Department of Economics

On Sanctions and Signals

How Formal and Informal Mechanisms Produce Compliance

Joël van der Weele

Thesis submitted for assessment with a view to obtaining the degree of
Doctor of Economics of the European University Institute

Florence, September 2009

EUROPEAN UNIVERSITY INSTITUTE
Department of Economics

On Sanctions and signals

How Formal and Informal Mechanisms Produce Compliance

Joël van der Weele

Thesis submitted for assessment with a view to obtaining the degree of
Doctor of Economics of the European University Institute

Jury Members:

Professor Frederick van der Ploeg, University of Oxford, Supervisor
Professor Fernando Vega-Redondo, EUI
Professor Steffen Huck, University College London
Professor Tore Ellingsen, Stockholm School of Economics

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

"Reward and punishment is the lowest form of education."

Zhuangzi

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Abstract

Why people comply with rules, why they contribute to public goods and why they behave prosocially in general is a fundamental question of social science. In the tradition of Gary Becker and the chicago school, economists have traditionally considered punishment by the authorities as the main or sole reason why people would comply with the law and contribute to public goods. In this thesis I argue that this model is importantly incomplete and leads to lopsided or even mistaken policy advice. I stress the importance of social interactions between agents and apply game theoric examples to show how the standard model can be enriched.

In the second chapter, I survey the empirical literature, both experimental and econometric, on the deterrence literature. From this review I conclude that the literature does not demonstrate a robust effect of deterrence. I then review theoretical work in which sanctions interact with social norms or long-term processes of preferences formation. In such models deterrence often does not have the straightforward effect that it has in standard theory. The chapter concludes with an example: a model of crime in neighborhoods where signaling is important. I show that in this case, the threat of police violence may be counterproductive on its own, but can be useful in combination with other, softer approaches.

The third chapter departs from the fact that the population of contributors to a public good consists of a mix of reciprocal and selfish agents, an assumption borne out by much experimental evidence. I then show that if there exists a government or authority that is superiorly informed about the fractions of these types in the population, a policy of harsh sanctions may convey that there are a lot of bad types in equilibrium. As a result, equilibrium sanctions will generally be lower then they would be under symmetric information.

In the fourth chapter, I report the results of a laboratory experiment aimed to test if sanctions can indeed have a signaling effect. In accordance with the signaling hypothesis I find that 'endogenous sanctions' tend to make people more pessimistic, especially those who were optimistic at the start of the game.

In the last chapter, I model an alternative approach to compliance. I consider the widely reported fact that the possibility to participate in a decision making procedure tends to raise voluntary compliance with authorities, even if the actual decision is not beneficial to the agent. I show that the introduction of a decision making procedure in which an agent can change a decision of the policymaker with some probability, can be a signal of altruistic motives of the policy maker towards the agent. This means that even if she does not change the outcome of the decision in practice, the agent trusts the policy maker to treat her well in the future, and will engage in more voluntary compliance.

In the Epilogue I add some remarks on the potential of participatory decision making as an alternative policy tool to the standard economic command and control framework.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Acknowledgements

I would like to thank my supervisor Rick van der Ploeg for his guidance, trust and timely motivations, and my advisor Fernando Vega-Redondo for his enthusiasm and his dedication to get to the bottom of things. I would like to thank my former advisor Karl Schlag for the selfless sacrifice of his time to help me with chapter 3. Karl Schlag and Roberto Galbiati are co-authors on chapter 4. I would like to thank Joel Sobel, Tobias Broer, Mark LeQuement, Sanne Zwart, Katherine Veie, Bastiaan Overvest, Uri Gneezy, Luis Izquierdo, Robert Dur, Christian Traxler, Jan Potters, Bauke Visser and James Andreoni for taking the time to comment on my work. I thank many seminar participants for patience and useful comments. All remaining errors are mine.

The research for this thesis was made possible by a research grant from the Dutch Ministry of Education. It would not have been possible without all the good people who make the European University Institute such a great environment to work. I want thank especially Jessica Spataro, Julia Valerio, Marcia Gastaldo, Lucia Vigna, Michela Pistolozzi, Martin Legner and Thomas Bourke.

Finally, I have been very fortunate to enjoy (moral) support from Stefania Milan, Tobias Broer, Jaap Jelsma, Cor van der Weele and many friends and colleagues, without which everything would have been much more difficult and lonely.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Contents

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Bibliography 123

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

List of Figures

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

List of Tables

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Chapter 1

Introduction

This thesis questions the simplistic model of social control that is predominant in economics, known as the deterrence model. The deterrence model holds that an authority can raise compliance with the rules by introducing official sanctions for non-compliance and/or rewards for compliance. The idea can be expressed in the following syllogism:

P1: People respond to incentives for compliance by complying more.

P2: Official rewards and punishments provide incentives for compliance.

⇒ Official rewards and punishments induce people to comply more.

This argument is so ingrained in economic theory that few economists, nor perhaps many others will question it. Nevertheless, I will quarrel with this logic in this thesis. Specifically, I will quarrel with premise P2. I will argue that the deterrence model is importantly incomplete, because it does not take into account the complex social lifeworld of individuals. My argument will be that the informal interactions between individuals provide myriads of motives for compliance and non-compliance. These motives are created by intrinsic forms of motivation, social norms, by internalized moral values and the desire for the esteem of others. These motives are sustained, reinforced and sometimes undermined by official incentives in ways that economists are only beginning to understand. They may resonate with formal policy or counteract it. In exceptional situations informal interactions may even lead official sanctions to be counterproductive.

The reason that economists have so far relied on such a simple account of social control is that underlying the deterrence theory is another flawed model which pervades all of economic theory. This is the Hobbesian model, which departs from the assumption that without authority, people would find themselves in a state of nature in which they make life miserable for one another.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Thus the role of government is to save a collection of individuals from the destructive pursuit of their self-interest trough the threat of force.

I do not contest the idea that a government can improve human coordination and cooperation, and also that the threat of force needs to be a part of the toolbox of a government. However, the idea of that the absence of government should be compared to an anarchic state of nature is quite mistaken. It is a caricature of human societies and human nature that was useful to the philosopher Hobbes to make the general points he was interested in, but it cannot be a basis for more specific social policies. Instead, much behavior is produced by social interactions between agents. Policy analyses that do not take into account these social interactions are impoverished and in many cases flawed.

This argument is elaborated in chapter 2. The rest of the thesis provides examples investigates what happens if we allow a specific type of interaction, the signaling of preferences, into the analysis of the exercise of authority. Thus, I move away from the traditional focus that investigates the impact of policies on payoffs. Instead, I argue that policies may induce or discourage compliance with the authority through their effect on the *beliefs* of agents. There are two ways in which the policies can transmit information. First, they can affect expectations that agents have about each other's behavior. This will affect their own behavior in horizontal interactions between group members. Second, they can affect beliefs that agents have about the authorities. This will affect behavior in vertical interactions between an authority and an agent.

I investigate an example of both horizontal and vertical interaction, and show how these information flows will influence the policy of the authorities. These examples will lead me to argue that although deterrence will always be an important tool of any authority, economists have neglected other tools that are perhaps equally more important.

1.1 Chapter by chapter overview of the thesis

Whether the deterrence model is a good model is eventually an empirical question. Thus, the second chapter opens with an investigation of the empirical performance of the deterrence model. Surprisingly perhaps, because of the intuitive nature of the model, the theory does not perform that well empirically. In general, econometric studies that examine the effect of deterrence on street crime and tax evasion show that deterrence seems to have an effect in the predicted direction, especially if the deterrence is strong. However, the effect is inconsistent and generally quite small. As a consequence, the deterrence model cannot explain the high level of variation in the level of cooperation with the law that is observed in many cases. Sometimes compliance is high despite low deterrence, while the opposite also occurs. Second, laboratory experiments that are specifically designed to test the deterrence hypothesis provide anything

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

but a consistent picture. On the contrary, they find that deterrence will sometimes *increase* deviant behavior.

The chapter continues with a survey of theoretical approaches that have been pioneered in economics in the last decades. These approaches investigate the effect of sanctions in a richer pshychological or social framework. I investigate the interaction of official sanctions with intrinsic motives for behavior, the diffusion of (moral) values in society through processes of socialization, and social norms of behavior. This survey shows that taking into account such mechanisms substantially weakens the link between sanctions and deterrence, without leaving the rational choice framework. The chapter finishes with a formal illustration of how different policy instruments interact in producing compliance. It shows how deterrence can be counterproductive on its own while it can be a very useful tool if combined with other, 'softer' approaches.

The third chapter looks at *horizontal relations* between group members. In the chapter, I model social dilemma or public good game in a large, heterogeneous population consisting of egoists and conditional cooperators. Each player is uncertain about the cooperative inclinations of the other players. A government or principal who has private information about the distribution of types may introduce sanctions if agents defect. I study the impact of such sanctions through the effect on the beliefs of the players about the distribution of types they are facing. In equilibrium, sanctions can crowd out trust between agents by sending a signal that there are many egoists around. This can lead the authority to set low sanctions to induce trust and 'crowd in' contributions from the conditional cooperators. In social dilemmas where conditional cooperation is an important factor, as is the case in tax compliance, the model provides a rationale for low observed penalties in the real world.

The mechanism in the third chapter is a theoretical exercise, and evidence that such a mechanism is at work is rather sketchy. The fourth chapter aims to remedy this by presenting the results of a laboratory experiment, designed to investigate the signaling role of sanctions in coordination environments[1]. I study a two-period minimum effort coordination game between two players, in the presence of a third player or 'principal'. This principal benefits from coordination on higher effort, and is the only one informed of pre-sanction coordination levels. I compare the effects of a mild sanction, when it is imposed exogenously by the experimenters and when it is imposed by the superiorly informed principal. The results indicate that exogenously introduced sanctions are effective in inducing optimistic beliefs about others and in raising effort levels. However, endogenously introduced sanctions are much less so. For subjects who play cooperatively in the first round, endogenous sanctions induce pessimism about the effort of the other player, and are

[1]The research presented in this chapter has been conducted together with Roberto Galbiati (EconomiX Nanterre) and Karl Schlag (Universitat Pompeu Fabra)

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

not effective in raising effort levels. The results supports the idea that the sanctions have an expressive dimension which can undermine their effectiveness by discouraging optimistic players.

The fifth and chapter considers *vertical relations* between the authority and an agent. Its point of departure is a large literature in social psychology on the importance of 'fair' procedures. Perhaps the most important component of fairness is the degree participation by agents in the decision making process. Participation has been shown to lead to increased compliance, *even* if the outcome of the decision is not favorable to the agent. This chapter presents a signaling model that explains these findings. I propose a stylized definition of participation as stochastic control of the agent on the outcome of the decision process. I then use this definition in a formal signaling model and show that an authority can use the level of participation as a signal of her benevolence. This explains why participatory procedures may increase cooperation in subsequent interactions.

Thus, while the dominant command and control approach embodied in the deterrence hypothesis will be qualified in chapter 2-4, the 5th chapter offers a glimpse of a completely different style of governing . This chapter shows how the authority, instead of exercising control, can increase cooperation and compliance through institutions that explicitly imply a *loss* of control. In the Epilogue I reflect briefly on the importance of the dispersion of decision making power in modern societies, and its implications for economic research.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Chapter 2

Deterrence in context: how formal and informal incentives for compliance interact.

> "[A] useful theory of criminal behavior can dispense with special theories of anomie, psychological inadequacies, or inheritance of special traits and simply extend the economist's usual analysis of choice."
>
> G. C. Becker (1968, p.170).

> "If we do not even bother to sort out the many different ways in which people (and other animals) are moved, how can we hope to have an adequate descriptive, much less a normative, theory?"
>
> M. Nussbaum (1997, p.1210).

2.1 Introduction

Of every 100 adults in the U.S., more than one is in jail according to a report by the Pew Center (2008). This represents a more than 6-fold increase since the early 70s, the result of an uninterrupted 36 year rise in the prison population. Total state spending on corrections topped $49 billion last year, up from $12 billion in 1987. By 2011, continued prison growth is expected to cost states an additional $25 billion. The trend of rising prison populations is present in most other OECD countries (OECD, 2007, p. 79), although on a far lower level of incarceration.

What drives this explosion in prison population over the last 30 years? Criminologists Blumstein and Beck (1999) investigated the near-tripling of the U.S. prison population during the period

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

1980-96 and conclude that changes in crime rates explained only 12% of the rise. Changes in sentencing policy on the other hand accounted for 88% of the increase. Policies to 'get tough' on crime are mostly responsible for the explosion in prison population.

In his book *The Culture of Control*, Criminologist David Garland (2001) has attributed the trend towards tougher policies in the U.S. and Britain to a renewed public confidence in the economic model of crime. This model is as simple as it is controversial: it states that a potential criminal will weigh the benefits of breaking the law with its cost, which consist of the probability of getting caught multiplied by the disutility induced by the penalty. Thus, the authorities in charge will be able to reduce crime by setting sufficiently high rates of deterrence. This theory is intuitive, simple and elegant. Its opponents claim it is also fatally flawed, precisely because crime is not simple, but depends on a complex interplay of social factors (e.g. Nussbaum 1997).

Is the deterrence theory a useful theory of crime, as Becker (1968) claims? Or is it mainly rhetoric proclaiming an illusion of social control, as criminologist David Garland (2001) argues? The stakes behind these questions are high, and a (largely fruitless) debate on the merits of the economic model has been raging for decades. In the midst of it, it is easy to forget that the validity of the deterrence theory is simply an empirical question, that should be judged on the basis of empirical evidence: can the economic model predict patterns of crime and non-compliance with the law more generally? The first part of this chapter is dedicated to a review of evidence on this question. This review shows that although the economic model has some empirical support, the overall evidence for it is rather inconsistent. Perhaps more importantly, variations in the levels of deterrence can not nearly explain the variation in level of compliance.

The uneven empirical record of the deterrence hypothesis suggests that economists are working with a model of social control that captures at best a small part of the reasons why people comply with the rules. Moreover, standard theory has no explanation for why deterrence works in some circumstances and not in others. Given that the deterrence hypothesis is such an important pillar of economic theory, does this disqualify the economic theory of incentives as being a 'useful' theory of crime and deviant behavior? I will argue that this conclusion is too quick.

Over the last decade, economic and legal theorist have begun to take their critics from other social sciences seriously and have started to incorporate models of social context in their analysis of crime. In Section 2.4 I show how this new economic literature can help account for the mixed press of deterrence. Generally, this literature distinguishes between a direct and an indirect effect of sanctions. The direct effect of sanctions is their standard effect; to provide incentives for compliance by changing economic payoffs. This is the effect that traditional economic theory in the tradition of Becker has focused on. The indirect effect operates trough the interaction of formal incentives with informal mechanisms in society. I have singled out a few such mechanisms. The first is what is known as 'motivational crowding out', the idea that external incentives have

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

an impact on individual preferences to engage in virtuous behavior. The second is the way in which sanctions disturb equilibria in games played between agents. These can be long-run evolutionary games that affect the formation of preferences for virtuous behavior. The impact of sanctions on such games is discussed in Section 2.4.2. Or they can more instantaneous games of signaling or coordination, in which equilibria may be associated with social norms or conventions, as discussed in Section 2.4.3. As we will see, taking into account the endogeneity of equilibria may cause deterrence to have very different effects than those predicted by standard theory.

The conclusion that emerges from these theoretical analyses is that both the short and long-run the effects of official sanctions are highly dependent on the social context. The impact of sanctions on behavior depends on whether they will crowd out virtuous motivations, are able to sustain and reinforce social norms of compliance and foster the existence and survival of preferences that favor compliance in the population. This has implications for economic theory, that generally prides itself for the generality of its models. In the discussion I argue that economists have been hampered by the 'Hobbesian' framework that usually underlies economic policy prescriptions. This framework perceives the actor as essentially individualistic and engaged in a never-ending 'war of all against all'. Although this Hobbesian view has its merits in terms of simplicity and rigor, it cannot generate specific policy advice. I argue that when we expand this narrow concept of economic man, not only does the effect of deterrence become more ambiguous, but other policy instruments that impact on different motivations become salient. What is needed is a clearer picture of how different policy instruments can be combined to increase compliance. To illustrate this last point, I present a simple model where an authority can use both a 'hard' deterrent policy and a 'soft' cultural policy aimed at reducing a norms of criminal behavior. I show that the hard policy may be counterproductive on it's own, but that it is an effective complement to the softer policy.

Thus, I will conclude, finding the optimal level of deterrence involves an analysis of the social context, something that economists have become increasingly good at doing. Moreover, deterrence is only one of a range of policy instruments to induce compliance, and much work is to be done to analyze the integrated effects of these different policy instruments.

2.2 The deterrence hypothesis and some evidence

The idea that authorities can reduce deviant or criminal behavior by changing the price of such behavior is one of the most basic building blocks of law and economics and economic policy more generally. The underlying model is that in deciding whether to commit an illegal act, criminals or deviants weigh the expected benefits and the expected costs of doing so. If one defines s as the (utility) cost of punishment, p as the probability of getting caught for a crime,

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

and b as the (utility) benefit of committing a crime, then the deterrence hypothesis holds that a person will commit a crime if:

$$b > p * s$$

Although this idea was already explicitly discussed by Beccaria (1770), Becker (1968) was the first to formally model the idea that criminal acts are the result of an expected utility maximization, and therefore highly predictable. Becker derived some important implications about optimal enforcement. An immediate implication is that punishment and probability of detection are substitutes in deterring crime, although the exact substitutability relation depends on the risk aversion of the potential criminal. This implies that the law-enforcers are flexible in choosing their instruments. Moreover, to be deterrent, punishments should increase in the benefits of the crime. Becker also argues that since raising the probability of detection by increased monitoring is costly, optimal deterrence should instead rely on high punishments. He specifically argues for the use of fines, since they are costless to administer and may provide compensation to the victims.

The deterrence hypothesis is very attractive because it is simple and intuitive. For these good reasons, it underlies an enormous literature in law and economics, surveyed by Polinsky and Shavel (2007) and Garoupa (1997). A prominent application is the Allingham and Sandmo model (1972) of tax evasion. They model tax evasion as a choice between a safe asset (declared income) with a low return, and an unsafe one (concealed income) with a potentially high return. They then show that an increase in the deterrence variables p and s make the risky option less attractive, and lead agents to conceal less income.

2.2.1 Does deterrence work?

A simple and elegant theory is not necessarily correct. Whether governments or authorities in general can effectively use deterrence to induce compliance in the population is an empirical question, that I will try to answer in this section. Given the size of the literature I can and do not aim to be exhaustive. Instead I rely on review studies conducted by others, and try to give a flavor of the literature by mentioning some specific examples that I think are instructive[1]. One immediate conclusion is that there is substantial disagreement among scholars about what the evidence says. Two quotes from different review-studies on deterrence will make this clear:

[1] I will take a very wide range of applications of deterrence. I will consider both criminal acts such as assault, theft and tax evasion, and mere anti-social behavior such as littering or not contributing to a public good in an experiment. There are many reasons to think that these are very different acts that warrant very different policy measures. However, from the point of the deterrence hypothesis, there is no fundamental difference between these acts or the way they should be counteracted. In this article I will not drop this particular generalization for reasons of space and time. However, I am confident that doing so will reinforce rather than diminish the conclusions of this chapter.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

> "[R]esearchers have enjoyed significant progress in recent years in testing the economic model. They have found that deterrence has a substantial but far from complete role in explaining observed patterns of criminal activity."
>
> Levitt and Miles (2007, p. 457).

> "Does criminal law deter? Given available behavioral science data, the short answer is: generally, no."
>
> Robinson and Darley (2004, p. 173).

Part of the difference between these conclusions can be explained by the fact that these two articles review partly different studies. However, a more important reason is that assessing the effects of deterrence in the real world is very hard indeed, and the evidence is open to different interpretations. For example, the econometric literature on real world data suffers from thorny identification problems and limited availability of data. For this reason I will also include data from field and lab experiments in this short (meta-)survey, since these methodologies can solve the problems associated with econometric studies. However, these studies generally cannot investigate real crimes, but rather milder forms of anti-social behavior. Moreover, given the artificial nature of laboratory experiments, the external validity of these studies is questionable[2].

Panel data and instrumental variables. Within the econometric literature that deals with real world data, I will discuss research on street, property and violent crime and tax evasion. The reason is simply that most evidence on the effect of deterrence has been gathered in these areas. The literature on crime has inspired several surveys (e.g. Eide 2000, Levitt and Miles 2007, Robinson and Darley 2004). There also exists a sizable literature on the determinants of tax evasion (e.g. Andreoni *et al.* 1998, Frazoni 1999, Alm 1998).

Most of the econometric literature on deterrence can be understood as the attempt to dodge two thorny identification problems. First, the amount of deterrence will often be a response to the level of crime, yielding a spurious positive correlation between deterrence and crime. For example, Dubin and Wilde (1988) find that tax audit rates are often endogenous. This may (but need not) explain the results found by Cameron (1988), who surveys 22 studies that investigate the relation between increase in the number of policemen and crime. Of these, only 4 find a

[2]Note that even though econometric studies rely on real world data, their generalizability cannot be taken for granted either. For example, Ayres and Levitt (1998) investigated the introduction of LoJack, a radio-tracking device for cars. LoJack greatly increases the possibility of the police to track down stolen cars and can not be detected from the outside. Ayres and Levitt (1998) find that it reduces auto thefts by as much 50% when it was implemented in the US. However, Gonzalez-Navarro (2008) studies the effectiveness of the device in Mexico, where it was only introduced in certain states. He shows that the reduction in thefts in those states where matched almost one for one by an increase in theft in neighboring states where LoJack was not introduced. This shows that a deterrent measure may be very successful in one situation and ineffective (on aggregate) in another.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

negative relationship, the others find either no relationship or a positive one. Dills et al. (2008) investigate simple correlations of time series and cross-country data, and find that police arrests, incarceration and the size of the police force are either not correlated, or positively correlated with crime. A second problem is that even if one finds an effect of punishments, the question is how to distinguish between deterrence and incapacitation. Do sanctions work because they deter, or because the potential criminals are behind bars?

