unequal with sellers making very small profits, since the buyers' distrust forces them to low pricing. Related market-embedded experiments obtained similar results (Brown et al., 2004, 2008). Slonim & Garbarino (2008) also finds clientelization and more cooperation in a repeated trust game with the possibility to form relationships (see also Slonim & Guillen, 2010). Chiang (2010) finds that more generous proposers and more tolerant responders are preferred as partners in an ultimatum game experiment. In a repeated public good experiment by Page et al. (2005), each subject was given a list (in random order and without further identifying information) of contributions of all other subjects in the session each third period, and the opportunity to express a preference order over those players, that is, with whom (s)he wishes to interact in the following three periods most. Rematching was implemented according to those preferences. They also found that cooperation is markedly higher under this kind of non-random than under random matching.

A plausible hypothesis about why this clientelization occurs is that long-term relationships are a response to the privateness of monitoring. Since each player observes only what happens in her or his own matches, and therefore only has information about players (s)he interacted with personally, switching partners carries an opportunity cost of lost *«informational capital»* (Bernanke, 1983).[49] The reputations (see section 1.2) acquired, based on the accumulated information about the common history, are a non-transferable, relation-specific asset.[50] Thus, the relationships that form in the setting with private monitoring can be viewed as «information islands» (Seabright, 2010, p. 251f), bonded by the accumulated informational capital. But if this is really the only glue that binds the relationships together, then disclosure of this information should mitigate such clustering. Falk et al. (2004) addressed this question by adding a condition with perfect public monitoring to the setup of Brown et al. (2004) in which all first movers could observe all previous outcomes in other matches as well. They find that clustering is indeed reduced in this condition, but the effect was surprisingly small; there was still a distinct clientelization into long-term relationships. However, the public monitoring condition turned out to produce more efficient results, presumably because there is now a more credible (implicit) threat to switch partners.

[49] See also the industrial organization literature on «switching costs» (see Klemperer, 1995, for an overview).

[50] Bernanke (1983) termed this asset «informational capital» and argued that its large-scale vaporization during the US banking crisis in the early 1930s through disruptions of long-term credit relationships was a major cause of the Great Depression.

However, in a related experiment (144 students from Penn State University) in which matching was fixed, Bolton et al. (2004) found the opposite. Subjects played 30 rounds of a trust game, and there were three conditions. In the «strangers market» subjects were randomly matched in the first period and then re-matched in a round robin fashion; subjects learned the outcomes of their own matches but no information about one another's history of action across matches was made available (i.e. monitoring was private). The «partners market» was identical except that each pair of subjects remained to be matched together for the entire duration. The «feedback market» was identical to the «strangers market» except that full previous histories of their coplayers were made available to the players. The results show a clear order it terms of efficiency, the «partners market» being most efficient, the «strangers market» very inefficient, and the «feedback market» in between. Thus, the differences seem to have to do with the matching procedure. However, more research to find out more about this and the fact that even under public monitoring many interactions take the form of enduring relationships would be interesting.

2.5.4 Conclusion

In the this section I considered studies that implement different conceptualizations of *private* monitoring. In section 1.4 I concluded that the theoretical literature points towards the hypothesis that information transmission, or the publication or disclosure of behavioral records may facilitate cooperation relative to conditions in which monitoring is private. The received evidence is affirmative. The frequency of cooperation is generally lowest in private monitoring conditions and information transmission in some from or the other increases cooperation. If matching is non-random, disclosure also reduces the clientelization effects observed under private monitoring. The evidence is more ambiguous with respect to the question whether random matching games with public monitoring provide for equally efficient results as repeated interaction among the *same* individuals, as suggested by the theorem of Kandori (1992a) reviewed in section 1.4.1.

2.6 Conclusion

What does the propensity of individuals in a given group to cooperate with one another has to do with the availability of information they have about each others' actions? To find answers to this question, I have considered a number of more specific questions that remained open at the end of the previous chapter: Do real people actually perform such tactics as envisaged in the theoretical models? Do real people *only* cooperate for tactical reasons? What strategies do they actually play in repeated games? What information about the coplayers' actions do they use? How do they respond if such information becomes limited? After having screened the existing experimental literature the following can be concluded.

In section 2.1 I focused on the first question (Do real people actually perform

such tactics as envisaged in the theoretical models?) by reviewing evidence of behavior in repeated games with perfect monitoring. Results show that a majority of subjects actually cooperates tactically. In indefinitely repeated cooperation games, cooperation is more likely to prevail if it is a supergame equilibrium (in particular if it is risk dominant) than when it is not. As predicted by the folk theorem, subjects respond to changes in the termination probability and the size of the gains from mutual cooperation. Most importantly with respect our guiding question, subjects use the available information about the coplayers' past behavior to condition their current behavior. The evidence suggests that the most commonly employed strategies are the two trigger strategies considered in section 1.2.3: TFT and GRIM.

The evidence also shows that there is typically significant cooperation in finitely repeated cooperation games. The average within-match dynamics are consistent with the strategic imitation model considered in section 1.2.2: significant cooperation over some fraction of the duration, with cooperation in the first period being increasing in the remaining duration, and a sharp decline as the terminal period approaches. However, while most people seem clearly to condition their behavior on past play, the individual level fine-structure of behavior is often inconsistent with the strategic imitation model, and subjects seem to have more difficulties to settle on a particular strategy. Finally, there is a persistent residuum of cooperation even in (approximated) one-shot interactions which cannot be accounted for by tactical considerations.

In section 2.2 I considered this residuum in more detail, and have put special emphasis on evidence that is informative with respect to our guiding question. A large body of evidence shows clearly that people do *not always* cooperate for tactical reasons, but there is also evidence that such cooperation is conditioned on coplayers' behavior anyway. This holds even for one-shot interactions for which the traditional behavioral model based on rational selfishness predicts no cooperation and therefore no conditioning at all (see section 1.1). While the evidence is not yet conclusive on the quantitative importance, typically around half of the subjects exhibits a tendency to return cooperation with cooperation even if a material return from doing so is ruled out.

If some subjects are willing to incur a cost to confer a benefit to someone else, are there also subjects who are willing to incur a cost to impose a cost on someone else? I reviewed the evidence on this question with a focus on the degree of conditioning on coplayers' past behavior, in which case it can serve as a punishment, in section 2.3. The evidence shows that such spiteful behavior is very common in various subject pools worldwide. Furthermore, it is frequently used as a punishment of non-cooperative behavior, even by unaffected third-parties, with strongly disciplining effects. It follows predictable patterns, despite the fact that pecuniary incentives seem to play a significantly smaller role than for cooperative behavior. But there are exceptions and strong evidence that the degree of conditioning differs significantly across cultures.

In sum, the evidence from experiments under perfect monitoring shows that if information about the other players' past behavior is readily available, this information

is used in order to play a discriminatory strategy, even in situations in which material returns from doing so is ruled out. As a result, there are material incentives to cooperate which is reflected in the notably higher frequency of cooperation in repeated games compared to (approximated) one-shot games. But there is also significant cooperation in the latter, in particular if costly punishment is available. In one sentence: if information about the other players' past behavior is readily available, cooperation generally occurs and can be very high if the interaction is repeated. But what happens if such information is limited?

While imperfect monitoring is an issue in the theoretical literature for quite some time (sections 1.3 and 1.4 in the previous chapter), there is only a very small (but growing) experimental literature, which I reviewed in sections 2.4 and 2.5. In section 2.4 I considered cooperation experiments in which the signals the players receive about their coplayers behavior are subject to some random disturbance (corresponding to the setting outlined in section 1.3). In PD experiments with an indefinite horizon and imperfect public monitoring, the available evidence shows that there is cooperation at any level of noise, but always less than maximal. Furthermore, the frequency of cooperation decreases as noise increases, and cooperation in conditions with very inaccurate information is not higher than under one-shot play. This supports the hypothesis derived in section 1.3, and poses empirical restrictions on the predictions of the folk theorem under imperfect public monitoring. Furthermore, subjects seem to resort to more lenient and forgiving variants of the classical TFT and grim trigger strategies, that wait for a coplayer to defect two or three times before reverting to punishment mode, and return to cooperation after a punishment phase has occurred. These result supports the claim I have made in the introduction to this chapter that one cannot just interpolate results obtained in a perfect monitoring setting to ones with informational constraints. This conclusion is reinforced by the available evidence on imperfect monitoring in cooperation games with an option to reduce coplayers' payoffs. Payoff reductions become *more* frequent and cooperation less frequent as monitoring gets more noisy.

Finally, in section 2.5 I considered studies that implement different conceptualizations of *private* monitoring. In section 1.4 I concluded that the theoretical literature points towards the hypothesis that information transmission, or the publication or disclosure of behavioral records may facilitate cooperation relative to conditions in which monitoring is private. The received evidence is affirmative. The frequency of cooperation is generally lowest in private monitoring conditions and information transmission in some from or the other increases cooperation. If matching is non-random, disclosure also reduces the clientelization effects observed under private monitoring. The evidence is more ambiguous with respect to the question whether random matching games with public monitoring provide for equally efficient results as repeated interaction among the *same* individuals, as suggested by the theorem of Kandori (1992a) reviewed in section 1.4.1. To summarize with respect to our guiding question, the available evidence suggest that the level of cooperation (and efficiency) is generally increasing in the accuracy and symmetry of information players have

about each others' actions.

I conclude this chapter with two remarks. First, the insights obtained in the reviews conducted in chapters 1 and 2 feed back on the two divergent views (outlined in the general introduction) about why people cooperate and what role information about coplayers' behavior might play. The evidence shows that both views have merit. People cooperate to a significant extent tactically, use reciprocity strategically to realize mutual benefits as envisaged by game theory, and thereby draw strongly on information about coplayers' past behavior. But people also forgo own material gains to cooperate or reduce incomes of others. A significant fraction of those latter behaviors are conditioned on information about coplayers' past play anyway. In sum, there typically is cooperation even in situations in which people have minimal or no information about others' behavior, but increasing such information generally facilitates cooperation significantly.

The second remark is the continuation of the second remark I have made in the conclusion of the previous chapter. Just as in the theoretical literature, the experimental literature considers information structures generally as an exogenous parameter. I explained in the previous chapter (and the general introduction) why this may narrow down our understanding of the guiding question of this thesis. I argue that an extension of the strategy set for opportunities to acquire and transmit information about others' behaviors might provide valuable insights not only in theory but empirical research as well. They might also qualify a number of results obtained under exogenous information structures. For instance, I noted above that the counter-intuitive results obtained in experiments with sanctioning opportunity under noisy monitoring may be an artifact exactly this design choice. The notion of viewing the information structure as endogenous is the core thrust underlying the contributions reported upon in the remaining chapters 3 through 7.

Chapter 3

Costly Monitoring and Non-Strategic Sanctioning: Experimental Evidence[1]

3.1 Introduction

Understanding the role of sanctions for maintaining cooperation in social dilemmas is a central theme in a rapidly growing experimental literature in economics. Since the seminal paper by Fehr & Gächter (2000a), the literature has examined many dimensions of this theme, for example the impact on punishment and cooperation of variations in the distance of the time horizon (Gächter et al., 2008), in the severity of punishment (Egas & Riedl, 2008; Nikiforakis & Normann, 2008), and in the cost of punishment (Anderson & Putterman, 2006; Carpenter, 2007a). One recent dimension within the literature concerns the impact of variations in the information structure of the social dilemma. Carpenter (2007b), for example, studies in a group setting how exogenous restrictions in the number of coplayers whose action a subject can observe affect that subject's punishment behavior. Ambrus & Greiner (2012) examine the impact of varying the severity of punishment on punishment and cooperation when there is a small amount of exogenous noise in the information on coplayers' actions. In a related paper, Bornstein & Weisel (2010) introduce exogenous noise in the endowments of players in a public good game. By introducing such exogenous information imperfections, these papers have not only added realism to the experimental designs, but have also delivered important insights into the role of different information structures in supporting, or weakening, the application of social sanctions in particular and mutual cooperation in general.

In this chapter, we build on the momentum towards more realistic settings. In particular, we extend the literature in a new direction by allowing important parts of the information structure of the game to be *endogenously* determined by the players. In many social dilemmas, information on coplayers' actions is not only imperfect, but players are frequently in a position to augment available information through costly

[1] This chapter is co-authored with Timo Goeschl.

effort. In other words, the player's choice whether to impose social sanctions on the coplayer is frequently preceded by a *choice* on whether to invest resources into monitoring his or her actions. Such monitoring effort to overcome imperfect information on coplayers' actions is well known in a variety of economically relevant contexts, such as shared resource management (Ostrom, 1990; Ostrom & Gardner, 1993; Seabright, 1993), production teams (Alchian & Demsetz, 1972; Kandel & Lazear, 1992; Dong & Dow, 1993; Craig & Pencavel, 1995) and alliances (Acheson, 1975, 1987, 1988; Palmer, 1991), labor relations (Shapiro & Stiglitz, 1984; Kanemoto & MacLeod, 1991; Lazear, 1993), micro-finance (Armendáriz & Morduch, 2005) or neighborhood watch (Sampson et al., 1997), to name just a few. Do individuals monitor actively and how does this interact with the supply of sanctions? Do subjects resort to «blind sanctioning», i.e. do they sanction without monitoring coplayer behavior? If present, both responses should, for different reasons, have problematic implications for maintaining cooperation through sanctions. Yet, there is little experimental evidence on how subjects respond to the option to overcome information imperfections at a cost. The experiment reported on in Carpenter (2007b), for example, links monitoring and punishment in the sense that it is infeasible to impose damages on coplayers whose previous actions were unknown. Ambrus & Greiner (2012) introduce noise into monitoring and vary the cost of punishment, but not the option of reducing noise at a cost.

The experiment we report upon in this chapter is designed to investigate the impact of information acquisition costs on mutual monitoring and punitive behavior in a setting in which sanctioning can only happen for subjective reasons. Our design extends a well studied dyadic cooperation game form that includes a mutual sanctioning stage (Fehr & Fischbacher, 2004b) by incorporating a monitoring decision into the sanctioning stage. In this *mutual cooperation-monitoring-punishment* (MCMP) game, subjects always have the option to impose punishment, at a cost to themselves, on their coplayer. Without payment of the information cost in the monitoring stage, however, their decision in the sanctioning stage remains uninformed about the coplayer's action at the cooperation stage of the game. By choosing to pay the fee they are able to condition their sanctioning on complete information regarding the coplayer's action. To the best of our knowledge, there exists no previous experimental evidence on *endogenous* mitigation of information asymmetries in games with an option to impose penalties. On the basis of this design, we address three main questions by comparing a baseline with zero information costs with two treatments with two different levels of information costs: First, what is the information acquisition behavior when observing coplayers' actions is costly? Second, how does information acquisition behavior respond to changes in the price of information? Third, how does the endogeneity of information impact on the occurrence and incidence of mutual sanctions? Having answered these three questions, we explore the link between the observed behavior to subjects' beliefs and risk postures and the impact of variations in monitoring costs on cooperation in a single round setting.

The key results are the following. First, positive information costs have the pre-

dicted impact of decreasing the supply of punishment. The supply of punishment responds to the presence of monitoring costs as if the costs of sanctioning had increased (Anderson & Putterman, 2006). Second, while all sanctioning is discriminate at zero information costs, in the presence of positive information costs a distinct share of subjects switches to «blind» punishment rather than refraining from sanctioning. The rise of *deliberately* blind punishment is an interesting new finding of this chapter. Third, positive information costs lead to defectors making up a smaller share of the punished and receiving smaller punishment compared to zero information costs. Fourth, we find that beliefs and risk aversion contribute to explaining individuals' monitoring and punishment behavior. Fifth, turning to cooperation, single-round cooperation under positive monitoring costs is maintained because subjects overestimate monitoring and punishment, in contrast to the baseline of zero information costs.

We proceed as follows. In section 3.2 we describe the aim, design, and procedures of the experiment. We proceed to present the main results in section 3.3. In section 3.4 we discuss the main results and likely mediating mechanisms. We conclude in section 3.5.

3.2 Experimental Design and Procedures

The aim of the design is to generate empirical evidence on social sanctioning behavior when information structures are endogenous. In particular, we are interested in the amount of costly information acquisition, the impact of price variation on information acquisition, and the relationship between these endogenized information structures and sanctioning behavior. The previous literature has been careful to identify and isolate non-strategic sanctioning. In keeping with this focus, we design the experiment such that strategic and altruistic motives for monitoring coplayers' actions are ruled out as far as possible and the observed monitoring and sanctioning should be driven only by subjective rewards to the individual carrying out these activities.

Different experimental game forms have been used to study costly punitive behavior experimentally (for an overview, see for example Gintis et al., 2005). The established game form involves a two-stage design consisting of a cooperation stage and a mutual sanctioning stage. The innovative step in our experiment consists of adding a monitoring stage between the cooperation and the sanctioning stage, resulting in a MCMCP game. At the monitoring stage, subjects decide whether to give up material rewards in order to receive full information about the actions of their coplayers at the cooperation stage. If they choose not to buy the information, they forgo the option of being able to discriminate between a cooperating and defecting coplayer at the sanctioning stage. If they do, sanctioning can be conditioned on observed coplayer behavior. We exogenously vary the magnitude of this fee as our treatment variable to observe how behavior is adjusted.

Among the experimental game forms used in the literature, we choose as starting point the simple cooperation game form by Fehr & Fischbacher (2004b), which is a one-shot two-person Prisoner's Dilemma in material payoffs. We do so for four

reasons. First, due to its simplicity, the game is easily understood even by inexperienced subjects. Second, the two-person design eliminates second-order coordination problems (Yamagishi, 1986) that are present in groups with more than two members. Third, the simplicity of the game form facilitates the use of Selten's strategy method for eliciting behavior in the monitoring stage. Finally, by drawing on established parameter constellations, our choice facilitates comparability with previous research. Specifically, our baseline condition constitutes an effective replication of the experiment by Fehr & Fischbacher (2004b).

3.2.1 Experimental game form

Subjects played a two-stage experimental game with one randomly matched and unknown coplayer. In the first stage, both players in a given group of two were endowed with 10 tokens and interacted with one another in a strategic game of the Prisoner's Dilemma type with monetary payoffs. Subjects' first-stage decision was binary: Players could either keep their tokens or transfer all of them to the other group member, in which case the experimenter tripled them. Thus, denoting a given player by i and his or her coplayer by j, the first stage decision by $a_i \in \{0,1\}$ where $a_i = 1$ means cooperation, the first stage payoff equal to $10\left(1 - a_i + 3a_j\right)$. The benefit of a binary choice at the cooperation stage is not only a reduction in complexity (Ambrus & Greiner, 2012), but also a gain in experimenter control over the value of information that subjects can obtain through monitoring.

At the beginning of the second stage, each player received an additional endowment of 40 tokens. Subsequently, both players in each group had the opportunity to observe the coplayer's action in the first stage («monitor») and could choose to punish the other. In order to avoid potential biases due to a variable cost-impact ratio (see Casari, 2005), we chose a linear sanctioning technology commonly employed in the literature (Fehr & Gächter, 2002; Fehr & Fischbacher, 2004b; Gächter et al., 2008; Ambrus & Greiner, 2012): Denoting by p_{ij} player i's expenditure for punishing player j, the latter's payoff is reduced by $3p_{ij}$ tokens. Expenditures were restricted to integers and damage was limited to sixty tokens, therefore $p_{ij} \in \{0,1,\ldots,20\}$. Subjects reported their choices in the monitoring and punishment stages jointly using a modified version of the strategy method. Without a monitoring stage, players would simply announce a contingent punishment plan,

$$p_{ij}\left(a_j\right) : \{0,1\} \to \{0,1,\ldots,20\}$$

indicating the number of deduction points p_{ij} for each of the coplayer's possible transfer decisions a_j, given their own decision in the first stage. The presence of a monitoring stage was implemented in the strategy method by allowing subjects a choice between two punishment plans. One was the standard contingent punishment plan $p_{ij}\left(a_j\right)$ that allowed the subject to condition their punishment on the coplayer's action a_j. This choice, denoted by $m_i = 1$, required payment of an exogenously set monitoring fee κ_m in addition to the applicable punishment costs p_{ij}. The alternative,

denoted $m_i = 0$, was a punishment plan p_{ij} in which the subject could not condition the punishment on the coplayer's action a_j. This plan did not involve a fee over and above the cost of punishment. After the punishment stage, the game ended.

In summary, the pecuniary payoff function of each subject i in this mutual monitoring game was

$$w_i = 10\left(1 - a_i + 3a_j\right) + 40 - m_i\kappa_m - p_{ij} - 3p_{ji}$$

When discussing the monitoring and punishment behavior at the individual level, the analysis will focus on three main behavioral types. One type are «non-punishing» subjects with $m_i = 0$ and $p_{ij} = 0$ that decide not to monitor and do not punish. These subjects are referred to as behavioral type N. Among the punishing subjects, there are two types. «Blind punishers» with $m_i = 0$ and $p_{ij} > 0$ devote no resources to monitoring, but still punish indiscriminately. This type is denoted as B. The other type T are «targeted punishers» with $m_i = 1$ and $p_{ij} > 0$ that devote resources both to monitoring and to punishment. The B-types and T-types together form the class of «punishers», denoted P. The forth possible type (monitoring without punishment) is excluded by design.

3.2.2 Additional tasks

We supplemented the MCMP game by incentivized elicitation of first-order beliefs and risk preferences. Both are potentially relevant in accounting for observed behavior (Keser & van Winden, 2000; Fischbacher et al., 2001; Fischbacher & Gächter, 2010). Since subjects receive no actual information in the course of play, they respond to their beliefs on their coplayer's behavior, and this response involves a decision under strategic risk. Thus, in addition to the MCMP game described above, every subject performed the following tasks, which were identical in each treatment and for every subject. Following their decision in the first stage, we asked subjects to state their belief about the transfer decision of the coplayer and, following their decision in the second stage, their belief about the punishment they expected from their coplayer. Those statements were incentivized with a opportunity to earn extra tokens in case of accurate predictions (see appendix).

At the end of each session we employed the Holt-Laury lottery choice method (Holt & Laury, 2002) to obtain an indication of each subject's risk preferences (see appendix). Subjects were only informed about this task after the MCMP game. The method may not capture the specific kind of risk involved in the situation, but we think it is worth investigating.

3.2.3 Design

Our primary focus in the design is on the relationship between information costs and monitoring and punishment behavior. The objective of the treatments is to introduce an exogenous variation in the cost of information acquisition with the aim of forcing a materially consequential choice by subjects regarding the desired information

status prior to the punishment decision. As a result, information structures become endogenous and observable.

We employ between-subjects variation in the monitoring fee κ_m and examine the changes in the shares of types N, B, and T as κ_m changes. For understanding the evolution of these shares of behavioral types, evidence on monitoring and sanctioning activity when monitoring is costless, i.e. when $\kappa_m = 0$, is an obvious starting point: It provides a direct replication of the PD experiment by Fehr & Fischbacher (2004b) and therefore a meaningful baseline treatment. This treatment is labeled $M0$. In the alternative treatments $M5$ and $M10$, the monitoring fee is set at $\kappa_m = 5$ and $\kappa_m = 10$. The choice of three fee levels $\kappa_m = \{0, 5, 10\}$ was on the basis that we had strong priors for a non-linear effect. In order to achieve efficient identification, we set the values at the extremes and the midpoint of the range and allocated approximately half of the subjects to the midpoint cell, a third to maximum cell, and a sixth to the minimum cell, respectively (see for example McClelland, 1997; List et al., 2011, for efficient sample arrangement methods).

On the basis of the design, we are able to test for responses in monitoring and punishment behavior with respect to changes in those costs by comparing the relative proportions of the three behavioral types in treatments $M0$, $M5$, and $M10$. In addition, we can exploit information about the resources spent on punishment and on behavior at the cooperation stage to understand more about the direction, intensity, and heterogeneity of social sanctions in the presence of costly monitoring.

3.2.4 Subjects and procedures

All experiments were conducted at the experimental laboratory of the Alfred-Weber-Institute (AWI-Lab) at Heidelberg University in late 2010. Participants were recruited from the general undergraduate student population using the online recruitment system ORSEE (Greiner, 2004). In total 134 subjects participated, of which 49.3 percent were female. A share of 64.2 percent had never participated in a laboratory experiment before. The mean age was 21.8 years. No subject participated in more than one session and treatment. Based on the sample arrangement described above, 22 subjects were assigned to $M0$, 66 to $M5$, and 46 to $M10$.

At the beginning of each session, upon entering the laboratory subjects were randomly assigned to the computer terminals. Direct communication among them was not allowed for the duration of the entire session. Booths separated the participants visually, ensuring that they made their decisions anonymously and independently. 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. The experiment was framed in a sterile way using neutral language and avoiding value laden terms in the instructions (see supplementary material). Participants had to answer a set of control questions individually at their respective seats in order to ensure comprehension of the rules. We did not start

the experiment before all subjects had answered all questions correctly.

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 each subject made her or his decisions and reported expectations. After being informed about the payoffs, the experimenter announced that another (and definitely final) experiment will now be conducted and distributed additional instructions that explained the supplementary lottery choice experiment. Thus, subjects did not know until this announcement that the experiment involved another task in order to rule out confounding effects of the lottery-choice task on the main experimental game. We cannot rule out reverse confounds which is, however, of minor significance. After being informed about the payoffs, subjects were asked to answer a short questionnaire while the experimenter prepared the payoffs. Subjects were then individually called to the experimenter booth, payed out (according to a random number matched to their decisions; no personal identities were used throughout the whole experiment) and dismissed.

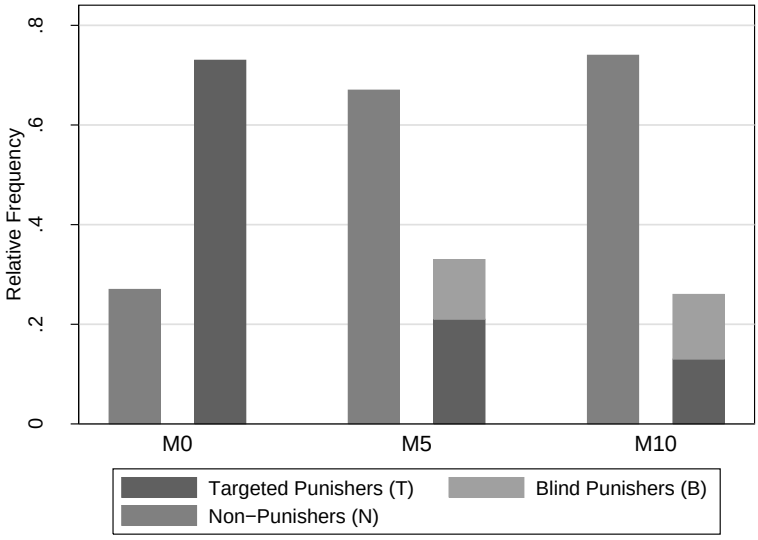
In every session subjects received a fixed show-up fee of €2, which was not part of their endowment. The whole experiment lasted approximately one hour and subjects earned an average of €10.99 (€0.10 per token earned), including the fixed show-up fee. Earnings exceed the local average hourly wage of a typical student job and can hence be considered meaningful to the participants.

3.3 Results

We organize our discussion of the experimental results in the following way. We begin by discussing our results from the baseline treatment that replicate previous experimental evidence on sanctioning behavior. We then describe the core result of the experiment, namely the relative prevalence of the three possible behavioral types N, B, and T for each of the monitoring cost treatments $M0$, $M5$, and $M10$. We relate the results from the additional treatments in which the monitoring cost was varied to the existing literature on variations in sanctioning costs and discuss the parallels and differences. In addition, we exploit the endogenous nature of information acquisition in our setting to move from the observable level of behavioral types to more sophisticated explanations of the observed evidence as a result of social preferences and beliefs. We believe that the evidence permits understanding more about the underlying nature of social sanctioning.

Figure 3.1 shows the relative frequencies of behavioral types N, B, and T for monitoring costs at three levels $\kappa_m = \{0, 5, 10\}$. Normalization is required on account of the between-subjects design with a variable number of subjects in each treatment. A readily apparent first diagnosis of the evidence is therefore that variations in the cost of information are associated with different patterns of monitoring and punishment behavior. The detailed nature of these differences is the subject of the following sections. In particular, our key result can also be highlighted: when information is imperfect and information costs are positive, a substantial share of subjects responds

Figure 3.1: Distribution of behavioral types at different monitoring cost levels.



with information acquisition prior to punishment, that is, a significant share of subjects does not take exogenously imposed information constraints as given in a social sanctioning situation as long as the costs of overcoming these constraints are not prohibitive.

3.3.1 Punishment with costless monitoring: Replication of previous findings

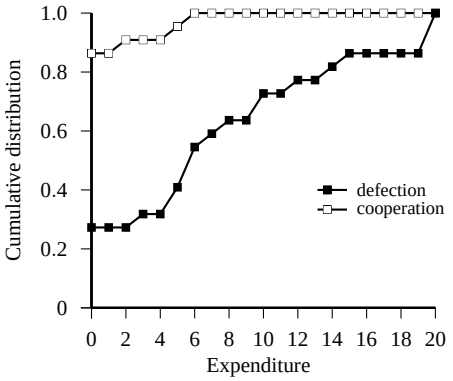
The baseline treatment with $\kappa_m = 0$ replicates earlier experiments on costly sanctions (see Camerer, 2003; Gintis et al., 2005; Gächter & Herrmann, 2009; Balliet et al., 2011, for overviews) and specifically the experiment run by Fehr & Fischbacher (2004b). These experiments find that when there are no hindrances to monitoring, the general propensity among subjects to punish coplayers in experiments involving social sanctions is high. Even after eliminating strategic motives for punishment, the share of subjects that apply costly sanctions consistently exceeds one half. Fehr & Fischbacher (2004b) find that 67 percent of subjects applied costly sanctions. Our results, displayed in the left-hand column in figure 1, are consistent with theirs.

Result 3.1. *When information costs are zero, most subjects make use of the non-strategic punishment option.*

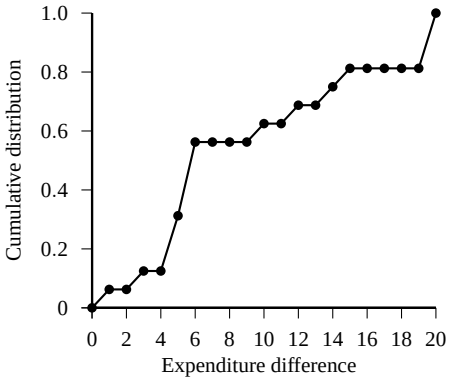
In particular, we find a large majority of 73 percent (16 out of 22) willing to impose costly sanctions on their coplayers for non-strategic reasons. The class of punishers *P* therefore dominates when there are no monitoring costs.

Previous experiments have also generated clear evidence that when information about coplayer behavior at the cooperation stage is perfect, then punishment is targeted. For example, Fehr & Gächter (2000a) find evidence of a clear relationship

Figure 3.2: Individual heterogeneity in punishment and targeting in the baseline condition.



(a) Distribution of punishment expenditures.



(b) Distribution of differences between punishment expenditure targeted on defection and cooperation, respectively, among the punishing individuals.

between deviation from average contribution in a public goods game. Fehr & Fischbacher (2004b) also find that behavior at the cooperation stage is a highly significant predictor of receiving punishment. Our experiment replicates these findings closely.

Result 3.2. *When information costs are zero, non-strategic punishment is always targeted.*

When monitoring was costless, *every* subject that imposed a punishment acquired the information whether her or his coplayer cooperated or defected. In other words, all sanctions were conditioned on the coplayer's first stage behavior. Specifically, of those 16 out of 22 subjects that imposed damages in at least one contingency, all 16 monitored their coplayer. The remaining six subjects neither monitored nor imposed any damages. In terms of our behavioral types, there are only T-type and N-types, but no B-types. The extent of targeting in the baseline treatment is reported in the right panel of figure 3.2. This depicts the cumulative distribution in the differences, at the level of individual punisher, between expenditures on punishing defectors and punishing cooperators. As evident, every punisher differentiates between cooperators and defectors and more than half the punishers differentiate their purchase of punishment by at least five tokens between defectors and cooperators. This may be viewed as a measure of the strength of the preference for targeting.

The third aspect of our replication concerns the direction of punishment. In previous experiments, the direction and intensity is clear in that cooperators dominate as punishers and defectors are their specific targets. Fehr & Fischbacher (2004b) do not report the composition of the group of punishers, but the following numbers are indicative: 69 percent of cooperators imposed sanctions on defectors as opposed to

50 percent of defectors imposing sanctions on other defectors. The two groups of punishers diverge even more in terms of the volume of sanctions. There, cooperators punishing defectors spent on average 9.2 tokens on sanctions while defectors punishing defectors spent 2.7 tokens. Previous experiments have also uncovered considerable heterogeneity among subjects with respect to sanctioning behavior (Charness & Rabin, 2002; Fischbacher & Gächter, 2010; Blanco et al., 2011). For example, alongside the targeted punishment of defectors by cooperators, punishment of cooperators is a less frequent but also robust empirical regularity (Cinyabuguma et al., 2006; Gächter et al., 2006; Herrmann et al., 2008; Ertan et al., 2009; Gächter & Herrmann, 2010).[2] In Fehr & Fischbacher (2004b), 8.3 percent of cooperators were targeted at the punishment stage.

Result 3.3. *When information costs are zero, the dominant direction and intensity of punishment is cooperators imposing high sanctions on defectors, but heterogeneity is high.*

19 punishment actions were carried out in the baseline treatment by 14 cooperators and 2 defectors.[3] Of these punishment actions, 15 (79 percent) were taken by cooperators (1 out of 14 cooperators decides to punish his co-player irrespective of his behavior) and 4 (21 percent) by defectors. Among the class of punishers, therefore, cooperators clearly supplied most of the punishment action. Taking intensity into account, cooperators accounted for 84 percent of total investment into punishment while defectors accounted for 16 percent. Whom were the punishments imposed on? No punisher targeted exclusively cooperators. In 16 out of 19 punishment decisions (84 percent) the target of a punishment action was a defector. Cooperators were the target of 14 percent of punishment actions. Again, taking intensity into account, defectors received 93 percent of all the punishment imposed while cooperators received 7 percent. Summarizing, punishment originates with cooperators and arrives at defectors.[4] At the same time, the general observation hides significant heterogeneity in individual-level behavior. The left panel of figure 3.2 provides an illustration by plotting the cumulative distribution of observed punishment expenditures by target. A majority of 84 percent chose zero punishment for those cooperating in the first

[2] In an experiment designed to differentiate between a variety of different motivations to impose costly sanctions, Leibbrandt & López-Pérez (2009, 2011) find substantial heterogeneity, with a combination of inequity aversion, self-interest, spite and direct reciprocal motivations accounting for observed patterns best.

[3] Recall that the strategy method allows each subject to punish up to two times, once for the case that the coplayer cooperates and once that he defects.

[4] To make our results comparable with Fehr & Fischbacher (2004b), 78 percent (14 out of 18) of cooperators imposed sanctions on defectors, with an average expenditure of 10.7 tokens, while only one (5.6 percent) of them also spent 2 tokens on sanctioning a cooperator (besides her or his spending of 8 tokens on sanctioning a defector). Two out of four defectors did not punish, while the remaining two imposed damages on both other defectors and cooperators: one expended 10 tokens for a sanction on a defector and 5 tokens for a sanction on a cooperator, the other expended 7 tokens for a sanction on a defector and 6 tokens for a sanction on a cooperator. Thus, defectors rather than cooperators are the prevailing target of sanctions.

stage. A minority of 16 percent of the subjects chose a positive punishment for cooperators, but no subject more than 6 tokens. A quarter of subjects chose to impose no punishment on defectors. Among the majority who imposed damages on defectors, the expenditures differ considerably between individuals, ranging from 3 to the maximum of 20 tokens.

A corollary of result 3.3 is that when monitoring was costless, defection was highly unprofitable: Defectors received much stronger punishment (22.8 tokens) on average than cooperators (1.8 tokens) such that the difference was sufficient for the average net punishment of defection (22.8 – 1.8 = 21 tokens) to exceed the gain (10 tokens) by a substantial margin. We return to this observation again when examining positive costs.

3.3.2 Monitoring and punishment with positive information costs

Results 3.1 through 3.3 underline the benchmark quality of the baseline treatment by providing additional evidence to support similar findings in the literature. We confirm the presence of non-strategic sanctioning, the dominance of targeted punishment, the direction and intensity of punishment, and the degree of behavioral heterogeneity. Against this benchmark, we now report in the following three results on the behavioral implication of endogenizing subject's information status at the time of the punishment decision through a costly choice.

When monitoring is costly, the relative shares of behavioral types diverge from the baseline: A higher overall cost of sanctioning is associated with a higher share of subjects that neither monitor nor punish, which is in line with previous evidence showing that increasing cost of punishment reduce the supply of sanctions (Suleiman, 1996; Oosterbeek et al., 2004; Anderson & Putterman, 2006; Carpenter, 2007a; Egas & Riedl, 2008).

Result 3.4. *In the presence of positive information costs, the propensity to punish decreases relative to the baseline.*

Figure 3.1 illustrates this first finding: The share of subjects that neither monitor nor punish (type N) increases from 27 percent at $\kappa_m = 0$ to 67 percent when $\kappa_m = 5$ and 74 percent when $\kappa_m = 10$. The hypothesis that the frequency of N-types and monitoring costs are independent can be clearly rejected (Kruskal-Wallis test, $p < .001$). Overall, the relationship is clearly negative (Kendall's $\tau_b = -.260$, $p = .002$), whereas pairwise, only the former change is statistically significant (Mann-Whitney test, $p = .001$), the latter is not ($p = .414$). The ratio of the frequency of punishers to the frequency of N-types decreases successively from 2.67 at $\kappa_m = 0$ to 0.5 at $\kappa_m = 5$ and to 0.35 at $\kappa_m = 10$. Total expected expenditures for sanctioning, given subjects' elicited priors about their coplayer's first stage behavior, remain constant or at most decrease slightly ($p = .062$) for higher monitoring costs. The observed pattern is consistent with the view that subjects interpret a positive monitoring cost as comparable to an increase in the cost of punishment.

In addition to the result on the propensity to punish at all, our design explicitly allows to study punitive behavior of those individuals that opt to remain ignorant about their coplayer's first stage behavior.

Result 3.5. *In the presence of positive information costs, the propensity to punish indiscriminately increases relative to the baseline.*

Figure 3.1 again reports the share of blind punishers (type B) that do not acquire information and yet impose sanctions. This share increases from zero (none of 22) in the baseline to 12 (8 out of 66) and 13 percent (6 out of 46) as the monitoring costs increase to 5 and 10 tokens, respectively. Here, the former change is statistically marginally significant (Mann-Whitney test, $p = .089$), the latter not ($p = .885$). However, both frequencies under positive monitoring costs are statistically significantly different from zero (Wilcoxon signed-rank tests, $p = .005$ and $p = .014$, respectively). The ratio of the frequency of indiscriminate punishers to the frequency of all punishers (the frequency of targeted punishers) increases successively from zero at $\kappa_m = 0$ to 0.36 (0.57) at $\kappa_m = 5$ and to 0.5 (1) at $\kappa_m = 10$.

Finally, the share of targeted punishers (type T) decreases from 73 percent for costless monitoring to 21 percent when $\kappa_m = 5$ and 13 percent when $\kappa_m = 10$. Again, the hypothesis that the frequency of T-types and monitoring costs are independent can be clearly rejected (Kruskal-Wallis test, $p < .001$), while pairwise only the former change is statistically significant (Mann-Whitney test, $p < .001$), the latter is not ($p = .269$).

Result 3.6. *At positive information costs, cooperators remain the dominant origin of the direction of punishment relative to the baseline, but defectors dominate less as its target.*

31 percent (15 out of 48) and 29 percent (10 out of 35) of cooperators imposed sanctions on defectors in $M5$ and $M10$, respectively, with respective average expenditures of 11.6 tokens and 8.2 tokens. Five and four of the cooperators in $M5$ and $M10$, respectively, also imposed punishments on other cooperators, with respective average expenditures of 6.8 tokens and 3 tokens. 11 out of 18 (61 percent) and 9 out of 11 (82 percent) of the defectors did not punish in $M5$ and $M10$, respectively. Of the remaining seven defectors in $M5$, four punished indiscriminately (B-type) with an average expenditure of 5.8 tokens, and three targeted (T-type) with an average expenditure of 6 tokens on defectors and 5 tokens on cooperators, plus the five tokens monitoring fee. Likewise, of the remaining two defectors in $M10$, both punished indiscriminately (B-type) with an average expenditure of 2 tokens.

Despite the rise in monitoring costs, defectors rather than cooperators remain the prevailing target of sanctions on average, although less pronounced. Defection is still punished significantly more often than cooperation in both $M5$ (32 vs. 17 percent, $p = .004$, Wilcoxon signed-rank test), and $M10$ (26 vs. 13 percent, $p = .014$, Wilcoxon signed-rank test). Also, defectors continue to receive stronger punishment on average than cooperators in both $M5$ (9.8 vs. 3.3 tokens, $p = .006$, Wilcoxon signed-rank test),

and $M10$ (5.6 vs. 1.1 tokens, $p = .014$, Wilcoxon signed-rank test). However, in both conditions defection was no longer unprofitable. We return to this below.

To sum up, a setting with a perfect, but costly monitoring option leads to three important differences compared with zero monitoring costs: (1) The supply of social sanctions decreases as if punishment costs increased. (2) A small, but significant share of subjects sacrifices resources to actively monitor their coplayer and thus target their punishment. (3) A small, but significant share of subjects chooses to punish without information on whether the coplayer cooperated. The first-order impact of exogenous information imperfections that can be overcome at a cost is therefore to reduce the supply of sanctions.[5] These imperfection also bring to light important heterogeneities across subjects with respect to the type of sanctioning they will supply, targeted or blind. Targeted sanctioning is conducive to maintaining cooperation because it rewards cooperative behavior in relative terms. The deliberately blind sanctioning observed in the experiment on the other hand is unproductive for maintaining cooperation.

3.4 Further results section

3.4.1 The role of beliefs and risk aversion at the cooperation and the monitoring stage

The comparison between results 3.1 through 3.3 on the one hand and results 3.4 through 3.6 on the other form the core of this chapter. Yet, exploiting the elicitation of beliefs and risk attitudes carried out in the course of the experiment provides further insight into drivers of behavior identified in earlier papers. Table 3.1 reports the results of system probit regression on both stages. The estimates derive from a Seemingly Unrelated Regression (SUR), with adjusted standard errors clustered at the session level.[6] The transfers reported in the left-hand column document subjects' behavior at the cooperation stage. The remaining columns report on behavior at the monitoring stage by explaining subjects sorting into the three behavioral types N, B, and T.

Before discussing the effects of the belief and risk attitudes, note that the top row of table 3.1 summarizes results 3.1 through 3.6: At the monitoring stage monitoring costs decrease the propensity to become a T-type, but increase the propensity to become an N- and B-type (all results statistically significant). At the cooperation stage, however, monitoring costs do not have a significant effect on the size of the transfer.

[5] This qualifies results reported in a related paper by Grechenig et al. (2010) in the sense that if it is the subject's *own choice* of whether to acquire information on their coplayer's behavior, the vast majority refrain from punishment altogether if they opted to remain ignorant.

[6] We use SUR since it is reasonable to assume that the error terms in the equations are correlated. As a robustness check, we also carried out a system probit regression estimated by maximum simulated likelihood (using the Geweke-Hajivassiliou-Keane simulator), and a system logit SUR that delivered equivalent results.

Table 3.1: Results of a system probit model estimated by Seemingly Unrelated Regression (SUR)

		Second stage		
	First stage transfer	Being a *P*-type	Being a *T*-type	Being a *B*-type
Monitoring fee	-0.036	0.133	-0.186	0.066
	(0.028)	(0.035)	(0.039)	(0.033)
	.201	*.000*	*.000*	*.047*
Pessimistic belief	-2.339	0.312	-0.259	-0.164
	(0.443)	(0.283)	(0.301)	(0.185)
	.000	*.271*	*.389*	*.375*
Optimistic belief	1.207	0.030	0.015	-0.196
	(0.411)	(0.222)	(0.253)	(0.318)
	.003	*.891*	*.952*	*.538*
Risk averse	-0.707	-0.654	0.451	0.970
	(0.361)	(0.170)	(0.224)	(0.428)
	.050	*.000*	*.045*	*.023*
Female	0.214	0.222	-0.415	0.196
	(0.330)	(0.215)	(0.250)	(0.209)
	.517	*.301*	*.097*	*.348*
Age	0.016	0.039	-0.084	0.011
	(0.031)	(0.023)	(0.068)	(0.036)
	.598	*.093*	*.215*	*.765*
Exp. net cost of defect	-0.003			
	(0.024)			
	.895			
Constant	0.810	-0.985	2.087	-2.679
	(0.757)	(0.775)	(1.469)	(0.997)
	.285	*.203*	*.155*	*.007*
Obs	134	134	134	134
Log likelihood	-47.599	-78.293	-63.507	-40.381
Prob > χ^2	0.000	0.002	0.000	0.175
Pseudo R^2	0.364	0.116	0.186	0.100

For each independent variable, the first row reports the estimated coefficient, the second row the standard errors in brackets), and the third row the *p*-value of a test with Null that the coefficient is equal to zero (in italics). Standard errors take into account clustering by session. Pessimistic (elicited belief of 0.3 or less) and optimistic (elicited belief of 0.7 or more) belief is relative to an indeterminate belief (elicited belief between 0.4 and 0.6), respectively. Risk averse (4 or less risky choices in the Holt-Laury task) is relative to risk neutral and risk seeking (5 or more risky choices), respectively.

Beliefs about coplayer behavior and risk aversion have been shown to play a role in determining subject's actions in many previous cooperation experiments (e.g. Eckel & Wilson, 2004; Houser et al., 2010; Fischbacher & Gächter, 2010). Our first result confirms these findings in our experiment.

Result 3.7. *At the cooperation stage, subjects' pessimistic (optimistic) beliefs about the coplayer's cooperation and their risk aversion (tolerance) inhibit (facilitate) cooperative behavior.*

The evidence from the cooperation stage of the game provides support for the broad notion of strong reciprocity among subjects. As the regressions in table 3.1 show, optimistic beliefs about coplayer behavior have a positive influence on subject's transfer and *vice versa*. Risk aversion interacts negatively with the size of first-stage transfers, as predicted in a first-stage transfer without information about the coplayer's type.

Result 3.8. *In contrast to a situation with zero monitoring costs, beliefs about coplayer behavior fail to explain the sorting into behavioral types at the monitoring stage when monitoring costs are positive. Risk aversion decreases the probability of sorting into non-punishment and increases the probability of sorting into targeted and blind punishment.*

In the baseline treatment with zero monitoring costs, the evidence suggest the prevalence of strong (negative) reciprocity among subjects, as result 3.2 makes clear. However, if this interpretation was correct, then beliefs about coplayer behavior should also be relevant in the second stage. Compared to subjects that declare to be uncertain about their coplayer's action, subjects with an optimistic belief about the coplayer's action should be more likely to sort into type N, pessimists into type B. As Table 3.1 makes clear, these predictions are not supported by the data: Both optimistic and pessimistic beliefs are statistically insignificant drivers at the monitoring stage.[7] Risk aversion, on the other hand, is associated in a statistically significant way with behavior in the monitoring stage: Risk averse individuals have a higher propensity to sort into type T and lower propensity to sort into type N. Blind punishment (type B), however, is also associated with higher risk aversion. To explain what to us is an unexpected finding, we draw attention to the fact that the informational imperfections of the MCMP game give rise to two separate types of risk that interact differently with a subject's other preferences. One risk of the informational imperfections is that a defecting coplayer «gets away with it», i.e. does not receive punishment. The net benefits associated with punishing defectors (net of punishment

[7] Among the 14 B-types, only one (7 percent) has a pessimistic belief (exactly 0.3), seven (50 percent) are completely uncertain (exactly 0.5), and the remaining five (36 percent) even stated an optimistic belief (0.7 or higher). Among the 78 N-types in the costly monitoring conditions, seven (9 percent) stated a pessimistic belief, 33 (42 percent) were uncertain, and 38 (49 percent) had an optimistic belief. Finally, among the 20 T-types in the costly monitoring conditions, one (5 percent) stated a pessimistic belief, 9 (45 percent) were uncertain, and 10 (50 percent) had an optimistic belief.

costs) would be affected by the presence of this type of risk. The other risk is that a coplayer gets punished whom the subject would not like to impose any losses on. The effect of this risk is typically associated with blind punishment being applied to a cooperating coplayer. It is then the *relative* strength of the two effects that determines, for a given level of risk aversion, whether subjects will sort into type T and spend resources to ensure that only intended targets get punished, or into type B that ensures that no intended target gets away unpunished. Our evidence demonstrates that both outcomes can be observed.

3.4.2 The emerging incentive structure

The non-repeated, between-subjects design of our experiment ensures the exclusive presence of non-strategic motivations for punishment and therefore enhances experimenter control. It constitutes an important step towards an exhaustive analysis of the equilibrium effects of endogenous information structures on cooperation that is an area of future research. Part of this first step is to assess what can be learned from first-round behavior about the emerging incentive structure for cooperation (see Ambrus & Greiner, 2012).

We start with the baseline treatment of costless monitoring. Here, in line with previous experiments, the credible threat of cheap targeted punishment for defectors favors cooperative behavior. Punishment was widespread and *every* subject that imposed a punishment did so discriminately. For positive monitoring costs, however, this was no longer the case and therefore changes the incentives for cooperation at the first stage.

Result 3.9. *If monitoring is costless, defection is highly unprofitable relative to cooperation. At positive monitoring costs, defection does pay.*

Table 3.2: Frequency of punishment and mean punishments received ($3 \cdot p_{ji}$).

	Relative frequency		Mean damages		
κ_m	Cooperation	Defection	Cooperation	Defection	Net Damages
0	.136	.727	1.8	22.8	21.0
5	.167	.318	3.3	9.8	6.5
10	.130	.261	1.1	5.6	4.6

The impact of endogenizing the information structure by varying the monitoring costs is summarized in table 3.2. We first observe that there is no significant variation in punishment of cooperation (Kruskal-Wallis test, $p = .756$). Punishment of defection, however, is significantly different between treatments ($p < .001$), exhibiting a diminishing trend as monitoring costs rise (Kendall's $\tau_b = -.266$, $p < .001$). The same is true for the mean net punishment of defection, which is the difference between damages imposed on defection and cooperation, respectively (Kendall's $\tau_b = -.299$,

$p < .001$). Hence, *on average* defecting gets less costly as monitoring costs rise. As a result, defection *did* pay when monitoring costs were positive.

Do subjects anticipate the effects of monitoring costs on punishment and the incentives to cooperate, as predicted previously?[8] There is little guidance to be derived from the literature since most papers do not report elicited expectations.[9] In the baseline condition of the present experiment, subjects' beliefs support the «punishment anticipation hypothesis». On average, subjects expected to receive punishment of 21.2 (1.0) tokens in case of defection (cooperation). This is a remarkably accurate prediction of the actual punishments of 22.8 (1.8) tokens. The rank correlation between expectations and realizations is significantly positive (Kendall's $\tau_b = .547$, $p = .001$, in case of defection, $\tau_b = .622$, $p = .004$, in case of cooperation). In the presence of positive monitoring costs, on the other hand, the punishment anticipation hypothesis fails in a non-repeated design.

Result 3.10. *While subjects correctly predict the direction and intensity when monitoring costs are zero, the intensity of punishment and mean net punishment are overestimated for positive monitoring costs. According to subjects' beliefs, defection is on average unprofitable under costly monitoring when, in fact, defection paid on average.*

Table 3.3: Frequency of subjects expecting to be punished and mean expected damages.

	Relative frequency		Mean damages		
κ_m	Cooperation	Defection	Cooperation	Defection	Net Damages
0	.136	.864	1.0	21.1	20.2
5	.212	.636	3.2	15.5	12.3
10	.196	.652	2.4	12.4	10.0

The evidence underpinning this result is shown in table 3.3. *Qualitatively*, subjects anticipate the actual impacts patterns as shown in table 3.2. *Quantitatively*, however, they significantly underestimate the effects of increasing monitoring costs on their coplayers' monitoring and punishment behavior. As in the case of actual punishments, expected punishment of cooperation (Kruskal-Wallis test, $p = .626$) does not vary significantly with monitoring costs while the expected punishment of defection does (Kruskal-Wallis test, $p = .045$). The difference is that in the subjects' expectations, positive monitoring costs have a much weaker impact on the punishment of defectors (and hence the net punishment of defecting) (Kendall's $\tau_b = -.156$, $p = .033$, and $\tau_b = -.166$, $p = .024$, respectively) than is actually the case (see test

[8] Defection «may cause strong negative emotions among the cooperators [which] may trigger their willingness to punish ... and *that most people expect these emotions*» (Fehr & Gächter, 2002, p. 139, emphasis added).

[9] A notable exception is López-Pérez & Kiss (2012) who explicitly address the issue of whether subjects anticipate positive and negative sanctions. They obtain similar results as ours. Notably, they find that punishments are better anticipated than rewards.

results above). For example, the *actual* net cost of defecting amounted to 6.5 tokens on average when $\kappa_m = 5$ and 4.6 on average when $\kappa_m = 10$. The *expected* net costs were 12.3 and 10.0, respectively, on average. As a result, cooperation still payed in subjects' expectation.

Turning finally to cooperation rates, there are no statistically significant differences between treatments. In the baseline treatment with $\kappa_m = 0$, 82 percent cooperated. When $\kappa_m = 5$ and $\kappa_m = 10$, the cooperation rate is73 percent and 76 percent, respectively. The differences are not statistically significant (Kruskal-Wallis test, $p = .688$). Monitoring costs did therefore not have a systematic impact on cooperation in the single round design. There are two, probably complementary, explanations for this. The first is the overestimation of punishments summarized in result 3.10. The second explanation is a second-order effect that the overestimation of punishment has on the beliefs of conditional cooperators that others will cooperate: This belief, in turn, makes conditional cooperators more willing to cooperate (Shinada & Yamagishi, 2007, see also Fehr & Gächter, 2000a; Gächter et al., 2008 for supporting evidence). We find, for example, that first stage cooperation correlates strongly with the belief about the coplayer's cooperation (Kendall's $\tau_b = .473$, $p = .000$). At the same time, those beliefs do not differ significantly between treatments (Kruskal-Wallis test, $p = .754$), which is consistent with this second explanation. Additional support comes from the regression results on behavior at the cooperation stage reported in Table 3.1. The coefficient of the net expected punishment of defection in the first stage cooperation equation was insignificant, while the estimated coefficients of the belief indicator variables were significant and consistent with a conditionally cooperative motivation.

The extent to which these single round results persist in a setting with repetition is a matter of future research.[10]

[10]There are several reasons why one might expect that cooperation rates in the costly monitoring treatments will deteriorate if the game is played repeatedly. First, one might expect that beliefs will be updated in the course of play so as to converge to the observable reality. However, note that this learning effect is, of course, dependent on the subjects' information acquisition behavior. We are currently experimenting on this issue. Second, as shown in previous research, cooperation appears to suffer if there is exogenous uncertainty about others' possible contributions or the reasons for contributing little (Bornstein & Weisel, 2010; Grechenig et al., 2010; Patel et al., 2010). Besides direct pecuniary incentives, this may be related to the hypothesis that cooperators and defectors respond differently to punishment because the former may feel anger when punished while the latter may feel shame (Bowles & Gintis, 2005a). If monitoring gets so costly that a large fraction of damages is inflicted untargeted, then sanctions may fail to induce feelings of shame in defectors, which seem to be an important deterrent besides any pecuniary harm (Barr, 2001; Fessler & Haley, 2003; Gintis, 2008; Hopfensitz & Reuben, 2009; Jacquet et al., 2011), and anger in cooperators who respond with withdrawing their cooperation. This is consistent with findings that mutual enforcement appears not to work well if punishment is perceived to be unfair or unjustified (Barclay, 2006; van Prooijen et al., 2008; Drouvelis, 2010).

3.5 Conclusion

We extend previous research about the effects of information imperfections on the role of sanctions for maintaining mutual cooperation in social dilemmas. To our knowledge, this chapter is the first to consider *endogenous*, rather than exogenously given, information structures by allowing for information acquisition on coplayer behavior in the context of a mutual-cooperation mutual-punishment game. This design enables both a close replication of previous evidence on the presence of non-strategic sanctioning and the observation of new types of heterogeneities among subjects. We show that a majority of subjects responds to positive monitoring costs in a manner that is broadly equivalent to a rise in punishment costs and obeys the Law of Demand. At the same time, we observe, on the one hand, active information acquisition for targeting sanctions that do not provide material returns either to the punisher or anybody else. On the other, the design also allows subjects to punish «blindly» if they so wish, and we find a significant share of subjects that chooses this course of action. Given the practical relevance of an endogenous information structure in many social dilemmas, we believe that there are three important implications. First, informational imperfections combined with perfect, but costly monitoring have a first-order negative effect on the supply of sanctions. Secondly, there is a second-order negative impact because some players choose to punish without monitoring. The productive type of sanctioning, namely that targeted at defectors, is provided, but only by a small share of subjects. Beliefs and risk aversion appear to determine how subjects sort into these latter two groups. Harnessing a group's propensity to supply sanctioning such as to maintain cooperation is therefore likely to hinge on the possibility to decrease monitoring costs and therefore the cost of targeting sanctions. Several avenues for future research along these lines suggest themselves, for example extending the setting to interactions with a longer horizon, increasing the group size, and varying the conditions and nature of information acquisition.

Appendix

Elicitation of beliefs

After subjects made their decisions in the first and second stage, respectively, we asked them to state their beliefs about the transfer decision of the coplayer as well as the damages they expect from their coplayer. Those statements were incentivized with a opportunity to earn extra tokens in case of good predictions. To elicit the subjects' beliefs at the transfer stage, we used an affine transformation of the one-dimensional Brier score (Brier, 1950). Let $\hat{a}^i_j \in [0,1]$ subject i's probability prediction that the target player j cooperated, then the Brier score is given by

$$s_{\text{Brier}} = \left(\hat{a}^i_j - a_j\right)^2 \in [0,1]$$

We implemented the specific transformation

$$s = 8 - 8s_{\text{Brier}} = 8 - 8\left(\hat{a}_j^i - a_j\right)^2 \in [0,8]$$

such that a perfect prediction was rewarded by eight tokens and the opposite by zero tokens. This scoring rule belongs to the class of quadratic score functions which are known to be proper (see for example Selten, 1998 and Gneiting & Raftery, 2007). Physically, we implemented the belief elicitation at the first stage by a screen with a slider with which subjects could specify their probabilistic belief that the target player transferred her or his endowment in ten-percent steps.

At the sanctioning stage we asked the subjects for point predictions of the damages assigned to them by their coplayers and incentivized statements with a distance score similar to those used, for example, by Gächter & Renner (2010) or Fischbacher & Gächter (2010). Here, each perfect prediction was rewarded by four tokens, a deviation of one point by two tokens, a deviation by two points by one token, and deviations of three of more points received no reward. This elicitation was physically implemented by a screen with input boxes shown at the end of the second stage.

Elicitation of risk preferences

At the end of each session we used the Holt-Laury lottery choice method (Holt & Laury, 2002) to obtain an indication of each subject's risk attitude. As described below in more detail, subjects were not told of this second task before the experimental game. Subjects faced a set of choices between two lotteries for each of ten decisions. The payoffs were identical for each lottery *A* (20 tokens or 16 tokens) and each lottery *B* (38.5 tokens and 1 token).[11] Probabilities for high and low payoffs are the same for both alternatives for each decision. Thus lottery *B* always has higher variance. As the subject moves down the ten decisions, the probability gradually shifts from the lower to the higher payoff. The expected return is higher for lottery *A* for the first four decisions, and for lottery *B* after that. We were interested in the point where subjects switch from the more risky option (lottery *B*) to the less risky option (lottery *A*) and how often they switch if applicable. A risk neutral agent is expected to switch in line three or four; the further down the switching point the higher the agent's degree of risk aversion. Physically, subjects made their choices on a screen with two check boxes for option *A* and *B*, respectively, for any of the ten decisions that were arranged row-wise on the screen. The lotteries had been resolved by throwing a ten-sided die, simulated by a random number generator program. One of the ten choices was actually paid out. Which one was determined by a second virtual ten-sided die.[12]

Table 3.4 contains summary data for the choices in the Holt-Laury task. The data is coded by the number of risky choices (option *B*).[13] Zero (ten) risky choices indicate

[11] These payoffs correspond to the payoffs in the low-stakes condition in Holt & Laury (2002).

[12] Laury (2005) provides evidence that supports the validity of using this random-choice payment method.

[13] Note that Holt & Laury (2002) report their data by the number of safe choices. We made the appropriate adjustments in comparing our data with theirs.

Table 3.4: Summary statistics of the Holt-Laury instrument

	Share by gender		Total share		
Risky choices	Male	Female	All	Consistent	Holt & Laury
0 – 1	.029	.106	.067	.076	.01
2	.074	.121	.097	.101	.03
3	.250	.197	.224	.235	.13
4	.250	.273	.261	.269	.23
5	.132	.167	.149	.135	.26
6	.132	.076	.105	.109	.26
7	.088	.046	.067	.059	.06
8	.015	.000	.008	.000	.01
9 – 10	.029	.015	.022	.017	.01
Mean	4.41	3.77	4.10	3.92	4.8

In the fifth column subjects whose choices were inconsistent with the typical one-crossover pattern (option *A* at the top, switching to *B* at some point) are removed.

extreme risk aversion (preference), six risky choices indicate risk neutrality. The subjects in our sample are somewhat risk averse, with an average of 4.1 risky choices. Consistent with most findings in the risk literature (see Eckel & Grossman, 2008), women are more risk averse than men, and this difference is marginally statistically significant in our sample (Mann-Whitney test, $p = .078$). The distribution of risk preference is similar to Holt & Laury (2002), however, the distribution in our sample is stronger skewed to the right indicating our subjects to be somewhat more risk averse. About 12 percent of the subjects had more than one crossover between the two options. The sub-sample of those inconsistent subjects is significantly different from the consistent subjects (Mann-Whitney test, $p = .009$), the former choosing on average more risky options (5.44) than the latter (3.92). In subsequent analysis, we include all subjects' choices.

Finally, since we conducted the lottery choice task at the end of each session, it may be possible that the prior tasks had some impact on the subjects' choices. In particular, choices in the lottery choice task may be different for subject's whose coplayer defected than for those whose coplayer cooperated in the first part of the session. However, a Mann-Whitney tests indicates that this is not the case: in terms of risky choices in the lottery choice task the sub-sample whose coplayer defected is not significantly different from the sub-sample whose coplayer cooperated ($p = .721$).

Chapter 4

Costly Monitoring and Third-Party Sanctioning: Experimental Evidence[1]

4.1 Introduction

Social norms, normative standards of behavior that are enforced by informal social sanctions, exist in every human society, and economists increasingly recognize their importance in accounting for a broad range of economic phenomena. Nonetheless, although they are frequently invoked as «all-round explanations» of behaviors that are inconsistent with standard behavioral models, their formation, content, and enforcement are still poorly understood (Fehr & Fischbacher, 2004a; Schram & Charness, 2011).

In this chapter we focus on the latter aspect, the underlying enforcement mechanisms. Social norms are enforced by the (justified) expectation that infringements will be sanctioned. Indeed, social enforcement is what social norms distinguish from moral norms, the latter being individually internalized and «enforced» internally (Elster, 2009). Thus, *observation* of behavior, Elster (2009) argues, is a core element of social norms, because this information is a critical conditioner of social sanctions. In fact, recent experimental research suggests that the efficacy of social sanctions is hampered by restrictions on observability (Carpenter, 2007b; Bornstein & Weisel, 2010; Grechenig et al., 2010; Ambrus & Greiner, 2012).

However, there are two important issues that have not been considered in combination so far. First, whether and to what extent a potential sanctioner is in the position to observe, or otherwise acquire information about the potential target individual's behavior, is rarely an exogenous matter of fact. In many situations, information on the target individual's actions is not immutably imperfect, but players are frequently in a position to augment available information through costly effort, either *ex ante* (e.g. putting oneself in a better «vantage point») or *ex post* (e.g. collecting evidence). In other words, a player's choice whether to impose social sanctions on

[1] This chapter is co-authored with Timo Goeschl.

another player is frequently preceded by a *choice* on whether to invest resources into monitoring his or her actions. Such monitoring effort to overcome imperfect information on coplayers' actions is well known in a variety of economically relevant contexts, such as shared resource management (Ostrom, 1990; Ostrom & Gardner, 1993; Seabright, 1993), production teams (Alchian & Demsetz, 1972; Kandel & Lazear, 1992; Dong & Dow, 1993; Craig & Pencavel, 1995) and alliances (Acheson, 1975, 1987, 1988; Palmer, 1991), labor relations (Shapiro & Stiglitz, 1984; Kanemoto & MacLeod, 1991; Lazear, 1993), micro-finance (Armendáriz & Morduch, 2005) or neighborhood watch (Sampson et al., 1997), to name just a few.

Second, while the vast majority of theoretical and experimental literature focuses on sanctions imposed by parties that are directly affected by the sanctioned behavior, so-called *second parties*, it has been argued that social sanctions by materially unaffected third parties are of particular importance to social norms in larger communities (Bendor & Mookherjee, 1990; Kandori, 1992a; Bendor & Swistak, 2001). There are some narrative accounts on third party sanctioning (e.g. Greif, 1993, 1994; Sober & Wilson, 1998; Fessler, 2002), and Fehr & Fischbacher (2004b) provided rigorous experimental evidence that third parties punish uncooperative behavior even at a cost to themselves.[2]

In this study we combine those two practically important aspects of social norm enforcement in a single experiment. In particular, we draw on the seminal experimental paradigm of Fehr & Fischbacher (2004b) that is designed to study third party punishment in the laboratory. This study was critical in the study of third party enforcement because the laboratory allows for control of many factors that may simultaneously drive punitive behavior in the field. Most importantly, since punishment is costly and there are certainly no pecuniary returns for third parties from sanctioning, we can be confident that any third party intervention is driven by subjective benefits alone. We augment this paradigm in a novel direction by allowing important parts of the information structure of the game to be *endogenously* determined by the players. Namely, while in Fehr & Fischbacher (2004b) third parties were automatically and freely informed about the target player's behavior before they had the opportunity to punish, third parties need to actively acquire this information, eventually at a cost, in our experiment. Without information acquisition, their decision in the sanctioning stage remains uninformed about the target player's action at the cooperation stage of the game. By choosing to acquire information they are able to condition their sanctioning on complete information regarding the coplayer's action. To the best of our knowledge, there exists no previous experimental evidence on *endogenous* mitigation of information asymmetries in a third party punishment game.

On the basis of this design, we address three main questions by comparing a baseline with zero information costs with two treatments with two different levels of

[2]This experiment has been replicated several times, sometimes with slightly differing designs (e.g. Charness et al., 2008; Ottone et al., 2008; Leibbrandt & López-Pérez, 2009; Kusakawa et al., 2012, see also chapter 4), and even in a number of very diverse non-standard subjects pools (Henrich et al., 2001, 2004, 2005, 2006; Marlowe et al., 2008).

information costs: First, what is the information acquisition behavior when observing target players' actions is costly? Second, how does information acquisition behavior respond to changes in the price of information? Third, how does the endogeneity of information impact on the occurrence and incidence of third party sanctions? Having answered these three questions, we compare third party behavior to second party behavior using another experimental condition that is identical except before-mentioned relation between sanctioner and sanctioned.[3]

The key results are the following. First, positive information costs have the predicted impact of decreasing the supply of third party punishment. Second, while virtually all sanctioning is discriminate at zero information costs, in the presence of positive information costs a distinct share of subjects switches to «blind» punishment rather than refraining from sanctioning. The rise of *deliberately* blind third party punishment is an interesting new finding of this chapter. Third, conditional on the decision to remain uninformed, it is risk averse subjects that tend to punish blindly, and more risk tolerant subjects that tend to refrain from punishment. Fourth, positive information costs lead to defectors making up a smaller share of the punished and receiving smaller punishment compared to zero information costs. Fifth, while third party punishment provides for weaker incentives to cooperate than second party punishment already when monitoring is costless, positive information costs render this difference even larger.

We proceed as follows. In section 4.2 we describe the aim, design, and procedures of the experiment. We proceed to present the main results in section 4.3. In section 4.4 we compare third party to second party behavior. We conclude in section 4.5.

4.2 Experimental Design and Procedures

The aim of the design is to generate empirical evidence on social sanctioning behavior when information structures are endogenous. In particular, we are interested in the amount of costly information acquisition, the impact of price variation on information acquisition, and the relationship between these endogenized information structures and sanctioning behavior. The previous literature has been careful to identify and isolate non-strategic sanctioning. In keeping with this focus, we design the experiment such that strategic and altruistic motives for monitoring coplayers' actions are ruled out as far as possible and the observed monitoring and sanctioning should be driven only by subjective rewards to the individual carrying out these activities.

We draw on the seminal study by Fehr & Fischbacher (2004b), who use a two-stage design consisting of a cooperation stage and a third-party sanctioning stage. The innovative step in our experiment consists of adding a monitoring decision into the sanctioning stage, in which third parties decide whether to give up material rewards in order to receive full information about the actions of their target player at

[3]We refer the reader to chapter 3 for a more thorough treatment of second party monitoring and punishment.

the cooperation stage. If they choose not to buy the information, they forgo the option of being able to discriminate between a cooperating and defecting target player at the sanctioning stage. If they do, sanctioning can be conditioned on observed target player behavior. We exogenously vary the magnitude of this fee as our treatment variable to observe how behavior is adjusted.

4.2.1 Experimental game form

Each subject played two two-stage experimental games, a second party monitoring game (SPMG) and a third party monitoring game (TPMG) with randomly matched and unknown coplayers in random sequence. Both games had two stages, wheres the first stage was identical in both games. In the first stage, both players in a given group of two were endowed with 10 tokens and interacted with one another in a strategic game of the Prisoner's Dilemma type with monetary payoffs. Subjects' first-stage decision was binary: Players could either keep their tokens or transfer all of them to the other group member, in which case the experimenter tripled them. Thus, denoting a given player by i and his or her coplayer by j, the first stage decision by $a_i \in \{0,1\}$ where $a_i = 1$ means cooperation, the first stage payoff equal to $10\left(1 - a_i + 3a_j\right)$. The benefit of a binary choice at the cooperation stage is not only a reduction in complexity (Ambrus & Greiner, 2012), but also a gain in experimenter control over the value of information that subjects can obtain through monitoring.

The second stage differed between the SPMG and the TPMG. At the beginning of the second stage, each player received an additional endowment of 40 tokens in both games. Subsequently, in the SPMG both players in each group had the opportunity to observe the coplayer's action in the first stage («monitor») and could choose to punish the other. In the TPMG, the first (second) player in a group had the opportunity to monitor and punish a player from another (a third) group. The matching protocol was identical to the one used by Fehr & Fischbacher (2004b) and rules are any reciprocity between third party monitors. In order to avoid potential biases due to a variable cost-impact ratio (see Casari, 2005), we chose a linear sanctioning technology commonly employed in the literature (Fehr & Gächter, 2002; Fehr & Fischbacher, 2004b; Gächter et al., 2008; Ambrus & Greiner, 2012): Denoting by p_{ij} player i's expenditure for punishing player j, the latter's payoff is reduced by $3p_{ij}$ tokens. Expenditures were restricted to integers and damage was limited to sixty tokens, therefore $p_{ij} \in \{0,1,\ldots,20\}$.

In both games, subjects reported their choices in the monitoring and punishment stages jointly using a modified version of the strategy method. Specifically, in the SPMG without a monitoring stage, players would simply announce a contingent punishment plan,

$$p_{ij}\left(a_j\right) : \{0,1\} \to \{0,1,\ldots,20\}$$

indicating the number of deduction points p_{ij} for each of the coplayer's possible transfer decisions a_j, given their own decision in the first stage. The presence of a monitoring stage was implemented in the strategy method by allowing subjects a

choice between two punishment plans. One was the standard contingent punishment plan $p_{ij}(a_j)$ that allowed the subject to condition their punishment on the coplayer's action a_j. This choice, denoted by $m_i = 1$, required payment of an exogenously set monitoring fee κ_m in addition to the applicable punishment costs p_{ij}. The alternative, denoted $m_i = 0$, was a punishment plan p_{ij} in which the subject could not condition the punishment on the coplayer's action a_j. This plan did not involve a fee over and above the cost of punishment.

Likewise, in the TPMG each third party, while being informed about the first stage transfer decision of the other member in their own group, had the choice between four punishment plans (since there are four possible transfer combinations in the other group). Specifically, denoting the first stage actions of the players in the other group by a_k and a_l, a generic punishment plan of third party i and target player k is given by

$$p_{ik}(a_k, a_l) : \{0,1\}^2 \rightarrow \{0,1,\ldots,20\}$$

where $m_i = 1$ if $p_{ik}(0, a_l) \neq p_{ik}(1, a_l)$ for all $a_j \in \{0,1\}$, else $m_i = 0$. Again, choosing $m_i = 1$ required payment of an exogenously set monitoring fee κ_m in addition to the applicable punishment costs. After the punishment stage, both games ended.

In summary, the pecuniary payoff function of each subject i in the SPMG was

$$w_i = 10(1 - a_i + 3a_j) + 40 - m_i\kappa_m - p_{ij} - 3p_{ji}$$

and in the TPMG

$$w_i = 10(1 - a_i + 3a_j) + 40 - m_i\kappa_m - p_{ik} - 3p_{hi}$$

where p_{hi} is the punishment i receives from h, the third party that is eligible to punish i.

When discussing the monitoring and punishment behavior at the individual level, the analysis will focus on three main behavioral types. One type are «non-punishing» subjects with $m_i = 0$ and $p_{ij} = 0$ or $p_{ik} = 0$ that decide not to monitor and do not punish. These subjects are referred to as behavioral type N. Among the punishing subjects, there are two types. «Blind punishers» with $m_i = 0$ and $p_{ij} > 0$ or $p_{ik} > 0$ devote no resources to monitoring, but still punish indiscriminately. This type is denoted as B. The other type T are «targeted punishers» with $m_i = 1$ and $p_{ij} > 0$ and $p_{ik} > 0$ that devote resources both to monitoring and to punishment. The B-types and T-types together form the class of «punishers», denoted P. The forth possible type (monitoring without punishment) is excluded by design.

4.2.2 Additional tasks

We supplemented the two games by incentivized elicitation of first-order beliefs and risk preferences. Both are potentially relevant in accounting for observed behavior (Keser & van Winden, 2000; Fischbacher et al., 2001; Fischbacher & Gächter, 2010). Since subjects receive no actual information in the course of play, they respond to

their beliefs on their coplayer's behavior, and this response involves a decision under strategic risk. Thus, in addition to the games described above, every subject performed the following tasks, which were identical in each treatment and for every subject. Following their decision in the first stage, we asked subjects to state their belief about the transfer decision of the target player and, following their decision in the second stage, their belief about the punishment they expected to receive. Those statements were incentivized with a opportunity to earn extra tokens in case of accurate predictions (see appendix).

At the end of each session we employed the Holt-Laury lottery choice method (Holt & Laury, 2002) to obtain an indication of each subject's risk preferences (see appendix). Subjects were only informed about this task after the two experimental games. The method may not capture the specific kind of risk involved in the situation, but we think it is worth investigating.

4.2.3 Design

Our primary focus in the design is on the relationship between information costs and monitoring and punishment behavior. The objective of the treatments is to introduce an exogenous variation in the cost of information acquisition with the aim of forcing a materially consequential choice by subjects regarding the desired information status prior to the punishment decision. As a result, information structures become endogenous and observable.

We employ between-subjects variation in the monitoring fee κ_m and examine the changes in the shares of types N, B, and T as κ_m changes, both in the TPMG in isolation and in comparison to the SPMG. For understanding the evolution of these shares of behavioral types, evidence on monitoring and sanctioning activity when monitoring is costless, i.e. when $\kappa_m = 0$, is an obvious starting point: It provides a direct replication of the PD experiment by Fehr & Fischbacher (2004b) and therefore a meaningful baseline treatment. This treatment is labeled $M0$. In the alternative treatments $M5$ and $M10$, the monitoring fee is set at $\kappa_m = 5$ and $\kappa_m = 10$. The choice of three fee levels $\kappa_m = \{0, 5, 10\}$ was on the basis that we had strong priors for a non-linear effect. In order to achieve efficient identification, we set the values at the extremes and the midpoint of the range and allocated approximately half of the subjects to the midpoint cell, a third to maximum cell, and a sixth to the minimum cell, respectively (see for example McClelland, 1997; List et al., 2011, for efficient sample arrangement methods). In order to counterbalance for possible order effects, one part of the subjects in each treatment cell played the SPMG before the TPMG, the other part the other way around.[4]

[4]It turned out that there were no systematic differences between sequences anyway. Using Mann-Whitney tests we find no significant differences among sequences in monitoring of second parties ($p = .160$) and third parties ($p = .775$), second party punishment of defectors ($p = .132$), second party punishment of cooperators ($p = .066$), third party punishment of defectors ($p = .122$ when the target's coplayer defected, $p = .116$ when the target's coplayer cooperated) and of cooperators when the target's coplayer cooperated ($p = .187$). Behavior was significantly different between sequences with respect to

On the basis of the design, we are able to test for responses in monitoring and punishment behavior with respect to changes in those costs by comparing the relative proportions of the three behavioral types in treatments $M0$, $M5$, and $M10$. In addition, we can exploit information about the resources spent on punishment and on behavior at the cooperation stage to understand more about the direction, intensity, and heterogeneity of social sanctions in the presence of costly monitoring.

4.2.4 Subjects and procedures

All experiments were conducted at the experimental laboratory of the Alfred-Weber-Institute (AWI-Lab) at Heidelberg University in late 2010. Participants were recruited from the general undergraduate student population using the online recruitment system ORSEE (Greiner, 2004). In total 134 subjects participated, of which 49.3 percent were female. A share of 64.2 percent had never participated in a laboratory experiment before. The mean age was 21.8 years. No subject participated in more than one session and treatment. Based on the sample arrangement described above, 22 subjects were assigned to $M0$, 66 to $M5$, and 46 to $M10$.

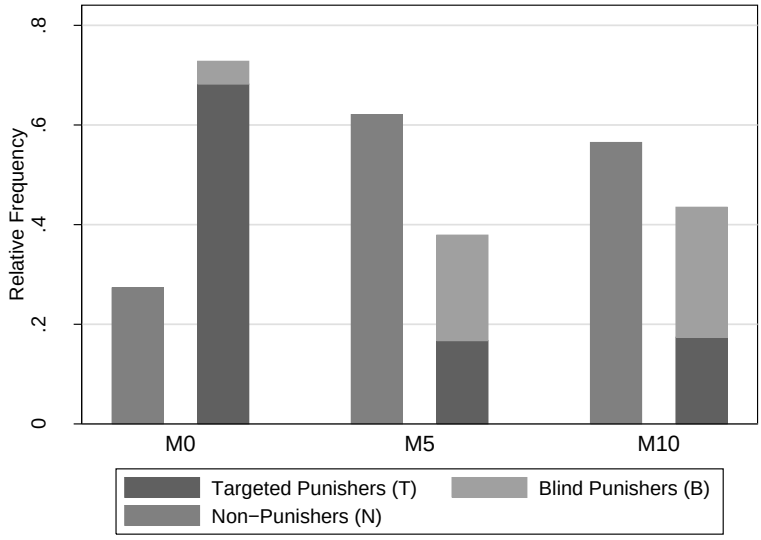
At the beginning of each session, upon entering the laboratory subjects were randomly assigned to the computer terminals. Direct communication among them was not allowed for the duration of the entire session. Booths separated the participants visually, ensuring that they made their decisions anonymously and independently. 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. The experiment was framed in a sterile way using neutral language and avoiding value laden terms in the instructions (see supplementary material). Participants had to answer a set of control questions individually at their respective seats in order to ensure comprehension of the rules. We did not start the experiment before all subjects had answered all questions correctly.

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 each subject made her or his decisions and reported expectations in the SPMG or TPMG, depending on the sequence. After being informed about the payoffs, the experimenter announced that second experiment will be conducted and distributed additional instructions that explained the differences to the first game. After being randomly re-matched into new groups, each subject made her or his decisions and reported expectations in the TPMG or SPMG, depending on the sequence. After being informed about the payoffs, the experimenter announced that another (and definitely final) experiment will now be conducted and distributed additional instructions that explained the supplementary lottery choice experiment. Thus,

third party punishment of cooperators when the target's coplayer defected ($p = .006$). Fisher exact tests yield similar results which are available on request.

Figure 4.1: Distribution of behavioral types at different monitoring cost levels.



subjects did not know until this announcement that the experiment involved another task in order to rule out confounding effects of the lottery-choice task on the main experimental games. We cannot rule out reverse confounds which is, however, of minor significance. After being informed about the payoffs, subjects were asked to answer a short questionnaire while the experimenter prepared the payoffs. Subjects were then individually called to the experimenter booth, payed 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 €2, which was not part of their endowment. The whole experiment lasted approximately 90 minutes and subjects earned an average of €18.46 (€0.10 per token earned), including the fixed show-up fee. Earnings exceed the local average hourly wage of a typical student job and can hence be considered meaningful to the participants.

4.3 Third Party Behavior

We organize our discussion of the experimental results in the following way. We begin by discussing our results from the baseline treatment in the third party monitoring condition that replicate previous experimental evidence on third party sanctioning behavior. We then describe the first core result of the experiment, namely the relative prevalence of the three possible behavioral types N, B, and T for each of the monitoring cost treatments $M0$, $M5$, and $M10$.

Figure 4.1 shows the relative frequencies of behavioral types N, B, and T for monitoring costs at three levels $\kappa_m = \{0, 5, 10\}$. Normalization is required on account of

the between-subjects design with a variable number of subjects in each treatment. A readily apparent first diagnosis of the evidence is therefore that variations in the cost of information are associated with different patterns of monitoring and punishment behavior. The key result of this section can already be highlighted: when information is imperfect and information costs are positive, a substantial share of third parties responds with information acquisition prior to punishment, that is, a significant share of subjects does not take exogenously imposed information constraints as given in a social sanctioning situation as long as the costs of overcoming these constraints are not prohibitive. The detailed nature of these differences is the subject of the following sections.

4.3.1 Third party punishment with costless monitoring: Replication of previous findings

The baseline treatment with $\kappa_m = 0$ replicates earlier experiments on third party punishment, specifically the experiment run by Fehr & Fischbacher (2004b). They find that when there are no hindrances to monitoring, the general propensity among third parties to punish target players is marked. Even after eliminating all strategic motives for punishment, through the matching scheme described in the previous section, the share of third parties that apply costly sanctions is somewhat above one half. Specifically, Fehr & Fischbacher (2004b) find that 59 percent of third parties applied costly sanctions on cheaters (unilateral defectors). Our results, displayed in the left-hand column in figure 1, are consistent with theirs.

Result 4.1. *When information costs are zero, most third parties make use of the non-strategic punishment option.*

In particular, we find a large majority of 73 percent (16 out of 22) willing to impose costly sanctions on the target players. The class of punishers P therefore dominates when there are no monitoring costs.

Previous experiments have also generated clear evidence that when information about coplayer behavior at the cooperation stage is perfect, then punishment is, on average, targeted. Fehr & Fischbacher (2004b) find that behavior at the cooperation stage is a highly significant predictor of receiving punishment from third parties. Our experiment replicates these findings closely.

Result 4.2. *When information costs are zero, third party punishment is targeted.*

When monitoring was costless, 15 out of 16 third parties that imposed a punishment acquired the information whether the target player cooperated or defected. In other words, almost all sanctions were conditioned on the target player's first stage behavior. In terms of our behavioral types, there are 68 percent T-types, 27 percent N-types, and 5 percent B-types.

Third parties also make a difference on whether defection of the target player is mutual or unilateral, that is, they also monitored the target player's coplayer. Among

the 15 T-types, 13 (or 87 percent) also conditioned punishment on the first stage behavior of the target player's coplayer.

The third aspect of our replication concerns the direction of punishment. In previous experiments, the direction and intensity is clear in that cooperators dominate as punishers and cheaters are their specific targets. Fehr & Fischbacher (2004b) do not report the composition of the group of punishers, but the following numbers are indicative: 68 percent (36 percent) of cooperating third parties imposed sanctions on unilateral (mutual) defectors as opposed to 40 percent (27 percent) of defecting third parties imposing sanctions on other unilateral (mutual) defectors. The two groups of punishers diverge even more in terms of the volume of sanctions. There, cooperating third parties punishing unilateral (mutual) defectors spent on average 3.7 tokens (1.7 tokens) on sanctions while defecting third parties punishing unilateral (mutual) defectors spent 1.9 tokens (0.9 tokens). While cooperating third parties sanctioning defectors, in particular unilateral ones, being the dominant direction of punishment, Fehr & Fischbacher (2004b) also uncovered that 15 percent of cooperators were targeted at the punishment stage (independently from the behavior of the target player's coplayer). Again, those patterns reappear in the baseline condition of our experiment.

Result 4.3. *When information costs are zero, the dominant direction and intensity of punishment is cooperating third parties imposing stiff sanctions on cheaters, but there is individual heterogeneity.*

36 punishment actions were carried out in the baseline treatment of whom 28 (78 percent) were taken by cooperating third parties and 8 (22 percent) by defecting third parties.[5] Among the class of third party punishers, therefore, cooperators clearly supplied most of the punishment action. Taking intensity into account, cooperating third parties accounted for 70 percent (166 of 237 tokens) of total investment into punishment while defecting third parties accounted for 30 percent.