The identification problems described above can be tackled by using instrumental variables: variables that correlate with the size of deterrence but not with crime. Alternatively, one can disentangle the direction of the causation with the use lagged variables. Panel data can help to correct for unobserved characteristics of particular communities of study.

Levitt has tried to tackle endogeneity problems by using a panel data set and instrumental variables. Levitt (1997) uses the fact that politicians tend to spend more resources on deterrence in electoral years to estimate the impact of deterrence across cities. He shows that a spurious positive correlation disappears when the instrument is used and turns into a modest negative relationship. Levitt estimates an elasticity of violent crime with respect to the number of police officers of -1.0. For property crime the elasticity is -0.3. However, the instrument is weak and the estimations are imprecise. McCrary (2002) also points out a computational error that leaves the results insignificant. In response to McCrary, Levitt (2002) uses the number of firefighters and municipal workers as instruments and finds smaller but more significant negative elasticities: Around -0.5 for both violent and property crime.

Some authors have taken other approaches to circumvent endogeneity. Moody and Marvell (1996) use a Granger-causal approach: using a panel data set they estimate whether bigger police forces precede drops in the crime rate. Corman and Mocan (2000) use monthly data to circumvent the simultaneity problem, arguing that a political response to rising crime rates takes long to materialize. Both studies find similar values to Levitt (2002).

The question how much of this is due to incapacitation and how much to deterrence is still largely open, although there is evidence that both phenomena play a role. For example, Kessler and Levitt (1999) focus on the short long term-impact of the effect of enhanced sentences for some offenses. They find that the short-term impact (due mainly to deterrence) is significant but lower than the long-term impact (due to both deterrence and incapacitation).

In their survey of the econometric literature, based on the studies cited here and other ones, Levitt and Miles (2007) conclude that there is rather consistent evidence that bigger police forces and more prisons reduce crime. However, not all studies using instrumental variables corroborate these results. Cornwell and Trumbull (2000) use a panel data set on counties in North Carolina. They use within estimators to correct for unobserved heterogeneity, which they find to be important. They find that the elasticities of crime to the probability of arrest are

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

significant but small (around -0.35 for the probability of arrest). They then use tax revenue as an instrumental variable for the number of police officers, and the ratio of 'face to face' crimes to other crimes as instrumental variable for the probability of arrest (because these crimes are more easily solved). When they use these two corrections, they find that significance of deterrent measures disappears. Labor market variables, such as the market wage, are more strongly (and inversely) correlated with crime.

In fact, general statements about the effect of deterrence are hard to make. Two debates in the empirical literature demonstrate this vividly. The first is the debate over the effects of the right to carry (concealed) handguns. In theory, allowing people to do so should have a deterrent effect, because criminals know that their victim may be armed. Lott and Mustard (1997) investigate the introduction from right to carry laws using a panel data set with county level data. They present evidence that concealed handguns have a significant deterrence effect on various crime categories. They estimate that at least 1,411 murders, 4,177 rapes, and more than 11,000 robberies could have been avoided if every state in the US would have introduced the legislation in 1992. However, their results have been sharply criticized by a number of authors. Dezhbakshs and Rubin (1998) attack the assumption of Lott and Mustard that right to carry legislation only affects the intercept of the relation between crime and the control variables and has no impact on the effect of individual controls. They show that in a more general model the effect is much smaller and no longer goes in one direction for all crime categories. Similarly, Black and Nagin (1998) expand the model of Lott and Mustard to allow the effect to be different across states that introduced the legislation. The effect of handgun regulation is very different across states and crime categories and no uniformly negative effect on crime is found.

A similar discussion rages over the deterrent effects of the death penalty. Donohue and Wolfers (2006) review research on this topic, and find that the empirical results are very sensitive to small model changes. They conclude that the literature has not demonstrated a robust effect of the death penalty, mainly because of a lack of variation in the available data.

These controversies show how difficult it is to get adequate measurements of deterrence effects. In the case of the handgun debate, it also shows how aggregating data on a high level tends to obscure the large variations between different communities. In the analysis below we will show that such heterogeneity is exactly what one would expect if sanctions interact with localized norms and values.

Moving from street and property crime to tax evasion, the results are rather similar. Franzoni (1999) and Andreoni *et al.* (1998) and Alm (1998) survey the theoretical and empirical literature on tax evasion. They all conclude that econometric studies indicate that penalties and audit probabilities seem to have some deterrent effects, where the typical elasticity of reported income to the audit rate is around 0.2. However, like in the case of crime, the estimated responses vary across studies.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Natural experiments. One of the most powerful identification strategies is the use of natural experiments. When using a natural experiment, the researcher investigates the effect of a (random) event or policy that causes an exogenous change in the deterrence policy. One can see this as a stronger variant of the instrumental variable techniques, because it does not rely on (the sometimes weak) correlation between the instrument and the explanatory variable. The art is to come up with suitable and clever natural experiments.

Lee and McCrary (2005) use the increase in the length of sentences at the age of 18 in the U.S. They do not find that adolescents reduce the amount of crimes at this age, a result they attribute to imperfect perception of the sentence length, or extreme short-sightedness of the offenders. Drago *et al.* (forthcoming) use an Italian clemency bill as a natural experiment. The bill released 22.000 criminals on the condition that if they were to commit a crime in the next five years, they would have to serve the residual jail sentence as well as the new sentence. This means that different people faced different penalties for comparable crimes. The authors find evidence for an effect of deterrence: an extra month of residual sentence reduces the probability of recidivism by 1.24 percent. Helland and Tabarrok (2007) use the three strikes legislation in California as a source of natural experiment. This legislation constitutes a harsh piece of deterrence: an offender is automatically given a life-sentence if he is convicted for the third 'strike-able' offense. The authors compare criminals who were convicted for a second strike, with those who were tried for a second strike-able offense but convicted of a non-strike-able offense. They find that the three-strike legislation significantly reduces felony arrest rates among the class of criminals with two strikes by 17-20 percent.

Field experiments. The problems in the econometric literature on crime can be resolved to a large extent using field experiments. Varying deterrence rates while controlling for other variables is a powerful method to test the deterrence hypothesis, and provide sharper insights into the effects of deterrence. Experiments with policies to combat crime tend to be controversial, so field experiment have typically focused on lighter forms of deviant behavior.

A study that deals a serious blow to the deterrence hypothesis is Gneezy and Rustichini (2000). In a field experiment in daycare centers in Haifa, the experimental condition consisted of the introduction of a small fine if for picking up one's child late. The results contradicted the deterrence hypothesis: when a fine was introduced late-coming went up significantly. Moreover, revoking the fine did not lead to a reversal in behavior; the post-fine level of late-coming was higher than before the introduction of a fine.

Cardenas *et al.* (2000) conducted an experiment among Columbian farmers, who were asked how much they would extract from a common resource. The farmers extracted more than the efficient level. After sanctions for extraction were introduced extraction levels initially went down, but after a few periods they rose again to almost the pre-sanction level. Bowles (2008)

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

surveys 24 experimental studies that found results of a similar nature. There is a consistent finding that small levels of incentives or fines reduce cooperative or compliant behavior, while large levels increase it.

Coleman (1997) reports the results of a large-scale field experiment on tax evasion amongst 47,000 taxpayers in Minnesota. Some 1700 of the taxpayers received a letter saying that the recipient was randomly selected for an audit. Coleman finds that this warning increases tax payments among low and middle income taxpayers, but not for high income taxpayers. The effect was most pronounced for a small group who had the most opportunity to evade taxes. For the rest of the taxpayers the effect was so modest that Coleman concludes that the benefits do not justify the cost of the audit.

Laboratory experiments. Laboratory experiments provide maximum possibility for control of environmental circumstances, and are the most effective method for solving endogeneity problems. On the down side, laboratory experiments cannot study real crimes and generally take place in artificial environments, so the external validity of experimental results is always questionable.

Surprisingly, a direct experimental test of the deterrence hypothesis has been conducted only recently, by Hörisch and Strassmair (2008). In their experiment, two subjects receive an endowment of money and are paired to play a simple game. The player with the smaller endowment is offered the possibility to steal part or all of the endowment of the other player. After she has made the decision whether or not to steal, the theft is detected with some probability and a fine is deducted from her wealth. The probability of detection and the size of the fine are varied over treatments. This way the authors can test whether deterrence works and whether the probability of detection and the fine are indeed substitutes. The authors find that in accordance with the deterrence hypothesis, the fine and the probability of detection seem to be substitutes in their effects on stealing. In accordance with the field evidence cited above, but in almost complete contrast to the deterrence theory, weak deterrence significantly *increases* the level of stealing relative to the no-deterrence case. Only the highest level of deterrence (of 6 levels) significantly decreases stealing. Also, strong deterrence creates a clear bipolar distribution in the amount stolen: while some steal everything, others steal nothing. Hörish and Strassmaier (2008) interpret this as evidence for the existence of different 'types' of people. There are selfish types who steal maximally if sanctions are low act as predicted by the deterrence theory. However, the (slight) majority consists of different or 'fair-minded' types, who do not steal maximally in the absence of deterrence but start doing so when weak incentives are in place. Fishbacher and Gächter (2006) explicitly test for the stability of such behavior across situations, and find clear evidence for the existence of different types of players. Around 25% is found to behave selfishly, 50% behaves as a conditional cooperator, and 25% displays more complicated behavior.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Conditional cooperators are people who condition their behavior on what they think others are doing. Indeed, the existence of conditional cooperators is a fact that is repeatedly confirmed in studies of dilemma games (Gächter, 2006), and plays an important role in the analysis of social norms in section 2.4.

Another interesting result from Hörish and Strassmeier (2008) relates to the temporal dimension of incentives. In the study all subjects participate in two treatments, and so within subject comparisons between several deterrence regimes are possible. Hörish and Strassmair (2008) find that subjects steal more if the treatment is preceded by a treatment with higher incentives. The hysteresis-results echoes the result of Gneezy and Rustichini (2000) cited above, and is also found in other studies, e.g. Frohlich and Oppenheimer (1995) and Gächter *et al.* (2007). Tax evasion has been the subject of many experimental studies. The earlier literature focuses on testing the Sandmo-Allingham expected utility model. Surveying this literature, Alm (1998) finds that elasticities of reported incomes to (random) audit rates generally have a small positive effect with elasticities in the range of 0.1 to 0.2. Fines have an even smaller effect, with an estimated elasticity of less than 0.1. Like Hörish and Strassmeier (2008), Alm also stresses the great heterogeneity between the behavior of subjects, and the importance of theory to replicate this fact.

Sanctions have been also studied in the context of public good games, although there are surprisingly few studies that directly test the introduction of a centralized sanction in a public good game in a within subject design (perhaps because the result is supposed to be obvious). Nevertheless, there is some evidence however that sanctions have a positive effect on contributions. Shinada and Yamagishi (2007) and Guillen *et al.* (2006) show in a between subject design that sanctions imposed by the experimenter raise contribution levels.[3]

2.2.2 Is deterrence important?

Most of the previous evidence relates to the question whether the effect of deterrence goes in the direction conjectured by the deterrence hypothesis. Perhaps a more important question is whether deterrence *matters*, in the sense that it explains the level and the variation in crime rates. There are studies that focus on the relative importance of different explanations, but these generally do not correct for the endogeneity problems mentioned above. Nevertheless, we can find many indications that variations in deterrence explain relatively little of the variance in crime and tax evasion.[4]

[3] In addition to these studies, there is by now a large experimental literature on the effect of performance incentives in principal-agent settings. This is surveyed in Fehr and Falk (2002). The gist of this literature is that a principal's use of incentives, such as fines, bonuses and enforced contracts, does sometimes crowd out voluntary effort provision.

[4] A caveat here is that most of this evidence relates to relatively small fluctuations in deterrence rates. I certainly do not want to extrapolate this result to say that the complete absence of a criminal justice system would not affect compliance with the law.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Glaeser *et al.* (1996) show that the variance in crime is staggering. Using data from 1980, they show that on a cross-country level, the United States has about 150 times the homicide rate of Japan. On a intra-country level, Atlantic City, New Jersey has about 40 times the crime rate of nearby Ridgewood Village. And on a intra-city level the 1st precinct in NYC has about 10 times the crime rate of the 123rd precinct. They also present some suggestive figures with respect to the relative importance of deterrence. There is no correlation between arrest rates and crime across NY precincts. Across cities arrest rates and convictions are slightly negatively correlated with crime (around -5%), which means that the arrest rates 'explain' less than 1% of the variation in crime. By contrast, the fraction of female headed households correlate 20% with crime across New York precincts and and slightly higher across cities. The authors also conduct a logit-regression analysis. They show that all the observable city-specific characteristics (education, age, income, unemployment, the property tax rate, ratio of households headed by females, and police per capita) explain less than 30% of the crime rate. Like I will do in the second half of this chapter, the authors argue that social interactions (of which they present a model in the paper) are responsible for the remaining variance.

Glaeser and Sacerdote (1999) attempt to decompose the causes of high crime in big cities. They first estimate that the elasticity of crime with respect to the city size is about .24, and the elasticity of the arrest rate with respect to city size is about -0.08. Using data from two different sources, they specify different models to get robust estimates for the elasticity of crime to arrest rates, which they show to be between -0.2 and -0.5 (i.e. similar to Levitt 1997). This means that the lower probability of arrest in big cities can explain between 8% and 20% of the increased crime in big cities. The presence of more female-headed households on the other hand can explain between one-third and one-half of the difference.

Fajnzylber *et al.* (2002) use panel data to estimate the cross-country determinants of violent crime. They include socioeconomic variables such as average educational attainment, unemployment, inequality, and output growth as regressors. The number of police personnel per capita and the existence of the death penalty proxy for the level of deterrence. They also include lags of the crime rate to correct for endogeneity. Using GMM estimation, they find that the deterrence variables have negative, marginally significant but very small impact on crime. Economic growth and income inequality are more important both in terms of significance and in terms of the size of the effect.

A nice and rare study of the long-run connection between deterrence and crime is provided by Lappi-Seppälä (2001), who describes the change in the Finnish penal regime over the last few decades. In the 1950s, the imprisonment rate in Finland was 4 times higher than in neighboring Scandinavian countries. Still, in 1975, Finland had one of the highest imprisonment rates in Europe. In the subsequent 20 years, Finland brought down the prison population to the same level of other Scandinavian countries, and to one of lowest in Europe (around 60 prisoners per

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

100.000 inhabitants). Lappi-Seppälä (2001) describes in detail the widespread reform of the criminal code that accomplished this, mainly through decriminalizing activities and reducing prison sentences for many others. The deterrence hypothesis would predict that crime rates would increase. In fact however, crime statistics in Finland in the same period have not deviated from those in other Scandinavian countries, and have remained lower than those of Sweden and Denmark. This raises doubt about any straightforward long-run relationship between deterrence and crime rates.

Turning to tax-evasion, there is a consensus that the real-world levels of deterrence for evasion cannot explain the observed levels of compliance. Andreoni *et al.* (1998, p.855) note that "The most significant discrepancy that has been documented between the standard economic model of compliance and real-world compliance behavior is that the standard model greatly overpredicts evasion". And on page 821: "For small amounts of evasion, [...] the expected cost of detection would appear to be extremely low for most tax-payers. So, we may ask, why are so many households honest and why don't cheaters cheat by more?" Alm (1998) concurs that the expected utility model greatly overpredicts evasion, and states in the conclusion of his survey on tax evasion that "there are significant limitations in the ability of the expected utility theory to explain major aspects of individual compliance behavior," (1998, p. 759). Alm (1998) concludes in his survey that social norms are one of the most important factors driving tax compliance. In section 2.4 we will see several economic approaches to the study of social norms.

2.2.3 Summary and stylized facts

The number of studies on the effect of deterrence is overwhelming, and the picture they present is far from consistent. Nevertheless, I will attempt to summarize the empirical results in a few stylized facts:

1. Real-world data show that if deterrence is strong (punishments and detection probabilities are high), such as under California's three-strikes law, the empirical literature generally supports the claim that deterrence decreases crime. However, estimates of the size of the effect are not very consistent.

2. Variations in deterrence explain only a relatively small portion of the real-world variation in crime and in tax evasion.

3. Experimental studies indicate that at low levels of deterrence, the direction of the effect of deterrence is ambiguous. A rising number of studies finds that modest amounts of deterrence can be counterproductive.

4. The experimental literature has shown some specific effects of deterrent measures. First, deterrence can have (adverse) effects that outlast the existence of the incentives themselves.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Second, different people react differently to deterrence measures. Specifically, only around one quarter of the population seem to be selfish agents that respond as economic theory would predict.

The conclusion from the empirical evidence must be that the deterrence hypothesis has a mixed press. Given that deterrence is such an important building block of economic and legal theory, this is both surprising and important. The next sections of this chapter are dedicated to the question why deterrence may work well in some circumstances and not in others. One answer is that the theory is sound, but that there are practical obstacles to the implementation of deterrence that often prevent deterrence levels to be sufficiently high to have an impact. We discuss this possibility in the next section. However, some of the evidence cited suggests that the theory itself is incomplete at best. Critics have taken this as a cue that economic models of incentives have nothing to contribute to policy debates about criminal and deviant behavior. However, I will argue that these critics throw out the child with the bath water. In section 2.4 I will introduce new economic models that analyze the interaction of deterrence with informal social mechanisms. These studies show that the straightforward link between deterrence and compliance disappears.[5]

2.3 Why deterrence may not work in practice

The theory of deterrence outlined above is clean and flawless. The practical implementation however is not so. One problem is that achieving probabilities of detection that are sufficiently high to make an impact may be prohibitively costly. Robinson and Darley (1997) estimated the objective probability of getting caught, convicted, and imprisoned for several offenses. For homicide the probability is 45%, for rape 12%, for robbery 4%, for assault, burglary, larceny, and motor vehicle theft, 1%. Robinson and Darley (2004) compute on the basis of data from the American Justice Department that the average probability of being sentenced for a criminal offense committed is 1.3%. Andreoni *et al.* (1998) report that in 1995, 1.7 % of all US taxpayers were audited. Of the people who's audit was reassessed, 4.1 % paid a fine.

For this reason, Becker (1968) advocated the use of heavy penalties, and especially fines, because these are cheap, or even profitable to administer. However, there are moral bounds to the level of sanctions that authorities can impose, given the widespread sentiment that the punishment must be proportional to the crime. The 'three strikes' law in California seems to be a good example of this. In the survey above, Helland and Tabarrok (2007) present perhaps the clearest

[5] With the exception of evolutionary models, I do not consider theories that suppose some form of irrationality on the part of potential offenders. I also do not consider theories of misperception of severity or arrest probabilities. Robinson and Darley (2004) give a convincing account that both of these factors are at work. My reason for focusing on informal social mechanisms is that so far, the assumptions of correct perception of deterrence and full rationality have been seen as generally sufficient for the deterrence hypothesis to hold.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

evidence around that the heavy sanctions of this legislation do indeed deter crime. The law is heavily criticized however, both for being too harsh (people have landed life sentences on the basis of rather innocent shoplifting crimes) and for straining the prison system too much.

The combination of costly monitoring and proportional sanctions, means that in practice it may be very difficult to attain levels of deterrence that are high enough to substantially influence behavior. This is especially true for small offenses that are hard to detect. One can think in this respect of small crimes or misdemeanors such as littering, small amounts of tax evasion and fare evasion. Improvements in monitoring technology may alleviate this problem in some cases. For example, the United States tax authorities employ a sophisticated computer algorithm that makes large underreporting of income taxes easy to detect. It is hard to imagine however that in a free society monitoring will eliminate crime. As an illustration, consider the problems encountered by the closed-circuit television (CCTV) system, the most elaborate monitoring system in the world, installed by the British police to solve and prevent crime. Britain has 4.2 million security cameras, and someone living in London is filmed an estimated 300 times a day. However, both the Home Office in 2005 and Scotland Yard in 2008 have concluded that the cameras are largely ineffective. "CCTV was originally seen as a preventative measure," Detective Chief Inspector Mick Neville, the officer in charge of the Metropolitan police unit has told the Security Document World Conference in London[6]. "Billions of pounds has been spent on the kit, but no thought has gone into how the police are going to use the images and how they will be used in court. It's been an utter fiasco: only 3% of crimes were solved by CCTV. There's no fear of CCTV. Why don't people fear it? [They think] the cameras are not working."

More recently, CCTV has also come under attack because it has been used by officials to prosecute small offenses like littering. Many people perceive as an invasion of privacy and a disproportional use of the technology, demonstrating again the limits to the social acceptability of raising deterrence.

Thus, in practice it turns out to be hard to generate high levels of deterrence. Even if such levels are technically possible, they may not always be socially acceptable. This means that results of experiment that document contradictions to the deterrence theory at low levels of deterrence become more salient. In the parlance of game theory: deterrence may often not succeed in making compliance a dominant strategy. What then, if any, is the effect of deterrence?