Whom were the punishments imposed on? No third party targeted exclusively cooperators. In 26 out of 36 punishment decisions (72 percent) the target of a punishment action was a defector, 16 actions against unilateral defectors and 10 actions against mutual defectors. Cooperators were the target of 28 percent of punishment actions. Again, taking intensity into account, defectors received 77 percent (182 out of 237) of all the punishment imposed, 60 percent (142 out of 237) unilateral defectors and 17 percent (40 out of 237) mutual defectors, while cooperators received 23 percent (55 out of 237). Summarizing, punishment originates predominantly with cooperating third parties and arrives predominantly at cheaters.

A corollary of result 4.3 is that when monitoring was costless, cheating was clearly unprofitable: Unilateral (mutual) defectors received on average $6.0 \cdot 3 = 18.0$ tokens ($2.3 \cdot 3 = 6.9$ tokens) of punishment, while cooperators received on average $1.6 \cdot 3 = 4.8$ tokens, such that the difference was sufficient for the average net punishment of cheating ($18.0 - 4.8 = 13.2$ tokens) to exceed the gain (10 tokens) by a

[5] Recall that the strategy method allows each subject to punish up to four times, once for each possible first stage action profile in the target player's group.

sufficient margin.

4.3.2 Third party monitoring and punishment with positive information costs

Results 4.1 through 4.3 underline the benchmark quality of the baseline treatment by providing additional evidence to support similar findings in the literature. We confirm the presence of non-strategic third party punishment, the dominance of targeted punishment, the direction and intensity of punishment, and the degree of behavioral heterogeneity. Against this benchmark, we now report in the following three results on the behavioral implication of endogenizing third party's information status at the time of the punishment decision through a costly choice.

When monitoring is costly, the relative shares of behavioral types diverge from the baseline: A higher overall cost of sanctioning is associated with a higher share of subjects that neither monitor nor punish, which is in line with previous evidence on second party punishment showing that increasing cost of punishment reduce the supply of sanctions (Suleiman, 1996; Oosterbeek et al., 2004; Anderson & Putterman, 2006; Carpenter, 2007a; Egas & Riedl, 2008).

Result 4.4. *In the presence of positive information costs, the propensity to punish decreases relative to the baseline.*

Figure 4.1 illustrates this first finding: The share of subjects that neither monitor nor punish (type N) increases from 27 percent at $\kappa_m = 0$ to 62 percent when $\kappa_m = 5$ and 57 percent when $\kappa_m = 10$. The hypothesis that the frequency of N-types and monitoring costs are independent can be clearly rejected (Kruskal-Wallis test, $p = .017$), whereas pairwise, only the change associated with increasing the fee from zero to five is statistically significant (Mann-Whitney test, $p = .005$), the change associated with increasing the fee from five to ten is not ($p = .554$). The ratio of the frequency of punishers to the frequency of N-types decreases from 2.67 at $\kappa_m = 0$ to 0.61 at $\kappa_m = 5$ and to 0.77 at $\kappa_m = 10$. Thus, part of the effect of positive monitoring costs in the supply of punishment is comparable to the effect of an increase in the cost of punishment.

In addition to the result on the propensity to punish at all, our design explicitly allows to study punitive behavior of those third parties that opt to remain ignorant about the target player's first stage behavior.

Result 4.5. *In the presence of positive information costs, the propensity to punish indiscriminately increases relative to the baseline.*

Figure 4.1 again reports the share of blind punishers (type B) that do not acquire information and yet impose sanctions. This share increases from 5 percent (one out of 22) in the baseline to 21 (14 out of 66) and 26 percent (12 out of 46) as the monitoring costs increase to 5 and 10 tokens, respectively. Here, the change associated with increasing the fee from zero to five is statistically marginally significant (Mann-Whitney test, $p = .073$), the change associated with increasing the fee from zero to

ten is statistically significant ($p = .036$), while the change associated with increasing the fee from five to ten is not significant ($p = .550$). However, both frequencies under positive monitoring costs are statistically significantly different from zero (Wilcoxon signed-rank tests, $p < .001$, respectively). The ratio of the frequency of indiscriminate punishers to the frequency of all punishers (the frequency of targeted punishers) increases successively from 0.06 (0.07) at $\kappa_m = 0$ to 0.56 (1.27) at $\kappa_m = 5$ and to 0.6 (1.5) at $\kappa_m = 10$.

Finally, the share of targeted punishers (type T) decreases from 68 percent for costless monitoring to 17 percent when $\kappa_m = 5$ and $\kappa_m = 10$, respectively. Again, the hypothesis that the frequency of T-types and monitoring costs are independent can be clearly rejected (Kruskal-Wallis test, $p < .001$). Pairwise the changes associated with increasing the fee from zero to five and zero to ten, respectively, are statistically significant (Mann-Whitney test, $p = .000$ and $p = .036$, respectively), while the change associated with increasing the fee from five to ten is clearly not ($p = .920$).

Result 4.6. *In the presence of positive information costs, cooperating third parties remain the dominant origin of the direction of punishment relative to the baseline, but unilateral defectors dominate less as its target.*

35 percent (15 out of 43) and 44 percent (14 out of 32) of cooperating third parties imposed sanctions on defectors in $M5$ and $M10$, respectively, with respective average expenditures for punishing unilateral (mutual) defectors of 2.9 tokens (1.3 tokens) and 3.2 tokens (2.0 tokens). Eight (19 percent) and ten (31 percent) of the cooperators in $M5$ and $M10$, respectively, also imposed punishments on other cooperators, with respective average expenditures for punishing unilateral (mutual) cooperators of 0.7 tokens (1.2 tokens) and 0.6 tokens (1.6 tokens). 13 out of 23 (57 percent) and 8 out of 14 (57 percent) of the defecting third parties did not punish in $M5$ and $M10$, respectively. Of the remaining ten defectors in $M5$, seven punished indiscriminately (B-type) with an average expenditure of 4.6 tokens in case the target player's coplayer defected, and 4.3 tokens in case the target player's coplayer cooperated. Three targeted (T-type) with an average expenditure of 6.7 tokens on unilateral defectors, 5 tokens on mutual defectors, and 1.7 tokens on cooperators, plus the five tokens monitoring fee. Likewise, of the remaining six defectors in $M10$, four punished indiscriminately (B-type) with an average expenditure of 4.3 tokens in case the target player's coplayer cooperated, and 1.8 tokens in case the target player's coplayer defected. Two targeted (T-type) with an average expenditure of 6 tokens on unilateral defectors, 2.5 tokens on mutual defectors, and between 0.5 and 1 tokens on cooperators, plus the ten tokens monitoring fee.

Despite the rise in monitoring costs, unilateral defectors rather than cooperators remain the prevailing target of sanctions on average, although less pronounced. In $M5$, unilateral defection is still punished significantly more often than mutual defection (36 vs. 26 percent, $p = .020$, Wilcoxon signed-rank test) or cooperation (24 percent, $p = .011$). The same is true for $M10$, where unilateral defection is punished 41 percent of the time, and mutual defection (30 percent) or cooperation (33 percent)

less often, although the differences are not statistically significant in this condition ($p = .059$ and $p = .103$, respectively). Also, unilateral defectors continue to receive stronger punishment on average than mutual defectors or cooperators in both $M5$ (2.6 vs. 1.6 vs. 1.4 tokens, $p = .038$ for the first and $p = .018$ for the second difference), and $M10$ (2.8 vs. 1.7 vs. 1.5 tokens, $p < .001$ for the first and $p = .055$ for the second difference). However, in both conditions unilateral defection was no longer unprofitable. We return to this below.

To sum up, a setting with a perfect, but costly monitoring option leads to three important differences compared with zero monitoring costs: (1) The supply of third party sanctions decreases if punishment costs increase. (2) A small, but significant share of third parties sacrifices resources to actively monitor the target player and thus target their punishment. (3) A small, but significant share of third parties chooses to punish without information on whether the coplayer cooperated. The first-order impact of exogenous information imperfections that can be overcome at a cost is therefore to reduce the supply of third party sanctions. These imperfection also bring to light important heterogeneities across third parties with respect to the type of sanctioning they will supply, targeted or blind. Targeted sanctioning is conducive to maintaining cooperation because it rewards cooperative behavior in relative terms. The deliberately blind sanctioning observed in the experiment on the other hand is unproductive for maintaining cooperation.

System binary response regressions, reported in tables 4.6 in the appendix, suggest that third parties' beliefs about the target player's behavior and their risk preferences play a role in accounting for observed assortment into behavioral types.[6] In particular, while the probability of being a T-type is clearly negatively related to the monitoring fee, a subject's degree of risk aversion is significantly positively related to the probability of being a B-type, and significantly negatively related to the probability of being a N-type. That is, relatively risk averse subjects tend to be B-types, and relatively risk tolerant subjects tend to be N-types. This suggests that third parties tend to weight the risk of type-II enforcement errors, that is, letting a cheater go unpunished, stronger than type-I errors (punishing cooperators). In addition, B-types tend to be more pessimistic about the target player's cooperation, but this relationship is only marginally significant.

Result 4.7. *Conditional on the decision to remain uninformed, it is relatively risk averse subjects that tend to punish blindly, and relatively risk tolerant subjects that tend to refrain from punishment.*

[6] We use SUR since it is reasonable to assume that the error terms in the equations are correlated. As a robustness check, we also carried out a system probit regression estimated by maximum simulated likelihood (using the Geweke-Hajivassiliou-Keane simulator), and a system logit SUR that delivered equivalent results.

4.3.3 The emerging incentive structure

As shown above, in the baseline treatment of costless monitoring third party sanctions were sufficiently strong and targeted on unilateral defection to render it unprofitable. Hence, incentives were such that cooperative behavior is favored. For positive monitoring costs, however, this was no longer the case.

Result 4.8. *If monitoring is costless, unilateral defection is unprofitable relative to cooperation. At positive monitoring costs, unilateral defection does pay.*

Table 4.1: Frequency of third party punishment by treatment condition.

	Cooperation		Defection	
κ_m	Unilateral	Mutual	Unilateral	Mutual
0	.227	.227	.727	.455
5	.200	.227	.364	.258
10	.174	.326	.413	.304

Table 4.2: Mean damages (punishment points times three) through third party punishment by treatment condition.

	Cooperation		Defection	
κ_m	Unilateral	Mutual	Unilateral	Mutual
0	3.4	4.1	18.0	6.8
5	3.0	4.0	7.9	4.8
10	1.7	4.6	8.5	5.0

The impact of endogenizing the information structure by varying the monitoring costs is summarized in tables 4.1 and 4.2. We first observe that there is no significant variation in punishment of cooperation (Kruskal-Wallis tests, $p = .796$ for unilateral and $p = .600$ for mutual cooperation) and mutual defection ($p = .291$). Punishment of unilateral defection, however, is significantly different between treatments ($p = .005$), its strength being markedly lower when monitoring is costly compared to the baseline condition. Hence, *on average* defecting gets less costly as monitoring costs rise. As a result, unilateral defection *did* pay when monitoring costs were positive.

Do subjects anticipate the effects of monitoring costs on third party punishment and the incentives to cooperate, as predicted previously? There is little guidance to be derived from the literature on third party punishment in a PD since we are not aware of papers who elicited expectations. In the baseline condition of the present experiment, subjects' beliefs support the «punishment anticipation hypothesis». On average, subjects expected to receive punishment of 17.9 tokens in case of unilateral defection, 8.7 in case of mutual defection, and at most 4.2 in case of cooperation. This is a remarkably accurate prediction of the actual punishments of 18.0 tokens in case of unilateral defection, 6.8 in case of mutual defection, and at most 4.1 in

case of cooperation. The rank correlation between expectations and realizations is significantly positive (Kendall's τ_b = .440 in case of unilateral defection, τ_b = .552 in case of mutual defection, τ_b = .426 in case of unilateral cooperation, τ_b = .453 in case of mutual cooperation, all p = .000). In the presence of positive monitoring costs, on the other hand, the punishment anticipation hypothesis fails.

Result 4.9. *While subjects correctly predict the direction and intensity when monitoring costs are zero, the intensity of third party punishment of defection is overestimated for positive monitoring costs.*

Table 4.3: Mean damages (punishment points times three) through third party punishment expected by the target players.

	Cooperation		Defection	
κ_m	Unilateral	Mutual	Unilateral	Mutual
0	2.6	4.2	17.9	8.7
5	4.4	2.9	11.2	7.5
10	3.1	2.7	11.4	6.0

The evidence underpinning this result is shown in table 4.3. *Qualitatively*, subjects anticipate the actual impacts patterns as shown in table 4.2. *Quantitatively*, however, they significantly underestimate the effects of increasing monitoring costs on their coplayers' monitoring and punishment behavior.

4.4 Third Party and Second Party Behavior Compared

In this section, we compare the subjects' behavior in the role of a third party with that in the role of a second party. Figure 4.2 depicts the same data as in figure 4.1 but in direct comparison to the relative frequencies of behavioral types N, B, and T in the SPMG.[7]

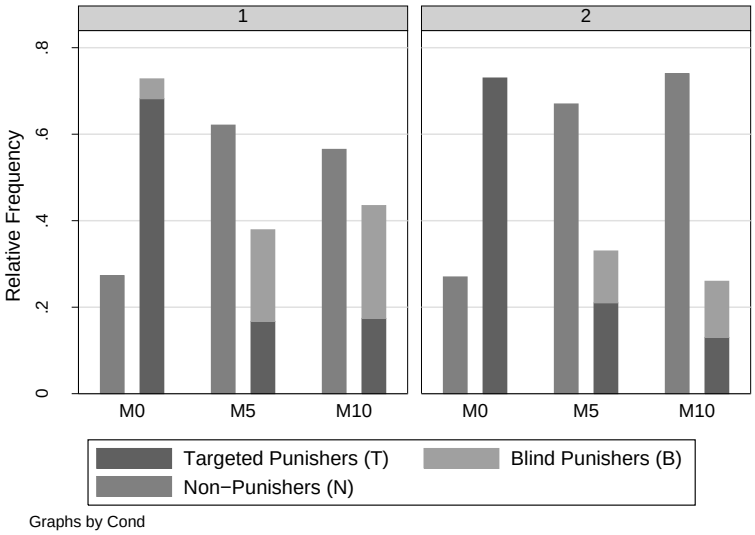
4.4.1 Aggregate Comparison

Result 4.10. *When information costs are zero, the frequency of second party punishment is equal to the frequency of third party punishment.*

In particular, 73 percent (16 out of 22) of the second parties were willing to impose costly sanctions, exactly the same relative frequency as for third parties. The 16 subjects are not identical in the two conditions, respectively. Two subjects who did not impose sanctions as a third party did so as a second party. Likewise, two subjects who did not impose sanctions as a second party did so as a third party.

[7]The results of the SPMG have been dealt with in detail in chapter3. Here we just compare the results to the TPMG results.

Figure 4.2: Distribution of behavioral types at different monitoring cost levels for third and second parties.



Result 4.11. *When information costs are zero, the share of targeted punishment is about equal in third and second parties.*

When monitoring was costless, all second parties (16 out of 16) that imposed a punishment acquired the information whether the target player cooperated or defected. In other words, all sanctions were conditioned on the target player's first stage behavior. In terms of our behavioral types, there are only T-types and N-types, but no B-types among the second parties. This distribution is approximately equal to the one in the third party monitoring condition.

Result 4.12. *When information costs are zero, the dominant direction and intensity of both third and second party punishment is cooperators imposing stiff sanctions on cheaters, but third party punishment is less targeted on defectors and weaker than second party punishment.*

19 punishment actions were carried out in the baseline treatment by second parties of whom 15 (79 percent) were taken by cooperators and 4 (21 percent) by defectors. Recall that among the third parties, 78 percent of punishment actions were taken by cooperators and 22 percent by defectors. Taking again intensity into account, cooperating second parties (third parties) accounted for 84 percent (70 percent) of total investment into punishment while defecting second parties (third parties) accounted for 16 percent (30 percent). Thus, the distributions do not differ much: Among the class of both second party and third party punishers, cooperators clearly supplied most of the punishment action.

Whom were the second party punishments imposed on? In 84 percent (72 percent) second parties' (third parties') punishment decisions the target of a punishment action was a defector. Cooperators were the target of 14 percent (28 percent) of second party (third party) punishment actions. Again, taking intensity into account, defectors received 93 percent (77 percent) of all the second party (third party) punishment imposed while cooperators received 7 percent (23 percent). Summarizing, both second and third party punishment originates predominantly with cooperators and arrives at defectors, but third parties are somewhat less targeting, imposing more punishment on cooperators.

Against the above benchmark comparisons, we now report on differential behavioral implication of endogenizing second and third parties' information status at the time of the punishment decision through a costly choice, respectively.

Result 4.13. *The propensity to punish decreases (relative to the baseline) slightly more strongly with information costs in second than in third parties.*

Figure 4.2 illustrates this first finding: The share of second parties (third parties) that neither monitor nor punish (type N) increases from 27 percent (27 percent) at $\kappa_m = 0$ to 67 percent (62 percent) when $\kappa_m = 5$ and 74 percent (57 percent) when $\kappa_m = 10$. Thus, the share of N-types is slightly less increasing in third than in second parties. The ratio of the frequency of second (third) party punishers to the frequency of N-types decreases from 2.67 (2.67) at $\kappa_m = 0$ to 0.5 (0.61) at $\kappa_m = 5$ and to 0.35 (0.77) at $\kappa_m = 10$.

Result 4.14. *The propensity to punish indiscriminately increases (relative to the baseline) more strongly in third than in second parties.*

Figure 4.2 again reports, in second and third parties, the share of blind punishers (type B) that do not acquire information and yet impose sanctions. In second (third) parties this share increases from zero (5 percent) in the baseline to 12 (21 percent) and 13 percent (26 percent) as the monitoring costs increase to 5 and 10 tokens, respectively. Analogously, the share of targeted second party (third party) punishers (type T) decreases from 73 percent (68 percent) for costless monitoring to 21 percent (17 percent) when $\kappa_m = 5$ and 13 percent (17 percent) when $\kappa_m = 10$, respectively. The ratio of the frequency of indiscriminate second party (third party) punishers to the frequency of all second party (third party) punishers increases successively from zero (0.06) at $\kappa_m = 0$ to 0.36 (0.56) at $\kappa_m = 5$ and to 0.5 (0.6) at $\kappa_m = 10$. The ratio of the frequency of indiscriminate second party (third party) punishers to the frequency of targeted second party (third party) increases successively from zero (0.07) at $\kappa_m = 0$ to 0.57 (1.27) at $\kappa_m = 5$ and to 1 (1.5) at $\kappa_m = 10$. Thus, third parties have clearly a stronger tendency to punish indiscriminately if monitoring is costly.

Result 4.15. *The incentives to cooperate are weaker in the third party than in the second party punishment condition at zero information costs. The presence of positive information costs increases this difference.*

31 percent (35 percent) and 29 percent (44 percent) of cooperating second parties (third parties) imposed sanctions on (unilateral) defectors in $M5$ and $M10$, respectively, with respective average expenditures of 11.6 tokens (2.9 tokens) and 8.2 tokens (3.2 tokens). 10 percent (19 percent) and 11 percent (31 percent) of the cooperating second parties (third parties) in $M5$ and $M10$, respectively, also imposed punishments on other cooperators, with respective average expenditures of 6.8 tokens (between 0.7 and 1.2 tokens) and 3 tokens (between 0.6 and 1.6 tokens). 61 percent (57 percent) and 82 percent (57 percent) of the defecting second (third) parties did not punish in $M5$ and $M10$, respectively. Of the remaining defecting second (third) parties in $M5$, 57 percent (70 percent) punished indiscriminately (B-type) with an average expenditure of 5.8 tokens (between 4.3 and 4.6 tokens), and 43 percent (30 percent) targeted (T-type) with an average expenditure of 6 tokens (between 5 and 6.7 tokens) on defectors and 5 tokens (1.7 tokens) on cooperators, plus the five tokens monitoring fee. Likewise, of the remaining defecting second (third) parties in $M10$, all (67 percent) punished indiscriminately (B-type) with an average expenditure of 5.8 tokens (between 1.8 and 4.3 tokens), and none (33 percent) targeted (T-type).

Despite the rise in monitoring costs, cheaters rather than cooperators remain the prevailing target of both second and third party sanctions on average, although less pronounced. Defection is still punished clearly more often by second parties than cooperation in both $M5$ (32 vs. 17 percent), and $M10$ (26 vs. 13 percent). The same has been shown above for third parties. Also, defectors continue to receive stronger punishment on average than cooperators by second parties in both $M5$ (9.8 vs. 3.3 tokens), and $M10$ (5.6 vs. 1.1 tokens). Again, the same has been shown above for third parties. The difference between average punishment of cheaters and average punishment of cooperators, however, decreases less strongly in second parties (from 20.1 tokens in $M0$ over 6.5 tokens in $M5$, a 68 percent decrease, to 4.5 tokens in $M10$, a further 31 percent decrease, for a total decrease of 78 percent) than in third parties (from 14.3 tokens in $M0$ over 1.2 tokens in $M5$, a 92 percent decrease, to 1.3 tokens in $M10$, for a total decrease of 91 percent). This reinforces the conclusion that while both second and third party punishment tends to become less targeted on cheaters through the introduction of monitoring costs, the effect is stronger in third parties. In other words, while third party punishment is provides weaker incentives to cooperate already at costless monitoring, introducing monitoring costs makes the difference even larger.

Result 4.16. *In both the second party and the third party monitoring games, cheating is only unprofitable if monitoring is costless.*

To sum up, while second parties respond to the introduction of monitoring costs more often than third parties with refraining from sanctioning altogether, third parties opt more often than second parties to indiscriminate punishment. Thus, third parties appear to be more willing to go for the risk of hitting a cooperator than second parties. In consequence, while both second and third party punishment gets less targeted, and third party punishment provides weaker incentives to cooperate than second party

punishment already when information costs are zero, introducing monitoring costs has a more deteriorating effect on the incentive structure emerging from third party punishment than from second party punishment.

4.4.2 Some fine-structure underlying the aggregate results

Consider punishment in the baseline treatment first. Previous and our own evidence shows that second party punishment of cheaters is stiffer than third party punishment *on average*. To what extent does this hold individually? Of the 18 subjects that imposed punishments on cheaters, eight (44 percent) punished strictly stiffer in the role of a second party than in the role of a third party, with an average expenditure difference of 6.6 tokens, and six (33 percent) made no difference. Thus, a majority of 14 subjects (78 percent) indeed punished at least as strong as a second party than as a third party. The remaining four subjects (22 percent), however, punished strictly stiffer in the role of a third party than in the role of a second party, with an average expenditure difference of 4.5 tokens. Although this individual level comparison reveals some additional heterogeneity, the conclusion that most people are willing to spend more on punishment when it is themselves who is cheated compared to when it is someone else is reinforced.

Nevertheless, what we are more interested in is the behavioral type, as defined above, subjects exhibited. In the baseline treatment $M0$, 17 out in 22 (77 percent) revealed the same behavioral type as both a second and a third party. The majority of those (13 subjects, or 76 percent) are universal T-types, that is, T-type as both a second party and as a third party. The rest (4 subjects, or 24 percent) are universal N-types, there exist no universal B-types when monitoring is costless. Of the remaining five subjects that revealed a different behavioral type as a second and a third party, respectively, two where type T as a second party and type N as a third party, two the reverse, and one subject was type T as a second party and type B as a third party.

Things were somewhat different when monitoring was costly. There, 80 out in 112 (71 percent) revealed the same behavioral type as both a second and a third party (49 out in 66 in $M5$, 31 out in 46 in $M10$). This is approximately equal to the baseline, but the composition of types differed: The clear majority are now the universal N-types (62 subjects, or 78 percent, 37 in $M5$ and 25 in $M10$). In addition, there is now a marked minority of universal B-types (8 subjects, or 10 percent, 5 in $M5$ and 3 in $M10$). The rest (10 subjects, or 12 percent, 7 in $M5$ and 3 in $M10$) are universal T-types, which were in the clear majority under costless monitoring.

Furthermore, the transition paths of those 32 subjects that revealed a different behavioral type as a second and a third party, respectively, are richer than in the baseline treatment. The complete transition matrix is shown in table 4.4 (a transition matrix including the baseline treatment can be found in 4.7 in the appendix). Apparently, transitions in all directions occur, but one trend can be highlighted: while only six subjects switch to B-type (from either N-type or T-type in the TPMG) in the SPMG, 18 subjects do so (from either N-type or T-type in the SPMG) in the TPMG. This is the fine-structure underlying the finding reported above that third party punishment

Table 4.4: Transition matrix of behavioral types across the SPMG and the TPMG for the costly monitoring treatments.

	TPMG Type			
SPMG Type	*N*	*T*	*B*	Margin
N	(62) .554	(6) .054	(10) .089	(78) .696
T	(2) .018	(10) .089	(8) .071	(20) .179
B	(3) .027	(3) .027	(8) .071	(14) .125
Margin	(67) .598	(19) .170	(26) .232	(112) 1.000

Relative frequencies. Absolute frequencies in parantheses.

becomes relatively less targeted than second party punishment when monitoring gets costly.

Of course, it is those subjects that change their behavioral type across the SPMG and the TPMG that drive the aggregate differences highlighted above. Thus, in order to understand those differences better, we take a closer look on them.

Pooling over all treatments, 37 out of 134 subjects change their behavioral type between the role of a second and a third party, respectively. There are two possible, not mutually exclusive explanations why subjects might behave differentially in the two roles. First, the kind of social or moral preferences applying to the situation are different. Second, the subjects' prior about the target player's behavior differs between the second and the third party monitoring condition.[8]

We have data to address the second hypothesis directly. Overall, the subjects' beliefs about the target player's behavior did not differ significantly between the second and the third party condition (Wilcoxon signed-rank test, $p = .340$): the average second party expected the target player to cooperate with probability .667, the average third party with probability .643.[9] For investigating the differential effects of monitoring costs on second and third party information acquisition and punishment, however, the relevant population is not the whole sample but only those 37 subjects that changed their behavioral type. But those subjects' beliefs about the target player's behavior did not differ significantly between the second and the third party condition either (Wilcoxon signed-rank test, $p = .766$): the average second party expected the target player to cooperate with probability .703, the average third party with probability .676.[10]

Thus, beliefs appear to be of minor significance in accounting for the differential effects of monitoring costs on second and third party information acquisition and punishment. This suggests that the key driver are differences in the kind of social or

[8] One might add that the shape of the risk preferences differs between conditions, but this appears not very plausible.

[9] Separated by treatment, only when monitoring is costless there is a significant difference ($p = .046$), the average second party expecting the target player to cooperate with probability .618, the average third party with probability .718.

[10] Separated by treatment, there are also no significant differences.

moral preferences applying in second and third parties, respectively.

4.5 Conclusion

We extend previous experimental research on social norms by combining two practically important aspects of their enforcement: third party sanctioning and the active mitigation of informational asymmetries. To our knowledge, this chapter is the first to consider *endogenous*, rather than exogenously given, information structures by allowing for information acquisition on target players' behavior in the context of a cooperation game with third party punishment. This design enables both a close replication of previous evidence on the presence of third party sanctioning and the observation of new types of heterogeneities among subjects. We show that a majority of third parties responds to positive monitoring costs in a manner that is broadly equivalent to a rise in punishment costs and obeys the *law of demand*. At the same time, we observe, on the one hand, active information acquisition for targeting sanctions that do not provide material returns either to the punisher or anybody else. On the other, the design also allows third parties to punish «blindly» if they so wish, and we find a significant share of subjects that chooses this course of action. Given the practical relevance of an endogenous information structure in many cooperation problems, we believe that there are three important implications. First, informational imperfections combined with perfect, but costly monitoring have a first-order negative effect on the supply of sanctions. Secondly, there is a second-order negative impact because some players choose to punish without monitoring. The productive type of sanctioning, namely that targeted at defectors, is provided, but only by a small share of subjects. The enforcement of cooperative social norms therefore likely hinges on the possibility to decrease monitoring costs and therefore the cost of targeting sanctions.

Appendix

Elicitation of beliefs

After subjects made their decisions in the first and second stage, respectively, we asked them to state their beliefs about the transfer decision of the coplayer as well as the damages they expect from their coplayer. Those statements were incentivized with a opportunity to earn extra tokens in case of good predictions. To elicit the subjects' beliefs at the transfer stage, we used an affine transformation of the one-dimensional Brier score (Brier, 1950). Let $\hat{a}^i_j \in [0,1]$ subject i's probability prediction that the target player j cooperated, then the Brier score is given by

$$s_{\text{Brier}} = \left(\hat{a}^i_j - a_j\right)^2 \in [0,1]$$

We implemented the specific transformation

$$s = 8 - 8s_{\text{Brier}} = 8 - 8\left(\hat{a}^i_j - a_j\right)^2 \in [0,8]$$

such that a perfect prediction was rewarded by eight tokens and the opposite by zero tokens. This scoring rule belongs to the class of quadratic score functions which are known to be proper (see for example Selten, 1998 and Gneiting & Raftery, 2007). Physically, we implemented the belief elicitation at the first stage by a screen with a slider with which subjects could specify their probabilistic belief that the target player transferred her or his endowment in ten-percent steps.

At the sanctioning stage we asked the subjects for point predictions of the damages assigned to them by their coplayers and incentivized statements with a distance score similar to those used, for example, by Gächter & Renner (2010) or Fischbacher & Gächter (2010). Here, each perfect prediction was rewarded by four tokens, a deviation of one point by two tokens, a deviation by two points by one token, and deviations of three or more points received no reward. This elicitation was physically implemented by a screen with input boxes shown at the end of the second stage.

Elicitation of risk preferences

At the end of each session we used the Holt-Laury lottery choice method (Holt & Laury, 2002) to obtain an indication of each subject's risk attitude. As described below in more detail, subjects were not told of this second task before the experimental game. Subjects faced a set of choices between two lotteries for each of ten decisions. The payoffs were identical for each lottery *A* (20 tokens or 16 tokens) and each lottery *B* (38.5 tokens and 1 token).[11] Probabilities for high and low payoffs are the same for both alternatives for each decision. Thus lottery *B* always has higher variance. As the subject moves down the ten decisions, the probability gradually shifts from the lower to the higher payoff. The expected return is higher for lottery *A* for the first four decisions, and for lottery *B* after that. We were interested in the point where subjects switch from the more risky option (lottery *B*) to the less risky option (lottery *A*) and how often they switch if applicable. A risk neutral agent is expected to switch in line three or four; the further down the switching point the higher the agent's degree of risk aversion. Physically, subjects made their choices on a screen with two check boxes for option *A* and *B*, respectively, for any of the ten decisions that were arranged row-wise on the screen. The lotteries had been resolved by throwing a ten-sided die, simulated by a random number generator program. One of the ten choices was actually paid out. Which one was determined by a second virtual ten-sided die.[12]

Table 4.5 contains summary data for the choices in the Holt-Laury task. The data is coded by the number of risky choices (option *B*).[13] Zero (ten) risky choices indicate extreme risk aversion (preference), six risky choices indicate risk neutrality. The subjects in our sample are somewhat risk averse, with an average of 4.1 risky choices. Consistent with most findings in the risk literature (see Eckel & Grossman, 2008),

[11] These payoffs correspond to the payoffs in the low-stakes condition in Holt & Laury (2002).

[12] Laury (2005) provides evidence that supports the validity of using this random-choice payment method.

[13] Note that Holt & Laury (2002) report their data by the number of safe choices. We made the appropriate adjustments in comparing our data with theirs.

Table 4.5: Summary statistics of the Holt-Laury instrument

	Share by gender		Total share		
Risky choices	Male	Female	All	Consistent	Holt & Laury
0 – 1	.029	.106	.067	.076	.01
2	.074	.121	.097	.101	.03
3	.250	.197	.224	.235	.13
4	.250	.273	.261	.269	.23
5	.132	.167	.149	.135	.26
6	.132	.076	.105	.109	.26
7	.088	.046	.067	.059	.06
8	.015	.000	.008	.000	.01
9 – 10	.029	.015	.022	.017	.01
Mean	4.41	3.77	4.10	3.92	4.8

In the fifth column subjects whose choices were inconsistent with the typical one-crossover pattern (option *A* at the top, switching to *B* at some point) are removed.

women are more risk averse than men, and this difference is marginally statistically significant in our sample (Mann-Whitney test, $p = .078$). The distribution of risk preference is similar to Holt & Laury (2002), however, the distribution in our sample is stronger skewed to the right indicating our subjects to be somewhat more risk averse. About 12 percent of the subjects had more than one crossover between the two options. The sub-sample of those inconsistent subjects is significantly different from the consistent subjects (Mann-Whitney test, $p = .009$), the former choosing on average more risky options (5.44) than the latter (3.92). In subsequent analysis, we include all subjects' choices.

Finally, since we conducted the lottery choice task at the end of each session, it may be possible that the prior tasks had some impact on the subjects' choices. In particular, choices in the lottery choice task may be different for subject's whose coplayer defected than for those whose coplayer cooperated in the first part of the session. However, a Mann-Whitney test indicates that this is not the case: in terms of risky choices in the lottery choice task the sub-sample whose coplayer defected is not significantly different from the sub-sample whose coplayer cooperated ($p = .721$).

Supplementary material

Table 4.6: Results of a system probit model estimated by Seemingly Unrelated Regression (SUR)

	Dependent variable		
	Being a *N*-type	Being a *T*-type	Being a *B*-type
Monitoring fee	0.072	-0.138	0.062
	(0.043)	(0.062)	(0.040)
	.097	*.027*	*.124*
Pessimistic belief	-0.064	-0.373	0.187
	(0.329)	(0.527)	(0.185)
	.845	*.479*	*.072*
Optimistic belief	-0.409	0.451	0.087
	(0.258)	(0.309)	(0.127)
	.113	*.144*	*.493*
Risk averse	-0.529	0.320	0.489
	(0.241)	(0.253)	(0.265)
	.028	*.205*	*.065*
Female	0.090	-0.142	0.041
	(0.183)	(0.229)	(0.285)
	.622	*.533*	*.885*
Age	0.038	-0.040	-0.016
	(0.022)	(0.024)	(0.031)
	.080	*.094*	*.595*
Constant	-0.616	0.614	-1.338
	(0.578)	(0.837)	(0.739)
	.287	*.463*	*.070*
Obs	134	134	134
Log likelihood	-86.288	-65.283	-63.126
Prob > χ^2	0.060	0.002	0.210
Pseudo R^2	0.066	0.140	0.062

For each independent variable, the first row reports the estimated coefficient, the second row the standard error (in brackets), and the third row the *p*-value of a test with Null that the coefficient is equal to zero (in italics). Standard errors take into account clustering by session. Pessimistic (elicited belief of 0.3 or less) and optimistic (elicited belief of 0.7 or more) belief is relative to an indeterminate belief (elicited belief between 0.4 and 0.6), respectively. Risk averse (4 or less risky choices in the Holt-Laury task) is relative to risk neutral and risk seeking (5 or more risky choices), respectively.

Table 4.7: Transition matrix of behavioral types across the SPMG and the TPMG pooled over all treatments.

	TPMG Type			
SPMG Type	*N*	*T*	*B*	Margin
N	(66) .493	(8) .060	(10) .075	(84) .627
T	(4) .030	(23) .172	(9) .067	(36) .269
B	(3) .022	(3) .022	(8) .060	(14) .104
Margin	(73) .545	(34) .254	(27) .201	(134) 1.000

Relative frequencies. Absolute frequencies in parantheses.

Chapter 5

Costly Monitoring and the Emergence of Blind Trust[1]

5.1 Introduction

The realization of mutual gains from cooperation is jeopardized by opportunism in a variety of economic interactions. For example, expert service providers, such as car mechanics, have plenty opportunity to cheat on their customers, *inter alia* by performing unnecessary fixes or charge for services that they did not actually perform (see Dulleck & Kerschbamer, 2006, for more examples).

It is well known that the prospect for repeat business, the «shadow of the future», can have, under appropriate conditions, a disciplining function in those situations (Rubinstein, 1979; Fudenberg & Maskin, 1986; Kreps et al., 1982), and the experimental evidence is clearly supporting (see section 5.2). One key component of those «appropriate conditions» is the perfectness of monitoring, that is, cheating is immediately and effortless detected. The introducing example illustrates that this is often not the case: most car owners have no idea how their car work, so they typically cannot assess the mechanic's service quality even after consumption; they often happen to trust the mechanic blindly. This monitoring imperfection is the key characteristic of «credence qualities» (Darby & Karni, 1973) as opposed to «experience qualities» (Nelson, 1970). While cheating on the latter is also possible, the consumer cannot perfectly monitor the occurrence or extent of fraud in the latter, and this persistent information asymmetry significantly heightens the temptation to cheat (Emons, 1997). Since credence qualities are ubiquitous in modern economies,[2] an important question, therefore, is how monitoring imperfections impact on relational enforcement.

[1] This chapter is co-authored with Timo Goeschl.

[2] For example, not only externally bought services involve credence qualities, the services bought from employees as well. It is also instructive to think of many forms of pollution as credence «bads» because the victims often not know how much they exposed to a particular pollutant nor how exposure translate into health states.

There is an increasing number of theoretical and experimental studies that introduce «noise» into the system (see section 5.2), that is, monitoring is subject to error. These papers have not only introduced an important practical feature of relational enforcement, but have also delivered important insights into the role of different information structures in supporting, or weakening, mutual cooperation in ongoing relationships. However, the entirely passive characterization of monitoring is somewhat at odds with the more *active* nature of monitoring in practice:[3] Information on coplayers' actions is routinely not only imperfect, but players are frequently in a position to augment available information through costly effort. In other words, the player's choice whether to cooperate is frequently accompanied by a *choice* on whether to invest resources into monitoring the coplayers' actions. Indeed, Darby & Karni (1973, p. 69) mentioned in their seminal article that «credence qualities are those which, although worthwhile, cannot be evaluated in normal use. Instead, the assessment of their value requires additional costly information ... Credence qualities are expensive to judge even after purchase.» Efforts to overcome imperfect information on coplayers' actions is well known in a variety of further economically relevant contexts as well, such as shared resource management (Ostrom, 1990; Ostrom & Gardner, 1993; Seabright, 1993), production teams (Alchian & Demsetz, 1972; Kandel & Lazear, 1992; Dong & Dow, 1993; Craig & Pencavel, 1995) and alliances (Acheson, 1975, 1987, 1988; Palmer, 1991), labor relations (Shapiro & Stiglitz, 1984; Kanemoto & MacLeod, 1991; Lazear, 1993), micro-finance (Armendáriz & Morduch, 2005) or neighborhood watch (Sampson et al., 1997), to name just a few. In this chapter, we account for this fact, and thereby extend the experimental literature in a new direction by allowing important parts of the information structure of the game to be *endogenously* determined by the players.

Specifically, we report upon an experiment in which subjects play finite horizon modified trust games with information structures that are endogenous in this respect. The trust game (Camerer & Weigelt, 1988; Kreps, 1990) is a generic representation of transactions such as those involving the opportunity to cheat on qualities, both of the experience and the credence types, and thus highlighting the need for the customer's trust for the exchange to occur: The latter may either purchase the service (cooperate) or not (defect), the provider may respond to a purchase by honest performance (cooperate) or fraud (defect). The novel feature in our experiment is that a cooperating first mover does not automatically learn her or his payoff at the end of a period, just as in the credence quality example, but may actively acquire information about the second mover's action in that period.

As indicated above, conceiving monitoring as an *action* also gives rise to a stronger notion of trust. In the experimental literature, «trust» is typically defined by a cooperative choice of the first mover in the trust (or related) game.[4] However, for Elster

[3] While those studies differ in the details of the specific monitoring imperfection considered, all of them share the assumption that those impediments are *exogenous* and *fixed*. Indeed, the commonly used definition of monitoring as «the *ability* of agents to observe each other's actions in the marketplace» (Holcomb & Nelson, 1997, p. 79, emphasis by us) underlines the implicit exogeneity assumption.

[4] Apart from the fact that there is no universally agreed upon definition of trust, one may question

(2007, p. 344) trust requires more, not only going for the risk of being exploited but «to *refrain from taking precautions against an interaction partner*» (emphasis in the original), to «lower one's guard». Thus, trust is, according to Elster, «the result of two successive decisions: *to engage in the interaction and to abstain from monitoring the interaction partner*» (p. 345, emphasis by us). Our experimental game design allows exactly for this succession. Elster's definition is narrow, indeed, but it certainly describes a particularly strong form or trust. To avoid confusion with broader definitions, we shall refer to this behavior as *blind trust*.

The primary questions we pose are: (i) Do first movers always make use of the monitoring option when there is a strategic reason to do so, or conversely, is blind trust an empirically relevant option? (ii) What are the consequences for cooperation, efficiency, and distribution? (iii) How do the answers to the previous questions change with the cost of monitoring? Another goal is to investigate the dynamic patterns of relationships under these conditions. As a working hypothesis, we draw on a proposition from the management literature which poses that ongoing relationships develop through a process starting with the control-driven stage and converging to a trust-based one (Lewicki & Bunker, 1996; Lewicki et al., 1998). In our experiment, this proposition predicts that, at least in the costly monitoring condition, monitoring will occur predominantly in the initial periods, whereas revealed trustworthiness eventually leads to increased blind trust over time.

To answer these questions we implement two experimental conditions in which we exogenously vary the cost of information acquisition: it is costless in one condition while information on the coplayer's action costs a fee in the other. The key results are the following. First and foremost, the introduction of information costs resulted in a decrease of monitoring but an emergence of blind trust as a new behavior type. The latter effect was quantitatively so strong that (i) blind trust turned out to be the dominant behavior under costly monitoring, (ii) such that first mover cooperation was significantly *more* frequent and payoffs higher under costly monitoring than under costless monitoring. Furthermore, 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. In sum, (i) blind trust is an empirically relevant phenomenon, and (ii) it is *caused* by monitoring costs, and (iii) seems to be a successful adaption to a setup in which monitoring is costly. We find that this static differences between conditions stem from some important differences in the respective dynamic patterns. In the costly monitoring condition, there is a dynamic shift from monitoring towards blind trust, fueled by remarkably high reciprocation rates in the initial rounds that shifted interactions on an very cooperative trajectory. Our preferred interpretation is a «second-order reputation building» hypothesis, according to which some second movers try to strategically exploit the

whether first mover behavior in the trust game, or the related investment game, actually measures trust. While it certainly measures cooperation, it may include *unconditional* cooperation which is distinct from trust if the latter is defined as cooperation conditional on the expectation of reciprocation (Cox, 2004).

costliness of monitoring by investing in a sufficiently favorable reputation in the initial periods in which they are likely to be monitored in order induce blind trust and reap larger gains from exploitation in later periods. We provide further results that support this interpretation.

In the remainder we proceed as follows. In section 5.2 we briefly review related work. In section 5.3 we describe the design of the experiment and report on procedures and implementation. The results are presented in section 5.4. We summarize and conclude in section 5.5.