2.4 How formal and informal mechanisms produce compliance

In this section I survey theoretical approaches within economics that help explain the mixed record of deterrence. The approaches I consider look at the interaction of deterrent measures

[6]See http://www.guardian.co.uk/uk/2008/may/06/ukcrime1.

with informal incentives resulting from social interactions. I will first briefly discuss the mechanisms I consider.

First I consider the interaction of sanctions with preferences in the context of *motivation crowding theory* (MCT). MCT distinguishes between intrinsic and extrinsic motivation. 'Intrinsic motivation' may consist amongst other things of the joy of engaging in the activity or the feeling that the activity is worth doing for moral reasons. 'Extrinsic motivation' refers to formal incentives applied by authorities as well as informal pressures from peers or other social actors. Motivation crowding theory allows that these two types of motivation are not independent from one another. More specifically, it holds that extrinsic motivation may reduce intrinsic motivation.

Second, I allow for the fact that sanctions may impact on games played between agents. One instance of this is in medium and long-run evolutionary games preference formation, or what I will call *internalization of values*. By this I mean the endogenous formation of preferences such as guilt, reciprocity, shame and preferences for fairness.

I also analyze the impact of sanctions on *social norms and conventions* in games. Social norms are hard to define, but here I will follow McAdams and Rasmusen (2007). They identify *conventions* with equilibria in different types of games, i.e. coordination games, signaling games, or repeated dilemma games. These are regularities that do not necessarily have normative content, they simply constitute what is normal. The fact that there are often multiple equilibria in these games means that conventions are to some extent accidental, such as the convention what side of the road to drive on. *Social norms* are conventions that are supported at least in part by normative attitudes. Such normative attitudes may create and sustain equilibria because they motivate people to provide informal punishments or rewards, such as (dis)approval or esteem (see McAdams (1997) for an elaborate account of an esteem-based account of social norms).

Naturally, there are intimate links between these three concepts. Intrinsic motivation may derive from internalized values. Similarly, the normative attitudes that underlie social norms depend on internalized values. Also, the evolved capacity for feeling guilt and shame makes people susceptible to the (dis)esteem of others, a fundamental issue in the conception of norms as it is used here.

Nevertheless, I believe there is reason to analyze these three issues separately. First, values are different from motivation, because 'motivation' applies to a specific action and context, whereas values are stable entities that persist over time. Values are also different from norms. The analysis of norms as Nash equilibria brings in many inter-personal considerations that are independent of the process of internalization. Again, even though values may underlie norms, they are thought to be deeper and more stable entities than these norms. Moreover, as McAdams (1997) points out, internalization usually occurs only for rather abstract values such as 'fairness'

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

or 'reciprocity'. Social norms are more practical rules of behavior that give meaning to these values in concrete social contexts.

2.4.1 Motivation crowding theory

Frey (1997a, 1997b) has been very active in popularizing the idea amongst economists that incentives do not just change relative prices. He has borrowed from the psychology literature to formulate what he calls 'motivation crowding theory' (MCT). MCT itself and the empirical evidence for it are surveyed by Frey and Jegen (2001). MCT holds that people have 'intrinsic motivation' for many activities. As Deci (1975, p. 105) puts it "one is said to be intrinsically motivated to perform an activity when one receives no apparent reward except the activity itself". Crucially, according to MCT, intrinsic motivation is not independent from exernal (monetary) incentives, or 'extrinsic motivation'. Instead, external incentives may increase ('crowd in') or reduce ('crowd-out') intrinsic motivation.

Crowding out is attributed to two psychological processes. The first is impaired self-determination. If the individual feels that the external incentive restricts her choice, the intrinsic motivation becomes redundant and she acts by reducing it. This is also called the over-justification effect. The second process is impaired self-esteem: if intervention signals that the agents motivation is not acknowledged or not good enough, the individual may feel less recognized or less competent and reduces her effort.

A useful theory of interdependence of preferences and incentives will have to make precise predictions on the relation between them. Frey (1997a, 1997b) sketches the relationship between extrinsic and intrinsic motivation. He argues that the crowding out effect is not gradual: no matter what the size of external incentives, they tend to replace the entire intrinsic motivation. It follows that when intrinsic motivation is large and extrinsic motivation is small, crowding out can even lead to the opposite effect of that predicted by economic relative price theory: discouraging behavior by monetary incentives can lead to more of that behavior.

MCT can account for the puzzling evidence generated by the (field) experiments above that small incentives have sometimes counterproductive effects, whereas stronger incentives seem to work. The main application of this idea in the economics literature so far is by Bowles and Hwang (2008), who build a model that investigates the consequences of MCT for (tax) incentives by policy makers. In their model people have 'values'; preferences to contribute to the public good, which depend on the size of the tax rate. They explicitly model non-separable preferences by assuming that a higher tax rate can augment or diminish preferences to contribute. They show that a social planner that does not take the crowding effect into account when making policy may either under or over-use incentives, depending on the social welfare objective and the direction of the crowding effect.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

While this paper is interesting, it is a reduced form exercise which sheds little light on when and where we can expect motivation crowding will occur. This is a general problem of MCT. Frey (1997a) writes that the psychological conditions for crowding out to appear are:

1. External incentives crowd out intrinsic motivation if the individual affected perceive them to be controlling.

2. External incentives crowd in intrinsic motivation if they are perceived as supportive.

However, MCT does not deliver much hints as to what is perceived as 'controlling' and 'supportive'. Instead, intrinsic motivation seems a blanket term that covers many potential motivations. Intrinsic motivation may consist of a sense of a Kantian moral duty or a preference for autonomous decision making, but also of self-esteem that comes from completing a task, or the desire to reciprocate the nice behavior of others. Each of these motivations may interact differently with external incentives. In this sense, intrinsic motivation is almost like a measure of things we don't understand. A more precise characterization of intrinsic motivation and the social situations in which it matters is therefore necessary.

2.4.2 Sanctions and Internalized Values

The idea that people have moral values is of course not new. However, the analysis of such values as endogenous to the environment of the agent is relatively recent. Bowles (1998) provides an excellent and wide-ranging survey of what is known about the effects of economic institutions, and especially markets, on preferences.

Recently, researchers have started to investigate the effects of deterrent strategies on values using evolutionary models. In general, these studies model how institutions affect the payoffs of 'cooperative' types vis-à-vis the payoffs of selfish types, and derive the evolutionary success of these types. That is, one uses version of evolutionary theory, the so-called indirect evolutionary approach, in which evolution impacts on preferences, rather than strategies. In an evolutionary approach, agents are rational optimizers but evolution selects the preferences that are best suited to the environment. Note that the term 'evolutionary' does not necessarily refer to a biological selection process, since this could take place only over a time span in which institutions cannot be reasonably be held constant. Rather it refers to a process of cultural evolution, where preferences spread by imitation and education.

Huck (1998) presents an evolutionary model where legal institutions have a *positive* effect on preferences for remorse from cheating in bilateral exchange. One party is in the position to cheat on the other party, which can observe cheating at a cost. Under exogenous preferences, penalties need to be very high to deter cheating if there is zero remorse. However, under

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

endogenous preferences, penalties on cheating hurt selfish individuals more than remorseful individuals, who will always comply. Thus remorse becomes an evolutionary stable trait. This in turn causes the optimal sanctions to be lower in the long run.

All other papers in this literature have focussed on *negative* impact of sanctions. Bar-Gill and Fershtman (2004) model the evolution of preferences for fairness in the population as a function of the contract enforcement strength of the legal system. They show first that fairness concerns will be widespread in an exchange economy if legal enforcement is weak. The reason is that preferences for fairness gives sellers a good bargaining position. If an unexpected rise in performance costs occurs, they can credibly threaten not to service the contract unless they get a higher price. However, when the buyer has legal options to enforce the original contract, she may prefer litigation to renegotiation with fair types. Thus, under strong enforcement, fair preferences provide less bargaining benefits, and evolution leads to lower fairness concerns.

Another mechanism by which legal enforcement can discourage the spread of virtuous character traits is given in Bohnet et al. (2001) and Bar-Gill and Fehrstmann (2005). The general idea is that there are two types, virtuous and selfish, which are perfectly observable. In the absence of enforcement, people will trustful only towards virtuous types. (Probabilistic) enforcement may make it worthwhile to trust also low types, so that the latter are better off and increase their share in the population.

In the model of Bohnet *et al.* (2001) the two types play a standard trust game, in which the receiver can decide to cheat or be trustworthy. Under endogenous preferences, there are potentially negative long run effects of a (probabilistic) enforcement of trustworthy behavior by third parties. If the contract is enforced by the third party, the cheater induces a cost (fine, legal costs). When the probability of enforcement is in an intermediate range, the trustors may be inclined to trust even if the trustee is a low type. This raises the payoffs of the cheaters and increases their share in the population. In the long run, only low types remain. Low enforcement on the other hand leads trustors to be more careful and only trust honest types. This means that honest types will eventually take over in the population. The authors test their result by means of an experiment, in which subjects are randomly matched to play the trust game. In all sessions, the last six interaction rounds featured low enforcement probabilities. In the first rounds, enforcement probabilities varied between high, low and medium. In accordance with their hypothesis, the authors find that trustees who interacted only in the low enforcement regime tended to be more trustworthy.

Bar-Gill and Fehrstmann (2005) model a similar logic in the context of a social dilemma game. In their model some agents care for status (high types) and others don't (low types). Agents match randomly in a 2×2 prisoners dilemma game. A decision to cooperate is a public good in the sense that everybody profits from the average amount of cooperation in society. Status can be acquired by contributing to the public good, and the amount of status increases in the

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

average cooperation level in society. The authors show that in the evolutionary equilibrium the high types cooperate with each other if their preference for status is strong enough. The low types defect. If a high type meets a low type, the low type defects and the high type is indifferent between cooperating or not. The unique evolutionary stable equilibrium is one where a fraction of high types cooperate with the low types. The authors now consider the effect of a small subsidy on cooperation. They show that in the short run, the subsidy raises the fraction of high types who cooperate. However, this raises the evolutionary payoffs of the low types, causing their presence to increase. In the long run, this effect dominates, and the subsidy lowers contributions in equilibrium.

This logic depends on the strong assumption that types are observable. Güth and Ockenfels (2005) show that when this assumption is dropped, legal institutions that punish non-cooperation become central to the evolution of cooperative preferences. Obviously, the reason is that cooperation can no longer be conditioned on type, and so private punishment of cheaters is impossible. As a consequence, cheaters will always be at least as well off as non-cheaters.

In summary, sanctions may decrease the relative payoffs of selfish types, which decreases the equilibrium level of such types and decreases cheating. However, countervailing dynamics exist. When types are observable, sanctions on defection or subsidies on cooperation may increase cooperation with selfish types. In the long run, the result is a larger share of such types which lowers aggregate cooperation levels. These are suggestive results with potentially important policy implications. However, they are largely derived in an empirical vacuum, given the almost complete lack of data on the long-run effect of sanctions.

2.4.3 Deterrence, norms and conventions

Following McAdams and Rasmusen (2007) we study conventions and norms as equilibria in coordination games, signaling games, or repeated dilemma games.

Deterrence in coordination games. Coordination games are games in which a player's best response is to mimic the other players' action. This implies that there are multiple equilibria in such games. These equilibria can typically be thought of as conventions, or mere regularities, such as driving on the right or left side of the road. However, normative attitudes may also play a role. Specifically, many games that on the surface look like social dilemmas, may have the character of coordination games due to the existence of people who behave in a reciprocal, or conditionally cooperative way. Such people are willing to cooperate and give up (modest) payoffs that come from defection, as long as they think others do so as well. The reason may be either that they have reciprocal preferences, or they fear informal repercussions of not conforming to standard behavior. The existence of such conditional cooperators is born out by experimental

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

findings. Gächter (2006)shows that conditional cooperation plays an important role in for example public good games, although the resulting cooperation is often fragile. Kahan (1997, 2005) argues that conditional cooperation plays a key a role in a number of social dilemma's such as tax evasion and not-in-my-back-yard problems, and even in the decision to commit crime. Thus, we may think of equilibria in coordination games as conventions or norms, depending on the exact preferences that sustain them.

Sanctions may influence behavior in coordination games in several ways. First, they may simply make one action so unattractive that nobody will play it anymore. Given the practical problems to raising deterrence discussed in section 3, such sanctions may often not be feasible. If sanctions do not make actions dominated, they can still influence behavior by shaping expectations or beliefs.

A first way in which sanctions may shape expectations, also called *the focal point theory of law*, is that law can make a particular equilibrium strategy salient (Cooter 1998, McAdams 2000a). By drawing attention to a specific equilibrium people expect that others will play it, that others think that they themselves will play it and so on. There is indeed experimental evidence that the mere introduction of a law can improve coordination on beneficial equilibria (Bohnet and Cooter, 2003), and that third-party cheap talk can help people achieve such coordination (see Devetag and Ortmann (2007) for a review). McAdams & Nadler (2005) show that third party cheap talk can foster coordination even in 'mixed-motive' games in which players rank the equilibria differently. Brandts and Cooper (2006) explicitly compare communication with small monetary sanctions as tools for a manager to raise coordination levels in a minimum effort coordination game. They find that communication that stresses the mutual benefit of exerting high effort is more effective in raising effort levels than sanctions.

The importance of beliefs has also been found in public good games. Shinada and Yamagishi (2007) conduct a multi person prisoner's dilemma experiment with a baseline treatment and two different punishment treatments. In one treatment, only one player was told that he would be (probabilistically) punished if he did not cooperate. This raised contributions of the threatened player substantially. In a second treatment it was made common knowledge that all participants faced punishment for defection. The authors find an additional effect of punishment that derives from increased expectations of contribution levels of others. Thus, in a population of conditional cooperators, sanctions can have a positive multiplier effect on cooperation. In a field study on tax evasion, Coleman (1997) finds that the most cost-effective way to increase tax payments is to send taxpayers a letter which explicitly states that almost nobody cheats on their taxes. Such a letter increases reported income among a large group of tax payers.

In the real world, this type of reasoning may underlie the famous but controversial 'broken window effect' (see Kahan 1997). The broken window effect is the name for the observed fact that combating small signs of disorder (i.e. broken windows) and minor crimes such as

panhandling can have a dramatic negative effect on the incidence of crime, both small and serious. Kahan (1997) argues that disorder tells residents that others don't care about the neighborhood. This causes them to also become more careless, which results in a downward spiral. In such an environment, zero-tolerance policies that minutely combat disorder may create a virtuous spiral by influencing people's beliefs about the attitudes and actions of others.

In summary, sanctions may have a positive avalanche effect on compliance in the population if they manage to raise expectations and create a self-fulfilling upward spiral. Thus, by it's second order effect on peoples' expectations, sanctions may actually be a more effective policy instrument than standard theory allows.

A second way in which sanctions may influence expectations in coordination games stems from the fact that sanctions are a reaction to the behavior they are supposed to regulate. The introduction of a law can signal that many people are not behaving well or efficiently, which may lower the expectation of future compliance by others.

In the next chapter I present a formal model of this phenomenon. In this model, there is a population of agents playing a public good game and a government that wants to induce cooperation between the agents trough the use of sanctions. The government has more information about the types of agents in the company or society than the agents themselves. In equilibrium, sanctions are introduced only when there is a large number of 'bad types' around, and therefore serve as a signal which makes people more cynical about their peers. Because of the negative signaling effect of sanctions, equilibrium sanctions are lower than they would be under symmetric information. Chapter 4 present experimental evidence that such a signaling effect may indeed occur.

In sum, sanctions may change behavior by influencing expectations about norms or conventions in coordination games. Such changes may be positive, if sanctions raise expectations that others will play the efficient equilibrium, or negative if sanctions indicate that the prevailing norm is to take inefficient or non-cooperative action.

Deterrence in signaling games. An increasingly rich literature focuses on the interaction of deterrence with equilibria in signaling games. We can distinguish between several variants. First, the use of official sanctions themselves can signal private information about the authorities to agents in society. I call this *vertical signals*. Second, sanctions can interfere with *horizontal signals* that are sent between agents in society. I discuss both these ideas in turn.

Gneezy and Rustichini (2000), discussed above, present an example of vertical signaling. They titled their paper on Israeli daycare centers "A fine is a price", to indicate that a fine reveals information about the cost of certain behaviors. They provide (amongst others) the following explanation for the result that higher sanctions lead more parents to pick up their children late

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

from the daycare center: in absence of any sanctions, parents are unsure about the price of coming late. For example, they may think that if they come late too often, the manager of the daycare centre may exclude them from the center's services altogether. Upon observing the sanction they are reassured that the price for coming late is only small, and therefore they will come late more often. Thus, the idea is that there is uncertainty about the 'toughness' of the authority. Small levels of sanctions then show that the authority is actually 'soft' which leads to more deviant behavior.

Ellingsen and Johannessen (2008) suggest an alternative mechanism of vertical signaling. They argue that agents care about gaining esteem of a partner in an exchange. However, the value of esteem to the receiver depends on the perceived character of the partner. People are more eager to earn the respect of an altruistic or 'nice' person than that of a selfish person. This implies what political philosopher Pettit (1995) calls 'the cunning of trust': trusting actions by the first mover can induce a reciprocal reaction, because the second mover wants to earn the esteem of the first mover who has revealed herself to be 'nice'. Bacharach *et al.* (2007) conduct an experiment which allows to distinguish between different reasons to be trustworthy. They do indeed find high levels of 'trust responsiveness': people behave in a trustworthy manner because they believe others trust them.

On the other hand, if the introduction of sanctions and other measures that restrict the choice of the second mover is a signal of selfishness or distrust on the side of the principal, the model of Ellingsen and Johannessen (2008) can explain why second movers will trust *less* when the first mover imposes sanctions that punish trust. Ellingsen and Johannessen (2008) use the result of an experiment by Falk and Kosfeld (2006) to motivate their model. In the experimental game, the first mover is a principal or employer, and the second mover is an employee or agent. The principal has the choice to restrict the choice set of voluntary effort provision of agents, by eliminating the choice of very low levels of effort. The results show that when principal exerts this form of control, the effort exerted subsequently by the agent is lower on average than when the principal does not control.

The two mechanisms identified by Gneezy and Rustichini (2000) and Ellingsen and Johannessen (2008) differ in the assumption of why people comply. Gneezy and Rustichini assume that people are motivated to comply out of fear of (off-equilibrium) repercussions if they don't comply. Thus, the authority can induce compliance by maintaining expectations of sufficiently severe penalties for deviance. Ellingsen and Johannesen (2008) assume that people comply because they want to be esteemed by the authority in case she is nice. The authority should therefore keep up the belief that she is a nice person. Whether these two strategies are mutually exclusive, and which is most effective in different contexts is still an open question.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

We now turn to the interaction of deterrence with *horizontal signaling* between agents. Posner (2000) discusses this issue at book-length. The most interesting example for our purposes is Posner's analysis of the effect of deterrence on the *social meaning* of actions. Posner proposes that people are motivated to signal their discount rate to others to establish a reputation as trustworthy partners. They can do so by engaging in (symbolic) actions, which Posner calls for concreteness 'saluting the flag'. Posner argues that there are separating equilibria in which those who engage in the time-consuming and costly rituals to salute the flag are rightly believed to have a low discount rate. Suppose now the government were to implement a law that obliged everybody to salute the flag. As a result, the separating equilibrium would be lost and no information could be gleaned from observing someone salute the flag. Thus, the law alters the meaning of the behavior. This may cause people to actually stop saluting the flag, because the incentives provided by the law do not make up for the lost signaling value[7].

A formal application of this reasoning is provided by Bénabou and Tirole (2006), who aim to explain how tax incentives can crowd out charitable giving. They assume that people like to be viewed by others as both altruistic and generous with money. In the absence of tax incentives for charitable contributions, the people who are most altruistic and care least about money are the ones that donate, making the signal sent by donating a strong one. When tax-breaks for donations are introduced, people who care about money will now also start donating, which weakens the signal. In equilibrium, this may lead the agents who do not have a strong altruistic motivation to stop donating, possibly resulting in net crowding out of donations.

This model has been tested experimentally by Ariely *et al.* (2009). In a lab experiment the subjects had to 'click for charity', i.e. they performed the boring task of hitting certain computer keys, which were then translated into donations for several charities. The authors contrast a private treatment and a public treatment. In the latter, the subjects had to publicly announce the amount of donations that they accumulated to the other subjects at the end of the session. They also interact these two treatments with a treatment in which clicks yielded additional private benefits. The results indicate that effort and donations are higher in the public treatment. However, when subjects get private benefits, donations in the private treatment go up while those in the public treatment go down. The authors conclude that monetary incentives for behaving pro-socially work better in private settings than in public ones.

Such logic may also apply to criminal acts. For example, Dur (2006) and Silverman (2004) argue that in many inner city subcultures there exists a preference to be seen as 'tough', autonomous and unafraid of others. Criminal acts in these communities are often signals, aimed at establishing the reputation of the perpetrator as a tough type. Dur (2006) builds a model where

[7]There are problems with Posner's general analysis, see McAdams (2000b) for a critique. In this particular case, the mechanism described seems contradictory: a necessary condition for the signal to become weaker when a law is implemented is exactly that *more* people have started signaling, so it is hard to see how deterrence could lead flag-saluting to fall. This is the reason that Bénabou and Tirole (2006) need a two-dimensional type space to generate crowding out.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

committing serious and highly risky crimes signals a high degree of toughness in equilibrium, while committing minor crimes merely signals some intermediate degree. Agents are engaged in a rat-race of wasteful signaling behavior. As in the analysis above, sanctions that affect the payoffs of such signals may cause unexpected shifts in criminal strategies. A zero-tolerance policy that cracks down on minor offenses may discourage the intermediate types from signaling. As a result, the really tough types will be able to signal their types by committing minor crimes rather than serious ones, and the incidence of violent crime goes down. This constitutes an explanation of the 'broken-window-effect', which like the one mentioned above, operates through modifying both the payoffs *and* the expectations of actors.