5.2 Related Literature

The idea that relational enforcement of cooperation can be a sub-game perfect or sequential equilibrium of a repeated game has been shown by Fudenberg & Maskin (1986) for the case of an indefinite horizon and by Kreps et al. (1982) for the case of a finite horizon.[5] Empirical support from experimental research is strong for both results: Findings based on a variety of specific stage games, such as the simultaneous PD (e.g. Andreoni & Miller, 1993; Cooper et al., 1996; Dal Bó, 2005; Dal Bó & Fréchette, 2011), the public good game (Andreoni & Croson, 2008, provide an overview), or the gift exchange game (Kirchler et al., 1996; Fehr et al., 1998a; Falk et al., 1999; Gächter & Falk, 2002) generally show that repeated play (with perfect monitoring) generates cooperation strictly above the one-shot Nash equilibrium level and below the first-best level. Specifically, there are three studies using a finite horizon trust game (Anderhub et al., 2002; Engle-Warnick & Slonim, 2004, 2006a), as we do, who find significantly more cooperation compared to a condition with random rematching in non-terminal rounds and a collapse in terminal rounds; Cochard et al. (2004) obtained similar results with an repeated investment game.

However, both theoretical results and related experiments are based on the assumption of *perfect public monitoring*, that is, each action taken by each player is observed by all other players automatically, immediately and without error. Specifically, monitoring is called *public* if the signal generated by one player's action is identical for all players. If the signal accurately confers a player's action, monitoring is called *perfect*. Likewise, if the signal is noisy in the sense that it may at times confer a different action than actually performed by the player, monitoring is called *imperfect*.

There is a growing literature attempting to extent the Folk Theorem to the case of imperfect monitoring. Fudenberg et al. (1994) extended their earlier result to imperfect public monitoring (see Fudenberg & Yamamoto, 2011, for generalizations). If different players receive different signals, including the case where some receive no signal at all, monitoring is called *private*. There are some approaches to prove

[5] The latter result 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, a belief that is indeed justified as demonstrated by a vast amount of recent evidence (Gintis et al., 2005; Henrich et al., 2004).

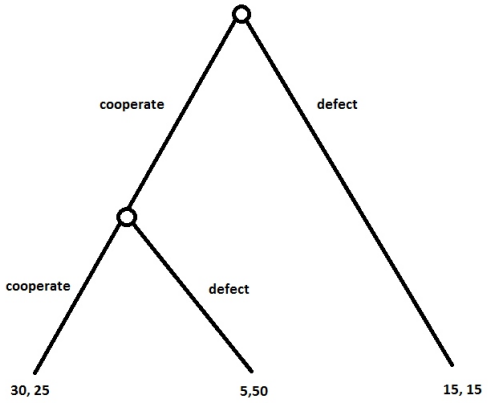
a Folk Theorem with private monitoring (e.g. Sekiguchi, 1997; Piccione, 2002; Ely & Välimäki, 2002; Bhaskar & Obara, 2002; Mailath & Morris, 2002; Hörner & Olszewski, 2006), but this case is technically challenging. Typically, those approaches are based on «close-to-public» monitoring, that is arbitrarily small disturbances from public monitoring, but it is difficult to assess how critical the informational requirements of these Folk Theorems are because generally a discussion of the actual order of magnitude of the disturbances from public monitoring is missing (Gintis, 2009). Furthermore, there are, to our knowledge, no general results extending the Kreps et al. (1982) result for finite horizon games to cases with imperfect monitoring.

For empirical purposes, however, those models suggest a simple intuitive prediction concerning the introduction of monitoring imperfections: Since under imperfect monitoring punishments are likely to occur along the trigger-strategy equilibrium path, payoffs will likely be lower compared to the perfect information case. This intuition has long been articulated in the industrial organization literature, beginning with Stigler (1964) who suggested that collusion in concentrated industries may be difficult to sustain in a repeated game with secret price cuts (see also Green & Porter, 1984). Hence, there are also some experimental studies that were framed in this context. Holcomb & Nelson (1997) study a repeated (Cournot) duopoly experiment in which information about the coplayer's quantity choice is randomly changed half of the time and find that such manipulation significantly hampers collusion relative to a perfect monitoring condition. Feinberg & Synder (2002) conducted a repeated (Bertrand) duopoly experiment with a similar monitoring imperfection and also found less collusive behavior in the noisy monitoring compared to a perfect information treatment. Specifically, while players did resort to punishments for undercutting under perfect monitoring, they appeared to refrain from trigger strategies and settle on the competitive stage game outcome under noisy monitoring.[6] Game theoretically, both studies examine imperfect *private* monitoring, that is, players do neither necessarily observe the same nor an accurate signal about the coplayers' actions. Using an imperfect *public* monitoring (the signal is not necessarily accurate but identical for all subjects), Aoyagi & Fréchette (2009) find that subjects' payoffs (i) decrease as noise increases, and (ii) are lower than the theoretical maximum for low noise, but exceed it for high noise.

There are also two studies introducing some specific monitoring imperfections into the experimental public good game. Sell & Wilson (1991) conducted a repeated public good experiment with three conditions: (1) no information about other members' contributions, (2) aggregated information about other members' contributions, and (3) individualized information about each member's contribution. They find that contributions in the individualized information condition are greater than contributions in the other two conditions, while contribution levels for no information and aggregated information do not differ. Also studying a repeated public good experiment, Cason & Khan (1999) compare standard perfect monitoring with perfect, but

[6]Note that both Holcomb & Nelson (1997) and Feinberg & Synder (2002) used finite horizon supergames but did not report what was told to the subjects about the horizon in the instructions.

Figure 5.1: Experimental «stage game». The complete game is the twelve-fold sequence with monitoring decisions in between.



delayed monitoring of past actions. They do not find any significant difference in the levels of contributions between the two treatments.

As noted in the introduction, common to all of those models and experiments is the operationalization of monitoring imperfections as «noise». To our knowledge, there are to date no studies that operationalize monitoring as an active decision. In this chapter, we extend the literature in this new direction.

5.3 Experiment

5.3.1 Design

We used the conventional (binary) trust game (Camerer & Weigelt, 1988; Kreps, 1990), as depicted in figure 5.1, as our point of departure. The first mover chooses between cooperation (option «pink» in the instructions) or 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 coplayer 30 tokens. Otherwise, he exploits the first mover by taking 50 tokens for himself while his coplayer gets 5 tokens. Every subject plays the game for 12 periods.

The novel feature in our experiment 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 coplayer 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 coplayer monitored them or not.[7]

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

The experimental game was manipulated along one dimension for a two factor between-subjects design. In the first experimental condition the acquisition of information on the coplayer'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 treatment conditions were exactly identical.

5.3.2 Subjects and Procedures

Participants were recruited from the general undergraduate student population of the University of Heidelberg using the online recruitment system ORSEE (Greiner, 2004). In total 48 subjects participated of which 47.9 percent have been female and 83.3 percent German citizen. The mean age was 23.7 years. Subjects were randomly assigned to treatment conditions, 24 subjects in each cell. 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 in spring 2012. 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.[8] 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

[7] This is an empirically accurate representation of many, but not all monitoring activities. If the second mover gets to know whether he or she is monitored, he or she may respond to this information in the subsequent periods (if any). We study this setup and evaluate a «crowding» hypothesis in a different paper.

[8] They were instructed to take this sheet with them after the experiment to ensure that nobody, including the experimenters, could observe their eventual notes.

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, payed 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 45 minutes and subjects earned €11.21 (€0.03 per token earned) on average, including the fixed show-up fee, with a minimum of €7.80 and a maximum of €14.25. Earnings exceed the local average hourly wage of a typical student job and can hence be considered meaningful to the participants.

5.4 Results

5.4.1 Main results

The key results of this chapter are illustrated in figures 5.2 and 5.3. The main result is the emergence of blind trust as monitoring becomes costly. As evident from the bottom panel of figure 5.2, cooperation accompanied by monitoring decreases by roughly 24 percent (from 50 percent to 37.8 percent) as one moves from the costless to the costly monitoring condition. Basing this and the following statistical tests on a cross-section in which each observation is an individual average taken over all 12 rounds,[9] this difference is statistically significant (Mann-Whitney rank sum test, $p = .023$). This is what one might expect. However, at the same time there is there is a new behavior type «blind trust», cooperating without monitoring, monitoring, when monitoring is costly, whose frequency is statistically significantly different from the costless monitoring condition (Mann-Whitney, $p < .001$). This behavior type is so dominant in the costless monitoring condition, that first mover cooperation rates are actually *higher* in the costly monitoring condition (Mann-Whitney, $p = .005$). First movers trusted blindly in remarkable 52.8 percent (76 out of 144) of the cases, resulting in a total cooperation rate of 80.6 percent (116 out of 144), which is much higher than in the costless monitoring condition (50 percent, 72 out of 144). Indeed,

[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 24 matches constitute one independent observation. The procedure used here is a common response to this fact (e.g. Vanberg, 2009).

Figure 5.2: Relative frequency of first mover cooperation by treatment condition averaged over time.

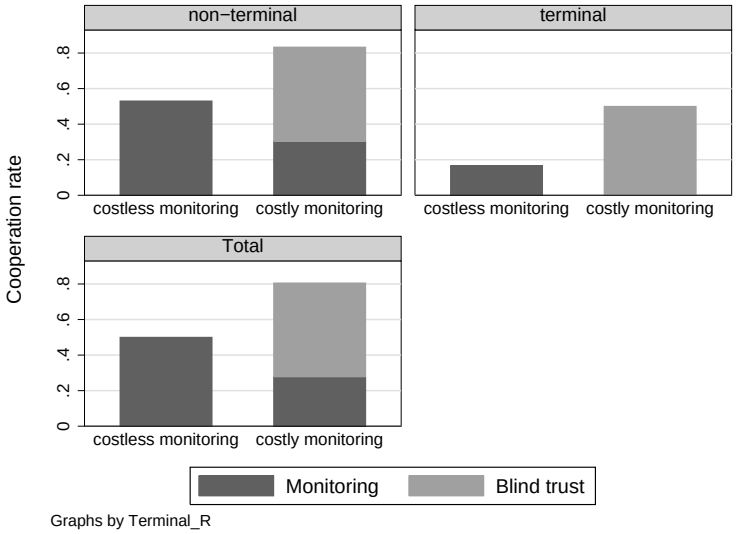


Figure 5.3: Relative frequency of reciprocation by treatment condition averaged over time.

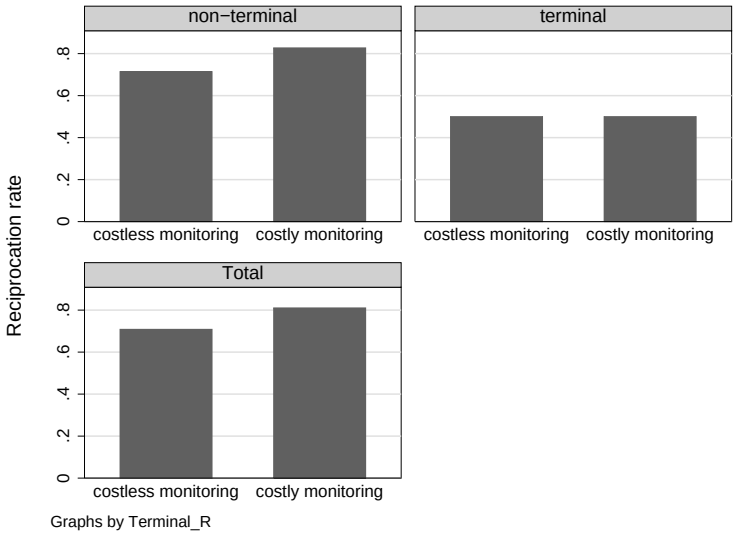
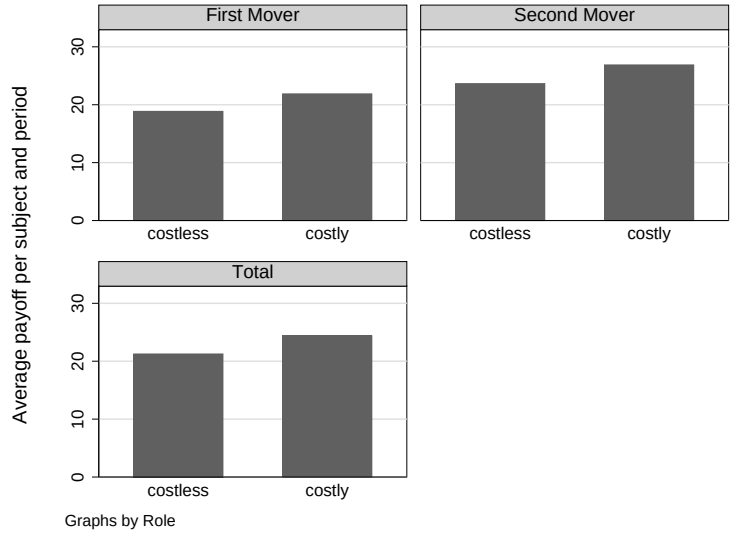


Figure 5.4: Average payoff per subject and round by treatment condition.



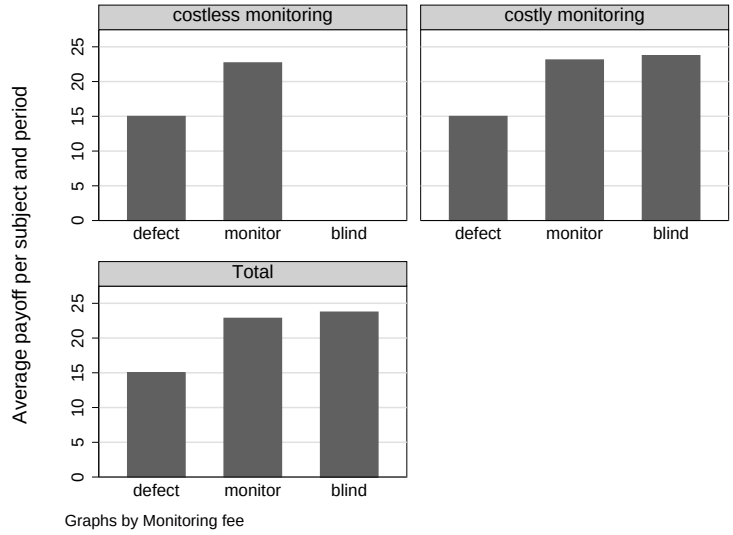
in the terminal round *all* instances of cooperation, after all 6 out of 12 cases, were blind trust (upper right panel of figure 5.2). Thus, in the costly monitoring condition blind trust turns out to be the dominant behavior.

Result 5.1. *There is less monitoring if it is costly then when it is costless. But there is a behavior type «blind trust», cooperating without monitoring, that is more frequent when monitoring is costly than when it is costless. Taken together, first mover cooperation is more frequent when monitoring is costly than when it is costless.*

But this raises the question why blind trust is so common, since one might expect that this option does not pay, because blind trustors are easily cheated upon, perhaps over multiple periods without noticing it. Figures 5.3 and 5.4 show a first indication that, on average, this was not the case. The reciprocation rate, depicted in figure 5.3, is actually slightly *higher* in the costly monitoring condition, although the difference is not statistically significant (Mann-Whitney, $p = .276$). Figure 5.4 shows the average payoffs earned per subject and round, including monitoring costs, by treatment.[10] Again, first movers earned on average actually *more* in the costly monitoring condition (21.9 tokens) than in the costless monitoring condition (18.9 tokens), even including monitoring costs, the difference being marginally significant (Mann-Whitney, $p = .082$). The average second mover also earned more in the costly

[10]Recall that, excluding monitoring costs, the minimum joint payoff was 30 tokens (i.e. 15 on average) and the maximum 55 tokens (27.5 on average). While expenditures on monitoring were, of course, zero in the costless monitoring condition, first movers spent per subject and period 1.39 tokens on monitoring in the costly monitoring condition (i.e. 16.7 tokens per match). Averaging only over those first movers who actually cooperated, they spent 1.72 tokens per period (i.e. 20.6 tokens per match).

Figure 5.5: Average first mover payoffs by behavioral type.



monitoring condition (26.9 tokens) than in the costless monitoring condition (23.6 tokens), the difference being significant (Mann-Whitney, $p = .049$), such that on average both types of player were, conservatively, not worse off through the introduction of monitoring costs. Jointly, the average pair earned significantly more in the costly than in the costless monitoring condition (Mann-Whitney, $p = .027$). Note that those marked behavioral effects result already at a relatively moderate cost of monitoring: 5 tokens are only 20 percent of the loss of betrayal, 33 percent of the gain from mutual cooperation, and about 26 percent of the average first mover per period payoff in the costless monitoring condition.

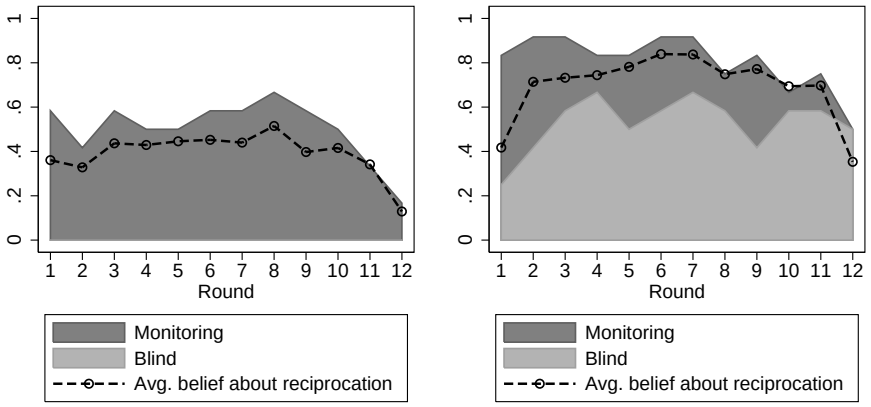
Of course, more appropriate to judge whether blind trust does or does not pay is to compare the payoff of blind trustors to the other behavioral alternatives.[11] Such comparison is depicted in figure 5.5. On average, cooperating clearly payed to first movers. Monitoring cooperators earned more than defectors, whereas blind trustors did not worse than the former. We thus can conclude that, on average, blind trust was *not* unprofitable.

Result 5.2. *The average first mover in the costly monitoring condition did not do worse than in the costless monitoring condition. Likewise, the average blind trustor did not do worse than the average monitor or defector.*

Summing up so far, three facts can be highlighted. First, blind trust is an empirically relevant option. In fact, in the costly monitoring condition, it is the dominant

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

Figure 5.6: 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.



behavior. Second, blind trust is *caused* by monitoring costs, as there is no blind trust if monitoring is costless. Finally, blind trust seems to be a successful adaption to a setup in which monitoring is costly, since neither did the average first mover in the costly monitoring condition worse than in the costless monitoring condition, nor did the average blind trustor worse than the average monitor or defector. In the remainder of this section we examine the data in more detail in order to get an idea how this works out.

5.4.2 Dynamics of first mover behavior under costless monitoring

We begin by examining the dynamics of first mover behavior. The upper panel of figure 5.2 depicts aggregate first mover behavior splitted up into terminal and non-terminal periods, and figure 5.6 depicts the full aggregate dynamics of first movers' behavior. In the latter, 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 left-hand column of figure 5.6 depicts the results for the costless monitoring condition. We already know that first movers never cooperated without monitoring the second mover's response, so the frequencies of cooperation coincide with the frequencies of monitoring. 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. In fact, the information about the second mover's behavior is clearly valuable in non-terminal periods of a given match. It is of least instrumental value in terminal rounds, but still, when the monitoring fee is zero, there is still no reason to refrain from monitoring contingent on cooperating. Con-

sistent with these considerations, none among the 72 instances of cooperation (out of 144 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.

As a first step, we observe that the average dynamic pattern of first mover cooperation in the costless monitoring condition fits well into the previous literature that studied repeated trust games (and related cooperation games) under perfect monitoring, referred to in section 5.2: moderate cooperation at the beginning, eventually slightly increasing over time before it collapses towards the terminal period. Cooperation rates are clearly higher in non-terminal rounds than in terminal rounds, first movers cooperating in 53.0 percent (70 out of 132) of the time in the former and in 16.7 percent (2 out of 12) of the time in the latter. We summarize sloppily:

Result 5.3. *If monitoring is costless, average first mover cooperation follows the typical pattern of finite horizon cooperation games with perfect monitoring.*

5.4.3 Dynamics of first mover behavior under costly monitoring

Based on this benchmark result, we now consider the costly monitoring condition in more detail. As evident from the right-hand column of figure 5.6, first mover cooperation rates exhibit a similar pattern than under costless monitoring, although at a notably higher level due to the emergence of blind trust.

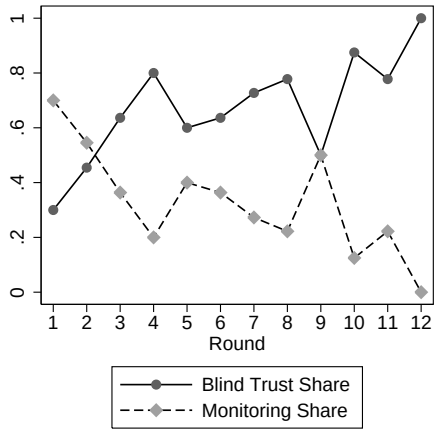
Result 5.4. *If monitoring is costly, average first mover cooperation follows a similar dynamic pattern as under costless monitoring, although at a markedly higher level.*

In this condition, the *value* of information about the second movers' response is unchanged, but the *cost of acquiring it* is now positive. Thus, by standard economic reasoning it can be expected that subjects acquire the information only if its subjective value amounts to at least five tokens, the monitoring fee. This leads to a number of predictions.

First, because for some players in some periods the subjective valuation of the information may be less than five tokens, the frequency of monitoring is predicted to be lower than in the costless monitoring condition. The lower panel of figure 5.2 shows that, averaged over time, this clearly turned out to be the case, and figure 5.6 illustrates that this also holds for each period individually.

Second, while there might be different motivations to monitor, considering the «instrumental value» of the information, the extent to which the information can be used to adjust one's behavior afterwards, might be a first guide to form an expectation about what one might observe: Since the instrumental value of information tends to decrease over time, or is at the very least lower in terminal than in non-terminal periods, the frequency of monitoring can be expected to decrease over time. To see this, consider the incentive to monitor in terminal periods. The instrumental value of the information is clearly zero in the terminal period, because there is no opportunity to

Figure 5.7: Relative dynamics of monitoring and blind trust in the costly monitoring condition.

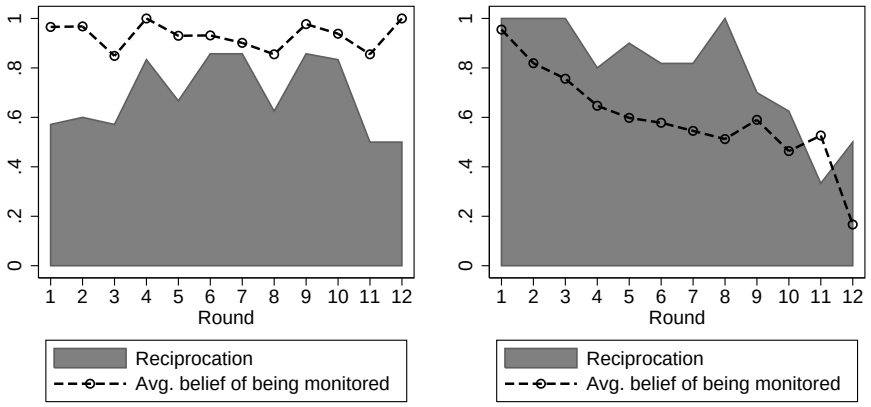


act on the information, such that it is reasonable to expect there to be little monitoring. In fact, in the terminal period, none out of 12 first movers monitored. In contrast to terminal periods, the instrumental value of the information about the coplayer's response is clearly positive in non-terminal periods, both for strategic and learning reasons. Thus, it is reasonable to predict higher monitoring rates in non-terminal than in terminal rounds. This prediction is supported by the data as well. In non-terminal rounds, first movers cooperated together with monitoring their coplayer in 30.3 percent (40 out of 132) of the time, while they never monitored their coplayer (0 out of 12) in terminal rounds. Figure 5.6 illustrates the decreasing trend of monitoring over time.

A first mover who does plan to not monitor the second mover has two alternatives: non-cooperation or blind trust. Figure 5.6 shows that the decrease in monitoring is only weakly accompanied by an increase in non-cooperation but by a marked increase in blind trust. While blind trust is already present in the first round, it is clearly the exception initially but strongly increases throughout the first four periods to become the dominant behavior.

This dynamic pattern is also informative about a proposition claiming that a relationship develops through a process starting with the control-driven stage and converging to a trust-based one (Lewicki & Bunker, 1996; Lewicki et al., 1998). In our experiment, this corresponds to the prediction that first movers predominantly resort to monitoring at the beginning of a match while tending to trust blindly towards the end. Putting the above results together, it is clear that this turns out to be the case: In the first period, the majority of cooperating first movers monitor, blind trustors are a minority. However, this changes over time, as blind trust increases notably in the initial four periods and then remaining approximately stable, while the frequency of monitoring successively decreases over time. The shift from monitoring towards

Figure 5.8: Aggregate dynamics of second mover reciprocation and beliefs about being monitored. The left column depicts the costless monitoring condition, the right column the costly monitoring condition.



blind trust over time can be seen even more clearly in figure 5.7, which depicts the frequency of blind trust and monitoring, respectively, as a fraction of all instances of first mover cooperation. However, our results also adds an important qualification to the above proposition, as it is only supported under costly monitoring. Thus, monitoring costs appear to be a key driver of convergence towards (blind) trust in ongoing relationships.

Result 5.5. *If monitoring is costly, monitoring gets less frequent and blind trust gets more frequent towards the terminal round, that is, there is a shift from monitoring towards blind trust over time.*

The implication is that much of blind trust is a dynamic phenomenon, that is, a great deal may likely be explained by the dynamics of the initial periods. We turn to an investigation of this conjecture next.

5.4.4 How does blind trust emerge? – Dynamics of second mover behavior

We begin by taking a closer look at the dynamics of second mover behavior. The upper panel of figure 5.3 depicts aggregate second mover behavior splitted up by terminal and non-terminal periods, and figure 5.8 depicts the full aggregate dynamics of second mover's behavior. The left-hand column of figure 5.8 depict the results for the costless monitoring condition. The time trend of the reciprocation rate is non-monotone since reciprocation is always contingent on first mover cooperation. Thus, a sorting effect tends to increase reciprocation rates over time, since exploiters are no longer trusted (at least for some time), while the strategic reciprocation incentive diminishes towards the terminal period, inducing a trend tendency into the opposite direction. Consistent with a strategically reciprocal strategy, cooperation rates

are clearly higher in non-terminal rounds than in terminal rounds. In non-terminal rounds, second movers reciprocated in 71.4 percent (50 out of 70) of the time, while in in terminal rounds, they returned cooperation in 50.0 percent (1 out of 2) of the time. As a result, overall second mover reciprocation rates were 70.8 percent (51 out of 72).

Consider now the costly monitoring condition depicted in the right-hand column of figure 5.8. As in the costly monitoring condition, cooperation rates are clearly higher in non-terminal rounds than in terminal rounds. In non-terminal rounds, second movers reciprocated in 82.7 percent (91 out of 110) of the time, while in in terminal rounds, they returned cooperation in 50.0 percent (3 out of 6) of the time. As a result, overall second mover reciprocation rates were 81.0 percent (94 out of 116), which is somewhat higher than under costless monitoring. Figure 5.8 illustrates why: while reciprocation rates are similar in both treatment conditions towards the terminal period, they are clearly higher in the costly monitoring condition than in the costless monitoring condition over the initial rounds, in fact at remarkable 100 percent in period 1 through 3. In addition, reciprocation rates have a markedly decreasing trend over time in the costly monitoring condition.

Result 5.6. *In both the costless and the costly monitoring condition the majority of first mover cooperation is reciprocated. However, reciprocation is markedly more frequent in the initial rounds of the latter compared to the former, and there is a more pronounced decreasing trend if monitoring is costly.*

The latter fact is quite consistent with the average second mover's belief about being monitored in the current period. The circles connected with the hatched line in figure 5.8 depict this belief. Evidently, the average second mover anticipates the average first mover's monitoring pattern (as summarized in result 5.5) remarkably closely.[12] While the average second mover's belief about being monitored is approximately stable close to unity in the costless monitoring condition, it exhibits a clearly decreasing pattern under costly monitoring. Thus, while the average second mover believes to be monitored with almost-certainty throughout the whole match in the costless monitoring condition, (s)he expects to be monitored predominantly at the beginning of a match, less so towards the end, under costly monitoring.

Result 5.7. *The average second mover anticipates the average first mover's monitoring pattern quite accurately. Specifically, while the average second mover believes to be monitored with almost-certainty throughout the whole match in the costless monitoring condition, (s)he expects to be monitored predominantly at the beginning of a match, less so towards the end, under costly monitoring.*

Did the average second mover responded consistently to this belief? Conventional economic theory suggests that the temptation to cheat is decreasing in the perceived likelihood of being detected, and *vice versa*. The belief and behavior patterns

[12] Recall that we deliberately kept them uninformed about whether they are currently monitored or not.

depicted in figure 5.8 are consistent with this intuition: while second movers respond in a standard way to their belief of being constantly monitored in the costless monitoring condition, reciprocation rates diminish with the belief of being monitored towards the end in the costly monitoring condition. In both conditions, the average cooperating second mover had a stronger belief of being monitored (.945 in the costless monitoring condition, .652 in the costly monitoring condition) than the average defecting second mover (.880 in the costless monitoring condition, .471 in the costly monitoring condition), where the difference is significant in the costly (Mann-Whitney, $p = .039$) but not in the costless monitoring condition ($p = .260$). Rank correlation between beliefs of being monitored and reciprocation is positive and significant when monitoring is costly (Kendall's $\tau_b = 0.168$, $p = .040$).

While this accounts neatly for the increasing trend of cheating in the costly monitoring condition, a simple response to the belief of being monitored in the current period cannot, however, account for result 5.6, namely that reciprocation is markedly more frequent in the initial rounds if monitoring is costly compared to the costless monitoring condition. Although speculative at this point, a possible explanation of this fact is that the active and costly nature of monitoring gives rise 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 our 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 our game there is the possibility of cheating over *multiple* periods once the first mover trusts blindly. This is a possible explanation why reciprocation rates are initially so much higher under costly monitoring. Intuitively, (some) second movers 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. Strategically acting subjects may do so in order to exploit their coplayers more easily and perhaps over multiple periods. This strategic incentive is missing in the costless monitoring condition, since second movers do not believe in blind trust anyway (and this belief is justified). Thus, we have the following conjecture

Conjecture 5.1. *Some second movers try to strategically exploit the costliness of monitoring by investing in a sufficiently favorable reputation in the initial periods in which they are likely to be monitored in order induce blind trust and reap larger gains from exploitation in later periods.*

We leave a direct test of this conjecture for future research, but the facts presented below already lend some preliminary support.

5.4.5 How does blind trust emerge? – Strategic reputation building and first mover responses

Support for this conjecture comes from the fact that the strategy is in some sense quite successful. To see this consider first mover behavior in more detail. First, note that (see figure 5.6) the average first mover's *prior* about their cooperation being reciprocated is approximately equal at .4 in both conditions. But while this belief does not change much over time in the costless monitoring condition, there is a marked increase of confidence in the initial periods of the costly monitoring condition. This increase may be viewed as justified given full reciprocation. In fact, rank correlation between first movers' belief about reciprocation in period $t+1$ and an indicator variable that is unity if a respective first mover detected her or his coplayer to cooperate in period t is strongly positive (Kendall's $\tau_b = .555$, $p = .000$),[13] that is, the average first mover became much more confident (pessimistic) in $t+1$ if (s)he observed her or his coplayer reciprocating (defecting) in t.

Figure 5.6 also indicates that first movers' confidence and cooperation go together. As evident from the left panel, in the costless monitoring condition the pattern of first mover cooperation rates corresponds quite closely to the average first mover's belief about reciprocation, depicted by circles, connected with the hatched line. The average first mover anticipated the coplayer's strategic incentive to defect towards the terminal period, and responds accordingly. Correlation between the first movers' belief about reciprocation and their own cooperation is strongly positive and significant (Kendall's $\tau_b = .618$, $p = .000$).

The same is true in the costly monitoring condition (Kendall's $\tau_b = .584$, $p = .000$).[14] Interestingly, however, separating first mover cooperation by cooperation coupled with monitoring and blind trust, it turns out that only the latter is correlated with first movers' beliefs about reciprocation (Kendall's $\tau_b = .410$, $p = .000$), the former is not (Kendall's $\tau_b = .059$, $p = .419$). The correlation is also positive at the individual level for all first movers who trusted blindly at least once, whereas most of the coefficients are not statistically significantly different from zero due to the small number of individual observations.[15] To put it differently, the average belief that the second mover will reciprocate was .874 for blind trustors, .788 for monitors, and .075 for defectors. Thus, a first mover's confidence in reciprocation appears to be a key determinant of blind trust.

Result 5.8. *While the average first mover's prior about reciprocation is approximately equal in both conditions, and does not change much over time in the costless monitoring condition, there is a marked increase of confidence in the initial periods of*

[13] We also estimated this correlation by means of various regression models that take into account period-specific effects which yield even stronger correlations. Those results are available upon request.

[14] Pooled over treatments, Kendall's rank correlation coefficient between beliefs about reciprocation and first mover cooperation is $\tau_b = .654$ with $p = .000$.

[15] Of the 12 first movers, four did never trust blindly, while eight did; among the latter, two correlations between beliefs and blind trust is marginally significant, six are not significant.

the costly monitoring condition, accompanied by a strong increase in blind trust. In fact, blind trust is positively correlated with first movers' belief about reciprocation.

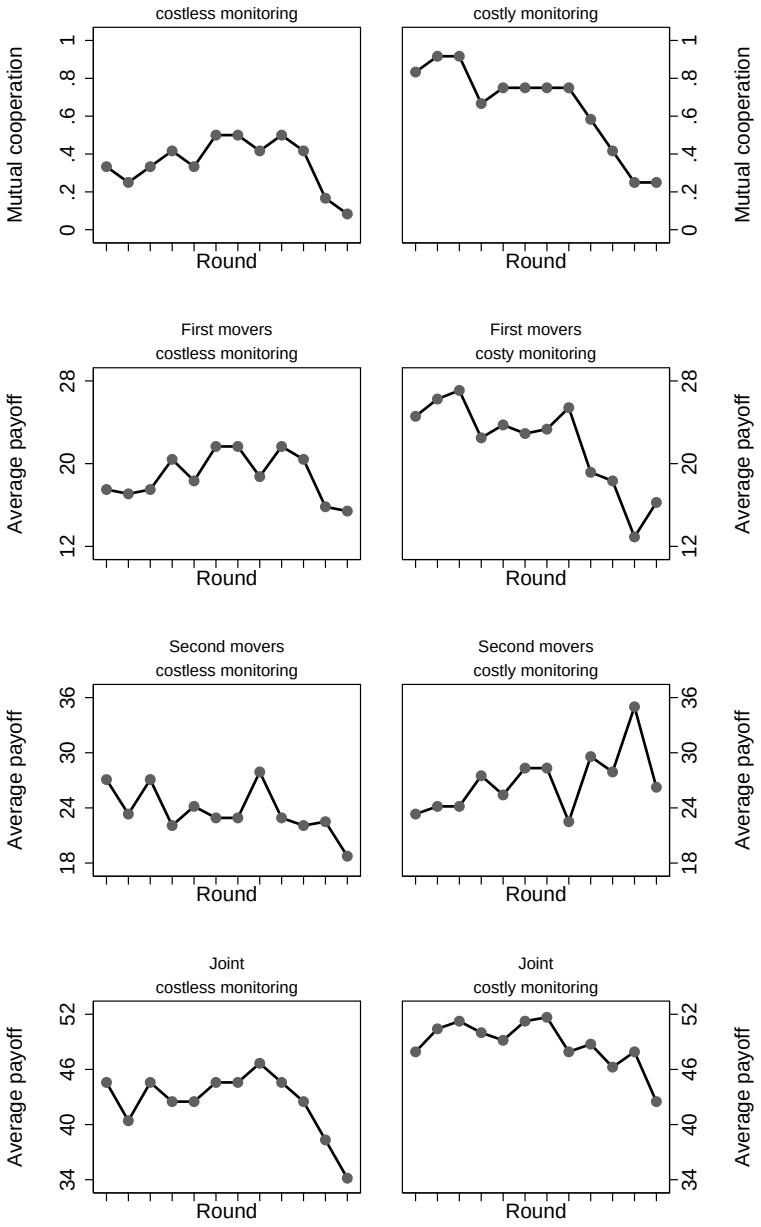
This kind of closes the circle and reinforces the interpretation that (i) (some) second movers try to «earn» being trusted blindly by cooperating in the initial periods, (ii) first movers become indeed more confident, and (iii) respond with trusting blindly.

As another piece of evidence in favor of this «second-order reputation building» interpretation, consider the dynamics of realized payoffs depicted in figure 5.9. In the costless monitoring condition (depicted in the left-hand column), the frequency of mutual cooperation is moderately positive throughout the initial periods, but strongly decreasing towards the end, partly through first movers stopping to cooperate (foreseeing the second movers' strategic incentive to defect), partly through second movers stopping reciprocating. This pattern is quite consistent with «conventional first-order» reputation building (Kreps et al., 1982). In consequence, participants realized on average 55 percent (13.8 tokens) of the surplus from cooperation throughout the initial ten periods, while efficiency sharply decreased to only about 33 percent (average surplus realized 8.3 tokens) and 17 percent (average surplus realized 4.2 tokens) in the penultimate and ultimate period, respectively.[16] The distribution of payoffs between first and second movers was quite stable with first movers reaping on average 44 percent of the joint payoff (or realized surplus), minimally 39 percent (in rounds 1 and 3) and maximally 49 percent (in rounds 6, 7, and 9). Most importantly, there is no round in which the average first mover made losses by cooperating: the lowest average payoffs were 15.8 and 15.4 tokens in the penultimate and ultimate periods, respectively, which is still slightly above the outside option payoff of 15 tokens.

The payoff efficiency and distribution dynamics are quite different in the costly monitoring condition. As evident from the right-hand column of figure 5.9, in the initial rounds the frequency of mutual cooperation is notably higher in the costly than in the costless monitoring condition, while dropping to equally low levels towards the end. However, this is evidently not entirely passed through to efficiency: average joint payoffs are higher in the costly monitoring condition in every single round, and drop somewhat less sharply towards the end. The reason is that in the costly monitoring condition a larger fraction of the diminishing mutual cooperation rates towards the terminal period is caused by second movers defecting, while in the costless monitoring condition it is predominantly caused by first movers' choice of the outside option. Since exchange efficiency is determined entirely by the first movers' action, less surplus gets lost in the costly monitoring condition. Specifically, participants realized on average 75 percent (78 percent, or 19.4 tokens, throughout the

[16] The surplus was 25 tokens per exchange, the difference between the joint payoff from mutual cooperation of 55 tokens and the outside option payoff of 30 tokens. Note that the efficiency of exchange is determined entirely by the first movers' action, the second movers' actions have (statically) only distributional consequences. But of course, reciprocation has indirect (dynamic) consequences for efficiency through its repercussions on future first mover behavior.

Figure 5.9: Relative frequency of mutual cooperation and average payoffs over time.



initial ten periods, 72 percent, or 17.9 tokens, and 50 percent, or 12.5 tokens, in the penultimate and ultimate period, respectively) of the surplus from cooperation in the costly monitoring condition, compared to only 50 percent in the costless monitoring condition.

However, as one might expect the above patterns leave their footprint on the *distribution* of payoffs between first and second movers. While the distribution was quite stable in the costless monitoring condition, it exhibits an interesting pattern in the costly monitoring condition. In the first eight periods, the average first mover reaps with 49 percent of the joint payoff (or realized surplus) a somewhat larger share in the costly than in the costless monitoring condition (44 percent), minimally 45 percent (in round 6) and maximally 53 percent (in round 8). In the remaining four periods the distribution strongly shifted towards second movers: in those periods, the average first mover received only 36 percent of the joint payoff (or realized surplus), minimally 27 percent (in round 11) and maximally 40 percent (in round 10). Interestingly, however, averaged over time those distributional differences between the conditions virtually cancel out, first movers receiving 45 percent of the joint payoff (or realized surplus) in the costly monitoring condition, and 44 percent in the costless monitoring condition. As a result, both first movers (21.9 vs 18.9 tokens) and second movers (26.9 vs 23.7 tokens) were on average better off in the costly compared to the costless monitoring condition.

Result 5.9. *Average exchange efficiency (including monitoring costs) is higher in in the costly monitoring condition than in the costless monitoring condition in every round, and declines less sharply towards the terminal period. The payoff distribution is much less stable in the costly monitoring condition, exhibiting a strong shift from first towards second movers towards the end.*

This suggest that strategic second movers are at least partially successful in reaping larger shares of payoffs towards the end, but while they invested in a reputation that merits of being trusted blindly, a much more cooperative trajectory than the one in the costless monitoring condition is entered.

Note, however, that strategic reputation building works only if there are some second movers (or first movers believe there are some second movers) in the population which are committed to cooperate (Kreps et al., 1982; Kreps & Wilson, 1982; Milgrom & Roberts, 1982; Wilson, 1985). We know from a large number of experiments that such types exist (see Fehr & Schmidt, 2006, for an overview). This justifies the first movers' try to cooperate, perhaps blindly. However, in doing so first movers who trust blindly nonetheless go for a significant risk of being betrayed (repeatedly). It is therefore not surprising that post-experimental surveys also suggest that it is especially the relatively risk tolerant and less betrayal averse first movers who trust blindly. Blind trust in our experiment is positively correlated with an experimentally validated survey measure of individual risk preference (Kendall's $\tau_b = .259$, $p < .001$),[17] and it is negatively correlated with a measure of negatively reciprocal in-

[17]The item contains the question «Are you generally willing to take risks, or do you try to avoid

clination (Kendall's $\tau_b = -.223$, $p = .003$), that has been argued to be a good proxy for betrayal aversion.[18] Thus, in addition to beliefs, social and risk preferences also appear to play a role in accounting for blind trust.[19]

Result 5.10. *Blind trust is negatively correlated with measures of risk aversion and betrayal aversion.*

This suggests also a possible explanation for the existence of blind trust in the first period of the costly monitoring condition. Those three subjects who trusted blindly in the first period are on average more confident (elicited belief about reciprocation 0.587 vs. 0.481), more risk tolerant (risk tolerance item score 6.67 vs. 4.84) and less betrayal averse (negative reciprocity item score 4.67 vs. 7.60) than all other subjects in the sample;[20] for them saving five tokens of information fee may already be enough compensation for bearing the risk of being exploited.

As a final step, 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 5.10 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. The majority of first movers start with a rather pessimistic prior regarding reciprocation (see in particular pairs 1–6, 11, and 12). 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

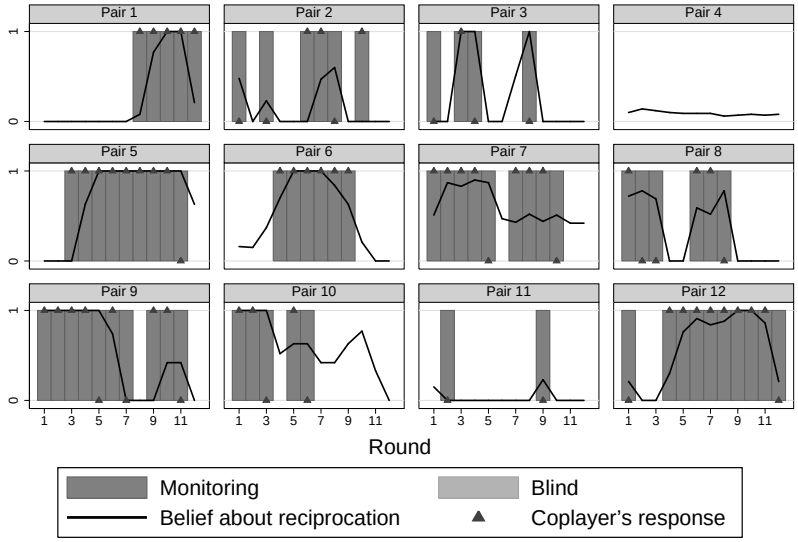
risks?», and respondents answer the question on a 11-point Likert scale ranging from 0 (very risk averse) to 10 (very risk seeking). The item is used in the German Socio-Economic Panel (SOEP) and has been shown to be good predictor of behavior in experiments with decisions under risk (Dohmen et al., 2011).

[18] 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. Fehr (2009a) argues that this measure is a good proxy for betrayal aversion.

[19] It is also positively correlated with the first generalized trust survey item used in the SOEP («In general, one can trust people», $\tau_b = .301$, $p < .001$), and negatively with the second («Nowadays, you can't rely on anybody», $\tau_b = -.268$, $p < .001$), and the third («When dealing with strangers, it is better to be cautious before trusting them», $\tau_b = -.342$, $p = .000$). Furthermore, blind trust is negatively correlated with a trust-vs-monitoring item created by us («Trust is good, control is better», $\tau_b = -.326$, $p = .000$).

[20] Those differences are of course not significant since there are only three 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 ($p = .066$), a test on the difference in risk aversion yields a $p = .161$.

Figure 5.10: Individual first mover dynamics in the costless monitoring condition.

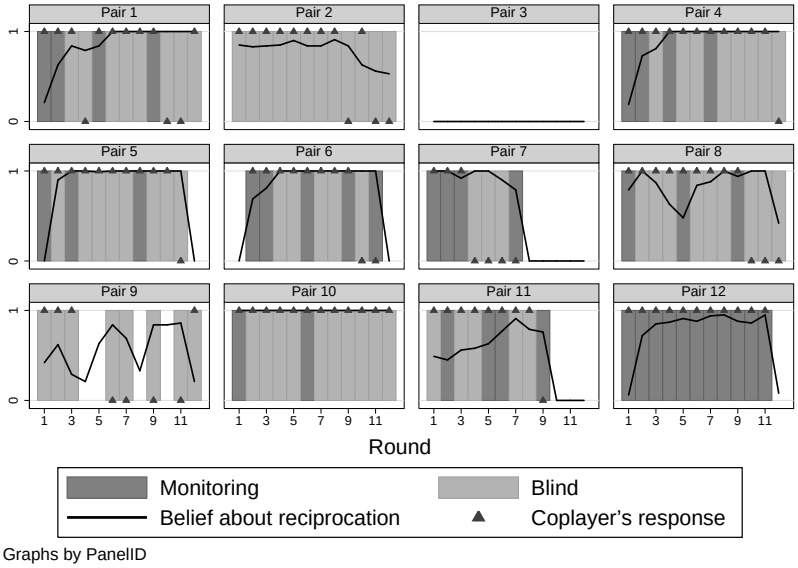


coplayer at least once over the course of interaction. Apparently, those whose cooperation is not exploited swiftly become more confident. The pessimistic testers and initially rather optimistic first movers (see pairs 7–10) who were disappointed became more, or remained, pessimistic and sometimes punished detected cheats (all cheats were detected) in non-terminal periods with at least one period of non-cooperation.[21] Note that virtually 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 audits in turn and responding with defection if the audit revealed a cheat (see pairs 1, 4–8, 11 and 12 for those patterns). Thus, at least 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.

[21] But note that punishment phases were always limited, that is, all of those who are initially cheated give their coplayers another chance, typically after one or a few punishment periods. For some this second chance works out well (see for example pair 12) for others not (see for example pair 3 or 11).

Figure 5.11: Individual first mover dynamics in the costly monitoring condition.



But again, there is apparently some individual heterogeneity as well. As in the costless monitoring condition, there is one first mover (pair 3) who refused to learn anything. 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). 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 preference-driven heterogeneity in the propensity to trust blindly. The results further above suggest that risk preferences and the degree of betrayal aversion are important here.

5.5 Conclusion

Some two decades ago, Hal Varian (1990, p. 153) commented that the 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.»

Indeed, the emergence of blind trust that we demonstrate in this chapter may be viewed as one of those phenomena. We showed this phenomenon in an experiment

in which subjects play finite horizon modified trust games in which cooperating first movers do not automatically learn their payoff at the end of a period, but may actively acquire information about the second mover's action in that period.

We find the following. First and foremost, the introduction of information costs resulted in a decrease of monitoring but an emergence of blind trust as a new behavior type. The latter effect was quantitatively so strong that (i) blind trust turned out to be the dominant behavior under costly monitoring, and (ii) it overcompensated the decrease in cooperation with monitoring such that first mover cooperation was significantly *more* frequent and payoffs higher under costly monitoring than under costless monitoring. Furthermore, neither did the average first mover in the costly monitoring condition worse than in the costless monitoring condition, nor did the average blind trustor worse than the average monitor or defector. In sum, (i) blind trust is an empirically relevant phenomenon, and (ii) it is *caused* by monitoring costs, and (iii) seems to be an a successful adaption to a setup in which monitoring is costly. This result is remarkable as it contrasts with the intuition that increasing monitoring costs hamper cooperation and diminish payoffs.

We find that this static differences between conditions stem from some important differences in the respective dynamic patterns. As a benchmark it is found that if monitoring is costless, first movers never cooperate without monitoring, and average first mover cooperation follows the typical pattern of finite horizon cooperation games with perfect monitoring. If monitoring is costly, monitoring gets less frequent and blind trust gets more frequent over time, that is, there is a dynamic shift from monitoring towards blind trust. In particular, while blind trust is the exception initially, it strongly increases to become the dominant behavior throughout the first four periods. The implication is that much of blind trust is a dynamic phenomenon.

In fact, we showed that in the initial rounds of the costly monitoring condition (i) reciprocation is markedly more frequent and (ii) first movers get much more confident compared to the costless monitoring condition. Associated with this increase in confidence is a marked increase in blind trust. In addition to beliefs, social and risk preferences also appear to play a role in accounting for blind trust, as we find blind trustors to be less risk and betrayal averse than other players.

While the average second mover anticipates the dynamic monitoring pattern well, and responds to a lower perceived likelihood of monitoring with more cheating, we argued that this alone cannot account for the remarkably high reciprocation rates at the beginning of the costly monitoring condition. Our preferred interpretation is a «second-order reputation building» hypothesis, according to which some second movers try to strategically exploit the costliness of monitoring by investing in a sufficiently favorable reputation in the initial periods in which they are likely to be monitored in order induce blind trust and reap larger gains from exploitation in later periods. We provided further results that support this interpretation. In particular, in later periods of the costly monitoring condition, there was a shift in the payoff distribution from first to second movers. However, high investments in reputation of strategically acting second movers and the existence of intrinsically trustworthy

responders shifted interactions in the costly monitoring condition on a sufficiently cooperative trajectory, that on average second *and* first movers were better off.

Various avenues fur further research suggest themselves. First and foremost, our suggested interpretation can be subjected to a more direct experimental test. Second, in addition to the *cost* of monitoring, one may systematically vary the instrumental *benefits* of information acquisition. We do so in in chapter 6 by shortening the length of interaction horizon. Another interesting question is how humans cope with the option to monitor at a cost in a indefinite horizon setup, similar to the ones assumed in the Folk Theorem literature. We leave this for further research.

Chapter 6

Costly Monitoring, Blind Trust, and the Length of the Horizon

6.1 Introduction

The realization of mutual gains from cooperation is jeopardized by opportunism in a variety of economic interactions. It is well known that the «shadow of the future», can have, under appropriate conditions, a disciplining function in those situations (Rubinstein, 1979; Fudenberg & Maskin, 1986; Kreps et al., 1982), and the experimental evidence is clearly supporting (e.g. Camerer & Weigelt, 1988; Andreoni & Miller, 1993; Dal Bó, 2005; Gächter & Falk, 2002; Anderhub et al., 2002; Engle-Warnick & Slonim, 2004). One key component of those «appropriate conditions» is the perfectness of monitoring, that is, cheating is immediately and effortless detected. In practice, however, this is routinely not the case. It is therefore an important question how monitoring imperfections impact on relational enforcement.

There is an increasing number of theoretical (e.g. Fudenberg et al., 1994; Kandori, 2002) and experimental (e.g. Holcomb & Nelson, 1997; Feinberg & Synder, 2002; Sell & Wilson, 1991; Cason & Khan, 1999) studies that introduce «noise» into the system, that is, monitoring is subject to error. These papers have not only introduced an important practical feature of relational enforcement, but have also delivered important insights into the role of different information structures in supporting, or weakening, mutual cooperation in ongoing relationships. However, the entirely passive characterization of monitoring is somewhat at odds with the more *active* nature of monitoring in practice: Information on coplayers' actions is routinely not only imperfect, but players are frequently in a position to augment available information through costly effort. Efforts to overcome imperfect information on coplayers' actions is well known in a variety of further economically relevant contexts as well, such as shared resource management (Ostrom, 1990; Ostrom & Gardner, 1993; Seabright, 1993), production teams (Alchian & Demsetz, 1972; Kandel & Lazear, 1992; Dong & Dow, 1993; Craig & Pencavel, 1995) and alliances (Acheson, 1975,

1987, 1988; Palmer, 1991), labor relations (Shapiro & Stiglitz, 1984; Kanemoto & MacLeod, 1991; Lazear, 1993), micro-finance (Armendáriz & Morduch, 2005) or neighborhood watch (Sampson et al., 1997), to name just a few.

In chapter 5 we account for this fact, and thereby extended the experimental literature in a new direction by allowing important parts of the information structure of the game to be *endogenously* determined by the players. Specifically, we reported upon an experiment in which subjects play finite horizon modified trust games with the novel feature 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. We exogenously varied the cost of information acquisition, and found that the introduction of information costs results in less monitoring and an emergence of *blind trust* as a new behavior type, where (i) blind trust is the dominant behavior under costly monitoring, (ii) first mover cooperation is *more* frequent, and (iii) payoffs are *higher* if information is costly. Furthermore, neither did the average first mover in the costly monitoring condition worse than in the costless monitoring condition, nor did the average blind trustor worse than the average monitor or defector.

In the present chapter, we extend on this paradigm. Standard game theoretic arguments suggest that (under perfect monitoring) the length of the game horizon plays a role in the decision to cooperate, and the experimental evidence is consistent with this. For example, in an experiment using a stage game very similar to ours, Anderhub et al. (2002) varied (within subjects) the length of matches between two and ten rounds, and found a tendency for less cooperation in matches with a shorter horizon compared to matches with a longer horizon. By very similar arguments one may expect that the length of the horizon also plays a role in the decision to monitor one's interaction partner. The intuition is that the shorter the horizon, the smaller the instrumental benefit from information acquisition. At the very extreme, if an interaction definitely terminates in the current period, there is no instrumental benefit at all, because there will be no opportunity to act on the information.

In this vain, we extend on chapter 5, in which the key exercise was an exogenous variation of the *cost* of information acquisition, by (i) replicating the key results qualitatively for matches with a considerably shorter horizon, and (ii) exogenously vary the instrumental *benefit* of information acquisition through a change in the length of the matches. Specifically, we supplement the data set of chapter 5 by a new experimental condition in which the experimental game is manipulated along two dimensions for a two-by-two factorial between-subjects design. The first factor is the size of the monitoring fee, which is either zero or five tokens. The second factor is the matching protocol: in the long horizon condition, subjects are randomly matched before the first period and then stay together for the entire 12 periods. In the short horizon condition, subjects are also randomly matched before the first period but are then re-matched after every two periods, such that the length of a match is 2 periods in the short horizon condition.

The questions we ask are to what extent our previous results are robust to the

length of the match horizon, and what are the eventual differences. Specifically, we (i) investigate the effect of introducing an information fee given that the horizon is short, and study given that information is (ii) costless or (iii) costly, respectively, the changes when the horizon is shortened from twelve to just two periods. The key results are the following: First, the key results of chapter 5 are qualitatively replicated for matches with a considerably shorter horizon: (i) the introduction of monitoring costs result in a decrease in the frequency of monitoring and an increase in the frequency of blind trust. (ii) The latter effect compensated for the former such that first mover cooperation was no less frequent under costly monitoring than under costless monitoring. (iii) In addition, neither did the average first mover in the costly monitoring condition worse than in the costless monitoring condition, nor did the average blind trustor worse than the average monitor or defector. This reinforces the findings that blind trust is a robust empirical phenomenon, and that it is caused by the costliness of monitoring. Still, in short horizon matches blind trust remained clearly the exception, which is the key difference to the long horizon condition.

Second, the average first mover's behavior is independent from the length of the horizon when monitoring is costless, while blind trust and overall first mover cooperation is significantly more frequent in long horizon than short horizon matches when monitoring was costly. Furthermore, while in long horizon matches first mover cooperation is significantly more frequent under costly than under costless monitoring, this difference vanishes in short horizon matches. In other words, while information costs *cause* the emergence of blind trust independently from the length of the match, the latter, as a proxy for the instrumental benefits of monitoring, significantly affects the *frequency* of blind trust. In sum, consistent with standard economic reasoning, a *ceteris paribus* increase of the cost of information acquisition decreases monitoring and increases blind trust, while a *ceteris paribus* increase in the instrumental benefit of information acquisition increases monitoring and decreases blind trust. In consequence, we found that efficiency (and the average first mover's share of it) is *ceteris paribus* higher in long horizon than in short horizon matches, and *ceteris paribus* not lower under costly than under costless monitoring. This is the main result of this chapter. We underpin this result by an investigation of the behavioral dynamics and a possible interpretation.

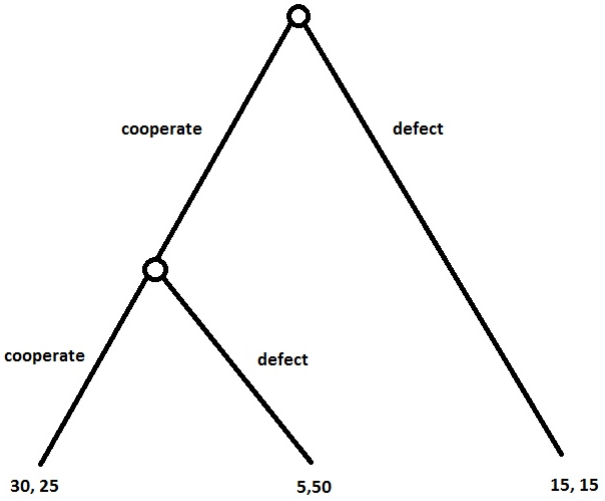
In the remainder we proceed as follows. In section 6.2 we describe the design of the experiment and report on procedures and implementation. The results are presented in section 6.3. We summarize and conclude in section 6.4.

6.2 Experiment

6.2.1 Design

We used the conventional (binary) trust game (Camerer & Weigelt, 1988; Kreps, 1990), as depicted in figure 6.1, as our point of departure. The first mover chooses between cooperation (option «pink» in the instructions) or an outside option («yel-

Figure 6.1: Experimental «stage game». The complete game is the twelve-fold sequence with monitoring decisions in between.



low»). 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 coplayer 30 tokens. Otherwise, he exploits the first mover by taking 50 tokens for himself while his coplayer gets 5 tokens. Every subject plays the game for 12 periods with one randomly matched coplayer.

The novel feature in our experiment 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 coplayer 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 coplayer monitored them or not.[1]

In order to learn more about the belief dynamics, we supplemented the experimental game by (non-incentivized) elicitation of the participants first-order beliefs about their current coplayer's behavior in a given period. In each period, before any decisions were made, first movers were asked to state their belief about whether their coplayer will respond with «brown» or «blue» to «pink», and second movers were asked to state their belief whether their coplayer will play «pink» or «yellow». Given

[1]This is an empirically accurate representation of many, but not all monitoring activities. If the second mover gets to know whether he or she is monitored, he or she may respond to this information in the subsequent periods (if any). We study this setup and evaluate a «crowding» hypothesis in a different paper.

that «pink» was played, second movers were asked after their decision to state their belief that their decision will be monitored.

The experimental game was manipulated along two dimensions for a two-by-two factorial between-subjects design. The first factor was the size of the monitoring fee, which was either zero or five tokens. The second factor was the matching protocol: in the long horizon condition, subjects were randomly matched before the first period and then stayed together for the entire 12 periods. In the short horizon condition, subjects were also randomly matched before the first period but then were re-matched after every two periods, that is, before periods 3, 5, 7, 9 and 11. The rematching was done in a way that each pair of subjects met no more than once. In sum, while the length of a match was 12 periods in the long horizon condition, it was 2 periods in the short horizon condition, while still having each subject play 12 periods in total.[2] Except for those variations, all treatment conditions were exactly identical.

6.2.2 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 96 subjects participated of which 45.8 percent have been female and 85.4 percent German citizen. The mean age was 23.1 years. Subjects were randomly assigned to treatment conditions, 24 subjects in each cell. 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 in spring 2012. 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.[3] 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

[2] We employed this matching procedure to assure that every subject played exactly 12 rounds, independently from the treatment condition. Note that the match length in the two conditions are the extreme points of the range of «meaningful» horizons; a length of one round only is outside this range because in this case it is impossible for a first mover to directly respond to the information acquired at the end of the round.

[3] They were instructed to take this sheet with them after the experiment to ensure that nobody, including the experimenters, could observe their eventual notes.

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, payed 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 45 minutes and subjects earned €10.56 (€0.03 per token earned) on average, including the fixed show-up fee, with a minimum of €5.85 and a maximum of €14.70. Earnings exceed the local average hourly wage of a typical student job and can hence be considered meaningful to the participants.

6.3 Results

As a first step we replicate the key results of chapter 5 qualitatively for matches with a considerably shorter horizon. That is, in section 6.3.1 we focus on the short horizon condition and study the differences between the costless and the costly monitoring condition. This reinforces the findings that blind trust is a robust empirical phenomenon, and that it is caused by the costliness of monitoring.

The second and core step of this chapter is a quantitative comparison of the long and the short horizon conditions. In section 6.3.2 we fix monitoring costs at zero and compare behavior in the short horizon condition to behavior in the long horizon condition. In section 6.3.3 we do the same with monitoring costs fixed at five tokens. In sections 6.3.4 and 6.3.5, respectively, we investigate the behavioral dynamics and a possible interpretation.

6.3.1 Costly monitoring in the short horizon condition

The key effects of introducing a monitoring fee are illustrated in figures 6.4 and 6.5. The main result of chapter 5, the emergence of blind trust as monitoring becomes costly, is qualitatively replicated for a considerably shorter horizon of just two periods. As evident from the bottom panel of figure 6.4, cooperation accompanied by monitoring decreases by roughly 24 percent (from 34.7 percent to 26.4 percent) as one moves from the costless to the costly monitoring condition. Following the previous chapter in basing this and the following statistical tests on a cross-section in

Figure 6.2: Relative frequency of first mover cooperation by monitoring fee condition averaged over time.

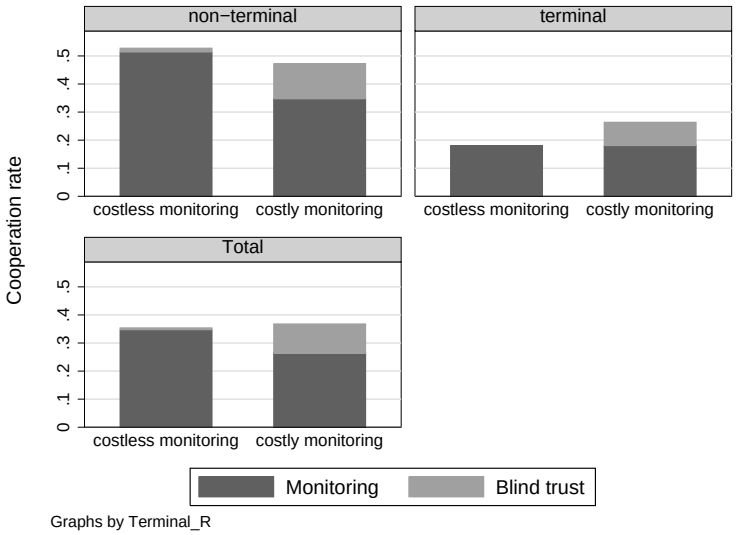
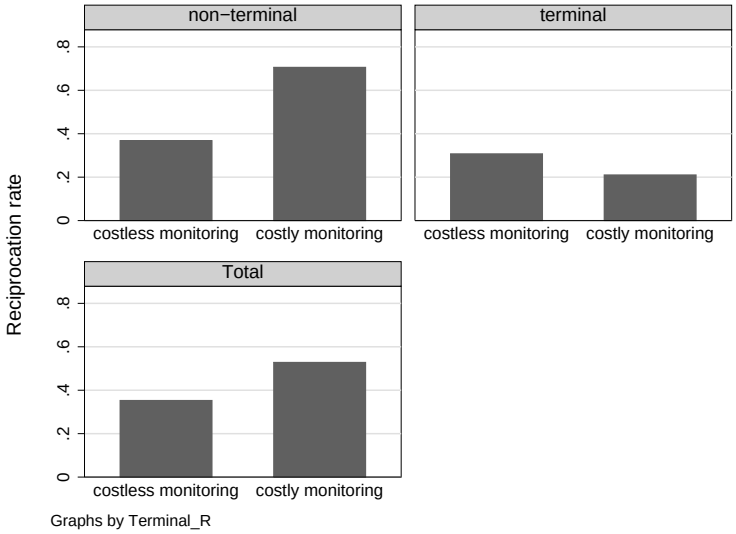


Figure 6.3: Relative frequency of reciprocation by monitoring fee condition averaged over time.



which each observation is an individual average taken over all 12 rounds, this difference is not statistically significant, however (Mann-Whitney rank sum test, $p = .337$). At the same time blind trust is more frequent if monitoring is costly than when it is costless, a difference that is statistically significant (Mann-Whitney, $p = .009$) and quantitatively sufficiently strong such that first mover cooperation rates are not lower in the costly monitoring condition (Mann-Whitney, $p = .954$). First movers trusted blindly in 10.4 percent (15 out of 144) of the cases, resulting in a total cooperation rate of 36.8 percent (53 out of 144), which is not lower than in the costless monitoring condition (35.4 percent, 51 out of 144). Still, blind trust is clearly the exception, monitoring being the dominant behavior in both the costless and the costly monitoring condition (this is a key difference to the long horizon condition, as shown below).

Result 6.1. *There is less monitoring if it is costly then when it is costless. But there is a behavior type «blind trust», cooperating without monitoring, that is more frequent when monitoring is costly than when it is costless. Taken together, first mover cooperation is no less frequent when monitoring is costly than when it is costless.*

Likewise, as in chapter 5, the reciprocation rate is *higher* in the costly monitoring condition, although the difference is not statistically significant (Mann-Whitney, $p = .209$). As a result, first movers earned on average no less in the costly monitoring condition (14.9 tokens) than in the costless monitoring condition (14.6 tokens), even including monitoring costs. The average second mover earned slightly less in the the costly monitoring condition (23.0 tokens) than in the costless monitoring condition (24.3 tokens), the difference being not statistically significant (Mann-Whitney, $p = .663$), such that both average players were, not notably worse off through the introduction of monitoring costs.

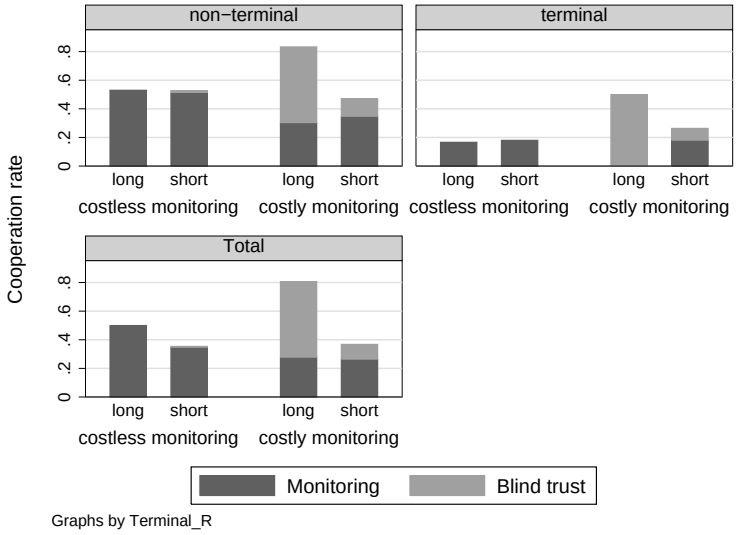
Of course, more appropriate to judge whether blind trust does or does not pay is to compare the payoff of blind trustors to the other behavioral alternatives.[4] In the costly monitoring condition, blind trustors clearly did best on average (18.3 tokens) compared to defectors (15 tokens) and monitors (13.2 tokens). Thus, as in Goeschl & Jarke (2012) we can conclude that, on average, blind trust was *not* unprofitable.

Result 6.2. *The average first mover in the costly monitoring condition did not do worse than in the costless monitoring condition. Likewise, the average blind trustor did not do worse than the average monitor or defector.*

Summing up so far, the key *qualitative* results reported in chapter 5 are upheld even in matches with a significantly shorter horizon: First, blind trust is an empirically relevant option even in matches with only a two-period horizon, and monitoring costs is the *cause* of its preponderance. Second, blind trust seems to be a successful adaption to a setup in which monitoring is costly, since neither did the average first mover in the costly monitoring condition worse than in the costless monitoring

[4]Note that we can do this because second movers were not informed about the monitoring decision. 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.

Figure 6.4: Relative frequency of first mover cooperation by condition averaged over time.



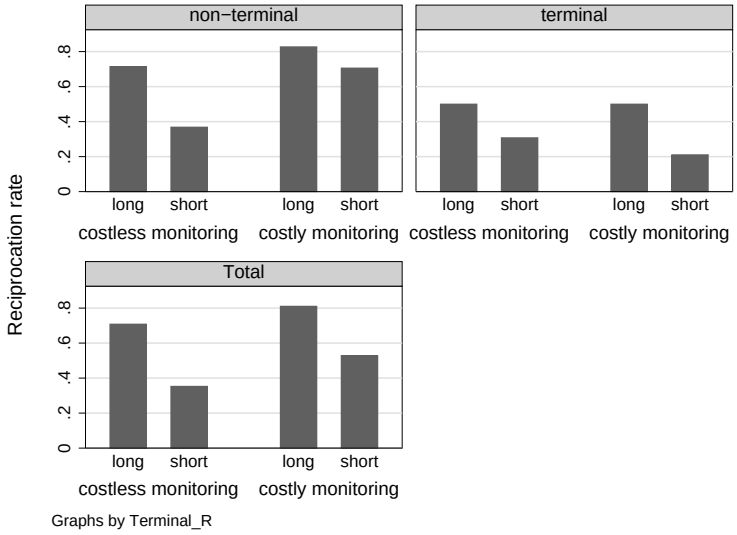
condition, nor did the average blind trustor worse than the average monitor or defector. However, there is also an important *quantitative* difference: While blind trust is the dominant behavior under costly information in the long horizon condition, it is much less frequent in the short horizon condition. In the remainder of this section, we investigate this difference more thoroughly.

6.3.2 Comparison of short and long horizon condition under costless monitoring

We now build on the results of the long horizon condition from chapter 5 and investigate the effects of shortening the length of the matches. The key results are illustrated in figures 6.4 and 6.5, which extend figures 6.2 and 6.3, respectively, by the long horizon condition.

We begin by fixing the information fee at zero. Comparing aggregate first mover cooperation rates across conditions, as depicted in the lower panel of figure 6.4, cooperation is less frequent in the short horizon that in the long horizon condition. But this comparison is not meaningful since there are more terminal periods in the former than in the latter. Thus, we need to compare non-terminal and terminal periods in isolation. In doing so, it is evident from the top panel of figure 6.4 that under costless monitoring first mover cooperation rates are approximately equal in the long and short horizon conditions, respectively. In non-terminal periods, first movers cooperated in 53 percent of the time (70 out of 132) in long horizon matches and in 52.8 percent percent of the time (38 out of 72) in short horizon matches. In non-terminal periods, first movers cooperated in 16.7 percent of the time (2 out of 12) in long horizon matches and in 18.1 percent percent of the time (13 out of 72) in short horizon

Figure 6.5: Relative frequency of reciprocation by condition averaged over time.



matches.

In addition, it is evident that blind trust is not an issue when monitoring is costless, independently from the length of the horizon. In the long horizon condition, first mover cooperation was never blind, while there was one instance of blind trust (out of 69 instances of cooperation) in the short horizon condition, which is in the range of performance error.[5]

Result 6.3. *If monitoring is costless, the average first mover's behavior is independent from the length of the horizon.*

6.3.3 Comparison of short and long horizon condition under costly monitoring

Things are quite different if monitoring is costly. As already shown in detail in chapter 5, blind trust emerges as a new behavior type if monitoring is costly, and in long horizon matches this increase is strong enough such that (i) blind trust becomes the dominant behavior and (ii) first mover cooperation is *more* frequent under costly monitoring despite cooperation accompanied by monitoring being less frequent. As evident from figure 6.4, these results are markedly mitigated but not eliminated by shortening the match horizon to two periods only.

Comparing again non-terminal and terminal periods in isolation, it is evident from the top panel of figure 6.4 that (with costly information) cooperation together with monitoring is slightly higher in the short horizon than in the long horizon condition,

[5]This instance occurred in the first round, and the same subject cooperated another two times in the second and the fifth round, these times together with monitoring, however.

both in non-terminal (34.7 vs 30.3 percent) and terminal rounds (18.0 vs zero percent). Continuing with basing our statistical tests on a cross-section of individual averages, only the latter difference is statistically significant (Mann-Whitney, $p = .015$). Blind trust, on the other hand, is clearly more frequent in the long horizon condition, both in non-terminal (53.0 vs 12.5 percent) and terminal rounds (50.0 vs 8.3 percent), with only the former difference being statistically significant (Mann-Whitney, $p = .002$). As a result, under costly monitoring blind trust is an exception in short horizon matches while it is the dominant behavior type in long horizon matches. Thus, while information costs *cause* the emergence of blind trust independently from the length of the match, the latter, as a proxy for the instrumental benefits of monitoring, significantly affects the *frequency* of blind trust. In brief: consistent with standard economic reasoning, a *ceteris paribus* increase of the cost of information acquisition decreases monitoring and increases blind trust, while a *ceteris paribus* increase in the instrumental benefit of information acquisition increases monitoring and decreases blind trust. This is the main result of this chapter.

Putting blind trustors and monitors together, figure 6.4 illustrates that first mover cooperation rates were clearly higher in the long horizon condition, both in non-terminal (83.3 vs 47.2 percent) and terminal rounds (50.0 vs 26.4 percent), again with only the former difference being (marginally) statistically significant (Mann-Whitney, $p = .067$).

Result 6.4. *If monitoring is costly, blind trust and overall first mover cooperation is significantly more frequent in long horizon than short horizon matches. While in long horizon matches first mover cooperation is significantly more frequent under costly than under costless monitoring, this difference vanishes in short horizon matches.*

A reason for this result is indicated by an investigation of the behavioral dynamics. In chapter 5 we showed at length that blind trust in long horizon matches is predominantly a dynamics phenomenon: If monitoring is costly, monitoring gets less frequent and blind trust gets more frequent towards the terminal round, that is, there is a shift from monitoring towards blind trust *over time*. Specifically, in non-terminal rounds of the long horizon condition with costly monitoring, first movers cooperated together with monitoring their coplayer in 30.3 percent (40 out of 132) of the time, while they never monitored their coplayer (0 out of 12) in terminal rounds. In the first period, the majority of cooperating first movers monitor, blind trustors are a minority. However, this changes over time, as blind trust increases notably in the initial four periods and then remaining approximately stable, while the frequency of monitoring successively decreases over time. Thus, blind trust is clearly the exception initially and needs some time to develop, about three to four periods.

This suggests why blind trust has a harder time to develop in the short horizon condition, because there are just two periods, that is, only one period for a first mover to «test» her or his coplayer. At the same time there are much weaker incentives for second movers to build a good reputation.

6.3.4 Strategic reputation building and the dynamics of blind trust in long matches

To investigate this further, we first review the key dynamics in the long horizon condition (see chapter 5 for a more extensive treatment). We begin by taking a closer look at the dynamics of second mover behavior. The upper panel of figure 6.5 depicts aggregate second mover behavior splitted up by terminal and non-terminal periods. Consistent with «standard» strategic reputation building, cooperation rates in the long horizon condition are, both under costless and costly monitoring, clearly higher in non-terminal rounds than in terminal rounds. When information was costless, second movers reciprocated in 71.4 percent (50 out of 70) of the time in non-terminal rounds, and in 50.0 percent (1 out of 2) of the time in terminal rounds.

In chapter 5 we conjectured that second movers have a stronger incentive to build a good reputation if monitoring is costly, because it may induce the first mover not only to cooperate but perhaps also to abstain from monitoring, that is, to trust blindly. We called this «second-order reputation building».[6] Consistent with this, reciprocation rates in non-terminal periods were with 82.7 percent (91 out of 110) indeed higher under costly monitoring than under costless monitoring (reciprocation rates in terminal periods were also at 50.0 percent, 3 out of 6). In fact, while reciprocation rates were similar in both treatment conditions towards the terminal period, over the initial rounds they are clearly higher in the costly monitoring condition, at remarkable 100 percent in period 1 through 3, than in the costless monitoring condition.

In addition, we showed that the within-treatment dynamics are quite consistent with the average second mover's belief of being monitored in the current period, but also that a simple response to this belief cannot account for the *differences* between treatments. In the long horizon condition the average second mover anticipated the average first mover's monitoring pattern (as reported above) remarkably closely: While the average second mover believes to be monitored with almost-certainty throughout the whole match in the costless monitoring condition, (s)he expects to be monitored predominantly at the beginning of a match, less so towards the end, under costly monitoring. In turn, the belief of being monitored and the likelihood of return cooperation went together (Kendall's $\tau_b = 0.168$, $p = .040$), that is, the average

[6] 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 our 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 our game there is the possibility of cheating over *multiple* periods once the first mover trusts blindly. Intuitively, this may lead (some) second movers 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. Strategically acting subjects may do so in order to exploit their coplayers more easily and perhaps over multiple periods. This strategic incentive is missing in the costless monitoring condition, since second movers do not believe in blind trust anyway (and this belief is justified).

cooperating second mover had a stronger belief of being monitored (.945 in the costless monitoring condition, .652 in the costly monitoring condition) than the average defecting second mover (.880 in the costless monitoring condition, .471 in the costly monitoring condition). As a result, reciprocation rates were rather stable in the costless monitoring condition, and markedly decreasing in time in the costly monitoring condition. However, since the average second mover's belief of being monitored was generally stronger in the costless monitoring condition, a simple response to this this belief cannot account for the treatment differences.

As a possible explanation of those differences, we advanced the «second-order reputation building» hypothesis. Support for this conjecture comes from the fact that the strategy is in some sense quite successful. First, rank correlation between first movers' belief about reciprocation in period $t+1$ and an indicator variable that is unity if a respective first mover detected her or his coplayer to cooperate in period t is strongly positive (Kendall's $\tau_b = .555$, $p = .000$), that is, the average first mover became much more confident (pessimistic) in $t+1$ if (s)he observed her or his coplayer reciprocating (defecting) in t. As a result, while the average first mover's *prior* about their cooperation being reciprocated is approximately equal at .4 in both conditions, there is a marked increase of confidence in the initial periods of the costly monitoring condition, which may be viewed as justified given full reciprocation, whereas the belief does not change much over time in the costless monitoring condition. This is illustrated by the circles connected with a hatched line in figure 5.6.

Second, correlation between the first movers' belief about reciprocation and their own cooperation is strongly positive and significant both in the costless monitoring condition (Kendall's $\tau_b = .618$, $p = .000$) and the costly monitoring condition (Kendall's $\tau_b = .584$, $p = .000$). Interestingly, however, separating first mover cooperation by cooperation coupled with monitoring and blind trust if information is costly, it turns out that only the latter is correlated with first movers' beliefs about reciprocation (Kendall's $\tau_b = .410$, $p = .000$), the former is not (Kendall's $\tau_b = .059$, $p = .419$). Thus, a first mover's confidence in reciprocation appears to be a key determinant of blind trust. In addition to beliefs, social and risk preferences also appear to play a role in accounting for blind trust. Post-experimental surveys suggest that it is especially the relatively risk tolerant and less betrayal averse first movers who trust blindly. Blind trust in our experiment is positively correlated with an experimentally validated survey measure of individual risk preference (Kendall's $\tau_b = .259$, $p < .001$),[7] and it is negatively correlated with a measure of negatively reciprocal inclination (Kendall's $\tau_b = -.223$, $p = .003$), that has been argued to be a good proxy for betrayal aversion.[8]

[7] 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).

[8] 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

Finally, the dynamics of realized payoffs depicted in in figure 5.9 also point toward second-order reputation building. In light of the above results it is no surprise that interactions were more efficient in the costly monitoring condition than in the costless monitoring condition. In the latter (depicted in the left-hand column), the frequency of mutual cooperation is moderately positive throughout the initial periods, but strongly decreasing towards the end, partly through first movers stopping to cooperate (foreseeing the second movers' strategic incentive to defect), partly through second movers stopping reciprocating. This pattern is quite consistent with «conventional first-order» reputation building (Kreps et al., 1982). In consequence, participants realized on average 53 percent (13.3 tokens) of the surplus from cooperation throughout the non-terminal periods, while efficiency sharply decreased to only 17 percent (average surplus realized 4.2 tokens) in the terminal period.[9] Under costly monitoring, the frequency of mutual cooperation is notably higher than in the costless monitoring condition in the initial rounds, while dropping to equally low levels towards the end. However, this is not entirely passed through to efficiency: average joint payoffs are higher in the costly monitoring condition in every single round, and drop somewhat less sharply towards the end. The reason is that in the costly monitoring condition a larger fraction of the diminishing mutual cooperation rates towards the terminal period is caused by second movers defecting, while in the costless monitoring condition it is predominantly caused by first movers' choice of the outside option. Since exchange efficiency is determined entirely by the first movers' action, less surplus gets lost in the costly monitoring condition. Specifically, participants realized on average 75 percent (77 percent, or 19.3 tokens, throughout the non-terminal periods, and 50 percent, or 12.5 tokens, in the terminal period) of the surplus from cooperation in the costly monitoring condition, compared to only 50 percent in the costless monitoring condition.

Interesting, however, are the distributional dynamics. In the costless monitoring condition, the distribution of payoffs between first and second movers was quite stable with first movers reaping on average 44 percent of the joint payoff (or realized surplus), minimally 39 percent (in rounds 1 and 3) and maximally 49 percent (in rounds 6, 7, and 9). Most importantly, there is no round in which the average first mover made losses by cooperating: the lowest average payoffs were 15.8 and 15.4 tokens in the penultimate and ultimate periods, respectively, which is still slightly above the outside option payoff of 15 tokens. In contrast, the payoff distribution is less stable in the costly monitoring condition, exhibiting a shift from first to second movers over time. In the first eight periods, the average first mover reaps with 49 per-

---

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 as measure of negatively reciprocal inclination. Fehr (2009a) argues that this measure is a good proxy for betrayal aversion.

[9] The surplus was 25 tokens per exchange, the difference between the joint payoff from mutual cooperation of 55 tokens and the outside option payoff of 30 tokens. Note that the efficiency of exchange is determined entirely by the first movers' action, the second movers' actions have (statically) only distributional consequences. But of course, reciprocation has indirect (dynamic) consequences for efficiency through its repercussions on future first mover behavior.

cent of the joint payoff (or realized surplus) a somewhat larger share in the costly than in the costless monitoring condition (44 percent), minimally 45 percent (in round 6) and maximally 53 percent (in round 8). In the remaining four periods the distribution strongly shifted towards second movers: in those periods, the average first mover received only 36 percent of the joint payoff (or realized surplus), minimally 27 percent (in round 11) and maximally 40 percent (in round 10). Interestingly, however, averaged over time those distributional differences between the conditions virtually cancel out, first movers receiving 45 percent of the joint payoff (or realized surplus) in the costly monitoring condition, and 44 percent in the costless monitoring condition. As a result, both first movers (21.9 vs 18.9 tokens) and second movers (26.9 vs 23.7 tokens) were on average better off in the costly compared to the costless monitoring condition. This suggest that strategic second movers are at least partially successful in reaping larger shares of payoffs towards the end, but while they invested in a reputation that merits of being trusted blindly, a much more cooperative trajectory than the one in the costless monitoring condition is entered.

6.3.5 Strategic reputation building and the dynamics of blind trust in short matches

Do we find similar patterns in the short horizon condition, and if so, what are the differences? Can those differences account for result 6.4?

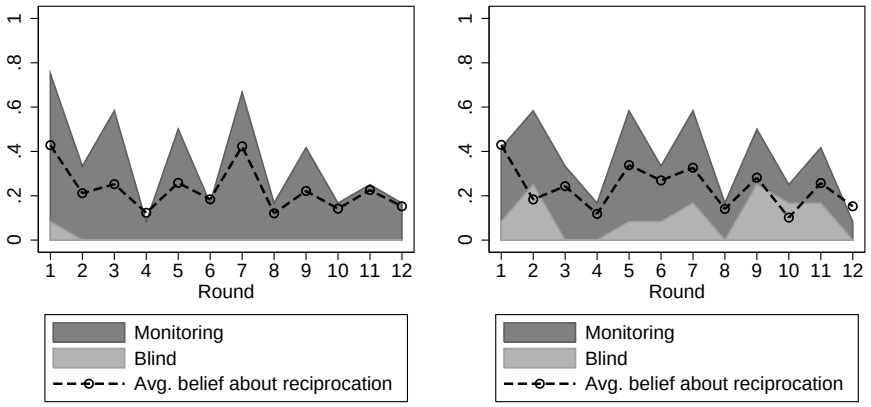
In fact, as evident from figure 6.5, comparing second mover behavior in terminal and non-terminal periods we find similar qualitative patterns in the short horizon condition. When information was costless, second movers reciprocated in 36.8 percent (14 out of 38) of the time in non-terminal rounds, and in 30.7 percent (4 out of 13) of the time in terminal rounds. Under costly monitoring, reciprocation occurred in 70.6 percent (24 out of 34) of the time in non-terminal rounds, and in 21.0 percent (4 out of 19) of the time in terminal rounds. Thus, strategic reputation building is an issue even if the match length is just two periods (the minimum), and, consistent with the second-order reputation building hypothesis, the incentive to do so appears to be much stronger in the costly monitoring condition.