In sum, by blurring the signals that are sent by virtuous actions, sanctions may end up merely replacing reputational motives for virtuous behavior. On the other hand, when people are engaged in a rat-race of wasteful or criminal activities to signal their types, sanctions can have a positive ripple effect. Deterrence of minor levels of such wasteful activities allow the whole hierarchy of signals to shift downwards, thus decreasing the more severe forms of signaling.

Deterrence and repeated dilemma games. A final source of norms identified by McAdams and Rasmusen (2007) are equilibria in repeated dilemma games. For example, in the infinitely repeated prisoners dilemma there are equilibria in which both players defect forever as well as equilibria in which both players cooperate forever. Either of the two equilibria (or one of the many others) are candidates for a social norm. To my knowledge there are no theoretical studies on the impact of sanctions in such games. However, one interesting study shows how such norms are important to policy makers. Mansour *et al.* (2006) try to explain the following stylized facts about the US: 1) sentences for drug trafficking and police activity to combat criminal gangs who do so went up manyfold in the last to decades, 2) the consumption of illegal drugs went up significantly in the same period, while 3) the price of cocain and heroin decreased by a factor 5. The authors explain these facts using a model of gang formation. The model assumes that in the first round of the game, coalitions are formed by gang members who agree (and can commit) to sharing all profits from drug trade. In the second round, gangs engage in Cournot competition on the market for drugs. Exogenous variables are the size of the market and the deterrence regime. An active assumption is that the detection probability increases with the size of the gang. The authors show that under these assumptions, stronger deterrence may increase drug productivity. The reason is that sanctions increase the cost of operating a large gang, which may cause gangs to split up. This increases competition and may result in more supply and a lower market price, even correcting for the probability that some gangs get detected. The authors argue that this story resembles what happened after increased deterrence dissolved the Mendellin and Cali drug cartels: the number of criminal organizations involved in the production of cocaine increased, and this was eventually translated into an increase in total production.

Although Mansour *et al.* (2006) do not explicitly rely on a repeated game, the lesson to draw from this study can easily be generalized to such situations. If criminals compete in the production of crime in cartel-like structures, deterrence functions as a crude form of anti-trust. Like anti-trust, deterrence may increase production by increasing competition. However, in the context of criminal activities, this is the exact opposite of the intentions of the policy maker.

2.5 Discussion

On first sight the lack of empirical support for the deterrence hypothesis seems like bad news for economic theory, which makes universal claims about the effects of incentives on behavior. However, in the previous section we have seen that the theory of incentives is more versatile than the simplest formulation of Becker's deterrence theory suggest. Observing these new approaches, it is perhaps surprising that it took rational choice theorists so long to face up to the rather obvious fact that informal institutions matter for the effect of formal policies.

The reason is that below the deterrence hypothesis is a deeper 'Hobbesian' paradigm of economic man that underlies much of standard economic theory. Economic man as traditionally conceived is selfish and individualistic. A society that results from having such agents live together can be called Hobbesian, after the 16th century philosopher Thomas Hobbes (1588-1679). Hobbes' argument, in a nutshell, is that in absence of authority or law, humans find themselves in a "natural condition of mankind". In this natural condition, "every man has a right to everything", and the result is a struggle for resources, and a continuous "bellum omnium contra omnes". Therefore, people are better off by reaching an agreement, or 'social contract', in which they yield their power to a central absolute authority, called the Leviathan. The Leviathan is both lawmaker, executive and judiciary, and creates order establishing and maintaining the law through corporal and pecuniary punishment. In short, the Leviathan solves what would otherwise be the worst case scenario in which everybody cheated everyone.

The Hobbesian view has much to recommend itself. It is a great improvement over theories of the good that rely on idealized and mistaken conceptions of human beings as intrinsically good and nice to others, or as being able to find their way to cooperative conduct through pure reasoning. Clearly though, Hobbes' metaphor of the state of nature is a caricature of human societies. Hobbes' assumption that if there is no authority or law, life will conform to the metaphorical "war of all against all" is wrong. People have numerous mechanisms of maintaining order in the absence of authorities. These mechanisms are based on social norms and reciprocal arrangements. Without a centralized government, people would not go around as atomistic individuals, dividing their time solely between killing and eating loot. This is not so much an argument against Hobbes' work, because the state of nature serves to drive home his point and defend the idea of a social contract. Problems start when the metaphor is used as

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

an anthropological assumption to recommend certain policies. Insisting on a Hobbesian view is harmful because, as we have seen from the preceding survey, it does not properly predict the effect of our policies in cases that social norms or reciprocal arrangements are present.

Insistence on the importance of social forces is of course not new. Many writers have commented that effective social control is based needs to rely on the informal forces in society. According to criminologist David Garland in his book *The Culture of Control* it is a 'basic sociological truth' that

> "[T]he most important processes producing order and conformity are mainstream social processes, located within the institutions of civil society, not the uncertain threat of legal sanctions. The project of establishing a sovereign state monopoly has begun to give way to a clear recognition of the dispersed, pluralistic nature of effective social control. In this new vision, the state's task is to augment and support these multiple actors and informal processes, rather than arrogate the crime control task to a single specialist agency."
>
> D. Garland (2001, p.126).

As we have seen in this survey, new approaches within economics recognize the importance of social processes for policy making. If we analyze deterrence in interaction with (social) preferences and as impacting on equilibrium behavior in games, played by people who maintain moral values, its effect becomes more ambiguous. Sanctions may destabilize equilibria in such games, and the behavior in the new equilibrium may be very different than the simple deterrence hypothesis would suggest.

However, expanding the concept of economic man has further reaching implications for policy making than just casting doubt on the effect of deterrence. It also raises questions about the set of policy instruments that are available. Economists know how policy makers can set incentives. But policy makers may also have tools to influence values in a more or less direct manner, such as educational campaigns. A more complicated model of human nature thus also requires examination of a broader palette of policy instruments. Moreover, policy instruments will interact with one another in producing results.

A real world example of this idea can be found in Lappi-Seppälä (2001). The author describes the remarkable change in the philosophy of the Finnish penal regime during the last 40 years when insights about the interdependence of different policy instruments replaced a system that was based predominantly on repression. One of the slogans of the new ideology was that "Criminal policy is an inseparable part of general social development policy". This slogan reflects that moral education and practiced at schools and other institutions is also a part of an integrated framework to shape values in society (Lappi-Seppälä, 2001, p. 110).

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Thus, deterrence should be analyzed in conjunction with other policy instruments, rather than separate from it. To make this point somewhat more concrete, the remainder of this chapter is dedicated to its illustration. The idea is that social policies that influence norms and values also affect the optimal level of deterrence and other social incentives. One intuition is that social policies that weaken norms promoting deviant behavior will reduce the need for deterrent strategies. This is a simple matter of substitution: if the deterrence becomes more effective, the policy maker needs less of it to direct the agent's behavior. However, the opposite may also be true. If deterrence becomes more effective, it may become optimal to use more of it. In the next section I illustrate this latter point in by analyzing the interaction between different policies in the presence of social norms.

2.6 A formal illustration: soft and hard approaches to crime

The model I present in this section is not aimed to be realistic enough to give policy advice, but rather to illustrate two of the central arguments of this chapter. First, I want to show that the effect of deterrence is potentially ambiguous in environments where there is a social norm. Second, I want to show that 'soft' policies that aim to change norms are compliments to deterrence policies. I will show that although deterrence is potentially counterproductive on its own, it is productive when it is part of a broader policy strategy to combat crime.

To illustrate the model, consider an environment which on first glance comes perhaps closest to a state of nature in Western society: poor American inner city neighborhoods. These neighborhoods often suffer from high juvenile crime rate. In such environments, it is very important to have the reputation of being tough, i.e. not afraid of conflict or fights, because this prevents one from being the victim of crime. Silverman (2004) surveys stylized facts that are consistent with this, such as the fact that many violent crimes are committed in front of witnesses. A reputation for toughness can be maintained by criminal behavior, violence against others, and defiance of the authorities. Dur (2006) collects ethnographic evidence from several sources supporting the importance of such norms. Topalli (2005, p. 797) writes

> "Traditional subcultural theorists maintain that offenders operate in an environment in which oppositional norms catering to ethics of violence, toughness and respect dominate the social landscape."
>
> V. Topalli (2005, p. 797).

From his own interviews with over 200 hardcore uncaught street offenders, Topalli is able to corroborate this view. If our analysis in the last section was correct, such oppositional norms will influence the effect of deterrent policies. A case study in this point is elaborated in Kahan (1997).

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

He considers the problem of combating gun control amongst youth in inner city neighborhood. Studies show that deterrent measures that aim to suppress gun ownership, e.g. metal detectors at schools, usually fail.[8] The reason lies in the motivation for possessing a gun. In a survey, 66% of the respondents listed that they own a gun to impress friends or peers, 56% said they wanted to be powerful or important, and 49% said it was for protection (see references in Kahan, 1997). Owning a gun confers status, because it shows you are tough and autonomous enough to defy authority. Not owning one shows that you are weak and this may make you a victim of aggression. A repressive strategy can reinforce the signal sent by gun ownership, because it raises the degree of defiance necessary to carry a gun. In Kahan's words "the crackdown strategy is at war with itself". It turns out that a policy that does seem to be effective is to pay others students to turn in gun possessors. This policy works because it changes the expressive value of having a gun. Suddenly, those with guns are vulnerable to betrayal from within their own communities, and do not appear so strong anymore. Also, showing guns in public becomes now less attractive, which defeats the status-building purpose of gun possession.

Authorities can use a host of alternative instruments to decrease the visibility and status of criminal acts in the community. Authorities may educate residents members about the effects of criminal conduct on the community, they may offer anti-violence courses, organize alternative pass-times for disenfranchised youth such as sporting events, they may emphasize role-models that have found success by socially acceptable means etc. To reduce visibility, they may enlist community members to report on crime to the police or to tell on criminals, as in the example above. We will now show that there are complementarities between both strategies in combating crime.

2.6.1 The model

The model consists of an authority or government, and a large population of agents. We do not explicitly model the decision of the authority; the aim of this example is not to derive conditions on optimal deterrence. Instead we simply examine the impact of the exogenous levels of two different policy instruments on the crime rate. One instrument is a *hard* policy, aimed at enforcing the law with deterrent measures, such as high level of police monitoring, or high penalties for offenders. The other is a *soft* policy, which tries to curb the social norm of respect that exists in the community for law offenders in the ways described above. The level of the hard policy is denoted by $h \geq 0$ and the level of the soft policy by $s \geq 0$.

There is a countably infinite population of agents indexed $i = 1, 2, ...$ Each agent is of a type $\theta \in [0, 1]$ that indicates her 'toughness' (defined below). The higher the type, the tougher the agent. Agents are distributed over the type space according to the continuous cdf $F(\theta)$. The

[8] See Kahan (1997, footnote 61, who cites several studies to this extent).

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

agents preferences are given by the following vNM utility function:

$$U(d_i, h, s) = B(\theta_i, h, d_i) + a(s)E\left[\theta_i \mid d_i\right]. \tag{2.1}$$

The first term in (3.2) is benefit of crime to the individual, the second term is the *respect* that the criminal earns. We assume that the benefit of crime takes the following form:

$$B(\theta_i, h, d_i) = \begin{cases} 0 & if\ d_i = 1 \\ \theta_i - h & if\ d_i = 0 \end{cases} \tag{2.2}$$

Naturally, the benefit of crime is 0 if the agent complies. If she does not comply, she incurs a benefit that decreases in the level of hard enforcement effort, and increases in the toughness θ of the agent. Thus toughness is defined in the model such that for any level of enforcement, tough agents gain more (lose less) from non-compliance than wimpy agents. One interpretation is that tough agents gain more from crime because they suffer less from the conflict situations in which crime may bring them; they are better able to handle rough treatment by the police and to survive difficult conditions jail than wimpy agents. One can also interpret this parameter as a moral stance: tough agents 'don't care' about breaking the law and/or inflicting damage on others, whereas wimpy agents' conscience suffers from engaging in criminal activities.

The second element of the utility function is the *respect* that the criminal earns by showing to be tough. $E\left[\theta_i \mid d_i\right]$ denotes the expectation of the other agents about the type of agent i. We assume that the agent derives positive utility from this. Reasons can be that he will be less likely to be attacked by other agents, secures resources through a better bargaining position, or that there are sexual benefits of being known as tough. The importance of being tough is measured by the function $a(s)$. We assume that a is continuous, decreasing and that $a(0) > 0$. The idea that the authority can influence the sources of esteem in the community is elaborated above.

2.6.2 The effect of deterrence in equilibrium

In this section we derive the existence of a perfect Bayesian equilibrium in the game and some implications for the effectiveness of soft and hard enforcement approaches. All proofs are in the appendix. I assume the tie-breaking rule that an indifferent agent complies.

Proposition 2.1 (Equilibrium compliance rate)**.** *If $\underline{h} \leq h < \overline{h}$, then there exist at least one partial pooling equilibrium in which all types lower than a threshold type θ^* comply and the others don't.*

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Proposition 2.1 says that if h has some intermediate level, it can give rise to a partial pooling equilibrium, in which only the types lower than θ^* comply. The most interesting implication of Proposition 2.1 is that for given levels of h and s, the compliance level may be *indeterminate*, i.e. there may be multiple equilibrium threshold types. The reason is that there may be different combinations of the threshold toughness θ^* and the respect premium $\delta(\theta^*)$ that satisfy the equilibrium conditions.

To see why this is the case, we need to examine the equilibrium conditions more closely. In a partial pooling equilibrium as described above, non-compliers receive a reputation premium, because non-compliance shows you are amongst the 'tough' guys. This premium is the *difference* between the signal sent by compliance and non-compliance, i.e. $\delta(\theta^*) = E[\theta \mid \theta^*, d = 0] - E[\theta \mid \theta^*, d = 1]$. We call $\delta(\theta^*)$ the *respect premium*.

Importantly, how the respect premium changes in θ^* depends on the shape of the distribution $F(\theta)$. For example, the uniform distribution yields a constant difference, $\delta(\theta^*) = \frac{1}{2}$. However, as a moment's thought will reveal, any other distribution will induce the respect premium to vary over the interval $[0, 1]$. The curved line in Figure 2.1 depicts a respect premium based on some ad-hoc distribution $F(\theta)$. This line thus shows the reputational benefits in any equilibrium characterized by θ^*.

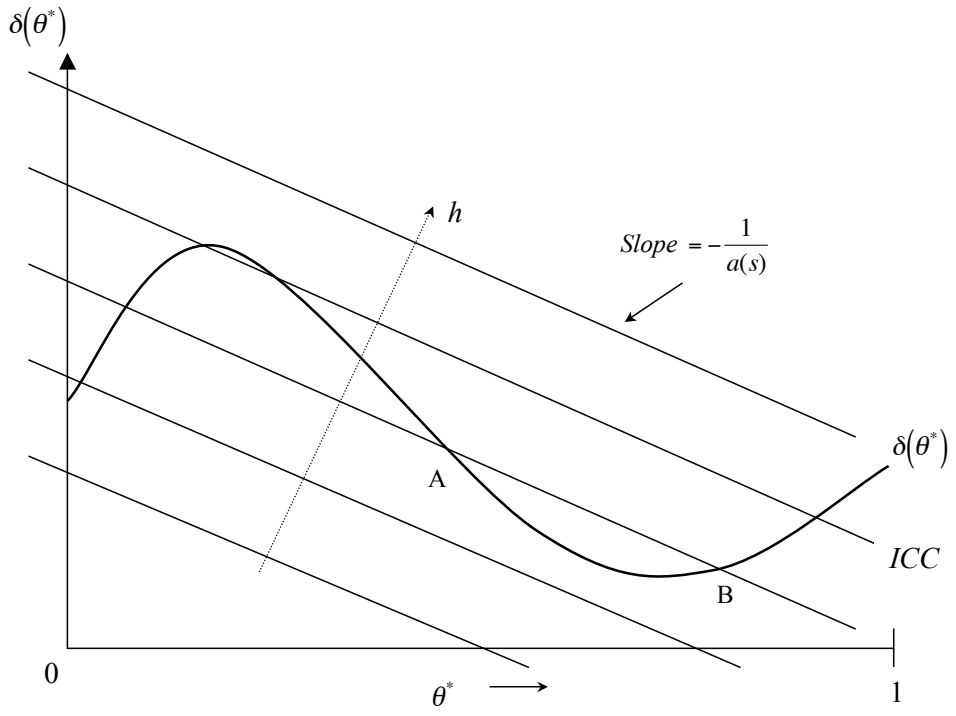

FIGURE 2.1: Equilibrium compliance levels.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

The straight lines in Figure 2.1 represent the incentive compatibility constraint (ICC) of the threshold type. Remember that the threshold type θ^* is the type who is just indifferent between incurring the punishment and incurring the respect premium $\delta(\theta^*)$, so the ICC is given by

$$h - \theta^* = a(s) * \delta(\theta^*)$$

Thus, the ICC line in Figure 2.1 represents all pairs $\{\theta^*, \delta(\theta^*)\}$ such that the threshold type is indifferent between complying and not complying for given levels of h and s. Above this line the respect premium is sufficiently high to motivate non-compliance, and vice versa below it. The position of the ICC is determined by h. A higher h shifts the ICC upwards, because a given threshold type will need a bigger respect premium to compensate for the loss brought about by the penalties. The ICC slopes downward, because a higher threshold type will need less respect for a given h. The slope is determined by the norm strength: a strong norm (high $a(s)$) means that a decrease in the respect premium leads to a large fall in utility which needs to be compensated by a big increase in toughness of the threshold type. This implies a flat ICC. Thus, a higher level of soft policy s (i.e. a lower $a(s)$) decreases the norm strength and makes the ICC steeper.

Equilibria of the game are found on the intersection of the ICC with the $\delta(\theta^*)$ curve. Here, the equilibrium respect premium $\delta(\theta^*)$ makes the threshold type exactly indifferent given the levels of h and $a(s)$. The fact that the respect premium is not constant, as depicted in the example in Figure 2.1, means that there is potentially more than one equilibrium pair $\{\theta^*, \delta(\theta^*)\}$. For a given level of sanctions, there can be equilibria with low levels of compliance (low θ^*) and high respect premia $\delta(\theta^*)$, or with high levels of compliance and a low respect premia. The reason is that respect premia may fall with the threshold type, so that sanctions aimed at raising the threshold type will be counteracted by the increase in the respect premium.

We can now evaluate the effect of a change in the level of deterrence h:

Proposition 2.2 (The effectiveness of deterrence)**.** *In the equilibrium characterized in Proposition 1,* $\frac{d\theta^*}{dh} < 0$ *if and only if* $\frac{d\delta(\theta^*)}{d\theta^*} < -\frac{1}{a(s)}$.

Proposition 2.2 tells us that the effect of a higher level of deterrence h in the separating equilibrium, can actually *increase* crime. This counterintuitive result can be explained as follows: raising h implies that either the equilibrium threshold type needs to be tougher, or the respect premium needs to rise to compensate the threshold type for the greater loss caused by deterrence. If in equilibrium the respect premium is decreasing in θ^* and the norm strength is high, a combination of a *lower* threshold type and a higher respect premium may constitute a new equilibrium. In other words, increased deterrence may coincide with increased informal incentives in favor of crime, resulting in a higher crime rate. Graphically, one can see Proposition 2 in Figure 2.1. Point A corresponds to a point where an increase in h will lead to a decrease in

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

θ^*. This happens when the ICC crosses the $\delta(\theta^*)$ line from below. Point B on the other hand represents an equilibrium where an increase in h lowers crime.[9]

For our purposes, an interesting implication of Proposition 2.2 is that if s is sufficiently high, h will always have a positive impact on crime. This is because s diminishes the importance of respect, and thus a high s rules out the possibility that increased respect incentives counteract deterrent incentives. Graphically, a higher s rotates the ICC in Figure 1, and makes sure the ICC crosses the $\delta(\theta^*)$ line from above.

In sum, while deterrence is potentially counterproductive when applied in isolation, it is productive in combination with a policy to disrupt oppositional social norms. Depending on the shape of the distribution of types, hard forms of deterrence may be 'at war with themselves', because they may increase the status that is associated with crime. Softer forms of deterrence that aim at weakening such effects may be useful complements to the standard deterrence approach. Because the policy implications depend on the exact distribution of types which is hard to observe, this model is not suitable for making policy recommendations. But it does illustrate two things. First, the 'social landscape', made up by social norms and values is crucial to determine the effect of deterrent polices. Second, policies that influence norms and values will have an impact on the effect of deterrence. Considering a comprehensive policy strategy that covers several instruments to combat crime is therefore indispensable.