However, the incentive for reputation building, both first-order and second-order, seems to be absolutely weaker in the short horizon condition, as reciprocation rates are lower under both costless and costly monitoring, respectively.

Result 6.5. *Reciprocation is more frequent in non-terminal than in terminal periods in all treatment conditions. In non-terminal periods, reciprocation is (i) more frequent under costly than under costless monitoring in both long and short horizon matches, and (ii) more frequent in long horizon than in short horizon matches under both costless and costly monitoring.*

Again, it can be shown that the within-treatment dynamics are quite consistent with the average second mover's belief of being monitored in the current period, but also that a simple response to this belief cannot account for the *differences* between treatments. In the short horizon condition the average second mover antic-

Figure 6.6: Aggregate dynamics of first mover cooperation and beliefs about reciprocation in the short horizon condition. The left column depicts the costless monitoring condition, the right column the costly monitoring condition.



ipated the average first mover's monitoring pattern (as reported above) remarkably closely: While the average second mover believes to be monitored with almost-certainty throughout the whole match in the costless monitoring condition (average belief of 0.940 in initial and 0.946 in terminal periods), (s)he expects to be monitored predominantly at the beginning of a match (average belief 0.726), less in the terminal period (average belief 0.567), under costly monitoring. This fits together with the fact that reciprocation rates were more stable over the two periods of a match in the costless monitoring condition, while there is a pronounced drop from the first to the second period in the costly monitoring condition. However, since again the average second mover's belief of being monitored is generally stronger in the costless monitoring condition, a simple response to this belief cannot account for the treatment differences.

What about second-order reputation building here? The dynamics of realized payoff distribution is again most illuminating. Interesting, however, are the distributional dynamics. In the costless monitoring condition, first movers reaped on average a larger share of the joint payoff (or realized surplus) in terminal periods (42.2 percent) than in non-terminal periods (33.8 percent).[10] In contrast, this is reversed under costly monitoring: the average first mover got a larger share of the joint payoff (or realized surplus) in non-terminal periods (42.1 percent) than in terminal periods (36.0 percent).[11]

How did first movers respond? As evident from figure 6.6, first movers were generally more confident that their cooperation will be reciprocated in the first than in the

[10]For comparison, in the long horizon condition first movers got on average 44.4 percent of the joint payoff (or realized surplus), 44.3 percent in non-terminal periods and 45.1 percent in terminal periods.

[11]For comparison, in the long horizon condition first movers got on average 44.9 percent of the joint payoff (or realized surplus), 45.4 percent in non-terminal periods and 38.2 percent in terminal periods.

second period of a match, that is, they anticipated the strategic incentive to defect at the end of the match from the very beginning. In addition, confidence and cooperation go together. Indeed, that the first movers' belief dynamics are crucial in accounting for observed behavior is suggested by the fact taht the average first mover's prior was approximately equal at 0.4 in all four treatment conditions, whereas the further path over time differs.

Result 6.6. *On average, first movers' confidence about reciprocation is never increasing in the course of a match in the short horizon condition, and generally lower than in the long horizon condition.*

This suggests a reason for the aggregate differences in the frequencies of cooperation and blind trust. First, first movers anticipate the strategic incentive to cheat in the terminal period quite well in all conditions. Second, the dynamics in the long horizon condition with costly monitoring suggests that blind trust needs some time to develop, a «testing phase» in which second movers have a very strong incentive to behave well. These two facts together suggest that a horizon of two periods is just too short for the latter dynamic to happen, since the second period is also the terminal period of a match, such that even if a first mover's cooperation is reciprocated, any eventual increase in confidence is countervailed by the anticipated strategic incentive to defect. In sum, while a long enough testing phase with high cooperation rates facilitates a shift from monitoring to blind trust over time in long horizon matches, a testing phase of just one period is apparently to short to enter the same cooperative trajectory.

It is therefore no surprise that efficiency is generally lower in short than in long horizon matches. Under costless monitoring, participants realized on average 53 percent (13.2 tokens) of the surplus from cooperation in non-terminal periods, while efficiency sharply decreased to only 18 percent (average surplus realized 4.5 tokens) in terminal periods.[12] Under costly monitoring, participants realized on average 32 percent (40 percent, or 10.1 tokens, throughout the non-terminal periods, and 23 percent, or 5.7 tokens, in the terminal period) of the surplus from cooperation in the costly monitoring condition, compared to 35 percent in the costless monitoring condition.[13] Interestingly, under costless monitoring the lower average efficiency is just due to the fact that there are more non-terminal periods in a match: average efficiency in non-terminal periods is exactly equal (53 percent), and average efficiency in terminal periods is approximately equal (17 and 18 percent, respectively) in long and short matches. Under costly monitoring, however, efficiency is in all respects higher in

[12] Mutual cooperation occurred in 19.4 percent of the cases in non-terminal periods and only 5.6 percent in terminal periods. For comparison, in the long horizon condition mutual cooperation occurred in 37.9 percent of the cases in non-terminal periods and 8.3 percent in terminal periods.

[13] The frequency of mutual cooperation is notably higher than in the costless monitoring condition in non-terminal rounds (33.3 percent), while dropping to equally low levels in terminal periods (5.6 percent). For comparison, in the long horizon condition mutual cooperation occurred in 68.9 percent of the cases in non-terminal periods and 25 percent in terminal periods.

long horizon (77 percent in non-terminal and 50 percent in terminal rounds) than in short horizon matches (40 percent in non-terminal and 23 percent in terminal rounds).

Result 6.7. *Efficiency is ceteris paribus higher in long horizon than in short horizon matches, and ceteris paribus not lower under costly than under costless monitoring. The average first mover's share of the realized surplus is also ceteris paribus higher in long horizon than in short horizon matches, and ceteris paribus not lower under costly than under costless monitoring. However, in both long and short horizon matches, the average first mover's share is higher in terminal than in non-terminal periods under costless monitoring, while reverse is true under costly monitoring.*

6.4 Conclusion

We reported upon an experiment in which subjects play finite horizon modified trust games in which cooperating first movers do not automatically learn their payoff at the end of a period, but may actively acquire information about the second mover's action in that period.

As a first step we replicated the key results of chapter 5 qualitatively for matches with a considerably shorter horizon. In matches with a length of just two periods, we studied the differences between a costless and a costly monitoring condition, and found that the introduction of monitoring costs resulted in a decrease in the frequency of monitoring and an increase in the frequency of blind trust. The latter effect compensated for the former such that first mover cooperation was no less frequent under costly monitoring than under costless monitoring. In addition, neither did the average first mover in the costly monitoring condition worse than in the costless monitoring condition, nor did the average blind trustor worse than the average monitor or defector. This reinforces the findings that blind trust is a robust empirical phenomenon, and that it is caused by the costliness of monitoring. Still, in short horizon matches blind trust remained clearly the exception, which is the key difference to the long horizon condition, a comparison that was the second and core step of this paper.

In comparing behavior in long and short horizon matches, we found that the average first mover's behavior was independent from the length of the horizon when monitoring was costless, while blind trust and overall first mover cooperation was significantly more frequent in long horizon than short horizon matches when monitoring was costly. Furthermore, while in long horizon matches first mover cooperation was significantly more frequent under costly than under costless monitoring, this difference vanished in short horizon matches. In other words, while information costs *cause* the emergence of blind trust independently from the length of the match, the latter, as a proxy for the instrumental benefits of monitoring, significantly affects the *frequency* of blind trust. In sum, consistent with standard economic reasoning, a *ceteris paribus* increase of the cost of information acquisition decreases monitoring and increases blind trust, while a *ceteris paribus* increase in the instrumental benefit of information acquisition increases monitoring and decreases blind trust. This is the main result of this paper.

We underpinned this result by an investigation of the behavioral dynamics and a possible interpretation. In chapter 5 we showed at length that blind trust in long horizon matches is predominantly a dynamic phenomenon: If monitoring is costly, monitoring gets less frequent and blind trust gets more frequent towards the terminal round, that is, there is a shift from monitoring towards blind trust *over time*: blind trust is clearly the exception initially and needs some time to develop, about three to four periods. This suggests why blind trust has a harder time to develop in the short horizon condition, because there are just two periods, that is, only one period for a first mover to «test» her or his coplayer. At the same time there are much weaker incentives for second movers to build a good reputation.

The latter is underpinned by the finding that reciprocation was more frequent in non-terminal than in terminal periods in all treatment conditions, while in non-terminal periods, reciprocation is (i) more frequent under costly than under costless monitoring in both long and short horizon matches, and (ii) more frequent in long horizon than in short horizon matches under both costless and costly monitoring. In addition, in both long and short horizon matches, the average first mover's share of the realized surplus is higher in terminal than in non-terminal periods under costless monitoring, while reverse is true under costly monitoring.

The former is supported by the dynamics of first mover behavior and beliefs. That first movers' belief dynamics are crucial in accounting for observed behavior is suggested by the fact the average first mover's prior was approximately equal at 0.4 in all four treatment conditions, whereas the further path over time differs: While the average first mover's confidence about reciprocation was strongly increasing over the initial periods, it is never increasing in the course of a match in the short horizon condition, and generally lower than in the long horizon condition.

This suggests a reason for the aggregate differences in the frequencies of cooperation and blind trust. First, first movers anticipate the strategic incentive to cheat in the terminal period quite well in all conditions. Second, the dynamics in the long horizon condition with costly monitoring suggests that blind trust needs some time to develop, a «testing phase» in which second movers have a very strong incentive to behave well. These two facts together suggest that a horizon of two periods is just too short for the latter dynamic to happen, since the second period is also the terminal period of a match, such that even if a first mover's cooperation is reciprocated, any eventual increase in confidence is countervailed by the anticipated strategic incentive to defect. In sum, while a long enough testing phase with high cooperation rates facilitates a shift from monitoring to blind trust over time in long horizon matches, a testing phase of just one period is apparently to short to enter the same cooperative trajectory.

In consequence, we found that efficiency (and the average first mover's share of it) is *ceteris paribus* higher in long horizon than in short horizon matches, and *ceteris paribus* not lower under costly than under costless monitoring.

Various avenues fur further research suggest themselves. For example, an interesting question is whether the incentive to monitor is altered by adding a costly

punishment option to the first mover's action set. Another interesting question is how humans cope with the option to monitor at a cost in a indefinite horizon setup, similar to the ones assumed in the Folk Theorem literature. We leave this for further research.

Chapter 7

Communication Technology and Generalized Trust: Evidence from the German Socio-Economic Panel

7.1 Introduction

The approach to view trust as a historically or culturally determined, short-term invariant variable has a large follwing in the social sciences (e.g. Arrow, 1971; Dore, 1987; Putnam, 1993b,a; Fukuyama, 1995), and a bulk of empirical studies has accumulated that uses some measure of trust, cooperativeness, or «social capital», within a population as an *independent* variable to predict various economic outcomes (e.g. La Porta et al., 1997; Knack & Keefer, 1997; Zak & Knack, 2001; Guiso et al., 2004, 2008). In economic theory, on the other hand, trust is entirely endogenous, completely determined by the structure of current interaction (e.g. Kreps et al., 1982; Fudenberg & Maskin, 1986; Fudenberg et al., 1994). In particular, the literature on matching games (e.g. Bendor & Mookherjee, 1990; Kandori, 1992a; Ellison, 1994a), and the related literature on indirect reciprocity in biology (see Nowak & Sigmund, 2005, for an overview), highlights the critical role of information transmission to sustain bootstrap notions of trust in a population. The central idea is that individuals condition their expectations about others' behavior, and hence their own behavior, on what they know about their interaction partners' past, their reputation. However, the literature is fairy unspecific about the exact mechanisms how such information might be transmitted; some degree of information dispersal generally enters the model as an exogenous assumption from the outset. One obvious candidate is communication technology. Distance communication technology renders it easier to acquire and spread information about others' in the relevant community. Can we expect, therefore, that people connected to such a communication network find it easier to trust

others?[1]

Empirically, very little systematic research exists along these lines. In more distantly related work, studies on feedback mechanisms embedded in online marketplaces suggest that they facilitate trust and exchange (e.g. Ba & Pavlou, 2002; Resnick & Zeckhauser, 2002; Resnick et al., 2006). There is also some research studying some fairly specific contexts, usually relying on quite limited data bases (e.g. Picot et al., 1996; Fandy, 2000; Ryssel et al., 2004). In a seminal study using a general data base, Fisman & Khanna (1999) constucted a fourty-country cross-sectional dataset from survey data taken from the 1990-93 Wold Values Survey (WVS) and the International Telecommunication Union (ITU) Yearbook 1994 to examine whether there is a relationship between telephones per capita and responses to the question «Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?», an item intended to measure the repondent's trust in a representative member of the relevant population. In fact, they find a positive relationship. However, they also point to two key limitations of their study. First, they admit that their coarse and highly aggregated data does not quite correspond to the theoretical micro-level approaches that speak to the relationship between information flows and trust. Second, they recognize the difficulties associated with the WVS survey item as a measure of trust.

The current study extends on this study, addressing both issues, and contributing to an assessment of the robustness of the relationship. First, I use a large cross-section of recent (the 2008 wave) individual level micro-data taken from the German Socio-Economic Panel (SOEP) that draws samples which are representative for the population in Germany. Second, the SOEP contains an improved, more fine-git, and experimentally validated measure of generalized trust. Third, I am able to control for a much larger set of confounding variables, comprising *inter alia* socio-demographics, education, social network inclusion, and even individual attitudes such as risk and time preferences. Fourth, I am able to draw on an experimentally validated measure of *beliefs* into others trustworthiness in order to investigate whether the effect of communication technology connectivity on trust is mediated through beliefs, as economic theory predicts. Finally, as the communication infrastructure underwent significant change in the last two decades, I extend the set of considered communication technologies to include cellular phones, internet (both narrow and broad bandwidth), and television. In sum, while the replication of the basic analysis of Fisman & Khanna (1999) is a first step of the present study, it significantly extends on it along a number of dimensions, providing for some original contributions to the literature.

I find the follwing. First, the results show a gap in measured trust between those repondents who have a landline phone in their household and those who do not,

[1]This might happen in two respects. First, connected individuals know that they are less easily exploited because they are more likely to be informed about the identity of cheaters, such that they find it easier to trust those with whom they interact. Second, connected individuals have means to spread information about cheaters to others, which may serve as a deterrent to cheat upon them in the first place, bootstrapping their trust.

confirming the results of Fisman & Khanna (1999). Second, the existence of this gap is robust to alternative estimation techniques and the specific construction of the dependent variable. Third, I find this gap also using the item intended to measure *beliefs* in others' trustworthiness more directly. Furthermore, including the variable into the original specification as a predictor, the gap vanishs, suggesting that the effect is indeed mediated by beliefs. Fourth, the gap narrows significantly as variables that capture the level of education and local network inclusion are controlled for. Since Fisman & Khanna (1999) did not control for those variables, their results might be biased upwards. Fifth, there are no signficant effects of celluar phone or narrow bandwidth interent connectivity, but there is a significantly positive effect of broad bandwidth internet connectivity, and a significantly negative effect of television posession. The order of magnitude of the latter effect is considerable, and may be an interesting input for further research.

The remaineder of the chapter is organized as follows. I describe the basic statistical model specification and the dataset in section 7.2 . The results are presented in section 7.3. I conlcude in section 7.4.

7.2 Specification and data

Generalizing on Fisman & Khanna (1999), the baseline specification is a linear model of the form

$$y = \alpha + \mathbf{x}\beta + \mathbf{z}\gamma + \varepsilon$$

where y is a survey measure of trust, $\mathbf{x}$ is a vector of dummies indicating the presence of various information technologies in the respondents' household, $\mathbf{z}$ is vector of control variables, α represents a constant, and ε the error term. The coefficient vectors β and γ are to be estimated, whereas the former is the one of focal interest in the present study.

The dataset is a cross-section constructed from the 2008 wave of the German Socio-Economic Panel (SOEP).[2] I have complete data vectors on a sample of $17,915$ individual respondents. I now describe the variables in more detail.

7.2.1 Dependent variables

The National Opinion Research Center's General Social Survey (GSS) item reads «Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?», with possible responses «most people can be trusted» and «can't be too careful». The same item is implemented in the World Values Survey (WVS). Since Fisman & Khanna (1999) conducted their study using this item, it has been subject to criticism. Most importantly, the two response categories are not mutually exclusive since respondents might consistently agree to both

[2] One item surveying membership in religious organizations were taken from the 2007 wave, because it was not included in the 2008 wave.

answers (Yamagishi et al., 2002). A possible consequence is that the interpretation of the item can differ widely among different societies (e.g. Miller & Mitamura, 2003).[3]

To address this problem, the SOEP contains an improved three-item battery (for a detailed description see Naef & Schupp, 2009b). The first two items just split the GSS/WVS question up into two parts: «On the whole one can trust people» and «Nowadays one can't rely on anyone», with four answer categories, respectively, «totally agree», «agree slightly», «disagree slightly», and «disagree totally». The third item is included to address a criticism of the GSS/WVS question that answers may signicantly depend on how people understand «most people», in particular, whether they include people they know personally or not (Reeskens & Hooghe 2008). The last item therefore explicitly asks about strangers: «If one is dealing with strangers, it is better to be careful before one can trust them», whith the same answer categories as before. I code the responses «totally agree» and «agree slightly» by 2 and 1, and the responses «disagree slightly» and «disagree totally» by -1 and -2, respectively.

Naef & Schupp (2009b) show that the above item battery has a high degree of reliability. However, economists are ultimately interested in behavior, so what about its behavioral relevance?[4] By now there are a number of studies that connected behavioral experiments with surveys in order to shed light on this issue. In general, the results are somewhat mixed. Using a sample of students at Harvard University, Glaeser et al. (2000) found that the GSS/WVS generalized trust item is not significantly related to first mover behavior but to second mover behavior in a trust game experiment. Using a large sample that was representative for the Dutch population, Bellemare & Kröger (2007) found that the variables predicting the responses the item and behavior in a trust game experiment were not identical. On the other hand, Sapienza et al. (2007) found significant correlation between survey responses and behavior in a trust game experiment with a sample of University of Chicago MBA students. Using a sizable and heterogenous sample of urban and rural dwellers in Russia and Belarus, Gächter et al. (2004) found the GSS/WVS generalized trust item positively correlated with contribution behavior in a public good game experiment, although not significantly so. In a study with a large sample that was representative for the Danish population, Thöni et al. (2012) found this correlation to be signficant.

Further research is certainly needed to explain this conflicting evidence. A possible explanation is that people with different cultural backgrounds have different situations in mind when confronted with the survey question (which is why I highlighted the sample populations of the above studies). Support for this conjecture comes from a study by Holm & Danielson (2005) who found that the survey measure is correlated with first mover behavior in a trust game experiment with undergraduates in Sweden but not in Tanzania. Most importantly for the present study, Fehr et al. (2002c) and

[3] Miller & Mitamura (2003) showed, for example, that Japanese students are more trusting than Americans measured with the above question from the GSS. Measuring trust and caution separately, they find that American students are more trusting than Japanese students but at the same time also more cautious.

[4] A discussion of this issue is missing in Fisman & Khanna (1999), so the following discussion is an additional contribution relative to their paper.

Naef & Schupp (2009b) found trust measured by the three-iteam SOEP battery to be significantly correlated with first mover behavior but not second mover behavior in a trust game experiment with a sample that was representative for the German population, and Naef & Schupp (2009b) find this to be robust to changes in the experimental protocol, such as stake sizes. We can therefore have some degree of confidence that the responses to the survey items have some behavioral relevance.

The main dependent variable that I will use throughout the analysis is a simple index built from the reponses of the three trust items by just adding them up («simple trust index» in what follows). Thus, this variable takes values between -6 and 6, and has a mean of -0.68 and a standard deviation of 2.50. To check whether results depend on this specific construction, I redo the anlysis with four alternative dependent variables. The second variable is obtained by performing a principal-component factor analysis with the three trust items as parameters, that indeed yields a unique factor that loads with 0.80, 0.83, and 0.60 on the three items, respectively, and using predicted rotated factor loadings as an alternative trust index («factor trust index» in what follows).[5] This variable is automatically normalized to have a mean of approximately zero (actually 0.004) and a standard deviation of approximately one (actually 1.001), and exhibits a range between -2.15 and 2.59. Finally, I will perform all analyses with each of the three items individually. A further dependent variable that is more directly targeted at measuring respondent's belief in others' trustworthiness will be introduced in section 7.3.2.

7.2.2 Focal predictors

The focal predictors in the present study are indicators of presence of a number of information technologies in the respondents' household. Fisman & Khanna (1999) used a single variable as a proxy for the average individual's access to two-way communication technology in a given country, the number of phones per capita. Correspondingly, the main variable in the present study indicates the presence of an ordinary landline telephone, which has a relative frequency of 0.937 in our sample. I have further data on variables that indicate the presence of a celluar phone (0.877), a narrow bandwidth internet connection (0.344), and a broad bandwidth internet connection (0.557). Furthermore, I consider the one-way communication technology television (0.974).

7.2.3 Control variables

One key contribution of the present study is the inclusion of a considerable set of individual-level control variables that allow for extensive checks of robustness. I will cluster the variables thematically into six batteries and include them in the analysis one at a time to obtain seven model specifications of increasing richness. The first battery is a set of federal state indicators to control for regional fixed effects. I will

[5] The complete results of the factor analysis are available upon request.

also check for effects of including an additional indicator that controls for former-GDR history fixed effects (but there will be no changes). I have chosen Saxony-Anhalt as the reference category, because it exhibits the lowest average trust index.

The second set consists of just one variable, household income. The major reason is a justified concern that posession of the various information technologies might merely pick up an income effect. I use the Infratest calculated household net income per month measured in Euro (mean $2,595.60$, standard deviation $1,888.61$, strongly skewed to the right) and take the natural logarithm for the regressions.

The third set consists of naturally exogenous (physical) socio-demographics, specifically height measured in centimeters (mean 171.43, standard deviation 9.41), weight measured in kilograms (mean 76.71, standard deviation 15.98), age measured in years (mean 50.04, standard deviation 17.22), and a gender dummy that indicates female with reference category male (relative frequency 0.522). Gender and age differences have been shown to exist in a variety of domains, and matter also for trust (e.g. Fehr et al., 2002c; Bellemare & Kröger, 2007; Dohmen et al., 2008; Garbarino & Slonim, 2009). But even physical variables such as height have been shown to matter for economically relevant behavior, such as risk taking (Dohmen et al., 2011), or trust and reciprocity (Dohmen et al., 2008).

The fourth set comprises variables on socialization and education. The first dummy indicates whether the respondent is a german citizen with natural reference category (relative frequency 0.942). It is intuitive that respondents with a recent migration background might have more difficulties in trusting others, and there is supporting evidence (Fehr et al., 2002c). The second dummy indicates whether the respondent is member of a hierarchically organized religion, with all other or no religion as reference category (relative frequency 0.316). This variable is also used by Fisman & Khanna (1999), and is motivated by claims of Putnam (1993a) that the Catholic Church has discouraged the formation of a «habit of trust» by their strongly hierarchical organization. La Porta et al. (1997) argued that this extends to any hierarchically religious organization, in particular Catholic and Orthodox Christianity and Islam. Thus, the above dummy indicates membership in those religions. The remaining variables are a set of five dummy variables that indicate the highest attained educational degree, specifically the completion of a lower secondary level degree (relative frequency 0.315), an intermediate secondary level degree (0.292), an upper secondary level degree (0.072), a tertiary level degree (0.218), or other degree (0.046), with reference category of no degree. There is also previous evidence indicating that the level of eduction matters for trust. Using pooled time-series and cross-sectional data from the US General Social Survey (GSS) from 1972 through 1996, and the DDB-Needham Life Style survey data from 1975 through 1997, Hellwell & Putnam (2007) find a significantly positive relationship between survey measured trust and education. In their large-scale experiment Bellemare & Kröger (2007) find similar evidence.

The fifth set consists of variables that describes the respondents' social network. The first two dummies indicate whether the respondent is in an active marriage (rela-

tive frequency 0.612), or in a permanent relationship (0.172), with being single as reference category. Another variable contains the number of self-reported close friends (mean 4.20, standard deviation 3.66). A dummy variable that indicates whether the respondent is gainfully employed is intended to capture integration in a social networks at the workplace. Finally, a set of four dummy variables indicates whether the respondent is integrated in spare time networks. Specifically, they indicate whether the respondent is active in local politics (relative frequency 0.108), attends church (0.475), volunteers (0.289), or is active in sports (0.664). In principle, all variables come with an intuitively natural (positive) prediction on trust. However, more importantly for the present purposes the variables may drive both trust and communication technology connectivity, since the marginal benefit of those technologies increase with network inclusion, such that their omission may result in a bias.

Finally, a set of trust-related personal preferences or attributes is included. First, it is well known from economic theory that impatience, discounting of the future, is a critical derteminant of the willingness to cooperate (e.g. Fudenberg & Maskin, 1986; Fudenberg et al., 1994). Thus, I include a self-reported measure of patience (mean 6.08, standard deviation 2.28), measured on a Likert scale from zero to 10 with increasing values representing increasing patience. By the above argument patience comes with a positive prediction. Second, I also include a self-reported measure of impulsiveness (mean 5.09, standard deviation 2.19), measured on the same scale with increasing values representing increasing impulsiveness. There are reasons while it might matter but no unambigous prediction: on the one hand, impulsive people may commit more «errors» which may disrupt trust relations more often, on the other hand it may also serve as a commitment device to act trustworthy against the own immediate self-interest more often (Frank, 1988), which may have repercussions (through others' reciprocity) on trust. Third, we also know from economic theory that risk aversion (tolerance) tends to inhibit (encourage) trust (see Fehr, 2009a, for a discussion of the evidence), since it involves a risk of betrayal. I therefore include a self-reported measures of risk preference (mean 4.44, standard deviation 2.30), also measured on a Likert scale from zero to 10 with increasing values representing increasing risk tolerance. The behavioral relevance of this item has been experimentally validated (Dohmen et al., 2011). By the above reasons, this variable also comes with a positive prediction.

7.3 Results

7.3.1 Trust and landline connectivity

Confirming the results of Fisman & Khanna (1999), we find a clear positive relationship between landline connectivity and survey measured trust. A raw comparison between those respondents with a landline phone and those without shows that the former exhibit a higher (simple) trust index than the latter: those with a phone score -0.640, those without a phone score -1.248, a difference that is significant using

Table 7.1: Results of linear model estimations with the simple trust index as dependent variable and landline connectivity as focal predictor.

	OLS estimates						
Trust index	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Landline	0.608	0.558	0.426	0.454	0.267	0.177	0.197
	(0.080)	(0.080)	(0.081)	(0.081)	(0.080)	(0.080)	(0.079)
	.000	*.000*	*.000*	*.000*	*.001*	*.026*	*.013*
State FE	*no*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Income	*no*	*no*	*yes*	*yes*	*yes*	*yes*	*yes*
Phys. demog.	*no*	*no*	*no*	*yes*	*yes*	*yes*	*yes*
Soc. & educ.	*no*	*no*	*no*	*no*	*yes*	*yes*	*yes*
Soc. netw.	*no*	*no*	*no*	*no*	*no*	*yes*	*yes*
Preferences	*no*	*no*	*no*	*no*	*no*	*no*	*yes*
Constant	-1.248	-1.711	-2.383	-7.085	-5.324	-5.362	-6.269
	(0.078)	(0.115)	(0.131)	(0.540)	(0.532)	(0.523)	(0.525)
	.000	*.000*	*.000*	*.000*	*.000*	*.000*	*.000*
Obs	17,915	17,915	17,915	17,915	17,915	17,915	17,915
Prob > F	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Adj. R^2	0.004	0.015	0.022	0.031	0.075	0.103	0.114

Robust Standard errors in parantheses, p-values in italics.

a Mann-Whitney rank-sum test (p = .000).[6] This difference is also captured by the coefficient of linear model 1 in table 7.1. Since this difference may be driven by regional-historic specifics that are correlated with both the dispersal of phones and trust, I include state dummies to control for those fixed effects. The most salient of such effects are differences between western and eastern states, the latter having a clearly different history during the existence of the former German Democratic Republic. As evident from model 2 in table 7.1, this reduces the difference only slightly.

A second concern may be that the posession of a phone merely picks up an income effect, so in model 3 I control for the logarithm of the net household income. This variable has an independent positive and highly significant effect on the trust index (see table 7.5 in the appendix). Furthermore, as evident from table 7.1 controlling for income further narrows the trust gap between phone owners and non-owners to about 0.426.

As a next step, model 4 includes the set of natually exogenous (physical) demographics, such as height, weight, gender, and age. All variables have an independent and statistically significant but modest effect on the trust index (see table 7.5 in the appendix). The difference between phone owners and non-owners remains approximately constant (it even increases slightly to 0.454).

In model 5, I include indicators of German citizenship, membership in hierarchi-

[6] Using the factor trust index, the former group scores 0.019 and latter −0.229. For the individual items, the differences are 0.256 vs. −0.008 for the first, 0.256 vs. −0.025 for the second, and −1.152 vs. −1.215 for the third. The markedly lower scores and the smaller difference suggest already that the third item measures something slightly different than the other two items. I will come back to this below.

cal religions, and education. Citizenship has a statistically significant and quantitatively noteworthy independented effect in the predicted direction (see table 7.5 in the appendix). In contrast to the predictions of Putnam (1993a) and La Porta et al. (1997), and the results of Fisman & Khanna (1999), I do not find a statistically significant effect of membership in a hierarchically organized religion in the dataset. Supporting the predictions (and earlier evidence) on the effect of education levels, respondents with an upper secondary level or tertiary level degree exhibit a markedly higher (simple) trust index than respondents without a degree. Those effects are the strongest among all variables included (table 7.5). Controlling for this cluster of variables narrowes the gap between phone owners and non-owners considerably, with a difference of 0.267 remaining. The vast majority of this narrowing stems from the inclusion of the education variables, suggesting that phone ownership picks up education effects to some extent.

I further extend the model by the set of variables that are aimed at capturing the respondents' social network. I do not find any independent effects of whether the respondent is in a permanent relationship (including marriage) or gainfully employed. However, the number of friends and whether the respondents participates in local social activities, such as politics, church, sports, or volunteering are all significant and positive predictors of measured trust (see table 7.5). As predicted, controlling for this cluster of variables narrowes the gap between phone owners and non-owners considerably further, with a difference of 0.177 remaining.

As a final step, I include the set of variables on trust-related personal preferences or attributes. As evident from table 7.5 in the appendix, patience and risk tolerance have the predicted independent effects on trust, which is interesting on its own. Impulsiveness also has a positive and significant coefficient, albeit somewhat weaker than the other two variables. Controlling for this cluster of variables widenes the gap between phone owners and non-owners somewhat, with a new difference of 0.197. Thus, the other control variables seem to absorb some effects of those individual characteristics if it is not controlled for them.

In sum, the (measured) trust gap between phone owners and non-owners clearly remains statistically significant at the five-percent level throughout all specifications, but diminishes somewhat through the intoduction of variables indicating a respondent's level of education and local social network. Fisman & Khanna (1999) did not control for neither, such that their results may be upwards biased.

In table 7.2 the results of some checks of robustness are shown. In the first row, I reestimate the models using Tobit and Ordered Probit methods. Evidently, the results are very robust to these variations. In the second row I use the factor trust index as an alternative dependent variable. By recalling that this index has a different range, it is evident that the results are robust to this alternative specification as well. In third, four, and fifth row, respectively, I reestimate the models with each of the three survey items separately, respectively (recall that each of them varies between -2 and 2). Apparently, the results become somewhat stronger with the first two items, in particular the second, while the third item seems to measure something else, which

Table 7.2: Robustness checks: Landline coefficients using different dependent variables and estimation methods.

	OLS		Tobit		Ordered Probit	
Dependent variable	Model 1	Model 7	Model 1	Model 7	Model 1	Model 7
Simple Index	0.608	0.197	0.635	0.208	0.243	0.081
	(0.080)	(0.080)	(0.083)	(0.082)	(0.033)	(0.034)
	.000	*.013*	*.000*	*.011*	*.000*	*.018*
Factor Index	0.248	0.082	0.259	0.087	0.248	0.085
	(0.032)	(0.032)	(0.033)	(0.033)	(0.033)	(0.034)
	.000	*.010*	*.000*	*.009*	*.000*	*.013*
Item 1	0.264	0.089	0.291	0.097	0.246	0.083
	(0.037)	(0.037)	(0.042)	(0.042)	(0.037)	(0.039)
	.000	*.016*	*.000*	*.019*	*.000*	*.034*
Item 2	0.281	0.119	0.342	0.142	0.248	0.107
	(0.039)	(0.039)	(0.047)	(0.003)	(0.035)	(0.036)
	.000	*.002*	*.000*	*.000*	*.000*	*.003*
Item 3	0.064	-0.012	0.124	-0.027	0.081	-0.017
	(0.030)	(0.031)	(0.054)	(0.055)	(0.035)	(0.037)
	.033	*.697*	*.021*	*.629*	*.020*	*.639*

Tabulated are the regression coefficients of the landline connectivity indicator. Robust Standard errors in parantheses, *p*-values in italics.

is perhaps not that surprising because it asks specifically about «strangers» while the others refer to people «as a whole».

The fact that the results are strongest using the second item («Nowadays one can't rely on anyone») alone is interesting, because one might argue that this item refers more to the respondents' *belief* about others' trustworthiness, while the other two refer more to the respondents' *behavioral* tendencies what one normally *does*. If this is correct, then this would be nicely consistent with standard economic theory which predicts that effects of information on behavior are completely mediated by beliefs. I now prusue this idea further.

7.3.2 Expectations

According to standard economic theory combined with recent experimental evidence, trust is driven by both beliefs about others' trustworthiness and preferences. Importantly, not only preferences towards risk or ambiguity but social preferences as well (Cox, 2004; Bohnet & Zeckhauser, 2004; Bohnet et al., 2008; Ashraf et al., 2006; Fischbacher & Gächter, 2010; Fehr, 2009a). The GSS/WVS contains an addional item that appears to be more directly asking about the beliefs in others' trustworthiness. It reads «Do you think most people would try to take advantage of you if they got a chance, or would they try to be fair?» with possible responses «would take advantage» and «would try to be fair». Gächter et al. (2004) found it significantly correlated with contribution behavior in the public good game experiment mentioned above, that is, if people believe that most others are fair, then they also contributed more to the experiment.

Thöni et al. (2012) conducted a large-scale public good experiment with a sample that was representative for the Danish population in which they connected contribution behavior in the experiment to survey measures. Consistent with previous studies, they find that the GSS/WVS generalized trust item has significant explanatory power for contribution behavior in the experiment. However, they go a step further in providing evidence that the GSS/WVS generalized trust item seems to measure preferences but not beliefs (i.e. responses to the question explain how much people contribute *given* their beliefs about others' contributions, but not how optimistic they are about other peoples' tendency to cooperate), whereas the expexted fairness item seems to measure beliefs but not preferences. They speculate that this is consistent with the way the questions are stated: the trust item evokes thoughts about what the respondent generally *does* («you can't be too careful») while the fairness item evokes thoughts about how *other* people generally behave («would they try to be fair?»).

Since, as argued above, any effect of information on behavior should be (theoretically) mediated by beliefs, the above results should be qualitatively no different if the dependent variable is replaced by the expected fairness item. The SOEP contains the item unaltered, so I am able to investigate this hypothesis. I code the affirmative response as one, respectively, and estimate the same model specifications as in the previous section. Since the response variable is binary (a relative frequency of 0.541 responded «would try to be fair»), I use logistic regression instead.[7]

The results are shown in table 7.3. Evidently, the results are remarkably similar to the previous ones. Furthermore, including the expected fairness item into the regressions reported above should absorb any direct effect of phone connectivity on the trust index. This turns out to be the case. Including the expected fairness item into the OLS estimation of model 7 with the simple trust index as dependent variable, the coefficient of the phone indicator becomes smaller and insignificant (0.107, $p = .134$), whereas the coefficent of the expected fairness item is huge and significant (2.225, $p = .000$). Sobel-Goodman mediation tests turn out to be significant at the 0.1 percent level. This reinforces the hypothesis that the expected fairness item measures beliefs about others' trustworthiness. Those results also render it more difficult, albeit not impossible, to argue that the phone indicator coefficient reflects reverse causality.

7.3.3 Further communication technologies

As final step, I expand model 7 to include the additional communication technology indicators, cellular phone, narrow bandwidth internet, broad bandwidth internet, and television. The results of this expandent model (model 8) using again three different estimation techniques are summarized in table 7.4. The coefficient of the landline phone indicator decreases a little, but remains significant at the five percent level.

[7] I also performed Probit regressions that yield approximately the same results, so I omit them here. They are, of course, available upon request. Note that 250 observations of respondents that did not answer the fairness item got lost in the regressions. However, restricting the previous analysis on this slightly smaller sample did not have any significant effect. Results on this checks are available upon request as well.

Table 7.3: Results of logistic regressions with the expected fairness item as dependent variable and landline connectivity as focal predictor.

	Logit estimates						
Trust index	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Landline	0.399	0.367	0.306	0.226	0.164	0.137	0.148
	(0.062)	(0.063)	(0.064)	(0.065)	(0.066)	(0.066)	(0.067)
	.000	*.000*	*.000*	*.000*	*.012*	*.039*	*.027*
State FE	*no*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Income	*no*	*no*	*yes*	*yes*	*yes*	*yes*	*yes*
Phys. demog.	*no*	*no*	*no*	*yes*	*yes*	*yes*	*yes*
Soc. & educ.	*no*	*no*	*no*	*no*	*yes*	*yes*	*yes*
Soc. netw.	*no*	*no*	*no*	*no*	*no*	*yes*	*yes*
Preferences	*no*	*no*	*no*	*no*	*no*	*no*	*yes*
Constant	-0.211	-0.552	-0.881	-4.135	-3.253	-3.360	-3.959
	(0.061)	(0.094)	(0.108)	(0.461)	(0.464)	(0.467)	(0.472)
	.000	*.000*	*.000*	*.000*	*.000*	*.000*	*.000*
Obs	17,665	17,665	17,665	17,665	17,665	17,665	17,665
Prob > χ^2	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Pseudo R^2	0.002	0.007	0.009	0.017	0.030	0.044	0.048

Robust Standard errors in parantheses, *p*-values in italics.

Table 7.4: Results of linear model estimations with the trust index as dependent variable and additional connectivity as focal predictors.

	Model 8		
Trust index	OLS	Tobit	Ordered Probit
Landline	0.171	0.180	0.069
	(0.082)	(0.085)	(0.035)
	.037	*.034*	*.051*
Cellular	0.020	0.024	0.007
	(0.063)	(0.065)	(0.027)
	.751	*.714*	*.808*
Internet narrow	-0.011	-0.007	-0.001
	(0.038)	(0.059)	(0.017)
	.780	*.885*	*.930*
Internet broad	0.187	0.193	0.084
	(0.042)	(0.043)	(0.018)
	.000	*.000*	*.000*
Television	-0.382	-0.383	-0.173
	(0.121)	(0.126)	(0.054)
	.002	*.002*	*.001*
Controls	*yes*	*yes*	*yes*
Constant	-5.948	-6.161	
	(0.536)	(0.550)	
	.000	*.000*	
Obs	17,915	17,915	17,915
Prob > *F*	.000	.000	
Adj. R^2	0.115		
Prob > χ^2			.000
Pseudo R^2		0.026	0.028

Robust Standard errors in parantheses, *p*-values in italics.

Owning a cellular phone on top has no additional independent effect. A narrow bandwidth internet connection (relative no internet connection) also has no significant effect. The coefficient of the broad bandwidth internet connectivity indicator, however, is positive, highly significant, and slightly larger than that for landline connectivity. This is interesting and is consistent with the fact that the internet has become a major communication medium, both two-way and public address.

Remarkable, however, is the quantitatively and statistically significant coefficient of the television indicator, which is strongly negative. Investigating the robustness and universality of this effect and possible explanations is perhaps an interesting avenue for further research. A speculative hypothesis would be that people without a TV set are less exposed to reports about opportunistic behavior, but base their beliefs more on their personal environment which can be expected to be more balanced with respect to instances of trustworthy and opportunistic behavior.

7.4 Conclusion

I have replicated the results of Fisman & Khanna (1999) (i) using a large cross-section of recent individual level micro-data taken from the German Socio-Economic Panel (SOEP), (ii) an improved, more fine-git, and experimentally validated measure of generalized trust, and (iii) using alternative estimation methods. I extended on the study by (i) controlling for a much larger set of confounding variables, (ii) drawing on an experimentally validated measure of *beliefs* into others trustworthiness in order to investigate whether the effect of communication technology on measured trust is mediated through measured beliefs, as predicted by economic theory, and (iii) extending the set of considered communication technologies to include cellular phones, internet (both narrow and broad bandwidth), and television.

I found that the follwing. First, the results show a gap in measured trust between those repondents who have a landline phone in their household and those who do not, confirming the results of Fisman & Khanna (1999). Second, the existence of this gap is robust to alternative estimation techniques and the specific construction of the dependent variable. Third, I find that this gap also using the item intended to measure *beliefs* in others' trustworthiness more directly. Furthermore, including the variable into the original specification as a predictor, the gap vanishs, suggesting that the effect is indeed mediated by beliefs. Fourth, the gap narrows significantly as variables that capture the level of education and local network inclusion are controlled for. Since Fisman & Khanna (1999) did not control for those variables, their results might be biased upwards. Fifth, there are no signficant effects of celluar phone or narrow bandwidth internet connectivity, but there is a significantly positive effect of broad bandwidth internet connectivity, and a significantly negative effect of television ownership. The order of magnitude of the latter effect is considerable, and may be an interesting input for further research.

There are, however, still important caveats remaining. First, I followed Fisman & Khanna (1999) in just *assuming* a causal direction. Economic theory provides good

reasons for this assumption, but empirically I cannot rule out reverse causality with the data at hand. Such a reversed relationship would boil down to the assumption with which I introduced this chapter, namely that trust is indeed (to some extent) exogenous. If one wants to believe this story (definite evidence is still lacking, see Fehr, 2009a) one could easily construct cases in which individuals' trust drives their technology adoption. The results around the expected fairness item are, of course, not easy to reconcile with this story. However, I cannot exclude such a possibility beyond doubt, such that any causal interpretation of the present results is to be made with due care.

Second, despite that I have included a large set of control variables, there may still be an omission bias such that the *magnitude* of the relationships is to be taken with care as well. In extreme, there is still the possibility that there is no relationship at all. Addressing these shortcomings, possibly complemented by experimental research, would be an interesting avenue for further research.[8]

Appendix

[8]The items on trust and on the household's technology equipment are included only on a irregular basis for time to time. The 2008 wave is currently the only one in which both item batteries were included at the same time. If this happens again in the future, it would also be interesting to construct a panel dataset in order to get even more robust estimates and study individual level dynamics.

Table 7.5: Additional results of linear model estimations with the simple trust index as dependent variable.

	Model 7		Model 8	
Trust index	OLS	Tobit	OLS	Tobit
Income (log)	0.057	0.059	0.056	0.058
	.000	*.000*	*.000*	*.000*
Female	0.192	0.205	0.197	0.209
	.000	*.000*	*.000*	*.000*
Age	0.005	0.005	0.007	0.007
	.001	*.001*	*.000*	*.000*
Height	0.015	0.016	0.015	0.015
	.000	*.000*	*.000*	*.000*
Weight	-0.008	-0.008	-0.008	-0.008
	.000	*.000*	*.000*	*.000*
German citizen	0.382	0.405	0.371	0.393
	.000	*.000*	*.000*	*.000*
Hier. religion	-0.025	-0.028	-0.020	-0.022
	.564	*.530*	*.653*	*.617*
Lower second. deg.	-0.249	-0.252	-0.240	-0.243
	.005	*.006*	*.006*	*.008*
Int. second. deg.	0.018	0.022	0.008	0.012
	.841	*.808*	*.923*	*.894*
Upper second. deg.	0.822	0.837	0.795	0.810
	.000	*.000*	*.000*	*.000*
Tertiary deg.	0.907	0.919	0.869	0.879
	.000	*.000*	*.000*	*.000*
Other deg.	-0.100	-0.107	-0.097	-0.103
	.398	*.385*	*.415*	*.403*
Married	-0.052	-0.047	-0.081	0.079
	.273	*.328*	*.093*	*.113*
Steady partner	0.043	0.055	0.028	0.039
	.471	*.367*	*.643*	*.525*
Close friends	0.068	0.070	0.068	0.070
	.000	*.000*	*.000*	*.000*
Employed	-0.010	-0.006	-0.022	-0.019
	.819	*.901*	*.611*	*.679*
Local politics	0.175	0.178	0.173	0.176
	.003	*.003*	*.004*	*.004*
Volunteering	0.223	0.227	0.220	0.224
	.000	*.000*	*.000*	*.000*
Church	0.368	0.382	0.369	0.382
	.000	*.000*	*.000*	*.000*
Sports	0.246	0.261	0.232	0.247
	.000	*.000*	*.000*	*.000*
Patience	0.086	0.089	0.086	0.089
	.000	*.000*	*.000*	*.000*
Impulsiveness	0.030	0.030	0.030	0.030
	.001	*.002*	*.002*	*.002*
Risk tolerance	0.065	0.067	0.064	0.066
	.000	*.000*	*.001*	*.000*
Constant	-6.269	-6.481	-5.948	-6.161
	.000	*.000*	*.000*	*.000*
Obs	17,915	17,915	17,915	17,915
Prob > F	.000	.000	.000	.000
Adj. R^2	0.114		0.115	
Pseudo R^2		0.026		0.026

In order to fit the table on a single page, the robust standard errors are not but only p-values are reported (in italics).

Final Remarks

With the studies in chapters 3 through 7 I have made steps toward extending the literature on cooperation in considering information structures as endogenous. I motivated this line of research by conjecturing that doing so might have a number of interesting implications, in particular that (i) individuals might adopt strategies that produce entirely different cooperative and sanctioning behaviors if we allow players to acquire more accurate information (at a cost), and (ii) that individuals also have the option to forgo the opportunity, to stay delibertely «blind». I found confirming evidence in chapters 3 through 6. Understanding more about the motivational fine-structure of the «blind punisher» and the «blind trustor» might be an interesting avenue for further research. Considering indefinite horizon settings would also be interesting to draw more explicit connections to the theoretical (folk theorem) literature. Another interesting route would be to extend the setting to study not only monitoring but also information transmission or pooling in a matching game, like those considered in sections 1.4 and 2.5. These steps would also underpin the results obtained in chapter 7 experimentally. In sum, there are vast opportunities to further extend the experimental (and theoretical) literature in ways to the study transparency as an endogenous («equilibrium») phenomenon, which is shaped by the players' information acquisition and transmission behavior.

Bibliography

Abbink, K. & Herrmann, B. (2011). The moral cost of nastiness. *Economic Inquiry*, 49(2), 631–633.

Abbink, K. & Sadrieh, A. (2009). The pleasure of being nasty. *Economics Letters*, 105(3), 306–308.

Abeler, J., Altmann, S., Kube, S., & Wibral, M. (2010). Gift exchange and workers' fairness concerns: When equality is unfair. *Economic Inquiry*, 48(6), 1299–1324.

Abreu, D., Pearce, D., & Stacchetti, E. (1986). Optimal cartel equilibria with imperfect monitoring. *Journal of Economic Theory*, 39(1), 251–269.

Abreu, D., Pearce, D., & Stacchetti, E. (1990). Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica*, 58(5), 1041–1063.

Acheson, J. M. (1975). The lobster fiefs: Economic and ecological effects of territoriality in the maine lobster industry. *Human Ecology*, 3(3), 183–207.

Acheson, J. M. (1987). The lobster fiefs revisited: Economic and ecological effects of territoriality in the Maine lobster industry. In B. McClay & J. M. Acheson (Eds.), *The Problem of the Commons: The Culture and Ecology of Communal Resources* (pp. 37–65). Tucson, Arizona, USA: University of Arizona Press.

Acheson, J. M. (1988). *The Lobster Gangs of Maine*. Hanover, NH, USA: University Press of New England.

Akerlof, G. A. (1982). Labor contracts as partial gift exchange. *Quarterly Journal of Economics*, 97(4), 543–569.

Akerlof, G. A. (1983). Loyalty filters. *American Economic Review*, 73(1), 54–63.

Akerlof, G. A. (1984). Gift exchange and efficiency-wage theory: Four views. *American Economic Review*, 74(2), 79–83.

Akerlof, G. A. & Yellen, J. L. (1990). The fair wage-effort hypothesis and unemployment. *Quarterly Journal of Economics*, 105(2), 255–283.

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

Alexander, R. D. (1987). *The Biology of Moral Systems*. New York, NY, USA: Aldine de Gruyter.

Alpizar, F., Carlsson, F., & Johansson-Stenman, O. (2008). Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica. *Journal of Public Economics*, 92(5-6), 1047–1060.

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

Ambrus, A. & Pathak, P. A. (2011). Cooperation over finite horizons: A theory and experiments. *Journal of Public Economics*, 95(7-8), 500–512.

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.

Anderson, C. M. & Putterman, L. (2006). Do non-strategic sanctions obey the law of demand? the demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior*, 54(1), 1–24.

Anderson, S. P., Goeree, J. K., & Holt, C. A. (1998). A theoretical analysis of altruism and decision

error in public goods games. *Journal of Public Economics*, 70(2), 297–323.

Andreoni, J. (1988). Privately provided public goods in a large economy: The limits of altruism. *Journal of Public Economics*, 35(1), 57–73.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *Economic Journal*, 100(401), 464–477.

Andreoni, J. (1995). Cooperation in public-goods experiments: Kindness or confusion? *American Economic Review*, 85(4), 891–904.

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., Erard, B., & Feinstein, J. (1998). Tax compliance. *Journal of Economic Literature*, 36(2), 818–860.

Andreoni, J. & Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2), 737–753.

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

Andreoni, J. & Petrie, R. (2004). Public goods experiments without confidentiality: A glimpse into fund-raising. *Journal of Public Economics*, 88(7-8), 1605–1623.

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

Apel, M., Friberg, R., & Hallsten, K. (2005). Microfoundations of macroeconomic price adjustment: Survey evidence from Swedish firms. *Journal of Money, Credit and Banking*, 37(2), 313–338.

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

Arrow, K. J. (1971). Political and economic evaluation of social effects and externalities. In M. D. Intriligator & D. A. Kendrick (Eds.), *Frontiers of Quantitative Economics: Papers Invited for Presentation at the Econometric Society Winter Meetings, New York, 1969 and Toronto, 1972*, volume 2 (pp. 3–23). Amsterdam, The Netherlands: North-Holland.

Arrow, K. J. & Debreu, G. (1954). Existence of an equilibrium for a competitive economy. *Econometrica*, 22(3), 265–290.

Arrow, K. J. & Hahn, F. (1971). *General competitive analysis*. San Francisco, CA, USA: Holden-Day.

Ashraf, N., Bohnet, I., & Piankov, N. (2006). Decomposing trust and trustworthiness. *Experimental Economics*, 9(3), 193–208.

Avery, C. & Zemsky, P. B. (1994). Money burning and multiple equilibria in bargaining. *Games and Economic Behavior*, 7(2), 154–168.

Axelrod, R. & Dion, D. (1988). The further evolution of cooperation. *Science*, 242(4884), 1385–1390.

Axelrod, R. & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211(4489), 1390–1396.

Axelrod, R. M. (1984). *The Evolution of Cooperation*. New York, NY, USA: Basic Books.

Ba, S. & Pavlou, P. (2002). Evidence of the effect of trust building technology in electronic markets: price premiums and buyer behavior. *MIS Quarterly*, 26(3), 243–268.

Baker, G., Gibbons, R., & Murphy, K. J. (2004). Relational contracts and the theory of the firm. *Quarterly Journal of Economics*, 117(1), 39–84.

Balliet, D., Mulder, L., & van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, 137(4), 594–615.

Banerjee, A. V., Besley, T., & Guinnane, T. W. (1994). The neighbor's keeper: The design of a credit cooperative with theory and a test. *Quarterly Journal of Economics*, 109(2), 491–515.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, 27(5), 325–344.

Bardhan, P. (2000). Water community: An empirical analysis of cooperation on irrigation in South India. In M. Aoki & Y. Hayami (Eds.), *Communities and Markets in Economic Development* (pp. 247–264). Oxford, UK: Oxford University Press.

Bardsley, N. (2008). Dictator game giving: altruism or artefact. *Experimental Economics*, 11(2), 122–133.

Bardsley, N. & Moffatt, P. G. (2007). The experimetrics of public goods: Inferring motivations from contributions. *Theory and Decision*, 62(2), 161–193.

Barr, A. M. (2001). *Social dilemmas and shame-based sanctions: Experimental results from rural Zimbabwe*. Centre for the Study of African Economies Series WPS/2001-11, University of Oxford, Oxford, UK.

Bartling, B., Fehr, E., Maréchal, A., & Schunk, D. (2009). Egalitarianism and competitiveness. *American Economic Review*, 99(2), 93–98.

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of beeing watched enhance cooperation in a real world setting. *Biology Letters*, 2(3), 412–414.

Beales, J. H. (1994). FDA regulation of pharmaceutical advertising: Economic analysis and the regulation of pharmaceutical advertising. *Seton Hall Law Review*, 24, 1370–1398.

Becker, G. S. (1974). A theory of social interactions. *Journal of Political Economy*, 82(6), 1063–1093.

Becker, G. S. (1976a). Altruism, egoism, and genetic fitness: Economics and sociobiology. *Journal of Economic Literature*, 14(3), 817–826.

Becker, G. S. (1976b). *The Economic Approach to Human Behavior*. Chicago, IL, USA: University of Chicago Press.

Beddow, R. & Sidwell, M., Eds. (2012). *Transparency International Annual Report 2011*. Berlin, Germany: Transparency International.

Bellemare, C. & Kröger, S. (2007). On representative social capital. *European Economic Review*, 51(1), 183–202.

Bellemare, C. & Shearer, B. (2009). Gift giving and worker productivity: Evidence from a firm-level experiment. *Games and Economic Behavior*, 67(1), 233–244.

Ben-Shakhar, G., Bornstein, G., Hopfensitz, A., & van Winden, F. (2007). Reciprocity and emotions in bargaining using physiological and self-report measures. *Journal of Economic Psychology*, 28(3), 314–323.

Bendor, J. (1993). Uncertainty and the evolution of cooperation. *Journal of Conflict Resolution*, 37(4), 709–734.

Bendor, J., Kramer, R. M., & Stout, S. (1991). When in doubt... cooperation in a noisy prisoner's dilemma. *Journal of Conflict Resolution*, 35(4), 691–719.

Bendor, J. & Mookherjee, D. (1990). Norms, third-party sanctions, and cooperation. *Journal of Law, Economics, and Organization*, 6(1), 33–63.

Bendor, J. & Swistak, P. (2001). The evolution of norms. *American Journal of Sociology*, 106(6), 1493–1545.

Benkler, Y. (2002). Coase's penguin, or, Linux and *The Nature of the Firm*. *Yale Law Journal*, 112(3), 369–446.

Benz, M. & Meier, S. (2008). Do people behave in experiments as in the field? evidence from donations. *Experimental Economics*, 11(3), 268–281.

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

Bernanke, B. S. (1983). Non-monetary effects of the financial crisis in the propagation of the Great Depression. *American Economic Review*, 73(3), 257–276.

Besley, T. & Coate, S. (1995). Group lending, repayment incentives and social collateral. *Journal of Development Economics*, 46(1), 1–18.

Bhaskar, V. & Obara, I. (2002). Belief-based equilibria in the repeated Prisoners' Dilemma with private monitoring. *Journal of Economic Theory*, 102(1), 40–69.

Binmore, K. (1998). *Game Theory and the Social Contract: Just Playing*. Cambridge, MA, USA: MIT Press.

Björkman, M. & Svensson, J. (2009). Power to the people: Evidence from a randomized field experiment on community-based monitoring in Uganda. *Quarterly Journal of Economics*, 124(2),

735–769.

Björkman, M. & Svensson, J. (2010). When is community-based monitoring effective? evidence from a randomized experiment in primary health in Uganda. *Journal of the European Economic Association*, 8(2-3), 259–267.

Blanco, M., Engelmann, D., & Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, 72(2), 321–338.

Blanton, T. S. (2006). The global openness movement in 2006: 240 years after the first freedom of information law, access to government information now seen as a human right. In J. Mustonen (Ed.), *The World's First Freedom of Information Act* (pp. 114–139). Kokkola, Finland: Anders Chydenius Foundation.

Blau, P. M. (1962). *Exchange and Power in Social Life*. New York, NY, USA: John Wiley & Sons.

Blinder, A. S., Canetti, E. R. D., Lebow, D. E., & Rudd, J. B. (1998). *Asking about Prices: A New Approach to Understanding Price Stickiness*. New York, NY, USA: Russell Sage Foundation.

Blount, S. (1995). When social outcomes aren't fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes*, 63(2), 131–144.

Bochet, O., Page, T., & Putterman, L. (2006). Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior & Organization*, 60(1), 11–26.

Boehm, C. (1993). Egalitarian behavior and reverse dominance hierarchy. *Current Anthropology*, 34(3), 227–254.

Boehm, C. (1999). *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Cambridge, MA, USA: Harvard University Press.

Boero, R., Bravo, G., Castellani, M., & Squazzoni, F. (2009). Reputational cues in repeated trust games. *Journal of Socio-Economics*, 38(6), 871–877.

Bohnet, I. & Frey, B. S. (1997). Rent leaving. *Journal of Institutional and Theoretical Economics*, 153(4), 711–721.

Bohnet, I., Greig, F., Herrmann, B., & Zeckhauser, R. (2008). Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *American Economic Review*, 98(1), 294–310.

Bohnet, I. & Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior & Organization*, 55(4), 467–484.

Bolle, F. (1998). Rewarding trust: An experimental study. *Theory and Decision*, 45(1), 83–98.

Bolton, G. E., Katok, E., & Ockenfels, A. (2004). How effective are electronic reputation mechanisms? an experimental investigation. *Management Science*, 50(11), 1587–1602.

Bolton, G. E., Katok, E., & Ockenfels, A. (2005). Cooperation among strangers with limited information about reputation. *Journal of Public Economics*, 89(8), 1457–1468.

Bolton, G. E. & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1), 166–193.

Boone, C., Declerck, C. H., & Suetens, S. (2008). Subtle social cues, explicit incentives and cooperation in social dilemmas. *Evolution and Human Behavior*, 29(3), 179–188.

Bornstein, G. & Weisel, O. (2010). Punishment, cooperation, and cheater detection in "noisy" social exchange. *Games*, 1(1), 18–33.

Bosman, R., Sutter, M., & van Winden, F. (2005). The impact of real effort and emotions in the power-to-take game. *Journal of Economic Psychology*, 26(3), 407–429. Tilburg Symposium on Psychology and Economics: Games and Decisions.

Bosman, R. & van Winden, F. (2002). Emotional hazard in a power-to-take experiment. *Economic Journal*, 112(476), 147–169.

Bowles, S. (1985). The production process in a competitive economy: Walrasian, Neo-Hobbesian, and Marxian models. *American Economic Review*, 75(1), 16–36.

Bowles, S. (2004). *Microeconomics: Behavior, Institutions, and Evolution*. The Roundtable Series in Behavioral Economics. New York, NY, USA / Princeton, NJ, USA: Russell Sage Foundation / Princeton University Press.

Bowles, S. & Gintis, H. (2002). Social capital and community governance. *Economic Journal*, 112(483), F419–F436.

Bowles, S. & Gintis, H. (2005a). Prosocial emotions. In L. E. Blume & S. N. Durlauf (Eds.), *The Economy as an Evolving Complex System*, volume III (pp. 307–326). Oxford, UK: Oxford University Press.

Bowles, S. & Gintis, H. (2005b). Social capital, moral sentiments, and community governance. In H. Gintis, S. Bowles, R. Boyd, & E. Fehr (Eds.), *Moral Sentiments and Material Interests. The Foundations of Cooperation in Economic Life* (pp. 379–398). Cambridge, Massachusetts, USA: MIT Press.

Bowles, S. & Gintis, H. (2011). *A Cooperative Species: Human Reciprocity and its Evolution*. Princeton, NJ, USA: Princeton University Press.

Boyd, R. & Richerson, P. J. (1988a). *Culture and the Evolutionary Process*. Chicago, IL, USA: University of Chicago Press.

Boyd, R. & Richerson, P. J. (1988b). The evolution of reciprocity in sizable groups. *Journal of Theoretical Biology*, 132(3), 337–356.

Boyd, R. & Richerson, P. J. (1995). Why does culture increase human adaptability? *Ethology and Sociobiology*, 16(2), 125–143.

Boyd, R. & Richerson, P. J. (2005). *The Origin and Evolution of Cultures*. Oxford, UK: Oxford University Press.

Brandts, J. & Charness, G. (2004). Do labour market conditions affect gift exchange? some experimental evidence. *Economic Journal*, 114(497), 684–708.

Brandts, J., Saijo, T., & Schram, A. (2004). How universal is behavior? a four country comparison of spite and cooperation in voluntary contribution mechanisms. *Public Choice*, 119(3-4), 381–424.

Brandts, J. & Schram, A. (2001). Cooperation and noise in public goods experiments: Applying the contribution function approach. *Journal of Public Economics*, 79(2), 399–427.

Brüggen, A. & Strobel, M. (2007). Real effort versus chosen effort in experiments. *Economics Letters*, 96(2), 232–236.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.

Brosnan, S. F. & de Waal, F. B. M. (2002). A proximate perspective on reciprocal altruism. *Human Nature*, 13(1), 129–152.

Brown, M., Falk, A., & Fehr, E. (2004). Relational contracts and the nature of market interactions. *Econometrica*, 72(3), 747–780.

Brown, M., Falk, A., & Fehr, E. (2008). *Competition and relational contracts: The role of unemployment as a disciplinary device*. IZA Discussion Paper Series 3345, Forschungsinstitut zur Erforschung der Arbeit, Bonn.

Browning, C. R., Feinberg, S. L., & Dietz, R. D. (2005). The paradox of social organization: Networks, collective efficacy, and violent crime in urban neighborhoods. *Social Forces*, 83(2), 503–534.

Buchanan, J. M. (1967). Cooperation and conflict in public-goods interaction. *Economic Inquiry*, 5(2), 109–121.

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron*, 60(5), 930–940.

Burlando, R. & Hey, J. D. (1997). Do Anglo-Saxons free-ride more? *Journal of Public Economics*, 64(1), 41–60.

Burlando, R. M. & Guala, F. (2005). Heterogeneous agents in public goods experiments. *Experimental Economics*, 8(1), 35–54.

Burnham, T. C. & Hare, B. (2007). Engineering human cooperation: Does involuntary neural activation increase public goods contributions? *Human Nature*, 18(2), 88–108.

Calvert, R. L. (1987). Reputation and legislative leadership. *Public Choice*, 55(1-2), 81–119.

Camera, G. & Casari, M. (2009). Cooperation among strangers under the shadow of the future. *American Economic Review*, 99(3), 979–1005.

Camera, G., Casari, M., & Bigoni, M. (2012). Cooperative strategies in anonymous economies: An experiment. *Games and Economic Behavior*, 75(2), 570–586.

Camerer, C. & Thaler, R. H. (1995). Ultimatums, dictators and manners. *Journal of Economic Perspectives*, 9(2), 209–219.

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

Camerer, C. F. (2003). *Behavioral Game Theory*. Princeton, NJ, USA: Princeton University Press.

Camerer, C. F. & Fehr, E. (2004). Measuring social norms using experimental games: A guide for social scientists. In J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, & H. Gintis (Eds.), *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies* (pp. 55–95). Oxford, UK: Oxford University Press.

Cameron, L. A. (1999). Raising the stakes in the Ultimatum Game: Experimental evidence from Indonesia. *Economic Inquiry*, 37(1), 47–59.

Carpenter, J. (2007a). The demand for punishment. *Journal of Economic Behavior & Organization*, 62(4), 522–542.

Carpenter, J., Connolly, C., & Knowles-Myers, C. (2008). Altruistic behavior in a representative dictator experiment. *Experimental Economics*, 11(3), 282–298.

Carpenter, J. & Seki, E. (2011). Do social preferences increase productivity? field experimental evidence from fishermen in Toyama Bay. *Economic Inquiry*, 49(2), 612–630.

Carpenter, J., Verhoogen, E., & Burks, S. (2005). The effect of stakes in distribution experiments. *Economics Letters*, 86(3), 393–398.

Carpenter, J. P. (2004). When in rome: Conformity and the provision of public goods. *Journal of Socio-Economics*, 33(4), 395–408.

Carpenter, J. P. (2007b). Punishing free-riders: how group size affects mutual monitoring and the provision of public goods. *Games and Economic Behavior*, 60(1), 31–51.

Carpenter, J. P., Matthews, H. P., & Org'ong'a, O. (2004). Why punish? social reciprocity and the enforcement of prosocial norms. *Journal of Evolutionary Economics*, 14(4), 407–429.

Casari, M. (2005). On the design of peer punishment experiments. *Experimental Economics*, 8(2), 107–115.

Cason, T. N. & Khan, F. U. (1999). A laboratory study of voluntary public goods provision with imperfect monitoring and communication. *Journal of Development Economics*, 58(2), 533–552.

Charness, G. (2004). Attribution and reciprocity in an experimental labor market. *Journal of Labor Economics*, 22(3), 665–688.

Charness, G., Cobo-Reyes, R., & Jiménez, N. (2008). An investment game with third-party intervention. *Journal of Economic Behavior & Organization*, 68(1), 18–28.

Charness, G., Frechette, G. R., & Kagel, J. H. (2004). How robust is laboratory gift exchange? *Experimental Economics*, 7(2), 189–205.

Charness, G. & Haruvy, E. (2002). Altruism, equity, and reciprocity in a gift-exchange experiment: an encompassing approach. *Games and Economic Behavior*, 40(2), 203–231.

Charness, G. & Kuhn, P. (2007). Does pay inequality affect worker effort? experimental evidence. *Journal of Labor Economics*, 25(4), 693–723.

Charness, G. & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117(3), 817–869.

Chaudhuri, A. & Gangadharan, L. (2007). An experimental analysis of trust and trustworthiness. *Southern Economic Journal*, 73(4), 959–985.

Cherry, T. L., Frykblom, P., & Shogren, J. F. (2002). Hardnose the dictator. *American Economic Review*, 92(4), 1218–1221.

Chiang, Y.-S. (2010). Self-interested partner selection can lead to the emergence of fairness. *Evolution and Human Behavior*, 31(4), 265–270.

Cinyabuguma, M., Page, T., & Putterman, L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics*, 9(3), 265–279.

Clark, A., Masclet, D., & Villeval, M. C. (2010). Effort and comparison income: Experimental and survey evidence. *Industrial and Labor Review*, 63(3), 407–426.

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

Clutton-Brock, T. H. & Parker, G. A. (1995). Punishment in animal societies. *Nature*, 373, 209–216.

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

Colman, A. M. (1982). *Game Theory and Experimental Games*. Oxford, UK: Pergamon Press.

Colman, A. M. (1999). *Game theory & its Applications in the Social and Biological Sciences*. New York, NY, USA: Routledge.

Cookson, R. (2000). Framing effects in public goods experiments. *Experimental Economics*, 3(1), 55–79.

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.

Cornes, R. & Sandler, T. (1984). The theory of public goods: non-nashbehaviour. *Journal of Public Economics*, 23(3), 367–379.

Cosslett, S. R. & Lee, L.-F. (1985). Serial correlation in latent discrete variable models. *Journal of Econometrics*, 27(1), 79–97.

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

Cox, J. C., Friedman, D., & Gjerstad, S. (2007). A tractable model of reciprocity and fairness. *Games and Economic Behavior*, 59(1), 17–45.

Craig, B. & Pencavel, J. (1995). Participation and productivity: A comparison of worker cooperatives and conventional firms in the Plywood industry. *Brookings Papers on Economic Activity: Microeconomics*, 1995, 121–174.

Cárdenas, J. C. & Jaramillo, C. R. (2010). *Cooperation in large networks: An experimental approach*. SFI Working Paper Series 10-03-009, Santa Fe Institute, Santa Fe, New Mexico, USA.

Croson, R. T. A. (1996). Partners and strangers revisited. *Economics Letters*, 53(1), 25–32.

Croson, R. T. A. (2007). Theories of commitment, altruism and reciprocity: Evidence from linear public goods games. *Economic Inquiry*, 45(2), 199–216.

Cubitt, R. P., Drouvelis, M., & Gächter, S. (2011). Framing and free riding: Emotional responses and punishment in social dilemma games. *Experimental Economics*, 14(2), 254–272.

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.

Dal Bó, P. & Fréchette, G. R. (2012). *Strategy Choice In The Infinitely Repeated Prisoners Dilemma*. Brown university mimeograph, Brown University, Providence, RD, USA.

Darby, M. R. & Karni, E. (1973). Free competition and the optimal amount of fraud. *Journal of Law and Economics*, 16(1), 67–88.

Darwin, C. (1871). *The Descent of Man and Selection in Relation to Sex*. London, UK: John Murray.

Dasgupta, S., Hong, J. H., Laplante, B., & Mamingi, N. (2006a). Disclosure of environmental violations and stock market in the republic of korea. *Ecological Economics*, 58(4), 759–777.

Dasgupta, S., Wang, H., & Wheeler, D. (2006b). Anders Chydenius and the origins of world's first freedom of information act. In T. Tietenberg & H. Folmer (Eds.), *International Yearbook of Environmental and Resource Economics* (pp. 93–119). Cheltenham, UK: Edward Elgar Publishing, 2006/2007 edition.

Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature*, 446, 794–796.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology*, 31, 169–193.

Dawes, R. M. & Thaler, R. H. (1988). Anomalies: Cooperation. *Journal of Economic Perspectives*,

2(3), 187–197.

de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, 305(5688), 1254–1258.

Debreu, G. (1959). *Theory of Value*. New York, NY, USA: John Wiley & Sons.

Denant-Boemont, L., Masclet, D., & Noussair, C. N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory*, 33(1), 145–167.

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. (2008). Representative trust and reciprocity: Prevalence and determinants. *Economic Inquiry*, 46(1), 84–90.

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.

Dore, R. (1987). *Taking Japan Seriously–A Confucian Perspective on Leading Economic Issues*. Stanford, CA, USA: Stanford University Press.

Drabek, Z. & Payne, W. (2002). The impact of transparency on foreign direct investment. *Journal of Economic Integration*, 17(4), 777–810.

Dranove, D. (2002). *The Economic Evolution of American Health Care*. Princeton, NJ, USA: Princeton University Press.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature*, 452, 348–351.

Drouvelis, M. (2010). *The behavioral consequences of unfair punishment*. Department of Economics Working Paper 10-34, University of Birmingham, Birmingham, UK.

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

Dufwenberg, M., Gächter, S., & Hennig-Schmidt, H. (2011). The framing of games and the psychology of play. *Games and Economic Behavior*, 73(2), 459–478.

Dufwenberg, M. & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2), 268–298.

Dugatkin, L. A. (1997). *Cooperation among animals: An evolutionary perspective*. Oxford: Oxford University Press.

Dulleck, U. & Kerschbamer, R. (2006). On doctors, mechanics, and computer specialists: The economics of credence goods. *Journal of Economic Literature*, 44(1), 5–42.

Durkheim, m. (1893). *De la division du travail social: Étude sur l'organisation des sociétés supérieures*. Paris, France: Félix Alcan.

Eckel, C. C. & Grossman, P. J. (2008). Men, women and risk aversion: Experimental evidence. In C. R. Plott & V. L. Smith (Eds.), *Handbook of Experimental Economics Results* (pp. 1061–1073). Amsterdam, The Netherlands: North-Holland.

Eckel, C. C. & Wilson, R. K. (2004). Is trust a risky decision? *Journal of Economic Behavior & Organization*, 55(4), 447–465.

Egas, M. & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Philosophical Transactions of the Royal Society: Biological Sciences*, 275(1637), 871–878.

Ellison, G. (1994a). Cooperation in the Prisoner's Dilemma with anonymous random matching. *Review of Economic Studies*, 61(3), 567–588.

Ellison, G. (1994b). Theories of cartel stability and the joint executive committee. *RAND Journal of Economics*, 25(1), 37–57.

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

Elster, J. (2009). Social norms and the explanation of behavior. In P. Hedström & P. Bearman (Eds.),

The Oxford Handbook of Analytical Sociology (pp. 195–217). Oxford, UK: Oxford University Press.

Ely, J. C. & Välimäki, J. (2002). A robust Folk Theorem for the Prisoner's Dilemma. *Journal of Economic Theory*, 102(1), 84–105.

Emons, W. (1997). Credence goods and fraudulent experts. *RAND Journal of Economics*, 28(1), 107–119.

Engelmann, D. & Fischbacher, U. (2009). Indirect reciprocity and strategic reputation building in an experimental helping game. *Games and Economic Behavior*, 67(2), 399–407.

Engelmann, D. & Ortmann, A. (2009). *The robustness of laboratory gift exchange: a reconsideration*. Royal holloway mimeograph, University of London, London, UK.

Engelmann, D. & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review*, 94(4), 857–869.

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.

Eroglu, S. (2010). Informal finance and the urban poor: An investigation of Rotating Savings and Credit Associations in Turkey. *Journal of Social Policy*, 39(3), 461–481.

Ertan, A., Page, T., & Putterman, L. (2009). Who to punish? individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, 53(5), 495–511.

Fafchamps, M. (2000). The role of business networks in market development in sub-Saharan Africa. In M. Aoki & Y. Hayami (Eds.), *Communities and Markets in Economic Development* (pp. 186–214). Oxford, UK: Oxford University Press.

Falk, A. (2007). Gift exchange in the field. *Econometrica*, 75(5), 1501–1511.

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, 73(6), 2017–2030.

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness - intentions matter. *Games and Economic Behavior*, 62(1), 287–303.

Falk, A., Fehr, E., & Zehnder, C. (2004). *Reputation and Performance*. Mimeograph, University of Zurich, Zurich, Switzerland.

Falk, A. & Fischbacher, U. (2006). A theory reciprocity. *Games and Economic Behavior*, 54(2), 293–315.

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., Meier, S., & Zehnder, C. (2010). *Did We Overestimate the Role of Social Preferences? The Case of Self-Selected Student Samples*. CESIfo Working Paper 3177, CESIfo, Munich, Germany.

Fandy, M. (2000). Information technology, trust, and social change in the Arab world. *Middle East Journal*, 54(3), 378–394.

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

Fehr, E. (2009b). Social preferences and the brain. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision Making and the Brain* (pp. 215–232). London: Academic Press.

Fehr, E., Bernhard, H., & Rockenbach, B. (2008). Egalitarianism in young children. *Nature*, 454, 1079–1083.

Fehr, E. & Camerer, C. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Science*, 11(10), 419–427.

Fehr, E. & Falk, A. (1999). Wage rigidity in a competitive incomplete contract market. *Journal of Political Economy*, 107(1), 106–134.

Fehr, E. & Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425, 785–791.

Fehr, E. & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Science*, 8(4), 185–190.

Fehr, E. & Fischbacher, U. (2004b). Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2), 63–87.

Fehr, E. & Fischbacher, U. (2005). The economics of strong reciprocity. In H. Gintis, S. Bowles, R. Boyd, & E. Fehr (Eds.), *Moral Sentiments and Material Interests. The Foundations of Cooperation in Economic Life* (pp. 151–192). Cambridge, Massachusetts, USA: MIT Press.

Fehr, E., Fischbacher, U., & Gächter, S. (2002a). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, 13(1), 1–25.

Fehr, E., Fischbacher, U., & Tougareva, E. (2002b). *Do high stakes and competition undermine fairness? Evidence from Russia*. Institute for Empirical Research in Economics Working Paper 120, University of Zurich, Zurich, Switzerland.

Fehr, E., Fischbacher, U., von Rosenbladt, B., Schupp, J., & Wagner, G. G. (2002c). A nation-wide laboratory–examining trust and trustworthiness by integrating behavioral experiments into representative surveys. *Schmollers Jahrbuch*, 122, 519–542.

Fehr, E., Fischbacher, U., von Rosenbladt, B., Schupp, J., & Wagner, G. G. (2003). *A nation-wide laboratory: Examining trust and trustworthiness by integrating behavioral experiments into representative survey*. IZA Discussion Paper Series 715, Forschungsinstitut zur Erforschung der Arbeit, Bonn.

Fehr, E. & Gächter, S. (2000a). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4), 980–994.

Fehr, E. & Gächter, S. (2000b). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 14(3), 159–181.

Fehr, E. & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415, 137–140.

Fehr, E. & Gächter, S. (2005). Human behaviour: Egalitarian motive and altruistic punishment (reply). *Nature*, 433, E1–E2.

Fehr, E., Gächter, S., & Kirchsteiger, G. (1997). Reciprocity as a contract enforcement device: Experimental evidence. *Econometrica*, 65(4), 833–860.

Fehr, E., Goette, L., & Zehnder, C. (2009). A behavioral account of the labor market: The role of fairness concerns. *Annual Review of Economics*, 1, 355–384.

Fehr, E. & Henrich, J. (2003). Is strong reciprocity a maladaptation? In P. Hammerstein (Ed.), *Genetic and Cultural Evolution of Cooperation*, Dahlem Workshop Reports (pp. 55–82). Boston, Massachusetts, USA: MIT Press.

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

Fehr, E., Kirchsteiger, G., & Riedl, A. (1993). Does fairness prevent market clearing? an experimental investigation. *Quarterly Journal of Economics*, 108(2), 437–459.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1998b). Gift exchange and reciprocity in competitive experimental markets. *European Economic Review*, 42(1), 1–34.

Fehr, E. & List, J. A. (2004). The hidden costs and returns of incentives–trust and trustworthiness among CEOs. *Journal of the European Economic Association*, 2(5), 743–771.

Fehr, E. & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3), 817–868.

Fehr, E. & Schmidt, K. M. (2003). Theories of fairness and reciprocity: Evidence and economic applications. In M. Dewatripont, L. P. Hansen, & S. J. Turnovsky (Eds.), *Advances in Economics and Econometrics, Theory and Applications, Eighth World Congress*, volume I (pp. 208–257). Cambridge, Massachusetts: Cambridge University Press.

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.

Fehr, E. & Schneider, F. (2010). Eyes are on us, but nobody cares: Are eye cues relevant for strong reciprocity? *Proceedings of the Royal Society: Biological Sciences*, 277(1686), 1315–1323.

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

Feinberg, R. M. & Husted, T. A. (1993). An experimental test of discount-rate effects on collusive behavior in duopoly markets. *Journal of Industrial Economics*, 41(2), 153–160.

Ferrary, M. (2003). The gift exchange in the social networks of silicon valley. *California Management Review*, 45(4), 120–138.

Fessler, D. M. T. (2002). Windfall and socially distributed willpower: The psychocultural dynamics of Rotating Savings and Credit Associations in a Bengkulu Village. *Ethos*, 30(1-2), 25–48.

Fessler, D. M. T. & Haley, K. J. (2003). The strategy of affect: Emotions in human cooperation. In P. Hammerstein (Ed.), *Genetic and Cultural Evolution of Cooperation*, Dahlem Workshop Reports (pp. 7–36). Boston, Massachusetts, USA: MIT Press.

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

Fischbacher, U. & Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, 100(1), 541–556.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, 71(3), 397–404.

Fisman, R. & Khanna, T. (1999). Is trust a historical residue? information flows and trust levels. *Journal of Economic Behavior & Organization*, 38(1), 79–92.

Fliessbach (2007). Social comparison effects reward-related brain activity in the human ventral striatum. *Science*, 318(5854), 1305–1308.

Flood, M. M. (1952). *Some experimental games*. Research Memorandum RM-789, RAND Corporation, Santa Monica, CA, USA.

Flood, M. M. (1958). Some experimental games. *Management Science*, 5(1), 5–26.

Forsythe, R., Horowitz, J. L., & Savin, N. E. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6, 347–369.

Fowler, J. H., Johnson, T., & Smirnov, O. (2005). Human behaviour: Egalitarian motive and altruistic punishment. *Nature*, 433, E1.

Frank, R. H. (1988). *Passions within reason: the strategic role of the emotions*. New York, NY, USA: Norton.

Frey, B. S. & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, 15(5), 589–611.

Frey, B. S. & Meier, S. (2004). Social comparisons and pro-social behavior: Testing 'conditional cooperation' in a field experiment. *American Economic Review*, 94(5), 1717–1722.

Frey, B. S. & Torgler, B. (2007). Tax morale and conditional cooperation. *Journal of Comparative Economics*, 35(1), 136–159.

Friedman, M. (1962). *Capitalism and Freedeom*. Chicago, IL, USA: Chicago University Press.

Fu, T.-T., Kong, W.-H., & Yang, C. C. (2007). *Monetary stakes and socioeconomic characteristics in ultimatum games: An experiment with nation-wide representative subjects*. Institute of economoics mimeograph, Institute of Economics, Academia Sinica, Nankang, Taiwan.

Fudenberg, D., Levine, D., & Maskin, E. (1994). The Folk Theorem with imperfect public information. *Econometrica*, 62(5), 997–1039.

Fudenberg, D. & Levine, D. K. (1991). An approximate folk theorem with imperfect private information. *Journal of Economic Theory*, 54(1), 26–47.

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

Fudenberg, D. & Maskin, E. (1990). Evolution and cooperation in noisy repeated games. *American Economic Review*, 80(2), 274–279.

Fudenberg, D., Rand, D. G., & Dreber, A. (2010). *Turning the Other Cheek: Leniency and Forgiveness*

in an Uncertain World. Mimeograph, Harvard University, Cambridge, MA, USA.

Fudenberg, D. & Yamamoto, Y. (2011). The folk theorem for irreducible stochastic games with imperfect public monitoring. *Journal of Economic Theory*, 146(4), 1664–1683.

Fukuyama, F. (1995). *Tust: The Social Virtues and the Creation of Prosperity*. New York, New York, USA: Free Press.

Garbarino, E. & Slonim, R. (2009). The robustness of trust and reciprocity across a heterogeneous U.S. population. *Journal of Economic Behavior & Organization*, 69(3), 226–240.

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

Gächter, S. & Herrmann, B. (2009). Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society: Biological Sciences*, 364(1518), 791–806.

Gächter, S. & Herrmann, B. (2010). The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia. *European Economic Review*, 55(2), 193–210.

Gächter, S., Herrmann, B., & Thöni, C. (2004). Trust, voluntary cooperation, and socio-economic background: Survey and experimental evidence. *Journal of Economic Behavior & Organization*, 55(4), 505–531.

Gächter, S., Herrmann, B., & Thöni, C. (2006). Cross-cultural differences in norm enforcement. *Behavioral and Brain Sciences*, 28(6), 822–823.

Gächter, S., Herrmann, B., & Thöni, C. (2010). Culture and cooperation. *Philosophical Transactions of the Royal Society: Biological Sciences*, 365(1553), 2651–2661.

Gächter, S., Nosenzo, D., Renner, E., & Sefton, M. (2012). Who makes a good leader? cooperativeness, optimism, and leading-by-example. *Economic Inquiry*, 50(4), 953–967.

Gächter, S. & Renner, E. (2010). The effects of (incentivized) belief elicitation in public goods experiments. *Experimental Economics*, 13(3), 364–377.

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science*, 322(5907), 1510.

Gelos, R. G. & Wei, S.-J. (2005). Transparency and international portfolio holdings. *Journal of Finance*, 60(6), 2987–3020.

Geraats, P. M. (2002). Central bank transparency. *Economic Journal*, 112(483), F532–F565.

Ghosh, P. & Ray, D. (1996). Cooperation in community interaction without information flows. *Review of Economic Studies*, 63(3), 491–519.

Giannakas, K. (2002). Information asymmetries and consumption decisions in organic food product markets. *Canadian Journal of Agricultural Economics/Revue Canadienne D'Agroeconomie*, 50(1), 35–50.

Gintis, H. (1976). The nature of labor exchange and the theory of capitalist production. *Review of Radical Political Economics*, 8(2), 36–54.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, 206(2), 169–179.

Gintis, H. (2008). Punishment and cooperation. *Science*, 319(5868), 1345–1346.

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.

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring trust. *Quarterly Journal of Economics*, 115(3), 811–846.

Gneezy, U. & List, J. A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, 74(5), 1365–1384.

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.

Goeree, J. K., Holt, C. A., & Laury, S. K. (2002). Private costs and public benefits: Unraveling the effects of altruism and noisy behavior. *Journal of Public Economics*, 83(2), 255–276.

Goeschl, T. & Jarke, J. (2012). *Costly Monitoring and the Emergence of Blind Trust*. Mimeograph, Department of Economics, University of Heidelberg, Heidelberg, Germany.