2.7 Conclusion

In their survey of optimal enforcement in Becker's tradition, Polinsky and Shavel (2000) find many aspects of law enforcement to be congruent with theory. Low probability of detection is often combined with high punishments and, punishments are higher for more serious offenses. However, they note that the general level of deterrence is often 'too low' from a theoretical point of view. The authors argue that

> "Given the ample opportunities that exist for augmenting penalties, as well as the possible desirability of increasing enforcement effort, society should probably raise deterrence in many areas of enforcement."
>
> M. A. Polinsky and S. Shavell (2000, p. 72).

[9] Doing such comparative statics when there are multiple equilibria may not be so convincing. However, if $\delta(\theta^*)$ is downward sloping over the whole range, there may be a unique equilibrium in which the effect of increasing h is to lower compliance.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

This may be a good recommendation in some cases, but the studies surveyed in this chapter suggest that the one-size-fits-all attitude embodied in this statement is mistaken. We have seen that the effect of deterrence is context dependent, and that we should not always expect it to be effective. Theory suggests that when deterrence works well, a large part of the effect may not come from its direct influence on payoffs. Rather, as in the case of the broken-window effect, it may come from it's indirect effect on beliefs about what others are doing or the disruption of harmful signaling behaviors in the population.

The argument in this chapter suggests several avenues for future research. Empirically, we need a focus on contextualized studies, preferably with data on community level or lower. Rigid models that aggregate statistical effects over large groups of people give us little information about how deterrence affects behavior, as the debate over the deterrent effect of handgun ownership (discussed in section 2.2) vividly demonstrates. Instead, carefully executed field studies in (semi-)controlled environments are needed to inform a more taylor-made policy approach to deterrence. On the other side of the spectrum, laboratory studies can also be made more contextualized by reconsidering the sacred dogma that framing in the laboratory should devoid of any references to social context. In fact, such neutral framing makes it even harder to apply the results outside of the lab. Although for example Hörisch and Strassmeier (2008) have demonstrated that deterrence can raise stealing in an abstract setting, it is not clear in which circumstances we may expect this result in the real world. Preferably, such result should be subjected to replication attempts in the field. Finally, studies into the long term effect of sanctions on values are almost entirely lacking. In order to be able to isolate the effect of deterrence, such studies could combine the output of value-surveys with the institutional arrangements in different countries or states.

Turning to theory, much can be gained by a further integration of the existing paradigms in sociology, law, psychology and economics, a research program that is already well underway. As I mentioned in the last section, specific attention should go to the effects of combining different policy instruments in order to formulate a more effective and integrated policy approach. Theoretical research on optimal deterrence could also learn from the field of industrial organization. In industrial organization the endogeneity of the interactions between firms, i.e. the market structure, is central to policy analysis. Successful antitrust measures, merger policies, r&d subsidies etc. operate explicitly in an environment where firms are interacting in both competitive and collaborative arrangements. Similarly, optimal deterrence should be calibrated to a society of individuals that play games of signaling, coordination and long term cooperation with each other. In this way, rational choice may finally fulfil Gary Becker's promise in the opening quote of this chapter, and serve as a 'useful' theory of crime.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Chapter 3

The signaling power of sanctions in collective action.

> "Laws are partly formed for the sake of good men, in order to instruct them how they may live on friendly terms with another, and partly for the sake of those who refuse to be instructed, whose spirit can not be subdued, or softened, or hindered from plunging into evil."
>
> Plato - The Laws

3.1 Introduction

What determines cooperation in social dilemmas has been a core problem for social scientists since the beginning of the discipline. Ever since Hobbes in the 17th century threatened the infamous 'war of all against all', the dominant strand of literature highlights the role of sanctions in coercing people to cooperate. But contemporary empirical research shows that people manage to find ways to cooperate even without the presence of government. There is substantial evidence that society has a large proportion of so called conditional cooperators: agents that condition the decision to cooperate on what they think others do. The existence of such agents means that collective action problems may be partly a matter of coordination, and substantial cooperation may be achieved without the need for much coercion. However, in the absence of high sanctions, a necessary condition for such cooperation is trust; the belief that others are willing to cooperate.

Thus, if society is indeed a heterogenous mix of egoists and conditional cooperators, a pressing and largely ignored question is how coercion and trust can be combined to induce cooperation. Specifically one may ask if trust between agents is independent of the use of sanctions? This chapter offers an answer to this question by presenting a model in which trust and coercion

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

interact in determining cooperation. It argues that there is a trade-off between sanctions and trust. High sanctions are necessary when there are many egoist around, but they can also 'crowd out' trust. This can happen when the sanctioning authority has superior information about the kind of people that make up a society, because in this case conditional cooperators will infer from the introduction of sanctions that others are likely to be selfish. This in turn decreases the willingness of conditional cooperators to cooperate. Conversely, the government can signal that cooperation is the norm in society by setting low sanctions, and thus 'crowd in' cooperation.

The point of departure of the model is a standard social dilemma or public good game. The game is played by a large population of heterogeneous agents: while some of them are selfish, others are conditional cooperators who don't mind contributing if sufficiently many others do so. Agents know their own type, but not that of the other players. It can thus be rational to either cooperate or defect, depending on a player's own type and the expectation of the type of the rest of the players. The model includes a government or principal, who knows the distribution of agents' types in society, and can alter the payoffs of the game by introducing sanctions for defection.

The main result is that the asymmetric information about the distribution of types can lead the government to set lower sanctions than it would do under complete information. I show that if conditional cooperators coordinate on mutual cooperation, there is a unique class of perfect Bayesian equilibria in which the government sets high sanctions if there are many egoists in society, and low sanctions if there are many conditional cooperators. This means high sanctions give a negative signal (to the conditional cooperators) and crowd out the belief that others are of a high type. Although this decreases the motivation of conditional cooperators to cooperate, there is no crowding out on the behavioral level, because the coercive power of the sanctions compensates for the effect of decreased trust in others. However, the signaling effect of sanctions leads the government to set lower sanctions in equilibrium to 'crowd in' trust between citizens.

The model has applications in social dilemmas in large scale societies or organizations. An application to tax evasion is discussed in the last section. The model asserts that the reason why real-world policies of tax evasion often feature low sanctions, is that governments rely on the reciprocal preferences of the tax-payers. The model suggest a rationale for evidence that raising sanctions on tax evasion sometimes has very little, or even a negative effect on tax evasion (Sheffrin and Triest, 1992). Being tough on tax evasion sends a mixed message: although evaders are being punished, they must be numerous to be taken so seriously. Thus, the article emphasizes a balancing act that the government must perform: It must deter those who are, to speak with Plato, inclined to 'plunge into evil', while maintaining the good men's motivation to live on friendly terms.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

3.2 Literature

There is an increasing amount of evidence for the existence of so-called conditional cooperators. A conditional cooperator is someone who will cooperate if she thinks others will do so as well. Fehr and Gächter (2000) and Gächter (2006) review the evidence on conditional cooperation from public good games and field experiments. They conclude that a large amount of studies finds much more cooperation than standard economic theory allows for, and that much of this cooperation is conditional on (expected) cooperation of others. However, there is substantial heterogeneity in these preferences for reciprocity or conditional cooperation. Fischbacher and Gächter (2006) among others, provide experimental evidence for the existence of a number of types whose behavior is stable across games. They find that close to 55% of their subjects act as conditional cooperators, 25% act as pure free riders, and the rest shows more complicated behavior, that often resembles conditional cooperation in the relevant range of play. Another source of evidence for conditional cooperation comes from field experiments that study contribution levels to charities. The results of four studies surveyed in Gächter (2006) are that those subjects who received information that others contributed a lot also contribute a lot. For example, Frey and Meyer (2004) find that students contribute significantly more to charity funds if they were told that others contributed more in the past.

The existence of conditional cooperators implies that trust is a crucial variable for cooperation. Without being overly sophisticated, we can define trust in a collective action setting as a person's *belief* that others in society are of a virtuous nature and therefore trustworthy (we provide a more detailed definition below). The literature on trust in economics has largely been concerned with the consequences of trust for the economy. However, the question of how beliefs are determined by institutional arrangements has received much less attention.

One strand of literature that does investigate the relation between beliefs and institutions are theories that combine the analysis of law and social norms (see for a survey McAdams and Rasmusen, 2007). These theories hold that official rules have an impact on behavior apart from their influence on payoffs. One of the ways in which they have such influence is by changing people's expectation of what others do. Cooter (1998) argues that non-deterrent laws may help people in this way to coordinate on efficient outcomes. For example, Tyran and Feld (2006) show in an experimental setup that mild, non-deterrent laws, can be effective in raising contributions in a public good game if they are the result of a public voting procedure. Such a procedure allows people to express their intentions to cooperate. However, Kahan (2005) emphasizes that an informational effect of the introduction of laws can also be negative. Official incentives express information about the dominant social values and norms in society. Consequently, a blanket crackdown on defection by the government in the form of high sanctions will give people the idea that non-cooperation is the prevailing social norm. To the extent that people are conditional cooperators, this reduces their own willingness to cooperate. This dual role of

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

incentives is the main message of this chapter. In our setup, incentives have the traditional motivational effect that economists take them to have, but they also shape the perceptions of people about the conduct of others in society.

In the next chapter I document some direct evidence for the signaling effect of sanctions. In general, the phenomenon falls into a category of studies that document crowding out effects of sanctions on cooperation. A number of experiments in psychology and economics, both in the laboratory and in the field document that sanctions for deviant behavior sometimes increase such behavior. This literature was briefly alluded to in chapter 2, which discussed the study by Gneezy and Rustichini (2000). In their field experiment they consider ten day-care centers in Haifa. In five of them they introduce a fine for parents who pick up their children late. In these five centers the number of late-comers went up significantly in the weeks after the introduction of the fines and stayed up relative to the control group even after the fines had been withdrawn.

An increasing amount of studies documents similar findings in social dilemma settings. Frey and Oberholzer-Gee (1997) find that people are *less* likely to accept siting of waste facilities in their neighborhood when they are offered substantial financial compensation compensation for it. They use several indicators of 'civic-mindedness' to predict individual choices whether to accept the facility. They find that when compensation is offered, civic mindedness is no longer a predictor of this choice. They conclude that the compensation reduces the feelings of civic duty of citizens. Ostmann (1998) provides experimental laboratory results that show that external enforcement financed by experiment participants only reduces 'harvests' in common pool problem by a small amount relative to a no-enforcement treatment. Frey and Jegen (2001) and Bowles (2008) present surveys of the rapidly expanding empirical literature in this field.

There is as yet little theoretical insight in the mechanisms that underlie these empirical results. Most explanations rely on a notion of 'intrinsic motivation', which is reduced when incentives are introduced. However, this notion does not help much in predicting the kind of circumstances in which crowing out will occur. In this study I show that standard rational inference upon observation of sanctions can generate crowding out of trust, which serves as an intrinsic motivation to cooperate. Although in the model this does not lead to net crowding out on the behavioral level, it does affect the optimal level of sanctions. Two theoretical papers present signaling models of crowding out. They both do so in a principal-agent context. In Bénabou and Tirole (2003) the principal has more information about the characteristics of a job and the ability of an agent to do it than the agent himself. The incentives that the principal chooses to introduce are therefore a signal to the agent that he might not be suitable, which diminishes his motivation for the job. Sliwka (2007) also considers a principal-agent context, in which there are three types of agents in a firm: altruists, who take into account the principal's payoff, egoists, who maximize their own material payoff, and conformists, who prefer to do whatever they think the majority does. Because preferences of conformists depend on their beliefs about others, this

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

is a psychological game. In this setting, the introduction of tight control by the principal may signal to the conformists that most people are selfish and this in turn will cause them to lower their effort. The principal may thus choose to trust rather than control the agents.

Like in Bénabou and Tirole (2003), the signaling effect in the present model arises because the government moves first and has more information than the agent, a reversal of a traditional assumption in the literature. I adapt the framework of Bénabou and Tirole (2003) by incorporating multiple agents and model strategic interaction between them. The incentives that the government uses convey information, but instead of learning something about their own type, the agent learns something about the type of the other agents. This signaling effect is similar to that in Sliwka (2007), but the models differ both in their focus and assumptions. Instead of focusing on the vertical principal-agent relation, we look at the effects of information transmission on the horizontal cooperation *between* agents in a public good game. In this context, the model is applied to a concrete technology of social control, namely official sanctions. My assumptions are more traditional than in Sliwka (2007). First, I do not use a psychological game. Beliefs in our model do not induce a preference change but serve the more traditional role of anticipating payoffs. Moreover, Sliwka (2007) assumes that there is a large proportion of unconditionally altruistic types in the population, an assumption which is rejected by the (experimental) evidence. I deviate from the standard homo economicus only by assuming the well-documented conditional cooperator.

3.3 The model

The model is a sequential game of costly signaling with three different kinds of players: agents, a principal and nature. The principal can be a government or a manager, and the agents correspondingly citizens or employees. Applications exist in both public and organizational context, but throughout this chapter I will frame the problem as a public one, and use the words 'government', 'citizens' and 'society'.

The central idea is the following: The citizens play a public good game with incomplete information. In contrast to standard assumptions, some of the citizens are conditional cooperators, who contribute only if they think a sufficient number of others does so. Whether mutual cooperation can be an equilibrium thus depends on the distribution of the types of the players. The citizens don't know the distribution of types, but have a common prior over the possible distributions.

Nature starts the game by determining the distribution of types (thus transforming the game into one of imperfect information). The government is the only player who observes this distribution. It's objective is to maximize contributions to the public good. To this end it chooses the level of sanctions for defection. The sanctions are observed by the citizens in the economy before

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

they choose their own action. Since the government has more information than the citizens, the citizens may make inferences from the sanctions about the distribution of types in society. There is thus double-sided asymmetric information: citizens have private knowledge of their type and the government has private knowledge of the distribution of types. In section 4 we derive the equilibria of the game and show that asymmetric information may lead the government to set lower sanctions in equilibrium.

Nature. At the beginning of the game, nature determines the types of all agents in society. With probability ω each agent is chosen to be a high type. This probability is itself a random variable Ω, of which nature determines the realization. The probability that nature picks a given ω is given by a uniform distribution with support on $[0, 1]$. Thus, ω is the proportion of conditional cooperators in society and $1-\omega$ is the proportion of egoists. We call the distribution characterized by ω *the state of society.*

The government. The government is the only player (apart from nature) to observe the state of society. Thus, ω is the 'type' of the government. The motivation for this assumption is that governments or managers have an advantageous position to collect information about their citizens or employees. Governments employ bureaucracies that collect statistics on the aggregate behavior of citizens and keep records of the amount of law-violations. By making policy they also gain information about the reaction of the citizens. Managers meet with employees in different departments of the firm and monitor productivity, working hours and indices of their corporate culture. Although the assumption of perfect knowledge of the type distribution is obviously extreme, it is likely that the combination of these information sources lead governments to have to superior knowledge about society than any individual would have.

On the basis of its knowledge, the government sets incentives $g \in \mathbb{R}$. The objective is to maximize cooperation by the citizens in the economy. The instrument to do so is the use of costly 'sanctions', a punishment on defection by the agents. (We will use the words 'sanctions', 'punishment' and 'incentives' interchangeably.) The government's objective function is:

$$W(m, g) = m - \alpha g \tag{3.1}$$

Here, m is the fraction of contributors in society, and $0 < \alpha < 1$ is a cost parameter. We offer two interpretations for the idea that higher sanctions carry higher cost. First, one can interpret these costs as the practical expenditures necessary to sustain a higher level of deterrence, such as putting police on the street or raising the probability of getting caught. Second, α can measure the moral cost of high sanctions, reflecting the idea that in a liberal society moral the punishment should be proportional to the crime. Although many people would agree that

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

stealing a bike is wrong, few would want to institute the death penalty for bike thieves, even if this were the most efficient way to deter them. It is of course an empirical question whether this moral constraint actually binds in a given application, but it is likely to limit the availability of 'cheap' deterrence strategies that combine harsh penalties with low enforcement.

Note that I do not necessarily interpret the sanctions as fines, and there are no revenues to the government from the sanctions. Although fines could be part of a sanctioning scheme, I want to focus purely on the deterring or Hobbesian effect of sanction and not on the revenue-raising aspect. Note also that sanctions (and their costs) are set before citizens choose their actions. This implicitly assumes commitment by the government to carry out the sanctions once they are in place. This is natural in a setting where sanctions are decided upon by politicians, and their execution and enforcement is subsequently carried out by the executive and judiciary branch of government.

Finally, the setup can easily be extended to include incentives in the form of subsidies or rewards. If the government has the possibility to reward cooperation with a costly subsidy, doing so would send the same signal as sanctioning defection: incentives are apparently necessary because there are many egoists. Any incentive scheme that is costly to the government and raises the citizens' expected utility of cooperation relative to that of defection sends such a signal.

The citizens. We assume that there is a countably infinite population of agents or citizens of measure 1, indexed $i = 1, 2, ...$ There are two types of citizens. A fraction ω is a so-called conditional cooperator or high type, the rests are egoists, or low types. After nature has determined the type of each agent (and thereby the distribution), and the government has set its policy, each agent chooses a contribution level $c \in \{0, 1\}$. The payoffs π_e of an egoistic agent i are as follows:

$$\pi_i^e(c_i, m) = h(m) - c_i - g(c_i) \tag{3.2}$$

Here, $h(m)$ is the individual payoff from the public good, financed by the contributions. We assume that $h(m)$ is increasing in the fraction of contributors m. Because the population consists of an infinite number of agents, the individual contribution is so small relative to the population size that we disregard its impact on m. This approximation simplifies things substantially. The second term c_i is the individual contribution and $g(c_i)$ is the government sanction, which is imposed only if the agent defects:

$$g(c_i) = \begin{cases} 0 & if\ c_i = 1 \\ g & if\ c_i = 0 \end{cases}$$

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

It is easy to see that (3.2) induces a social dilemma, because in the absence of sanctions it is a dominant strategy for the egoists not to contribute. Egoists will only contribute if the sanctions that the government sets for non-contribution are high enough, that is, if $g \geq 1$.

The payoffs π_c of a conditional cooperator are given by:

$$\pi_i^c(c_i, m) = \begin{cases} h(m) - c_i - g(c_i) & if\ m < \overline{m} \\ h(m) - \theta c_i - g(c_i) & if\ m \geq \overline{m} \end{cases} \tag{3.3}$$

Here $\theta \in (0, 1]$ and $0 < \overline{m} < 1$. If aggregate contribution levels are low, conditional cooperators have the same cost of contributing as egoists. If aggregate contribution levels are high, the cost of contributing for an conditional cooperator is lower than that of an egoist. In fact, egoists are a special case of conditional cooperators with $\theta = 1$. The type space can thus be written $\Theta = \{1, \theta\}$.

We can interpret the parameter θ as a 'warm-glow' from contributing that only arises when others contribute. The strength of this warm glow decreases in θ. When others do not contribute, the warm-glow disappears because one rather feels like the only 'sucker' who contributes. Such a conditional feeling of warm glow is also interpretable as a reciprocal preference. In any case, the particular specification of preferences is not intended as being especially realistic, but rather as a simple or reduced form that generates conditional cooperation. As such, it is consistent with that of models that have more structural pretensions such as the ECR model of Bolton and Ockenfels (2000) and the inequality aversion model of Fehr and Schmidt (1999).

To see that these preferences generate conditional cooperation, we let $p = P(m \geq \overline{m})$ denote the subjective belief that at least the threshold fraction of people contributes, and compute the expected utilities of contributing and defecting:

$$\begin{aligned} E_m\left[\pi^e(1, m)\right] &\geq E_m\left[\pi^e(0, m)\right] \\ p\left(h(m) - \theta\right) + (1 - p)\left(h(m) - 1\right) &\geq h(m) - g \\ p &\geq \frac{1 - g}{1 - \theta} \end{aligned} \tag{3.4}$$

In words, (3.4) says that in order for a conditional cooperator to contribute, the subjective belief that at least a fraction $\overline{m}$ will contribute will have to be high enough. The the stronger the warm glow (the lower is θ) and the stronger the sanctions g, the lower such expectations need to be to induce contributions from the high types. Throughout the analysis we apply the tiebreaking rule that an indifferent agent complies.

In sum, the game the agents are playing is a standard public good game with two twists. The first twist is that the government can introduce sanctions that punish defection. The second

twist is that a fraction ω of the players have no dominant strategy. Instead, their best response depends on what they think other players will do.

Timing. Reiterating, the timing of the game is as follows:

1. Nature chooses the state of society characterized by the proportion of high types ω.
2. The government observes ω and decides on its policy g.
3. The citizens learn their own type and the government policy g, update their prior, and choose their contribution level $c \in \{0, 1\}$.

Trust. In this chapter, we talk about crowding out of trust. However, the definition of trust is a notorious source of conflict, so we take some time to get the definition right. In the introduction, we defined trust in passing as the *belief* that the other is a high type. A trusting act (in this case, contributing to the public good) is performed on the basis of this belief. One thinks the other will cooperate because her intentions or character are virtuous. Other definitions, like Hardin's notion of 'encapsulated interest' (Hardin, 1991), define trust more broadly as a situation where the trustor has reason to think that the trustee cooperates because her interests are aligned with her own. This definition includes situations where the trustee is expected to cooperate because of external enforcement. In this article we stick with the first definition because we are interested in how people assess the likelihood that others cooperate when sanctions are low. That is, trust can exist only in a situation in which the trustor is at risk precisely because she does not know the character of the people she is facing. By contrast, we define as 'confidence' the belief that the other will cooperate out of self-interest.