Golan, E., Kuchler, F., Mitchell, L., Greene, C., & Jessup, A. (2001). Economics of food labeling. *Journal of Consumer Policy*, 24(2), 117–184.

Goldstein, J. & Freeman, J. R. (1990). *Strategic Reciprocity in World Politics*. Chicago, IL, USA: University of Chicago Press.

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25(2), 161–178.

Grechenig, K., Nicklisch, A., & Thöni, C. (2010). Punishment despite reasonable doubt - a public goods experiment with sanctions under uncertainty. *Journal of Empirical Legal Studies*, 7(4), 847–867.

Green, E. J. & Porter, R. H. (1984). Noncooperative collusion under imperfect price information. *Econometrica*, 52(1), 87–100.

Greif, A. (1989). Reputation and coalitions in medieval trade: Evidence on the Maghribi traders. *Journal of Economic History*, 49(4), 857–882.

Greif, A. (1993). Contract enforceability and economic institutions in early trade: The Maghribi traders' coalition. *American Economic Review*, 83(5), 1281–1302.

Greif, A. (1994). Cultural beliefs and the organization of society: A historical and theoretical reflection on collectivist and individualist societies. *Journal of Political Economy*, 102(5), 912–950.

Greif, A. (2006). *Institutions and the path to the modern economy: Lessons from medieval trade*. New York, New York, USA: Cambridge University Press.

Greig, F. & Bohnet, I. (2008). Is there reciprocity in a reciprocal-exchange economy? evidence of gendered norms from a slum in Nairobi. *Economic Inquiry*, 46(1), 77–83.

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.

Güth, W. (1995). On ultimatum bargaining experiments - a personal review. *Journal of Economic Behavior & Organization*, 27(3), 329–344.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4), 367–388.

Güth, W. & Tietz, R. (1990). Ultimatum bargaining behavior: A survey and comparison of experimental results. *Journal of Economic Psychology*, 11(3), 417–449.

Guala, F. & Mittone, L. (2005). Experiments in economics: External validity and the robustness of phenomena. *Journal of Economic Methodology*, 12(4), 495–515.

Guiso, L., Sapienza, P., & Zingales, L. (2004). The role of social capital in financial development. *American Economic Review*, 94(3), 526–556.

Guiso, L., Sapienza, P., & Zingales, L. (2008). Trusting the stock market. *Journal of Finance*, 63(6), 2557–2600.

Haley, K. J. & Fessler, D. M. T. (2005). Nobody's watching? subtle cues affect generosity in an anonymous economic game. *Evolution and Human Beahvior*, 26(3), 245–256.

Hameed, F. (2005). *Fiscal transparency and economic outcomes*. IMF Working Paper WP05/225, International Monetary Fund, Washington, DC, USA.

Hamilton, J. T. (2005). *Regulation through revelation: the origin, politics, and impacts of the Toxics Release Inventory program*. Cambridge, UK: Cambridge University Press.

Hamilton, W. D. (1964a). The genetical evolution of social behaviour. i. *Journal of Theoretical Biology*, 7(1), 1–16.

Hamilton, W. D. (1964b). The genetical evolution of social behaviour. ii. *Journal of Theoretical Biology*, 7(1), 17–52.

Hammerstein, P., Ed. (2003). *Genetic and Cultural Evolution of Cooperation*. Dahlem Workshop Reports. Boston, Massachusetts, USA: MIT Press.

Hannan, R. L., Kagel, J. H., & Moser, D. V. (2002). Partial gift exchange in an experimental labor market: Impact of subject population differences, productivity differences, and effort requests on behavior. *Journal of Labor Economics*, 20(4), 923–951.

Harrington, W. (1988). Enforcement leverage when penalties are restricted. *Journal of Public Economics*, 37(1), 29–53.

Harrison, G. W., Lau, M. I., & Rutström, E. E. (2007). Estimating risk attitudes in Denmark: A field experiment. *Scandinavian Journal of Economics*, 109(2), 341–368.

Harrison, G. W., Lau, M. I., & Williams, M. B. (2002). Estimating individual discount rates in Denmark: A field experiment. *American Economic Review*, 92(5), 1606–1617.

Harsanyi, J. C. & Selten, R. (1988). *A General Theory of Equilibrium Selection in Games*. Cambridge, MA, USA: MIT Press.

Heldt, T. (2008). *Informal sanctions and conditional cooperation: A natural experiment on voluntary contributions to a public good.* Paper presented at the 16th annual conference of the european association of environmental and resource economists, Dalarna University, Falun, Sweden.

Hellwell, J. F. & Putnam, R. D. (2007). Education and social capital. *Eastern Economic Journal*, 33(1), 1–19.

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.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of Homo Economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review*, 91(2), 73–78.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Ensminger, J., Smith Henrich, N., Hill, K., Gil-White, F., Gurven, M., Marlowe, F. W., Patton, J. Q., & Tracer, D. (2005). 'economic man' in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, 28(6), 795–815.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & John, Z. (2006). Costly punishment across human societies. *Science*, 312(5781), 1767–1770.

Henrich, N. & Henrich, J. (2007). *Why humans cooperate: A cultural and evolutionary explanation*. New York, NY, USA: Oxford University Press.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, 319(5868), 1362–1367.

Hirshlifer, J. (1978). Competition, cooperation, and conflict in economics and biology. *American Economic Review*, 68(2), 238–243.

Hirshlifer, J. (1985). Expanding the domain of economics. *American Economic Review*, 75(6), 53–68.

Hirshlifer, J. (1987). On the emotions as guarantors of threats and promises. In J. Dupré (Ed.), *The Latest on the Best: Essays on Evolution and Optimality* (pp. 307–326). Cambridge, MA, USA: MIT Press.

Hobbes, T. (1651). *Leviathan, or, The Matter, Form, and Power of a Common-wealth Ecclesiastical and Civil*. London: Printed for Andrew Crooke, at the Green Dragon in St. Pauls Church-yard. Available online at Early English Books Online (EEBO).

Hoffman, E., McCabe, K. A., & Smith, V. L. (1996). On expectations and the monetary stakes in ultimatum games. *International Journal of Game Theory*, 25(3), 289–301.

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

Holm, H. J. & Danielson, A. (2005). Tropic trust versus nordic trust: Experimental evidence from Tanzania and Sweden. *Economic Journal*, 115(503), 505–532.

Holt, C. A. & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5), 1644–1655.

Hong, K. & Bohnet, I. (2007). Status and distrust: The relevance of inequality and betrayal aversion. *Journal of Economic Psychology*, 28(2), 197–213.

Hopfensitz, A. & Reuben, E. (2009). The importance of emotions for the effectiveness of social punishment. *Economic Journal*, 119(540), 1534–1559.

Hossain, M. (1988). *Credit for alleviation of rural poverty: The Grameen Bank in Bangladesh*. Washington, DC, USA: International Food Policy Research Institute.

Houser, D. & Kurzban, R. (2002). Revisiting kindness and confusion in public goods experiments. *American Economic Review*, 92(4), 1062–1069.

Houser, D., Schunk, D., & Winter, J. (2010). Distinguishing trust from risk: An anatomy of the investment game. *Journal of Economic Behavior & Organization*, 74(1-2), 72–81.

Hörner, J. & Olszewski, W. (2006). The Folk Theorem for games with private almost-perfect monitoring. *Econometrica*, 74(6), 1499–1544.

Hörner, J. & Olszewski, W. (2009). How robust is the folk theorem? *Quarterly Journal of Economics*, 124(4), 1773–1814.

Isaac, R. M. & Walker, J. M. (1988). Group size effects in public goods provision: The voluntary contributions mechanism. *Quarterly Journal of Economics*, 103(1), 179–199.

Isaac, R. M., Walker, J. M., & Williams, A. W. (1994). Group size and the voluntary provision of public goods: Experimental evidence utilizing large groups. *Journal of Public Economics*, 54(1), 1–36.

Jacquet, J., Hauert, C., Traulsen, A., & Milinski, M. (2011). Shame and honour drive cooperation. *Biology Letters*, doi: 10.1098/rsbl.2011.0367.

Jensen, K. (2010). Punishment and spite, the dark side of cooperation. *Philosophical Transactions of the Royal Society: Biological Sciences*, 365(1553), 2635–2650.

Jensen, K., Hare, B., Call, J., & Tomasello, M. (2006). What's in it for me? self-regard precludes altruism and spite in chimpanzees. *Proceedings of the Royal Society: Biological Sciences*, 273(1589), 1013–1021.

Jin, G. Z. & Leslie, P. (2003). The effect of information on product quality: Evidence from restaurant hygiene grade cards. *Quarterly Journal of Economics*, 118(2), 409–451.

Johansson-Stenman, O., Mahmud, M., & Martinsson, P. (2005). Does stake size matter in trust games? *Economics Letters*, 88(3), 365–369.

Johnson, T., Dawes, C. T., Fowler, J. H., McElreath, R., & Smirnov, O. (2009). The role of egalitarian motives in altruistic punishment. *Economics Letters*, 102(3), 192–194.

Jøsang, A., Ismail, R., & Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2), 618–644.

Kagel, J. H. & Roth, A. E., Eds. (2011). *The Handbook of Experimental Economics*. Princeton: Princeton University Press, 2nd edition.

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

Kandori, M. (1992a). Social norms and community enforcement. *Review of Economic Studies*, 59(1), 63–80.

Kandori, M. (1992b). The use of information in repeated games with imperfect monitoring. *Review of Economic Studies*, 59(3), 581–593.

Kandori, M. (2002). Introduction to repeated games with private monitoring. *Journal of Economic Theory*, 102(1), 1–15.

Kandori, M. (2011). Weakly belief-free equilibria in repeated games with private monitoring. *Econometrica*, 79(3), 877–892.

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.

Kappeler, P. M. & van Schaik, C. P., Eds. (2006). *Cooperation in Primates and Humans. Mechanisms and Evolution*. Berlin / Heidelberg, Germany: Springer.

Kawagoe, T. (2000). Middlemen in a peasant community: Vegetable marketing in Indonesia. In M. Aoki & Y. Hayami (Eds.), *Communities and Markets in Economic Development* (pp. 129–152). Oxford, UK: Oxford University Press.

Kelley, H. H. & Stahelski, A. J. (1970). Social interaction basis of cooperators' and competitors' beliefs about others. *Journal of Personality and Social Psychology*, 16(1), 66–91.

Kennedy, D. & Norman, C. (2005). What don't we know? *Science*, 309(5731), 75.

Keser, C. & van Winden, F. (2000). Conditional cooperation and voluntary contributions to public goods. *Scandinavian Journal of Economics*, 102(1), 23–39.

Kikuchi, M., Fujita, M., & Hayami, Y. (2000). State, community, and market in the deterioration of a national irrigation system in the Philippines. In M. Aoki & Y. Hayami (Eds.), *Communities and Markets in Economic Development* (pp. 265–294). Oxford, UK: Oxford University Press.

Kirchgässner, G. (2008). *Homo Oeconomicus: The Economic Model of Behaviour and its Applications in Economics and other Social Sciences*. New York, NY, USA: Springer.

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.

Kirchsteiger, G. (1994). The role of envy in ultimatum games. *Journal of Economic Behavior & Organization*, 25(3), 373–389.

Klein, E. (1971). *A Comprehensive Etymological Dictionary of the English Language*. Amsterdam, The Netherlands: Elsevier.

Klemperer, P. (1995). Competition when consumers have switching costs: An overview with applications to industrial organization, macroeconomics, and international trade. *Review of Economic Studies*, 62(4), 515–539.

Knack, S. & Keefer, P. (1997). Does social capital have an economic payoff? a cross-country investigation. *Quarterly Journal of Economics*, 112(4), 1251–1288.

Knauft, B. M. (1991). Violence and sociality in human evolution. *Current Anthropology*, 32(4), 391–428.

Knoch, D., Gianotti, L. R. R., Baumgartner, T., & Fehr, E. (2010). A neural marker of costly punishment behavior. *Psychological Science*, 21(3), 337–342.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, 314(5800), 829–832.

Kocher, M. G. (2008). *Robustness of conditional cooperation in public goods experiments*. Mimeograph, University of Munich, Munich, Germany.

Kolstad, I. & Wiig, A. (2009). Is transparency the key to reducing corruption in resource-rich countries? *World Development*, 37(3), 521–532.

Krebs, D. L. (1970). Altruism: An examination of the concept and a review of the literature. *Psychological Bulletin*, 73(4), 258–302.

Kreps, D. M. (1990). Corporate culture and economic theory. In J. E. Alt & K. A. Shepsle (Eds.), *Perspectives on Positive Political Economy* (pp. 90–143). Cambridge, UK: Cambridge University Press.

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.

Kreps, D. M. & Wilson, R. (1982). Reputation and imperfect information. *Journal of Economic Theory*, 27(2), 253–279.

Kube, S., Maréchal, M. A., & Puppe, C. (2006). *Putting reciprocity to work – Positive versus negative responses in the field*. Department of Economics Discussion Paper 2006-27, University of St. Gallen, St. Gallen, Switzerland.

Kube, S., Maréchal, M. A., & Puppe, C. (2011a). *The currency of reciprocity – Gift-exchange in the workplace*. Institute for Empirical Research in Economics Working Paper 377, University of Zurich, Zurich, Switzerland.

Kube, S., Maréchal, M. A., & Puppe, C. (2011b). *Do Wage Cuts Damage Work Morale? Evidence from a Natural Field Experiment*. Institute for Empirical Research in Economics Working Paper 471, University of Zurich, Zurich, Switzerland.

Kusakawa, T., Ogawa, K., & Shichijo, T. (2012). An experimental investigation of a third-person enforcement in a prisoner's dilemma game. *Economics Letters*, 117(3), 704–707.

La Porta, R., Lopez-De-Silanes, F., Shleifer, A., & Vishny, R. W. (1997). Trust in large organizations. *American Economic Review*, 87(2), 333–338.

Laury, S. K. (2005). *Pay one or pay all: Random selection of one choice for payment*. Andrew Young

School Policy Studies Research Paper 06-13, Department of Economics, Georgia State University, Atlanta.

Lave, L. B. (1962). An empirical approach to the prisoners' dilemma game. *Quarterly Journal of Economics*, 76(3), 424–436.

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

Ledyard, J. O. (1995). Public goods: A survey of experimental research. In J. H. Kagel & A. E. Roth (Eds.), *The Handbook of Experimental Economics* (pp. 111–194). Princeton: Princeton University Press.

Lee, L.-F. & Porter, R. H. (1984). Switching regression models with imperfect sample separation information–with an application on cartel stability. *Econometrica*, 52(2), 391–418.

Lehmann, L. & Keller, L. (2006). The evolution of cooperation and altruism - a general framework and a classification of models. *Journal of Evolutionary Biology*, 19(5), 1365–1376.

Leibbrandt, A. & López-Pérez, R. (2009). *An exploration of third and second party punishment in ten simple games*. Mimeo, Universidad Autonoma de Madrid, Madrid, Spain.

Leibbrandt, A. & López-Pérez, R. (2011). *Individual heterogeneity in punishment and reward*. Economic Analysis Working Paper Series 1/2011, Universidad Autonoma de Madrid, Madrid, Spain.

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.

List, J. A. (2006). The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. *Journal of Political Economy*, 114(1), 1–37.

List, J. A. & Cherry, T. L. (2008). Examining the role of fairness in high stakes allocation decisions. *Journal of Economic Behavior & Organization*, 65(1), 1–8.

List, J. A., Sadoff, S., & Wagner, M. (2011). So you want to run an experiment, now what? some simple rules of thumb for optimal experimental design. *Experimental Economics*, 14(4), 439–457.

Loewenstein, G. F., Thompson, L., & Bazerman, M. H. (1989). Social utility and decision making in interpersonal contexts. *Journal of Personality and Social Psychology*, 57(3), 426–441.

López-Pérez, R. (2007). Introducing social norms in game theory. In A. Innocenti & P. Sbiglia (Eds.), *Games, Rationality and Behavior. Essays in Behavioral Game Theory and Experiments* (pp. 26–46). Houndmills, England: Palgrave Macmillan.

López-Pérez, R. (2008). Aversion to norm-breaking: A model. *Games and Economic Behavior*, 64(1), 237–267.

López-Pérez, R. (2010). Guilt and shame: An axiomatic analysis. *Theory and Decision*, 69(4), 569–586.

López-Pérez, R. & Kiss, H. J. (2012). Do people accurately anticipate sanctions? *Southern Economic Journal*, forthcoming.

MacLeod, W. B. & Malcomson, J. M. (1989). Implicit contracts, incentive compatibility, and involuntary unemployment. *Econometrica*, 57(2), 447–480.

MacLeod, W. B. & Malcomson, J. M. (1998). Motivation and markets. *American Economic Review*, 88(3), 388–411.

Mailath, G. J. & Morris, S. (2002). Repeated games with almost-public monitoring. *Journal of Economic Theory*, 102(1), 189–228.

Manninen, J. (2006). Anders Chydenius and the origins of world's first freedom of information act. In J. Mustonen (Ed.), *The World's First Freedom of Information Act* (pp. 18–53). Kokkola, Finland: Anders Chydenius Foundation.

Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Ensminger, J., Gurven, M., Gwako, E., Henrich, J., Henrich, N., Lesorogol, C., McElreath, R., & Tracer, D. (2008). More 'altruistic' punishment in larger societies. *Proceedings of the Royal Society: Biological*

Sciences, 275(1634), 587–592.

Martin, R. & Randal, J. (2008). How is donation behaviour affected by the donations of others? *Journal of Economic Behavior & Organization*, 67(1), 228–238.

Marwell, G. & Ames, R. E. (1979). Experiments on the provision of public goods. i. resources, interest, group size, and the free-rider problem. *American Journal of Sociology*, 84(6), 1335–1360.

Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic theory*. New York, New York, USA: Oxford University Press.

Masclet, D., Noussair, C., Tucker, S., & Villeval, M.-C. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. *American Economic Review*, 93(1), 366–380.

Masclet, D. & Villeval, M.-C. (2008). Punishment, inequality, and welfare: A public good experiment. *Social Choice and Welfare*, 31(3), 475–502.

Matsushima, H. (2004). Repeated games with private monitoring: Two players. *Econometrica*, 72(3), 823–852.

Maximiano, S., Sloof, R., & Sonnemans, J. (2007). Gift exchange in a multi-worker firm. *Economic Journal*, 117(522), 1025–1050.

Mazur, J. E. (2002). *Learning and behavior*. Upper Saddle River, New Jersey: Prentice Hall, 5th edition.

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods*, 2(1), 3–19.

Meier, S. (2005). Does framing matter for conditional cooperation? evidence from a natural field experiment. *B.E. Journal of Economic Analysis & Policy*, 5(2), Article 1.

Meier, S. (2007). A survey of economic theories and field evidence on pro-social behavior. In B. S. Frey & A. Stutzer (Eds.), *Economics and Psychology. A Promising New Cross-Disciplinary Field* (pp. 51–88). Cambridge, Massachusetts, USA: MIT Press.

Milgrom, P. & Roberts, J. (1982). Predation, reputation, and entry deterrence. *Journal of Economic Theory*, 27(2), 280–312.

Milgrom, P. R., North, D. C., & Weingast, B. R. (1990). The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs. *Economics & Politics*, 2(1), 1–23.

Milinski, M. & Rockenbach, B. (2007). Spying on others evolves. *Science*, 317(5837), 464–465.

Milinski, M., Semmann, D., & Krambeck, H. (2002a). Donors to charity gain in both indirect reciprocity and political reputation. *Proceedings of the Royal Society: Biological Sciences*, 269(1494), 881–883.

Milinski, M., Semmann, D., & Krambeck, H.-J. (2002b). Reputation helps solve the 'tragedy of the commons'. *Nature*, 415, 424–426.

Miller, A. S. & Mitamura, T. (2003). Are surveys on trust trustworthy? *Social Psychology Quarterly*, 66(1), 62–70.

Miller, J. H. (1996). The coevolution of automata in the repeated prisoner's dilemma. *Journal of Economic Behavior & Organization*, 29(1), 87–112.

Monin, B. (2007). Holier than me? threatening social comparison in the moral domain. *Revue Internationale de Psychologie Sociale*, 20(1), 53–68.

Morduch, J. & Sicular, T. (2000). Risk and insurance in transition: Perspectives from Zouping County, China. In M. Aoki & Y. Hayami (Eds.), *Communities and Markets in Economic Development* (pp. 215–246). Oxford, UK: Oxford University Press.

Mueller, D. C. (2003). *Public choice*, volume III. Cambridge, UK: Cambridge University Press.

Mui, V.-L. (1995). The economics of envy. *Journal of Economic Behavior & Organization*, 26(3), 311–336.

Muller, L., Sefton, M., Steinberg, R., & Vesterlund, L. (2008). Strategic behavior and learning in repeated voluntary contribution experiments. *Journal of Economic Behavior & Organization*, 67(3-4), 782–793.

Munier, B. & Zaharia, C. (2002). High stakes and acceptance behavior in ultimatum bargaining: A contribution from an international experiment. *Theory and Decision*, 53(3), 187–207.

Murnighan, J. K. & Roth, A. E. (1983). Expecting continued play in prisoner's dilemma games–a test of several models. *Journal of Conflict Resolution*, 27(2), 279–300.

Naef, M. & Schupp, J. (2009a). *Can we trust the trust game? A comprehensive examination*. Department of Economics Discussion Paper 2009-05, Royal Halloway College, University of London, London, UK.

Naef, M. & Schupp, J. (2009b). *Measuring Trust: Experiments and surveys in contrast and combination*. SEOP Papers on Multidisciplinary Panel Data Research 167, Deutsches Institut für Wirtschaftsforschung, Berlin, Germany.

Nelson, P. (1970). Information and consumer behavior. *Journal of Political Economy*, 78(2), 311–329.

Neugebauer, T., Perote, J., Schmidt, U., & Loos, M. (2009). Selfish-biased conditional cooperation: On the decline of contributions in repeated public goods experiments. *Journal of Economic Psychology*, 30(1), 52–60.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: can we really govern ourselves? *Journal of Public Economics*, 92(1-2), 91–112.

Nikiforakis, N. (2010). Feedback, punishment and cooperation in public good experiments. *Games and Economic Behavior*, 68(2), 689–702.

Nikiforakis, N. & Normann, H.-T. (2008). A comparative statics analysis of punishment in public-good experiments. *Experimental Economics*, 11(4), 358–369.

Normann, H.-T. & Wallace, B. (2012). The impact of the termination rule on cooperation in a prisoner's dilemma experiment. *International Journal of Game Theory*, 41(3), 707–718.

Noussair, C. & Tucker, S. (2005). Combining monetary and social sanctions to promote cooperation. *Economic Inquiry*, 43(3), 649–660.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, 314(5805), 1560–1563.

Nowak, M. A. & Sigmund, K. (1992). Tit for tat in heterogeneous populations. *Nature*, 355, 250–253.

Nowak, M. A. & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, 364, 56–58.

Nowak, M. A. & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, 393, 573–577.

Nowak, M. A. & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437, 1291–1298.

Ohno, A. (2000). Market integrators for rural-based industrialization: The case of the hand-weaving industry in Laos. In M. Aoki & Y. Hayami (Eds.), *Communities and Markets in Economic Development* (pp. 153–185). Oxford, UK: Oxford University Press.

Okun, A. M. (1981). *Prices & quantities: A macroeconomic analysis*. Washington, D.C., USA: The Brookings Institution.

Okuno-Fujiwara, M. & Postlewaite, A. (1995). Social norms and random matching games. *Games and Economic Behavior*, 9(1), 79–109.

Olson, Jr, M. (1965). *The logic of collective action: Public goods and the theory of groups*. Cambridge, Massachusetts, USA: Harvard University Press.

Ones, U. & Putterman, L. (2007). The ecology of collective action: A public goods and sanctions experiment with controlled group formation. *Journal of Economic Behavior & Organization*, 62(4), 495–521.

Onions, C. T., Ed. (1966). *The Oxford Dictionary of English Etymology*. Oxford, UK: Oxford University Press.

Oosterbeek, H., Sloof, R., & van de Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7(2), 171–188.

Orbell, J. M. & Dawes, R. M. (1993). Social welfare, cooperators' advantage, and the option of not playing the game. *American Sociological Review*, 58(6), 787–800.

Osterloh, M., Rota, S., & Kuster, B. (2008). *Open Source Software Production: Climbing on the Shoulders of Giants*. Institute for research in business administration mimeograph, University of Zurich, Zurich, Switzerland.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cam-

bridge, UK: Cambridge University Press.

Ostrom, E. & Gardner, R. (1993). Managing local commons: Theoretical issues in incentive design. *Journal of Economic Perspectives*, 7(4), 113–134.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: self-governance is possible. *American Political Science Review*, 86(2), 404–417.

Otsuka, K. & Tachibana, T. (2000). Evolution and consequences of community forest management in the Hill Region of Nepal. In M. Aoki & Y. Hayami (Eds.), *Communities and Markets in Economic Development* (pp. 295–317). Oxford, UK: Oxford University Press.

Ottone, S., Ponzano, F., & Zarri, L. (2008). *Moral sentiments and material interests behind altruistic third-party punishment*. Department of Economics Working Paper 48, University of Verona, Verona, Italy.

Page, T., Putterman, L., & Unel, B. (2005). Voluntary association in public goods experiments: Reciprocity, mimicry and efficiency. *Economic Journal*, 115(506), 1032–1053.

Palfrey, T. R. & Prisbrey, J. E. (1997). Anomalous behavior in public goods experiments: How much and why? *American Economic Review*, 87(5), 829–846.

Palfrey, T. R. & Rosenthal, H. (1994). Repeated play, cooperation and coordination: An experimental study. *Review of Economic Studies*, 61(3), 545–565.

Palmer, C. T. (1991). Kin-selection, reciprocal altruism, and information sharing among Maine lobstermen. *Ethology and Sociobiology*, 12(3), 221–235.

Parco, J. E., Rapoport, A., & Stein, W. E. (2002). Effects of financial incentives on the breakdown of mutual trust. *Psychological Science*, 13(3), 292–297.

Patel, A., Cartwright, E., & van Vugt, M. (2010). *Punishment cannot sustain cooperation in a public good game with free-rider anonymity*. Discussion Paper in Economics 451, University of Gothenburg, Gothenburg.

Peisakhin, L. V. (2012). Transparency and corruption: Evidence from India. *Journal of Law and Economics*, forthcoming.

Pencavel, J. (2002). *Worker participation: Lessons from the worker co-ops of the Pacific Northwest*. New York, NY, USA: Russell Sage Foundation.

Pennisi, E. (2005). How did cooperative behavior evolve? *Science*, 309(5731), 93.

Percival, R. V., Schroeder, C. H., Miller, A. S., & Leape, J. P. (2009). *Environmental Regulation: Law, Science, and Policy*. New York, NY, USA: Aspen Publishers, 6th edition.

Pereira, P. T., Silva, N., & Andrade e Silva, J. (2006). Positive and negative reciprocity in the labor market. *Journal of Economic Behavior & Organization*, 59(3), 406–422.

Piccione, M. (2002). The repeated Prisoner's Dilemma with imperfect private monitoring. *Journal of Economic Theory*, 102(1), 70–83.

Picot, A., Ripperger, T., & Wolff, B. (1996). The fading boundaries of the firm: The role of information and communication technology. *Journal of Institutional and Theoretical Economics*, 152(1), 65–79.

Pigou, A. C. (1932). *The Economics of Welfare*. New Jersey, NJ, USA: Macmillan and Company.

Pillutla, M. M. & Murnighan, J. K. (1997). Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational Behavior and Human Decision Processes*, 68(3), 208–224.

Pinker, S. (2011). *The Better Angels of Our Nature*. London, UK: Allen Lane.

Platteau, J.-P. & Seki, E. (2000). Community arrangements to overcome market failures: Pooling groups in Japanese fisheries. In M. Aoki & Y. Hayami (Eds.), *Communities and Markets in Economic Development* (pp. 344–402). Oxford, UK: Oxford University Press.

Porter, R. H. (1983a). Optimal cartel trigger price strategies. *Journal of Economic Theory*, 29(2), 313–338.

Porter, R. H. (1983b). A study of cartel stability: The joint executive committee, 1880–1886. *Bell Journal of Economics*, 14(2), 301–314.

Porter, R. H. (1985). On the incidence and duration of price wars. *Journal of Industrial Economics*, 33(4), 415–426.

Powell, R. A., Symbaluk, D. G., & MacDonald, S. E. (2002). *Introduction to learning and behavior*.

Belmont, California: Wadsworth.

Prosser, W. L. (1971). *Handbook of the Law of Torts*. Eagan, MN, USA: West Publishing Co., 4th edition.

Putnam, R. D. (1993a). *Making democracy work: Civic traditions in modern Italy*. Princeton: Princeton University Press.

Putnam, R. D. (1993b). The prosperous community: Social capital and public life. *American Prospect*, 13(1995), 65–78.

Rapoport, A. (1974). The prisoner's dilemma–recollections and observations. In A. Rapoport (Ed.), *Game Theory as a Theory of Conflict Resolution* (pp. 17–34). Dordrecht, The Netherlands: Reidel Publishing.

Rapoport, A. & Chammah, A. M. (1965). *Prisoner's dilemma: A study in conflict and cooperation*. Ann Arbor, Michigan, USA: University of Michigan Press.

Rege, M. & Telle, K. (2004). The impact of social approval and framing on cooperation in public good situations. *Journal of Public Economics*, 88(7-8), 1625–1644.

Reinikka, R. & Svensson, J. (2005). Fighting corruption to improve schooling: Evidence from a newspaper campaign in Uganda. *Journal of the European Economic Association*, 3(2-3), 259–267.

Renner, E. & Tyran, J.-R. (2004). Price rigidity in consumer markets. *Journal of Economic Behavior & Organization*, 55(4), 575–593.

Resnick, P. & Zeckhauser, R. (2002). Trust among strangers in internet transactions: Empirical analysis of eBay's reputation system. In M. R. Baye (Ed.), *The Economics of the Internet and E-commerce*, Advances in Applied Microeconomics (pp. 127–157). Bingley, UK: Emerald Group Publishing Limited.

Resnick, P., Zeckhauser, R., Swanson, J., & Lockwood, K. (2006). The value of reputation on eBay: A controlled experiment. *Experimental Economics*, 9(2), 79–101.

Reuben, E. & Riedl, A. (2009). Public goods provision and sanctioning in privileged groups. *Journal of Conflict Resolution*, 53(1), 72–93.

Reuben, E. & van Winden, F. (2010). Fairness perceptions and prosocial emotions in the power to take. *Journal of Economic Psychology*, 31(6), 908–922.

Rigdon, M., Ishii, K., Watabe, M., & Kitayama, S. (2009). Minimal social cues in the dictator game. *Journal of Economic Psychology*, 30(3), 358–367.

Rigdon, M. L. (2002). Efficiency wages in an experimental labor market. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20), 13348–13351.

Roberts, R. D. (1984). A positive model of private charity and public transfers. *Journal of Political Economy*, 92(1), 136–148.

Rosenthal, R. W. (1979). Sequences of games with varying opponents. *Econometrica*, 47(6), 1353–1366.

Roth, A. E. (1995). Introduction to experimental economics. In J. H. Kagel & A. E. Roth (Eds.), *The Handbook of Experimental Economics* (pp. 3–110). Princeton: Princeton University Press.

Roth, A. E. & Murnighan, J. K. (1978). Equilibrium behavior and repeated play of the prisoner's dilemma. *Journal of Mathematical Psychology*, 17(2), 189–198.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, 2(3), 808–849.

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

Russell, C. S., Harrington, W., & Vaughan, W. J. (1986). *Enforcing pollution control laws*. Washington, DC, USA: Resources for the Future.

Ryssel, R., Ritter, T., & Gemünden, G. (2004). The impact of information technology deployment on trust, commitment and value creation in business relationships. *Journal of Business & Industrial Marketing*, 19(3), 197–207.

Sahlins, M. (1972). *Stone Age Economics*. Chicago, IL, USA: Aldine-Atherton.

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

Samuelson, P. A. (1954). The pure theory of public expenditure. *Review of Economics and Statistics*, 36(4), 387–389.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, 300(5626), 1755–1758.

Sapienza, P., Toldra, A., & Zingales, L. (2007). *Understanding Trust*. NBER Working Paper 13387, National Bureau of Economic Research, Cambridge, Massachusetts.

Schachtman, T. R. & Reilly, S. S., Eds. (2011). *Associative Learning and Conditioning Theory: Human and Non-Human Applications*. New York, New York, USA: Oxford University Press.

Schelling, T. (1978). *Micromotives and macrobehavior*. New York, New York, USA: W. W. Norton & Company.

Schram, A. (2005). Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology*, 12(2), 225–237.

Schram, A. & Charness, G. (2011). *Social and moral norms in the laboratory*. Creed mimeograph, Amsterdam School of Economics, Amsterdam, The Netherlands.

Schwartz, S. T., Young, R. A., & Zvinakis, K. (2000). Reputation without repeated interaction: A role for public disclosures. *Review of Accounting Studies*, 5(4), 351–375.

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

Seabright, P. (2010). *The Company of Strangers: A Natural History of Economic Life*. Princeton, NJ, USA: Princeton University Press, revised edition.

Sefton, M., Shupp, R., & Walker, J. M. (2007). The effect of rewards and sanctions in provision of public goods. *Economic Inquiry*, 45(4), 671–690.

Seinen, I. & Schram, A. (2006). Social status and group norms: Indirect reciprocity in a repeated helping experiment. *European Economic Review*, 50(3), 581–602.

Sekiguchi, T. (1997). Efficiency in repeated Prisoner's Dilemma with private monitoring. *Journal of Economic Theory*, 76(2), 345–361.

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

Selten, R. (1967). Die strategiemethode zur erforschung des eingeschränkt rationalen verhaltens im rahmen eines oligopolexperiments. In H. Sauermann (Ed.), *Beiträge zur experimentellen Wirtschaftsforschung*, volume I (pp. 136–168). Tübingen: Mohr Siebeck.

Selten, R. (1978). The chain store paradox. *Theory and Decision*, 9(2), 127–159.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1(1), 43–62.

Selten, R. & Hammerstein, P. (1984). Gaps in Harley's argument on evolutionarily stable learning rules and in the logic of "tit for tat". *Behavioral and Brain Sciences*, 7(1), 115–116.

Selten, R. & Stoecker, R. (1986). End behavior in sequences of finite prisoner's dilemma supergames–a learning theory approach. *Journal of Economic Behavior & Organization*, 7(1), 47–70.

Semmann, D., Krambeck, H.-J., & Milinski, M. (2005). Reputation is valuable within and outside one's own social group. *Behavioral Ecology and Sociobiology*, 57(6), 611–616.

Seymour, B., Singer, T., & Dolan, R. (2007). The neurobiology of punishment. *Nature Reviews Neuroscience*, 8, 300–311.

Shabman, L. & Stephenson, K. (1994). A critique of the self-interested voter model: The case of a local single issue referendum. *Journal of Economic Issues*, 28(4), 1173–1186.

Shang, J. & Croson, R. (2009). A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods. *Economic Journal*, 119(540), 1422–1439.

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

Shinada, M. & Yamagishi, T. (2007). Punishing free riders: Direct and indirect promotion of cooperation. *Evolution and Human Behavior*, 28(5), 330–339.

Siang, C., Requate, T., & Waichman, I. (2011). On the role of social wage comparisons in gift-exchange experiments. *Economics Letters*, 112(1), 75–78.

Slonim, R. & Garbarino, E. (2008). Increases in trust and altruism from partner selection: Experimental evidence. *Experimental Economics*, 11(2), 134–153.

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

Slonim, R. & Roth, A. E. (1998). Learning in high stakes Ultimatum Games: An experiment in the Slovak Republic. *Econometrica*, 66(3), 569–596.

Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*. London, England: W. Strahan and T. Cadell.

Sobel, J. (2005). Interdependent preferences and reciprocity. *Journal of Economic Literature*, 43(2), 392–436.

Sober, E. & Wilson, D. S. (1998). *Unto others: The evolution and psychology of unselfish behavior*. Cambridge, Massachusetts, USA: Harvard University Press.

Sonnemans, J., Schram, A., & Offerman, T. (1999). Strategic behavior in public good games: When partners drift apart. *Economics Letters*, 62(1), 35–41.

Stanca, L. (2009). Measuring indirect reciprocity: Whose back do we scratch? *Journal of Economic Psychology*, 30(2), 190–202.

Stanca, L. (2010). How to be kind? outcomes versus intentions as determinants of fairness. *Economics Letters*, 106(1), 19–21.

Stephan, M. (2002). Environmental information disclosure programs: They work, but why? *Social Science Quarterly*, 83(1), 190–205.

Stigler, G. J. (1964). A theory of oligopoly. *Journal of Political Economy*, 72(1), 44–61.

Sugden, R. (1986). *The Economics of Rights, Co-Operation, and Welfare*. Oxford, UK: Basil Blackwell.

Suleiman, R. (1996). Expectations and fairness in a modified ultimatum game. *Journal of Economic Psychology*, 17(5), 531–554.

Sunstein, C. R. (1999). Informational regulation and informational standing: Akins and beyond. *University of Pennsylvania Law Review*, 147(3), 613–675.

Sutter, M. (2007). Outcomes versus intentions: On the nature of fair behavior and its development with age. *Journal of Economic Psychology*, 28(1), 69–78.

Takahashi, S. (2010). Community enforcement when players observe partners' past play. *Journal of Economic Theory*, 145(1), 42–62.

Tangney, J. P. (1999). The self-conscious emotions: Shame, guilt, embarrassment and pride. In T. Dalgleish & M. J. Power (Eds.), *Handbook of Cognition and Emotion* (pp. 541–568). Chichester, England: John Wiley & Sons.

Tangney, J. P. & Dearing, R. L. (2003). *Shame and Guilt*. New York, New York, USA / London, UK: Guilford Press.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology*, 58, 345–372.

Thöni, C. (2011). *Inequality aversion and antisocial punishment*. Department of Economics Discussion Paper 2011-11, University of St. Gallen, St. Gallen, Switzerland.

Thöni, C., Tyran, J.-R., & Wengström, E. (2012). Microfoundations of social capital. *Journal of Public Economics*, 96(7-8), 635–643.

Tietenberg, T. (1998). Disclosure strategies for pollution control. *Environmental and Resource Economics*, 11(3-4), 587–602.

Tietenberg, T. & Wheeler, D. (2001). Empowering the community: Information strategies for pollution control. In H. Folmer, H. L. Gabel, S. Gerking, & A. Rose (Eds.), *Frontiers in Environmental Economics* (pp. 85–120). Cheltenham, UK: Edward Elgar Publishing.

Tilly, C. (1981). Charivaris, repertoires and urban politics. In J. M. Merriman (Ed.), *French Cities in*

the Nineteenth Century (pp. 73–91). New York, NY, USA: Holmes & Meier.

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für Tierpsychologie*, 20(4), 410–433.

Titmuss, R. M. (1970). *The Gift Relationship*. London, UK: Allen & Unwin.

Torgler, B., Frey, B. S., & Wilson, C. (2009). Environmental and pro-social norms: Evidence on littering. *B.E. Journal of Economic Analysis & Policy*, 9(1), 1–39.

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

van Damme, E. (1983). *Refinements of the Nash Equilibrium Concept*. Berlin/Heidelberg, Germany: Springer.

van Damme, E. (1998). On the state of the art in game theory: an interview with Robert Aumann. *Games and Economic Behavior*, 24(2), 181–210.

van Prooijen, J.-W., Gallucci, M., & Toeset, G. (2008). Procedural justice in punishment systems: Inconsistent punishment procedures have detrimental effects on cooperation. *British Journal of Psychology*, 47(2), 311–324.

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.

Vasconcelos, H. (2004). Entry effects on cartel stability and the joint executive committee. *Review of Industrial Organization*, 24(3), 219–241.

Vega-Redondo, F. (2003). *Economics and the Theory of Games*. Cambridge, UK: Cambridge University Press.

Walker, J. M. & Halloran, M. A. (2004). Rewards and sanctions and the provision of public goods in one-shot settings. *Experimental Economics*, 7(3), 235–247.

Wedekind, C. & Milinski, M. (2000). Cooperation through image scoring in humans. *Science*, 288(5467), 850–852.

Weibull, J. W. (1996). *Evolutionary Game Theory*. Cambridge, MA, USA: MIT Press.

Weimann, J. (1994). Individual behaviour in a free riding experiment. *Journal of Public Economics*, 54(2), 185–200.

West, S. A., Griffin, A. S., & Gardner, A. (2007a). Evolutionary explanations of cooperation. *Current Biology*, 17(16), R661–R672.

West, S. A., Griffin, A. S., & Gardner, A. (2007b). Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology*, 20(2), 415–432.

Whitford, W. C. (1973). The functions of disclosure regulation in consumer transactions. *Wisconsin Law Review*, 400, 400–470.

Wiepking, P. & Heijnen, M. (2011). The giving standard: Conditional cooperation in the case of charitable giving. *International Journal of Nonprofit and Voluntary Sector Marketing*, 16(1), 13–22.

Williams, G. C. (1966). *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton, New Jersey, USA: Princeton University Press.

Willinger, M., Keser, C., Lohmann, C., & Usunier, J.-C. (2003). A comparison of trust and reciprocity between France and Germany: Experimental investigation based on the investment game. *Journal of Economic Psychology*, 24(4), 447–466.

Willinger, M. & Ziegelmeyer, A. (2001). Strength of the social dilemma in a public goods experiment: An exploration of the error hypothesis. *Experimental Economics*, 4(2), 131–144.

Wilson, R. (1985). Reputation in games and markets. In A. E. Roth (Ed.), *Game-Theoretic Models of Bargaining* (pp. 27–62). Cambridge, UK: Cambridge University Press.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, 51(1), 110–116.

Yamagishi, T., Kikuchi, M., & Kosugi, M. (2002). Trust, gullibility, and social intelligence. *Asian Journal of Social Psychology*, 2(1), 145–161.

Yamamoto, Y. (2007). Efficiency results in n player games with imperfect private monitoring. *Journal*

of Economic Theory, 135(1), 382–413.

Zak, P. J. & Knack, S. (2001). Trust and growth. *Economic Journal*, 111(470), 295–321.

Zelmer, J. (2003). Linear public goods experiments: A meta-analysis. *Experimental Economics*, 6(3), 299–310.

Zizzo, D. J. (2003). Money burning and rank egalitarianism with random dictators. *Economics Letters*, 81(2), 263–266.

Zizzo, D. J. (2004). Inequality and procedural fairness in a money burning and stealing experiment. In F. Cowell (Ed.), *Inequality, Welfare and Income Distribution: Experimental Approaches*, Research on Economic Inequality (pp. 215–247). Bingley, UK: Emerald Group Publishing Limited.

Zizzo, D. J. & Oswald, A. J. (2001). Are people willing to pay to reduce others' incomes? *Annales D'Économie et de Statistique*, 63/64, 39–65.