So defined, we interpret trust as an 'intrinsic motivation' for cooperation that can sustain cooperation when 'extrinsic motivation', i.e. sanctions, is low or absent. In the model, a certain amount of trust defined in this way is a necessary condition for a conditional cooperator to cooperate if $g < 1$. Thus, by 'crowding out of trust', we mean that higher sanctions are associated with lower trust, i.e. with a lower posterior probability of each agent that the other agents are of a high type.

3.4 Crowding out of trust

This section is structured as follows: We start by introducing some notation and terminology. To clear the way for the analysis of asymmetric information, we first derive equilibria in the simpler but instructive case of symmetric information. Proposition 3, the main result, characterizes the equilibrium under asymmetric information. All proofs are in the Appendix.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Let $g(\omega)$ denote the government policy, and $s(\Theta, g)$ the strategy of a citizen of type Θ. Denote $\mu(\omega \mid \theta, g)$ the posterior probability distribution of a citizen of type Θ about the state of society ω, and by $U(s, m, g, \Theta)$ the expected utility to a citizen of playing strategy s. We define an equilibrium as follows:

Definition 3.1. *An equilibrium consists of a government strategy* $g : [0,1] \rightarrow \mathbb{R}$*, a posterior belief of each agent about the true state of society* $\mu : [0,1] \times \Theta \times \mathbb{R} \rightarrow [0,1]$ *and a strategy for each citizen* $s : \Theta \times \mathbb{R} \rightarrow \{0,1\}$*, such that:*

$$g(\omega) \in \arg\max_{g \in \mathbb{R}} W(m, g)$$

$$s(\Theta, g) \in \arg\max_{s \in S} U(s, m, g, \Theta)$$

$$\mu(\omega \mid \Theta, g) \text{ is updated by Bayes' rule whenever possible}$$

This definition corresponds to that of a perfect Bayesian equilibrium (pBe). We restrict the analysis to pure strategy equilibria and require that the equilibrium satisfy the Cho and Kreps (1987) 'intuitive criterion'.

3.4.1 Symmetric information

Before we tackle the asymmetric information case, it will be instructive to discuss the case in which the citizens know ω. We solve the game backwards, and start with the reaction function of the citizens. In the absence of high sanctions and if $\omega > \bar{m}$, conditional cooperators face a coordination game amongst themselves. There is an equilibrium in which they all contribute, and one in which they all defect. The equilibrium of the larger game depends on the equilibrium in this coordination game. We will see that when high types coordinate on contribution, there is an unique equilibrium, which features two pooling regions. We develop some terminology for this partial-pooling (or semi-separating) equilibrium. In this equilibrium there are two regions of realizations of ω, in each of which the government plays the same policy. We call the threshold value between the regions ω^*. We call a region where $\omega \in [0, \omega^*)$ (i.e. where society consists of relatively many egoists) a 'bad state of society', and those where $\omega \in [\omega^*, 1]$ a 'good state of society'. We label the government policy for this partial pooling equilibrium as follows: the policy that is set in the bad state of society is called g_1, and the policy in the good state of society is called g_2. We define off-equilibrium beliefs μ_{oe} as the beliefs that citizens have about ω when they see a policy that is not part of the equilibrium profile.

Proposition 3.1. *Under symmetric information, there are two pBe:*

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

1. *When high types coordinate on not contributing, the unique equilibrium is a 'Hobbesian' pooling equilibrium in which the government sets $g = 1$, and all citizens contribute. If the government were to set $g < 1$, all citizens would defect.*

2. *When high types coordinate on contributing, the unique equilibrium features a threshold ω^*. A government that observes $\omega < \omega^*$ sets a sanction $g_1^* = 1$ and all citizens cooperate. A government that observes $\omega \geq \omega^*$ sets a sanction $g_2^* = \theta < 1$ and only the high types cooperate.*

As explained above, the conditional cooperators face a coordination game amongst themselves if $g < 1$. In this coordination game there are multiple equilibria. Either the conditional cooperators can coordinate on mutual contribution, or on mutual defection. We can interpret these equilibria as being associated with a social norm of contribution, or a social norm of defection. The fraction of conditional cooperators determines the amount of norm adherence. The first part of Proposition 1 describes the 'Hobbesian' pooling equilibrium, in which high types coordinate on defection. In this case, high types are behaviorally equivalent to egoists, and it is perhaps is unsurprising that the model generates a 'Hobbesian' conclusion, which says that only strong punishment will induce agents to contribute.

The second part of the Proposition tells us that when the high types coordinate on contribution, the government strategy has a threshold ω^*. The intuition is again straightforward: government types below ω^* will never set low sanctions (< 1), because there are too many egoists around. Inducing cooperation only from the high types generates so few contributions that it pays to set a high sanctions. Government types above ω^* will set low sanctions: because there are few egoists, low sanctions are a cheap way to induce a high level of contributions. Thus, when there are many conditional cooperators, and those conditional cooperators follow a norm of contributing, the government does best to implement low sanctions and tolerate a few defectors. Social norms are such that there is no reason for the government to use costly coercive strategies.

This simple setup captures two extremes in political thinking. On the one hand, when social norms of cooperation are absent we are led to a Hobbesian conclusion. On the other hand, it shows that when there is a sufficient amount of people who follow a cooperative social norm, sanctions can be low. The latter is a simple consequence of the existence of conditional cooperators, and something we seem to observe in many real-world social dilemmas.

3.4.2 Asymmetric information

We now turn to the case of asymmetric information, in which the government is the only player who knows ω. To start with, we can immediately verify the existence of 'Hobbesian' equilibrium, just as in the symmetric information case. The proof of the existence of this equillibrium did

not depend on the information conditions. The reason is that when high types coordinate on defection, their beliefs about ω are irrelevant.

Proposition 3.2. *Under asymmetric information, there is a 'Hobbesian' pooling pBe in which the government sets $g = 1$, and everyone contributes. If the government were to set $g < 1$, everyone would defect.*

In the remainder of the chapter we focus on equilibria in which high types coordinate on cooperation, i.e. there is a norm for contribution. It turns out that under asymmetric information, the analysis is substantially more complicated if high types coordinate on cooperation. Before we characterize the equilibria of the game we collect some useful results that serve to narrow down the search.

Lemma 3.1. *In any pBe in which high types coordinate on cooperation there are at most two different levels of sanctions g.*

Lemma 1 narrows down the search substantially. It implies that there are only two possible types of equilibria in which high types coordinate on contributing: pooling equilibria, and semi-separating (or partial pooling) equilibria with two pooling regions. The following lemma rules out the former:

Lemma 3.2. *In a pBe in which high types coordinate on cooperation there are no pooling equilibria.*

The intuition behind this lemma is the following. Governments that observe a very bad state of society will always set a high sanction. If they did not, the egoists who are a substantial part of the population, would defect. On the other hand, governments that observe a very good state of society will always want to set a low sanction, because this is a cheap way to induce cooperation of the great majority of people. This is not immediately obvious: one might think that there exist pooling equilibria on $g = 1$ supported by very pessimistic off-equilibrium beliefs. However, we can rule out such equilibria by applying the 'intuitive criterion' (Cho and Kreps 1987), a standard refinement of Bayesian Nash equilibrium. An equilibrium fails the intuitive criterion (IC) if it requires off-equilibrium beliefs that place positive probability on types for whom deviation payoffs are dominated by equilibrium payoffs. The idea is that it is 'unreasonable' to believe that such types would have deviated. Applied to the present model, we can show that ruling out deviations to sanctions below $g = 1$ of governments that observed a very high ω, requires off-equilibrium beliefs that are 'unreasonable' (as judged by the intuitive criterion). To rule out such deviations, off-equilibrium beliefs would have to be very pessimistic. However, a deviation to a low sanction is only attractive for the governments that observe a very good state of society, so equilibria based on such pessimistic beliefs don't survive the IC.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Summing up the results of our two Lemmas, we know that an equilibrium should feature two pooling regions. We are now in position to state the main result of this study:

Proposition 3.3. ***Crowding out of trust***

1. *If high types coordinate on cooperation, the unique class of pBe has two pooling regions characterized by the parameter ω^*. A government that observes $\omega < \omega^*$ sets a sanction $g_1^* = 1$ and all citizens cooperate. A government that observes $\omega \geq \omega^*$ sets a sanction $g_2^* < 1$ and only the high types cooperate.*
2. *If $\overline{m} \geq 1 - \alpha(1 - \theta)$, then under asymmetric information there exist equilibria in which the equilibrium threshold ω^* is strictly lower than under symmetric information.*

The first part of Proposition 3.3 repeats the result of Proposition 3.1 that when there are many conditional cooperators, government does best to implement low sanctions and tolerate a few defectors. The intuition is straightforward: the government will punish heavily when it knows that there are a lot of egoists around, because this is the only way to insure substantial amounts of cooperation in such an environment. It will punish less heavily when it expects many citizens to follow a norm of conditional cooperation, because cooperation can be induced cheaply in such an environment by setting lower sanctions. However, in contrast to the symmetric information case, such a government strategy implies crowding out of trust in equilibrium, because higher sanctions transmit information about the state of society to the citizens. This means that sanctions are 'bad news'.

The second part of the proposition states the implication of this signaling effect for government policy. It says that there is a continuum of equilibria under asymmetric information in which the government plays low sanctions for values of ω where it would not do so under symmetric information. The intuition behind this result is that when there is a norm of contribution between the high types, the government induces trust of citizens by setting a low sanction. To see how this works, consider a government under symmetric information that observes a state of society $\omega < \overline{m}$. Under symmetric information, the citizens know that ω is the state of society and the high types will not be motivated to cooperate. However, under asymmetric information, agents are more optimistic in the sense that upon observing high sanctions, they attach positive probability to states of society that are higher than $\overline{m}$. Because beliefs and sanctions are complements in generating compliance from the high types, this allows the government to set lower sanctions. Lower sanctions make inducing cooperation cheaper, which expands the region in which the government plays low sanctions. Thus, low sanctions induce citizens to trust each other more and thereby they 'crowd in' cooperation between the citizens.

Figure 3.1 shows the region in which the authorities play low sanctions under both symmetric and asymmetric information. In the grey area low sanctions are played under symmetric information. The border of this area is the unique equilibrium threshold ω^* for each level of $\overline{m}$. If

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

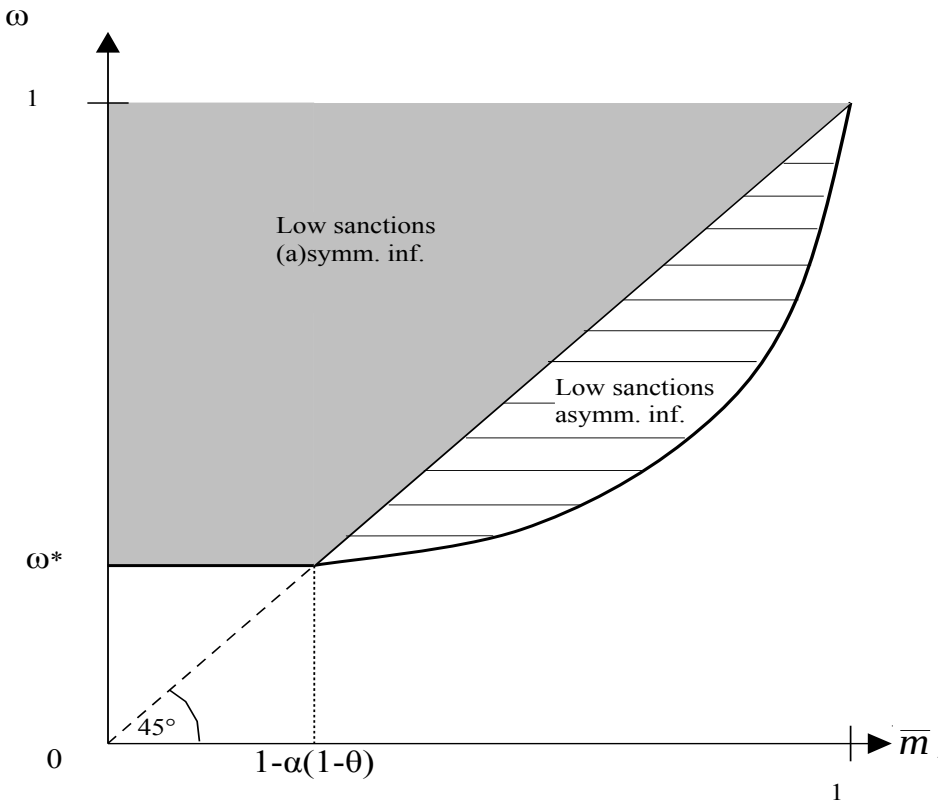

FIGURE 3.1: Equilibrium region with low sanctions under (a)symmetric information.

we turn to asymmetric information, we see that if $\overline{m} < \omega$, i.e. the threshold cooperation level to experience a warm glow is relatively low, the equilibria under symmetric and asymmetric information coincide. The reason is that when $\overline{m}$ is low, beliefs about m are very optimistic under both forms of information. Thus, low sanctions do not make people more optimistic than they would be if they knew ω. Note that equilibria with high levels of ω^*, supported by negative off-equilibrium beliefs, cannot exist. The intuitive criterion puts a lower bound on the off-equilibrium beliefs, and this lower bound is too high to rule out deviations to the lowest sanction that induces cooperation from the high types.

However, if $\overline{m} > \omega$, then the region where low sanctions are played expands under asymmetric information. In the hatched area all values of ω^* can be equilibrium values, and ω^* can be lower than under symmetric information. As mentioned above, this is because under asymmetric information low sanctions have a positive effect on beliefs. The reason that there exist multiple equilibria when $\overline{m}$ is relatively high, is that the lower bound of the reasonable (as judged by the intuitive criterion) off-equilibrium beliefs is now lower than $\overline{m}$. This means that we can find off-equilibrium beliefs such that no-one will cooperate when they see a deviation to a sanction lower than the equilibrium sanction. This supports the existence of many equilibria.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

The comparative statics of α, the cost of sanctions, and θ, the strength of the warm glow is intuitive. If the cost of sanctions increases, the region in which low sanctions are played increases. The same is true if the strength of the warm glow increases (a lower θ).

A final, rather subtle effect of asymmetric information is that high types are always more positive about the state of society than low types. An agent's own type gives her information about the state of society, because the probability that each agent is a high type is given by ω. Thus, being a high type implies that others are more likely to be a high type.

In sum, asymmetric information enlarges the region where low sanctions are can be played, because low sanctions are 'good news'. The signaling effect is a by-product of the fact that coercion is necessary only in bad states of society. In the terminology of Kahan (2005), it is truly the 'expressive dimension' of sanctions.

3.5 Implications and discussion

The model in this chapter can incorporate two extreme views of society. When $\theta = 1$ (no warm glow) and/or $\overline{m} = 1$, the agents in the model are all egoists, and the model generates standard Hobbesian predictions. If $\overline{m}$ is low and $\theta = 0$, the model admits a rather romantic equilibrium in which equilibrium sanctions are zero: the government relies completely on social norms of cooperation. Realistically, the truth will be somewhere in between, so even if there are many conditional cooperators, the government still has a role to play. Although citizens' behavior is partly driven by trust, conditional cooperators will still need a 'push in the back' from a sanctioning scheme, because they are aware that there are some egoists around which reduces their desire to cooperate.

Second, there is no net crowding out of cooperation by sanctions. As in Bénabou and Tirole (2003), incentives are what they call 'short term reinforcers'. In both models, higher sanctions 'override' the effect of diminished beliefs. Thus, an econometrician looking solely at the relation between sanctions and cooperation, would support the standard Becker-Stigler results. However, there is crowding out on the level of trust which influences the optimal sanction level. This brings us back to the definitions of 'trust' and 'confidence' as defined in Section 3. It should be clear that in contrast to trust, confidence increases with sanctions, because high sanctions make it in everybody's interest to cooperate.

Even though in this model sanctions compensate for the behavioral effects of decreased trust of agents, it suggests ways in which decreased trust may affect behavior. The model is consistent with the experimental observations of crowding out of cooperation: a drop in contributions if sanctions are raised to an off-equilibrium level. We must not forget that the sanctions implemented in (field) experiments are always off-equilibrium sanctions. They may thus interact

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

with off-equilibrium beliefs. In this model, equilibria on low sanctions are supported by negative off-equilibrium beliefs. Thus, implementing a deviation to a (higher) off-equilibrium sanction may lead to less contributions.

Moreover, trust is an attitude that determines behavior in many social situations. The crowding out of trust by incentives in one area could therefore have spill-over effects in other policy areas and into the future. Suppose that besides playing the public good game described above, agents are matched privately with each other to play another dilemma or trust game. In each of those games agents face partners drawn from the state of society. A government that sets a high sanction may improve cooperation levels in the public good game, but will induce negative beliefs that may cause agents to defect in private interactions. Thus, a raise in sanctions in one policy area may cause a drop in cooperative behavior in other areas. As an example, consider the stigmatizing effect of police crackdowns on immigrant populations. This may lead people to think that immigrants must be criminal to have merited such police action. This may make them less willing to cooperate with immigrants in private interactions.

Sanctions may also have spillover effects into the future. Since the government cannot undo an information transmission, trust may not easily return. For example, when high sanctions are exogenously lowered (for reasons not described in the model) after they have been introduced, cooperation may see a large drop, as even the by now cynical high types will refuse to cooperate. This is consistent with experimental evidence as in Gneezy and Rustichini (2000) or Gächter *et al.* (2007). These studies show that when incentives are withdrawn, cooperation does not return to pre-incentive levels.

Finally, as remarked by Bénabou and Tirole (2003), one can imagine a situation where people think they would be able to get away with defection, e.g. when non-cooperative behavior is very hard to detect. In this case, only the negative signaling effect remains, whereas the coercive effect of incentives disappears. A proper analysis of these cases is a task for future research. Souvorov (2003) has worked in this direction, and shows an intertemporal 'addiction to rewards' in a two-period model of a principal and a single agent. In the context of our model, spillover effects will result in an 'addiction to sanctions' as principals will need to maintain controlling measures to compensate for the reduced trust.

3.6 An application to tax evasion

The potential applications of the model described in this paper are various. In fact, they are everywhere where the conditions of the model are met: The principal has more information than the agents, some agents behave as conditional cooperators, and sanctions are costly. Kahan (2005) suggests applications in the public realm including not in my back yard (NIMBY)

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

problems and tax evasion (discussed below). One can also think of fare evasion in public transport, where the size of the penalty is an indication the norm of free riding. In the context of organizations and personnel economics, one can apply the model to incentive structures in large organizations and teams. In the context of sports one can think of the doping-dilemma, where harsh sanctions are indicative of a norm of widespread use of doping.

The example of tax evasion fits the model well because it is a private activity: any single taxpayer has very limited information on how honestly others pay their taxes. Tax offices on the other hand estimate evasion rates. This makes tax-enforcement policies a vehicle of signals on how widespread tax evasion is. Moreover, there is overwhelming evidence that conditional cooperation is a prevalent attitude in tax compliance. Econometric studies conducted both on an individual level (Scholz 1998) and on an aggregate level (Frey and Torgler 2008), show that the decision to evade taxes is in large part based on dispositional attitudes. Especially important are the belief that fellow taxpayers evade and the perceived legitimacy of the use of tax revenue.

The model in this paper can explain some puzzling facts about tax evasion. Andreoni et al. (1998, page 821) remark that "For small amounts of evasion, [...] the expected cost of detection would appear to be extremely low for most taxpayers. So, we may ask, why are so many households honest, and why don't cheaters cheat by more?". The model in this paper readily provides an answer to this question: people are conditionally cooperative, and as a consequence the government's best response is to apply mild (and cheap) sanctions instead of relying on heavy deterrence.
Another prediction of the model is that in equilibrium, high sanctions on tax evasion only make a difference for low types. High types will pay their taxes for any equilibrium sanction. Wenzel (2004) shows in the context of tax evasion that official sanctions are effective only for those that have a weak personal norm of paying taxes. People with strong personal norms on the other hand also cooperate for low sanctions.

Evidence from (field) experiments also give some indications that a signaling effect of sanctions is at work. Coleman (1997) reports the results of the Minnesota tax experiment, amongst 47,000 tax payers in Minnesota. Some 1700 of them received a letter announcing that they had been randomly selected for an audit. The responses with respect to reported income were mixed: middle and low income taxpayers increased their reported income (although most of them by small amounts), but high-income taxpayers did not. In one treatment, the experimenters sent another letter to 20,000 tax-payers saying that the numbers of cheating tax-payers was much lower than commonly assumed. This significantly increased reported income. Sheffrin and Triest (1992) find that highly publicized campaigns against tax evasion often fail to have the desired effect, and that some campaigns may increase distrust in other citizens.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

3.7 Concluding remarks

In the last chapter, I quoted Polinsky and Shavel (2000). In their survey on theory of law enforcement, they note that from a theoretical perspective sanctions often are too low. This paper gives an explanation why sanctions may be 'too low'. It asks whether Hobbesian coercion in social dilemma problems remains optimal when society is a mix of conditional cooperators and egoists. What is the optimal policy to promote cooperation if the situation in question is a prisoners' dilemma for some and a coordination game for others? The paper shows that the optimal level of sanctions depends on the relative proportions of the two agents in society. When there are many egoists, the high sanction or Hobbesian solution is optimal. When there are many conditional cooperators, a policy of low sanctions may be more efficient. If the government knows more about the composition of types in society, this implies that high sanctions are 'bad news'. Thus, its superior information allows government to induce or crowd in cooperation by setting low sanctions. The paper thus shows that sanctions may have a dual role. They both change economic payoffs and alter agents' perception of the environment. The government has to perform a balancing act: it has to punish the deviators, while keeping the conditional cooperators optimistic.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Chapter 4

Can sanctions induce pessimism? An experiment.

4.1 Introduction

In the standard economic view, sanctions are effective because they change economic payoffs and modify individuals' incentives to engage in certain actions. In the last chapter I explored a mechanism through which sanctions can be effective, not just by their effects on payoffs, but also by their effects on beliefs. The signaling effect of sanctions explored there is relevant in environments where the authority with the ability to introduce sanctions is more informed than those that are sanctioned. In this chapter, two collaborators[1] and me provide direct experimental evidence that such an effect can occur.

The experiment is not exactly a test of the model of chapter 3, because instead of a dilemma game, we investigate behavior in a coordination environment. There are several reasons for this. First, the use of a coordination game instead of a social dilemma means we do not need to verify assumptions about the exact distribution of types, which may require a more specialized setup. Instead, the complementarity of the actions of different players is already incorporated in the payoffs. Second, the setup with many Pareto ranked equilibria allows us to analyse the effect of sanctions in more detail than a setup with merely two actions. Third, in the current setup, we are able to investigate how 'small' sanctions affects play in coordination games, something which the literature has not yet considered.

Specifically, we consider the following research questions related to both the positive and negative effects of sanctions:

[1] This chapter is joint work with Roberto Galbiati (EconomiX Nanterre) and Karl Schlag (Universitat Pompeu Fabra)

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

1. Can the incentives associated with non-deterrent sanctions induce desired behavior and make agents more *optimistic* about other players' actions?

2. In situations of imperfect information about the past behavior of other group members, can the introduction of sanctions make agents more *pessimistic* about the actions of others by giving a signal that other players do not behave well? If so, does this reduce the effectiveness of sanctions?

We investigate these questions in an experiment based on the minimum effort game. The minimum effort game is a coordination game with many Pareto ranked equilibria. Each player chooses a level of costly effort, and is rewarded according the minimum of the efforts of all players in the group. The more efficient equilibria result only if all players play individually risky strategies. Doubt about the other player's willingness to play such a strategy may result in inefficient outcomes. Because there are multiple equilibria and players' efforts are strategic complements, the game is particularly suitable as a workhorse to answer our questions.

Consider first Question 1. Sanctions have a direct effect by providing incentives to choose higher effort. They also have an indirect or *forward looking* belief effect due to efforts being strategic complements. Anticipating that opponents are similarly affected by the sanctions and thus are expected to choose higher efforts reinforces one's own incentive to choose a higher effort.

Question 2 addresses the signaling or *backward looking* belief effect of sanctions. When past behavior is not directly observable, sanctions may carry a signal that things are not going so well. After all, why introduce a sanction to suppress socially undesirable behavior when everybody behaves saintly? In other words, sanctions may be perceived as 'apparently necessary'. Thus, the signaling effect of introducing sanctions may reduce the willingness to play a high and risky level of effort, and decreases the effectiveness of sanctions.

To answer the questions above we describe the results of a laboratory experiment, in which we focus on the effects of mild, non-deterrent sanctions in a coordination game. In particular, we look at the differences between the effects of 'exogenous sanctions', and the effects of 'endogenous sanctions' (defined below). Our workhorse game is the minimum effort coordination game with many Pareto-ranked equilibria as introduced by Goeree and Holt (2001, 2005). In all treatments agents were matched in groups of three, where the third player was a "principal" who benefitted proportionally to the minimum effort chosen by the other two in the group. The subjects played the minimum effort game twice, but the third player was the only one to be informed of the outcome of the first round before the second round was played. This information structure was common knowledge. Before the second round of the minimum effort game was played, the principal could decide whether to introduce a sanction F to both players in the group, that lowered the earnings of a subject if she selected low effort. The sanction F came at a small cost to the principal's own earnings. We call this the *endogenous* sanction, because it was

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

introduced by a third party in a reaction to the behavior of the subjects. The sanction was 'mild' in the sense that it made playing low effort a more costly, but not a dominated strategy. In another treatment, the same sanction F was introduced automatically. We call this the *exogenous* sanction, because it was introduced by the experimenter unconditional on past effort choices by the subjects. Across these treatments we compare the effect of sanctions on effort choices and reported beliefs about what the other player will do.

Our results show that exogenously introduced sanctions increase beliefs about the effort that the other player will play. As a result they effectively increase coordination on more efficient equilibria. However, our answer to the second question reveals a significant difference between endogenously and exogenously introduced sanctions. In our analysis of the data we distinguish players on the basis of their behavior in the first round. The signaling hypothesis leads us to expect that people who played high effort in the first round and are confronted with a sanction, will infer that the effort of the other person must have been low. By contrast, someone who was pessimistic and played low effort will not be able to make such an inference, because she also played low, and thus a sanction may have been introduced as a reaction to her *own* behavior. We thus expect a difference between the effects of endogenous and exogenous sanctions for high effort players, but not for low effort players. In accordance with this hypothesis, we find that there is a significant difference in the effectiveness of the two kinds of sanctions for players who exerted high effort in the first round. For these players, the exogenous sanction has a substantial positive effect on effort and beliefs about the other player's effort. By contrast, the effect of an endogenous sanctions is not distinguishable from not introducing a sanction at all. As the signaling explanation predicts, the way in which the sanction was introduced did not matter for those who played low effort in the first round.

To our knowledge this is the first study that looks empirically at the effects of sanctions on beliefs in a minimum effort game. Moreover, it is the first paper that empirically studies the signaling effect that the introduction of sanctions may have. Its main message is that the effectiveness of sanctions depends on the context in which they are introduced. On the one hand, people recognize the incentive effects that sanctions will have on others, which multiplies their effectiveness. On the other hand, when information about the behavior of others is limited, as is the case in modern large-scale societies, the introduction of sanctions may cause pessimism by drawing attention to past misbehaviors. This is especially true for those that are optimistic and behave cooperatively. This finding implies a difficult balancing act that a government or principal must perform: it must try too keep the optimist optimistic, while at the same time encouraging the pessimists to change their behavior. The results of this experiment suggest that 'mild law' may not be the optimal way to do so, because it induces pessimism with little compensation in the way of material incentives.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

A further contribution of this chapter is the use of novel statistic tests. We use a new test developed by Schlag (2008) based on a so-called stochastic inequality (Cliff, 1993). This is an exact test designed to assess the direction of a treatment effect, without making (parametric) assumptions about the distribution of the samples.[2] Instead of comparing means in the underlying distribution one compares a random observation from each distribution. Note that the Wilcoxon-Mann-Whitney (WMW) test can only reject the hypothesis that two samples are drawn from identical distributions. Thus, it can identify the existence of a treatment effect, but is not informative about *why* the two distributions differ significantly. For instance, without additional assumptions, one cannot draw conclusions about whether and how the means of the samples differ. Although the results of WMW test are completely in line with our results, its different null hypothesis would have only allowed us to conclude *that* sanctions influenced behavior, we would not be able to draw conclusions how about *how* sanctions influenced behavior.

4.2 Literature

Our experimental analysis of the effects of sanctions is related to several strands of literature. The empirical literature on crowding out is already discussed in chapters 2 and 3. 'Crowding out' refers to the tendency of material or monetary incentives to diminish the internal motivation to engage in the desired behavior. In extreme circumstances this can lead to less of the desired behavior. This phenomenon has been empirically documented in many economic settings (see Frey and Jegen (2001) and Bowles (2008) for surveys). For our purposes, the most interesting cases involve sanctions to members of a group or a society. In a well-known experiment, Gneezy and Rustichini (2000) show that introducing a fine for picking up children late from a day-care centre resulted in an increased number of people who picked up their children late. This effect endured even after the sanctions had been withdrawn. Ostmann (1998) provides experimental results showing that external enforcement financed by experiment participants only reduces harvests in common pool problem by a small amount relative to a no-enforcement treatment. Frey and Oberholzer-Gee (1997) conducted a survey on willingness to have nuclear waste repository built in their community. Without compensation, 50.8% of the respondents answered positively, but when the request was accompanied by an offer of (substantial) monetary compensation, the acceptance rate dropped to 24.6%.

Most existing explanations for the crowding effect focus on a notion of 'intrinsic motivation', which can be diminished by sanctions under certain circumstances (Frey and Jegen, 2001). Kahan (2005) suggests another explanation based on the idea that the situations in which crowding out occurs can be viewed as coordination games. Although on the face of them, settings like the ones mentioned above seem to resemble dilemma games, there is much evidence

[2] An exact test is a test where the statement about its level can be proven, in contrast to a level that is derived from an asymptotic approximation as the sample size tends to infinity.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

that coordination plays a large role in the outcome. This is due to the existence of so-called conditional cooperators or reciprocal agents (Fehr and Gächter, 2000). In determining their behavior in social dilemmas, conditional cooperators condition their behavior on their beliefs of what others do. Gächter (2006) surveys the evidence on the existence of conditional cooperation. Insofar as people are conditionally cooperative, their belief that others will cooperate will turn out to be a crucial variable in determining the outcome of collective action problems.

The experiment aims to test an explanation for crowding out that was suggested in the last chapter. That is, sanctions may provide conditional cooperators with a signal that others do not behave well, and this will diminishes their own willingness to cooperate. Chapter 3 as well as Sliwka (2007) provide formal models of this phenomenon.

Our research is also related to a well established strand of literature in legal scholarship: the *focal point theory of expressive law* (McAdams, 2000). This view holds that laws express values and attitudes, that can shape individual behavior. Cooter (1998) argues that the expressive character of sanctions can be used to coordinate expectations on a beneficial equilibrium. People expect others to follow the law, and so a self-fulfilling equilibrium can be induced by a sanction that penalizes behavior pertaining to other equilibria. The core idea is that for this to happen, laws do not necessarily have to be fully deterrent (i.e. they can be mild), because their role is merely to create focal points. Bohnet and Cooter (2001) and McAdams and Nadler (2003) provide evidence that mild sanctions can lead to better coordination in coordination games with two equilibria. These result is in line with results about experimental coordination games showing that in coordination environments, even advisory cheap talk by an external party or coordinator can help to bring about coordination on efficient equilibria (Chaudhuri and Bangun 2007, Van Huyck *et al.* 1992).

Finally, we relate to the experimental literature on coordination games. The specific game that we use was introduced by Goeree and Holt (2001, 2005) who also foreshadow our answer to question 1. Unlike the present chapter, they do not introduce sanctions between rounds, but investigate the behavior of different subject populations under high and low costs of effort. They show that over multiple periods, convergence to more efficient equilibria gradually takes place. Devetag and Ortmann (2007) provide a comprehensive survey of experimental results in coordination games.

In a recent paper, Brandts and Cooper (2008) compare the effectiveness of cheap talk and monetary incentives in an experimental design close to ours. Groups consist of five: four agents play a minimum effort game, and a manager profits from the degree of coordination that they reach. The manager can use financial incentives or communication messages to try to increase the level of cooperation. The authors find that communication is more effective in increasing coordination than are incentives. However, in contrast to our setup, incentives in this game cannot give any signals since the minimum effort of the previous round is known to each player.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

In the paper that is perhaps closest to ours, Tyran and Feld (2006) explicitly compare the effects of endogenously and exogenously introduced mild incentive or 'law'. In their experiment, subjects allocated to groups of three can first vote on whether sanctions for defectors should be introduced. They then play a public good game with or without the sanctions. The authors find that mild sanctions are effective when they result from the voting procedure, but not when imposed exogenously (by the experimenters). The authors show that voting for mild law raises expectations that others cooperate, and this in turn raises cooperation.

4.3 Discussion of the experimental setup

The study of sanctions comes up in settings that can often be described as either a coordination game or a Prisoners' dilemma. We chose a coordination game as an object of study, because in such games the rational choice depends only on the beliefs about the actions of the other player(s) in the game. This allows us to isolate the sanctions' effects on behavior that derive from the change in a subject's belief, and we can disregard issues to do with social preferences and/or dominant strategies that usually play a role in Prisoners' dilemmas.

4.3.1 The experimental game

We use as a workhorse the minimum effort game by Goeree and Holt (2001 , 2005), because it has large action spaces that allow players to express rather precisely their preferences and beliefs. The structure of the game is as follows: two players simultaneously choose an effort level between 110 and 170 (the bounds are chosen such that there are no clear focal points). Subjects' payoffs are determined by the minimum of these two efforts, minus the cost of their own effort times a parameter $k \in [0, 1]$, which is the same for both players. In each period we also elicit from each player an interval in which he believes the other will play his effort (see below). In contrast to the original setting by Goeree and Holt (2001) in which the game is played only once, in our experiment the game is played twice. Moreover, in some treatments (see below) a sanction F was introduced in the second round, where $F = 0.5 \cdot (170 - e_i)$. Thus, F implements a subtraction to the payoffs that is proportional to the deviation of the chosen second round effort from the maximum effort (170). Although this sanction decreases the riskiness of playing higher effort, the game remains a coordination game. The sanction is applied to both players in the group, although the actual subtraction may differ between the players depending on their second round effort choice. Another difference with the game of Goeree and Holt (2001) is the presence of a third player in the group. Depending on the treatment, this third player is either active or inactive. When she is active, she can choose before the start of the second round whether to introduce a sanction for both players in the group. Player 3 receives a payoff proportional to the minimum effort chosen by the other two players.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

In sum, payoffs in round 1 are determined as follows:

$$\pi_i(e_i, e_{-i}) = \min\{e_i, e_{-i}\} - 0.85 \cdot e_i, \quad \text{for } i = 1, 2;$$
$$\pi_3(e_1, e_2) = 0.25 \cdot \min\{e_1, e_2\}.$$

where $\pi_i(e_1, e_2)$ is the payoff of player i in tokens, $e_i \in [110, 170]$ is the effort level chosen either by player 1 or player 2, and k is the cost of effort. In the second round the sanction F may be implemented by either player 3 or the experimenters. Payoffs in round 2 are given by the following equations:

$$\pi_i(e_i, e_{-i}) = \min\{e_i, e_{-i}\} - 0.85 \cdot e_i - 0.5 \cdot (170 - e_i), \quad \text{for } i = 1, 2;$$
$$\pi_3(e_1, e_2, s) = 0.25 \cdot \min\{e_1, e_2\} - s \cdot c_s,$$

where c_s is the cost of introducing a sanction for the third player and $s \in \{0, 1\}$ is the choice to introduce a sanction (1) or not (0).

An important element of the experimental design is the information structure. The third player is the only one to be informed of the effort levels of players 1 and 2 when the first round is concluded. That is, at the beginning of the second round, players 1 and 2 do not know the effort levels of the other player, nor their own payoffs from the first round. However, before making any choices in the second round, players 1 and 2 know whether a sanction has been applied to their group. Note that players did not know before the first round that there would be a second round. They were informed of this only after the first round had concluded.

4.3.1.1 Parameters, treatments, and procedures

We chose to set the cost of effort at 0.85, i.e. close to 1. The evidence reported in Goeree and Holt (2001) indicates that in the presence of high costs of effort, individuals tend to coordinate on lower effort levels. We wanted effort choices to be not too high in order to give player 3 an incentive to introduce a sanction in the treatments in which she is active. We set $c_s = 4$, a level calibrated to induce roughly half of the players 3 to introduce a sanction.

We now describe the treatments. In all treatments, the first round is the same: players 1 and 2 play the minimum effort game and player 3 is inactive. In the baseline treatment there is no sanction in the second round, and player 3 is inactive. That is, the second round is conducted exactly as the first, and no mention of a sanction was made. We refer to this treatment as the exogenous no-sanction (ExNS) treatment. By *exogenous* we mean that the choice to (not) introduce a sanction was not conditional in any way on previous decisions by the subjects. This was clear to the subjects because the choice was made by the experimenters in a centralized fashion for all groups in the session. In the second treatment, sanction F is implemented in

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

the second round. The sanction was communicated to the players before they reported their effort level and their beliefs about others' actions. They then played the second round with the sanction in place. In spirit of the experimental economic literature, we refer to the sanction in neutral terms, i.e. as a "subtraction". In the remainder, we refer to this treatment as the exogenous sanction treatment (ExS).

Although player 3 is present in all treatments, she is only active in the third treatment. After player 3 has observed the chosen effort levels of players 1 and 2 in the first round, she is asked to decide whether to a) change both player 1's and 2's payoff structure in the second round by introducing a sanction F, or b) leave the payoff structure unaltered with respect to the first round. After player 3 has taken her decision, players' 1 and 2 are informed of it. They then play the second round with payoff structure decided by the principal. We refer to this treatment as either the endogenous sanction treatment or EnS (if a sanction is introduced by player 3) or the endogenous no-sanction treatment or EnNS (if no sanction was introduced).

Because the experiment features just two rounds of play and no possibility of learning, it was very important that people understood the game correctly from the start. To this purpose we ran a tutorial before the start of the first round. In the tutorial, participants had 5 minutes to come up with hypothetical effort choices of players 1 and 2 and to calculate their payoffs resulting from these choices. The tutorial took place before assigning subjects to a role, so that also players 3 could practice with the calculation of payoffs of players 1 and 2. In addition to this tutorial, the input screens in the actual experiment provided subjects with the possibility calculate their payoffs from a given choice. That is, after entering and before confirming their choices, subjects could enter a hypothetical choice of the other player and let the computer calculate their payoffs resulting from these choices.

The experiment was conducted in several sessions at the economics lab of the university of Siena, Italy. The first sessions took place in May and June 2007. Another series of sessions was conducted in November 2007. Subjects entered their effort and belief choices on a computer that was running on the software z-Tree (Fischbacher, 2007). The number of subjects in an experimental session varied between 18 and 30. The subjects earnings were in tokens as specified above, which were converted into euro's at the end of the experiment at an exchange rate of 10 tokens = 0.75 euro. The instructions were read out loud to make them public knowledge. The instructions and the input screen are provided in appendix B.

4.3.1.2 Elicitation of a belief interval

Apart from the effort choices, we are interested in the effect of sanctions on players' anticipation of what the other will do. Therefore, in the same input screen in which players 1 and 2 enter their effort choice, we asked them to enter beliefs about the other player's effort choice in that

round. Rather than elicit a point belief, we decided to elicit an *interval*. More precisely, players have to specify a range (i.e. a lower bound L and its upper bound U) in which the other player's choice is believed to fall. In order to increase accuracy in belief reporting we reward a correct guess[3]. The earnings from a guess are determined as follows:

$$\pi_i(L,U) = \begin{cases} 0 & \text{if } e_{-i} \notin [L,U] \\ 0,15 \cdot (60 - (U-L)) & \text{if } e_{-i} \in [L,U] \end{cases}$$

That is, a wrong guess (the actual number chosen by the other player falls outside the specified range) yields no payoff. A correct guess (the actual number chosen by the other player lies within the specified range) yields 15% of difference between the length of the interval $[110, 170]$ and the width of the interval $[L,U]$. Thus, the smaller the specified range, the higher the earnings if the guess is correct. However, a smaller range also increases the risk that the guess is not correct, in which case no tokens are earned.

Eliciting an interval has the advantage that it gives information not only about the location of the belief distribution, but also about its dispersion. Schlag and Van der Weele (2009) show that provided the belief distribution is single peaked, this *interval scoring rule* will induce rational decision makers to include both the median and the mode of their belief distribution in the chosen interval. Moreover, the width of the interval increases if the beliefs of the decision maker are more noisy. This makes the width of the interval a proxy for how 'sure' the decision maker is. These results hold for both risk neutral or risk averse decision makers. Note that the alternative quadratic scoring rule is only guaranteed to reveal the mean when the decision maker is risk neutral.

4.4 Non-parametric tests of stochastic inequality

One contribution of this study is the use of new non-parametric tests that have been designed for small samples (Schlag 2008). The disadvantage of existing tests is that they either add distributional assumptions (e.g. assuming normality or restricting the parameter space so that the alternative hypothesis is no longer the complement of the null hypothesis) or that they can only establish that a treatment changes the distribution of outcomes, not how. Specifically, the standard non-parametric test in the experimental literature for comparing samples has been the Wilcoxon-Mann-Whitney (WMW) test. The null hypothesis of this test is that the two samples are drawn from the same population. Thus, unless one is willing to make further assumptions on the underlying distributions (i.e. that all other moments of the probability distributions except

[3] Gächter and Renner (2006) show that incentivizing beliefs' reporting has a positive impact on beliefs accuracy.

the mean are equal), the WMW test cannot identify the direction of the treatment effects. It can only establish that they are different.

We analyze the effect of sanctions by testing 'stochastic inequality'. In order to identify the direction of a treatment effect we compare the likelihood that one variable realizes a higher outcome than the other. We measure this degree by the so-called *stochastic difference* which ranges from -1 to 1. Specifically, given two random variables Y_1 and Y_2, $\delta = Pr(Y_2 > Y_1) - Pr(Y_2 < Y_1)$ is called the stochastic difference of Y_1 versus Y_2. δ is estimated by taking the sample average across all pairings of the data. One says that Y_2 tends to realize higher outcomes than Y_1 if $\delta(Y_1, Y_2) > 0$. To establish a treatment effect in this direction, we test the null hypothesis that $\delta \leq 0$. When $Pr(Y_1 = Y_2) = 0$ then this is equivalent to testing that $Pr(Y_2 > Y_1) \leq 1/2$. When the data is given as matched pairs then the appropriate test is a sign test. When data is given by two independent samples, we implement the test of Schlag (2008). Appendix A gives a more extensive formal treatment of these procedures.

It is worth noting that there are no other exact nonparametric tests for comparing means or testing stochastic inequality given independent samples. In particular, the WMW test is not an exact test for comparing the underlying means given two independent samples (e.g. see simulations of Forsythe et al., 1994). Neither are there other exact nonparametric tests for correlation; the Spearman rank correlation test can only identify non-identical distributions. Non-exact tests of stochastic inequality have appeared in the biostatistical applications (Brunner and Munzel, 2000). One innovation of the tests we use here is that they are exact, in the sense of having the level that they are claimed to have, and do not rely on asymptotic approximations. They are the first exact tests for this stochastic inequality based on independent samples. Unlike tests for means, the ordinal nature of tests of stochastic inequality makes them less sensitive to outliers and hence they are very well suited to uncover significant differences given small samples.

We want to emphasize that the results of the more traditional Wilcoxon-Mann-Whitney test support our analyses. All significant results that we present are also significant, often more so, in the corresponding WMW test (these results are available on request). However, as explained above, without further assumptions the null hypothesis of the WMW test does not allow us to draw conclusions about the direction of the effect. Because the WMW test is rather powerful, we will use it when we want to gather support for a claim that two samples have similar distributions. In this case we are not primarily interested in the direction of the effect. Rather, we want to have the strongest possible test to falsify the claim that two samples are similar.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

4.5 Hypotheses and results

In this section we present the results of our experiment. We present our analysis by testing conjectures that are based on the research questions mentioned in the introduction. These conjectures are specific enough to provide us with the null hypotheses necessary for classical statistical analysis.

4.5.1 Statistics for the entire sample

The number of participants in the experiment was 243: 45 in treatment 1, 51 in treatment 2, and 147 in treatment 3. In treatment 3, the principal decided to introduce a sanction in 29 out of 49 groups. Each experimental session lasted roughly 35 minutes and the subjects earned 7.5 euros on average[4]. In the tutorial 82% (199 out of 243) correctly computed the payoffs from hypothetical choices. As another indication of whether people understood the game, we also checked whether there were 'anomalous observations': people who specified an effort choice above the upper bound of their belief interval. We found just 6 such observations.

We observe a high correlation between beliefs and effort in the first round of each treatment, as you would expect in a minimum effort game. The correlation coefficient between the lower bound of beliefs and the effort choice is 0.85^{***}, which is highly significant[5]. The correlation with the upper bound was somewhat lower (0.81^{***}), because many subjects specified an upper bound at, or close to 170 in the first round. They were thus restricted in moving this upper bound in the second round. For most subjects this was not true for the lower bound of the belief interval. For this reason we take the lower bound of the interval as our indicator of beliefs throughout this chapter.

Figure 4.1 shows a histogram of first period effort choices, aggregated over all treatments. We see a large clustering of observations around 170, a smaller cluster around 110 and an otherwise fairly uniform distribution[6]. We want to analyze the effect of the introduction of a sanction in the second round, and hence in the remainder we focus on the *changes* of effort and beliefs between rounds. We compute for each subject the change in beliefs and effort levels, and compare these changes across treatments.

There are a few complications to analyzing changes between rounds. First of all, the observations for the group members in the third treatment are not independent. The effort decision of one

[4] If this seems little, remember that the incentives were concentrated on only two (effort) choices. At each of these choices there was thus relatively a lot at stake.

[5] The significance is based on a test with the null hypothesis that the covariance is less than 0 (Schlag 2008)

[6] The effort levels are higher than those in Goeree and Holt (2001) with a cost of effort of 0.9. Reasons may be that the cost of effort is slightly lower in our setup and that in the instructions we did not use the word "cost" when referring to k.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

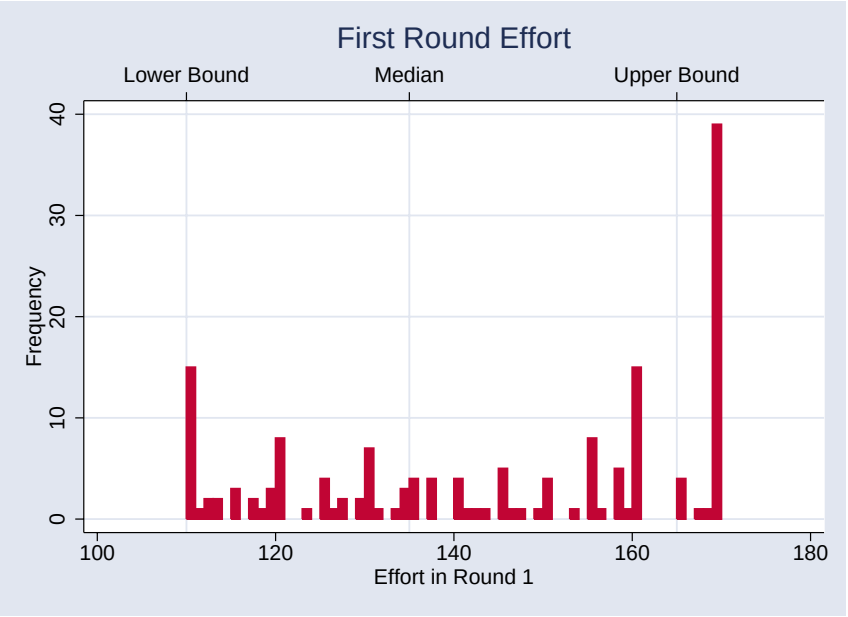

FIGURE 4.1: Histogram of first round effort choices of all subjects.

subject in the first round will influence the decision to implement a sanction by the third player. This in turn may influence the effort and beliefs of the other subject in the second round. When we do statistical testing, we correct for this dependence by taking the average of two observations whenever the subjects come from the same group, and treating it as one observation.

Second, interpreting changes in efforts and belief intervals as reaction to the experimental setting is not straightforward. Subjects that specified an effort level or a lower bound on beliefs of (or close to) 170 in the first round are unable to adjust this level upwards, and subjects who chose close to 110 cannot adjust it further downwards. This will generate observations of zero changes that may not reflect the actual preferences or adjustment of beliefs of the participants.

As we will see, the general trend in the experiment was for subjects to adjust their beliefs and efforts upwards in the second round. Thus, the problem is not severe for those who are initially on the lower bound. Specifically, there were no subjects who chose low effort (below 135 but above 110) and subsequently moved their effort downwards, and only three who chose a (small) downward adjustment of beliefs. Therefore we do not consider those who chose 110 to be severely constrained. However, the matter is different for those who chose effort or belief levels on (or close to) the upper bound of 170. It is likely that most of those subjects would have liked to change their behavior if they had been able to move upward further, but were constrained to do so. We believe that the fact that these people do not change their behavior does not give us accurate information about their actual change in beliefs and their preferences over effort levels. Therefore, we restrict ourselves to analyzing the choices of those subjects who actually had a choice. We focus on comparing the behavior across treatments of subjects who reported beliefs or effort lower than or equal to 165. In practice this means that for the

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

analysis of the beliefs, we excluded subjects who chose first round belief levels strictly higher than an upper bound of 165. This resulted in excluding 11 observations. For the analysis of the efforts, we excluded subjects who chose first round effort levels strictly higher than 165. This resulted in the exclusion of 39 observations. The median first round effort of the sample thus obtained is 135. The values of the upper and lower bound that we applied are indicated in Figure 4.1. In the remainder, we define high effort players as those who play first round effort in $e \in \{135, ..., 165\}$ (i.e. above the median), and low effort players as those who play first round effort in $e \in \{110, ..., 134\}$ (i.e. below the median).

4.5.2 Effort and beliefs in the baseline treatment (ExNS)

While our analysis will focus on comparing behavior across treatments it is of interest to consider what happens in the baseline case, where there are no exogenous sanctions. Recall that there is no feedback between rounds in the treatment without sanctions. One might conjecture that in the absence of feedback there is no change in effort and yet it is not clear whether behavior should not change over time simply due to the fact that a choice is made a second time. We present the evidence in Table 4.1. We denote by Mean ExNS1 the mean of first round variables in the exogenous no-sanction treatment, and by ExNS2 the second round variables. The last column presents the estimated stochastic difference of the first round versus the second

	n	Mean ExNS1	Mean ExNS2	Stochastic Difference ExNS1 vs ExNS2
Effort	23	133	137	0.17
Belief	29	134	138	0.15**

TABLE 4.1: Mean efforts, mean beliefs, and stochastic difference between round 1 (ExNS1) and 2 (ExNS2) in the exogenous no-sanction treatment (ExNS). * Denotes significance at 10%, ** denotes significance at 5%, *** denotes significance at 1%.

round in treatment ExNS (remember from section 4 that this is the estimate of $\delta = Pr(Y_2 > Y_1) - Pr(Y_2 < Y_1)$). It is worthwhile to note that testing for stochastic inequality for matched pairs is equivalent to performing a sign test. We find insignificant differences in the effort (confirmed by the Wilcoxon rank sum test). On the other hand we find significant evidence that the lower belief level tends to be higher in the second round. Apparently people move up their belief levels, but as we can see from Table 1, changes are small so people are not sufficiently optimistic to change their effort levels by much.

4.5.3 The effect of exogenous sanctions (question 1)

Our first question relates to the effects of exogenous sanctions on efforts. In the case of exogenous sanctions we can abstract from any signaling considerations because the sanction is

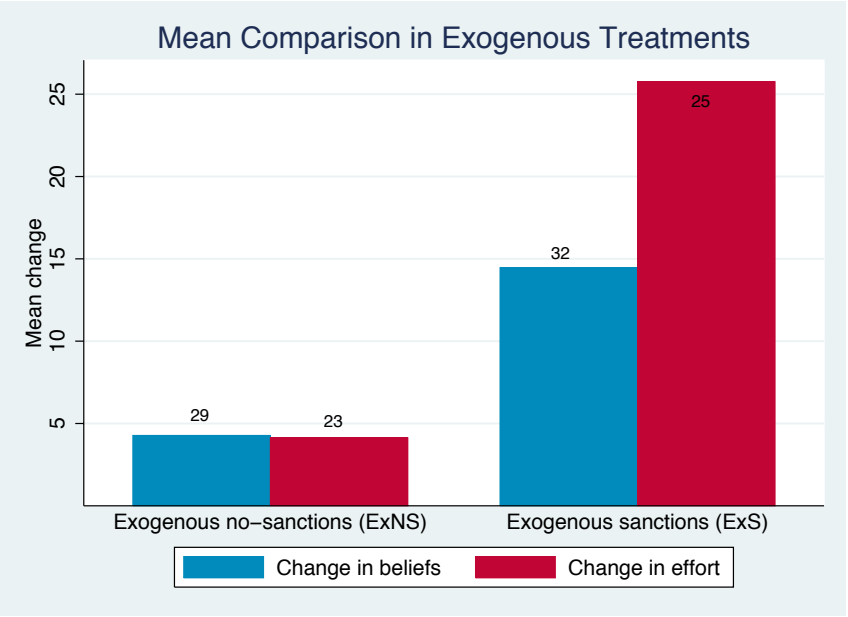

FIGURE 4.2: The change in beliefs and sanctions for the whole sample, except those who chose first round efforts $\in \{166, 167, ..., 170\}$ or first round beliefs $\in \{166, 167, ..., 170\}$. (Number of independent observations for each sample at the top of the bar).

unconditionally imposed by the experimenters. Sanctions are modeled in our experiment by an additional cost of making efforts below the maximum 170. Mathematically this translates into a reduced cost of effort. Under a given belief distribution such a change in the cost of effort causes a rational agent to increase effort.

If the subject anticipates that the other player also increases effort, her beliefs about opponent effort become more optimistic, which makes it rational to increase effort even more. Thus, we expect that introducing sanctions causes an increase in beliefs but an even stronger increase in effort. If we compare behavior in round one and round two in the sanction treatment, we cannot separate this anticipated effect of sanctions from other effects that we observed in the case of no sanctions. The appropriate benchmark for comparison is the treatment without sanctions. We formulate the following conjecture about this comparison:

Conjecture 4.1. *The change in effort and belief levels between rounds 1 and 2 is larger when there are exogenous sanctions than when there are no exogenous sanctions in period 2. This effect is more pronounced for efforts than it is for belief levels.*

This conjecture can also be motivated with the results of Goeree and Holt (2001), who find that a lower cost of effort increases effort levels in a between-subject design. Since our sanction effectively lowers the cost of effort, it is reasonable to conjecture that (exogenous) sanctions will increase beliefs and effort. We now gather evidence for our conjecture. Figure 4.2 presents the change in the means between round 1 and round 2 for the ExS and ExNS treatments. In

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Table 4.2 we report the results of our statistical analysis of conjecture 1. We estimate the stochastic difference of the change in effort under exogenous sanctions (ExS) versus the change in effort under exogenous no-sanction (ExNS). To indicate changes between the two rounds of a treatment X we use the notation dX. Similarly we consider the changes in the lower bound of the belief intervals, comparing the change under exogenous sanctions, and the change under no exogenous sanction. Comparing the significance levels we indeed observe a more pronounced

	Stochastic Difference dExNS vs. dExS
Effort	0.64^{***}
Belief	0.31^{*}

TABLE 4.2: Values of stochastic difference between *changes* in the exogenous no-sanction (ExNS) treatment and *changes* in the exogenous sanction (ExS) treatment. * Denotes significance at 10%, ** denotes significance at 5%, *** denotes significance at 1%.

difference in terms of effort than in terms of beliefs, as is also apparent from Figure 4.2. To formally test this finding would involve designing a new test which is outside the scope of this chapter. However we do note that the 20% equi-tailed confidence intervals overlap; by this crude method at least this difference is not found to be significant.

Summary 4.1. *We confirm our conjecture that changes in efforts and beliefs tend to be higher when there are exogenously imposed sanctions in the second round than when there are no sanctions in the second round. The data lend support to the claim this effect is stronger for effort than for beliefs.*

4.5.4 The signaling effect of sanctions (question 2)

We now investigate the effects of endogenous sanctions. We compare subjects' choices under exogenous sanctions to subjects' choices under endogenous sanctions. Note however that there are at least two differences between these two groups. One difference is that in the exogenous case the sanction was imposed by the experimenter while in the other case it was imposed by a subject in the experiment. A second difference arises from the fact that the choice of a sanction by the subject need not be unconditional (like the experimenter's sanction) or random. The choice of a sanction may reflect the observations of particular first round effort choices. It is exactly this kind of information transmission we wish to analyze, and the experiment is designed to isolate the signaling effect from the incentive effects of sanctions, by comparing ExS and EnS. Before we analyze the reactions of the subjects to the imposition of an endogenous sanction, we investigate the choice of sanction by the third player.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

4.5.4.1 The choice of endogenous sanctions

To see why player 3 would decide to implement a sanction, consider her monetary incentives. The third player is rewarded proportionally to the minimum group effort. However, imposing sanctions carries a small cost. A maximizing principal will implement a sanction if she expects to recoup these costs through an increased minimum effort level. When initial effort is low, there is a large potential range for effort increases, and changing behaviors can be very profitable. Moreover, if effort is low in the first round, there is no clear reason to think that it will rise without a sanction. Thus we can formulate the following conjecture:

Conjecture 4.2. *In the endogenous sanction treatment, the likelihood of sanctions being imposed by the 'principal' is decreasing in the minimal effort chosen in the first round.*

In order to test this conjecture we compare the minimum first round effort in the sanctioned groups to the minimum first round effort of non-sanctioned groups. We use the Wilcoxon-Mann-Whitney test because we are interested in any difference between the samples. However, we cannot find marginally significant evidence that the distributions of minimal effort are different in the groups where sanctions are imposed as compared to the group without sanctions imposed (the p-value is 0.63). Of course the samples are small, so the test is not very powerful. However, as Table 4.3 shows, the descriptive data do not point at large differences either. Note that

	Mean of Min. Group Effort	# Below 165	# Above 166
No Sanction	138	17	3
Sanction	135	28	1

TABLE 4.3: Descriptive data on first round minimum effort of sanctioned and non-sanctioned groups. The columns show the mean, and the number of groups with minimum effort below 165 and above 166.

sanctions were also introduced occasionally when minimum effort was high. Note that this need not contradict equilibrium behavior. To see this, assume that there are some subjects that always choose low effort ('low' types) while others choose high effort as they believe that the others that think like them also choose high effort. There can be equilibria in which a sanction is imposed only if minimum effort is high, and therefore are a signal that the group consists of high types. Observing no sanction be a signal that the other subject is of type low and hence it would be best to choose low. Thus the principal will impose sanctions on high types to preserve coordination. This behavior is optimal for all players, provided there are sufficiently few low types to make play of high effort in the first round an equilibrium. Obviously there are other equilibria in which coordination on high effort is not sanctioned. This multiplicity may be a reason why there is no clear pattern when sanctions are imposed. For all practical purposes

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

however, we can just assume that the behavior is random. This leads us to the following conclusion:

Summary 4.2. *We have no significant evidence that sanctioned groups had lower minimum effort. The descriptive statistics similarly indicate a lack of a clear pattern. Sanctions seem to be randomly imposed in our data set.*

This result implies that there is no endogeneity problem that could have arisen if only low-effort players had been sanctioned. To the extent that people who play low effort react different to sanctions than others, this would have made the comparison with exogenous sanction treatment more difficult. To this comparison we turn now.

4.5.4.2 The effect of endogenous sanctions

Although the apparently random imposition of sanctions means that there is no clear informational content of sanctions, subjects may still believe that sanctions were imposed systematically. Specifically, subjects may follow the same reasoning that led us to formulate Conjecture 4.2. If this is the case, sanctions may still influence beliefs about the other group member. A small thought exercise teaches us that the inference that can be made depends on a subjects' own effort in the first round. Consider a subject who believes Conjecture 4.2 to be true. Assume first that this subject chose high effort in the first round. When she observes that the principal imposes no sanction, the subject infers that the opponent chose a high effort because otherwise they would have been sanctioned. This may give her cause for optimism, and a reason to keep choosing high effort. On the other hand, if the high-effort subject is sanctioned, she infers that it is likely that the opponent made a low effort. The high effort player will face the following questions: Will the opponent react to the sanction with a sufficient increase in effort such that I should increase my own effort too? Or is the opponent simply someone with a tendency to make low efforts even under sanctions, in which case I should lower my own effort? Compared to the case of exogenous sanction, the observation of a sanction induces uncertainty that the other subject chose low effort and will do so again. Now assume that the subject played low effort in the first round. A sanction no longer has any informational content as long as the subject believes in Conjecture 4.2. Specifically, any sanction can always be interpreted as being aimed at the subject himself. Thus, there is no reason to assume his beliefs about the opponent will change, and we expect him to behave much like someone under exogenous sanctions would behave.

Note that higher order expectations that the players may have about each other may complicate this pattern. For example, the low-effort player who observes a sanction may think that if his opponent is a high-effort player, she will now be discouraged. We content ourselves with trying to identify first-order patterns. We summarize these patterns in two conjectures (remember

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

from Section 5.1 that by low first round effort we mean effort $\in \{110, ..., 134\}$, and by high effort we mean effort $\in \{135, ..., 165\}$).

Conjecture 4.3. *a) For those that chose a low effort in the first round, the change in efforts and beliefs under endogenous sanctions will be similar to the change under exogenous sanctions.*

b) For those that chose a high effort in the first round, the change in efforts and beliefs will be larger under exogenous sanctions than under endogenous sanctions (signaling effect).

We first consider Conjecture 3*a*). Figure 4.3 presents the mean changes in beliefs and effort for people who played low effort in the first round. Figure 4.3 reveals no large differences

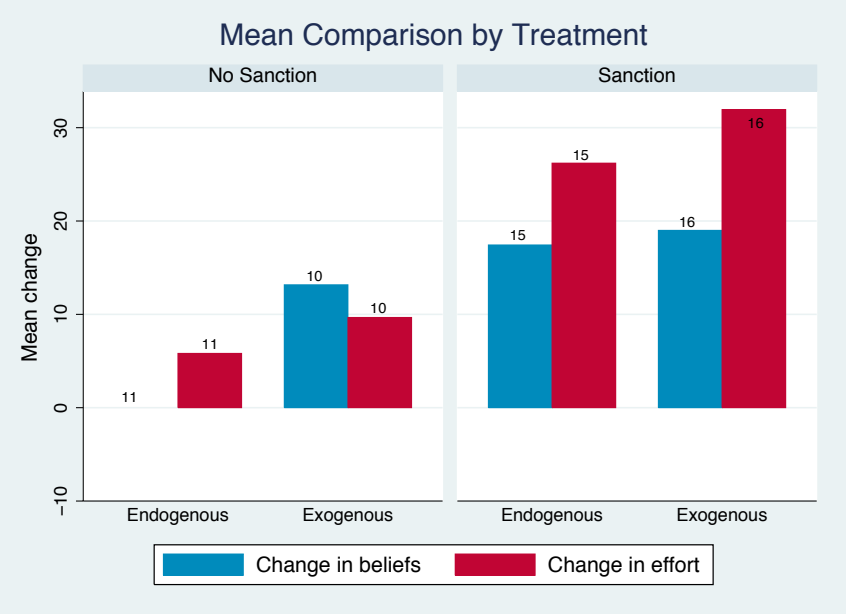

FIGURE 4.3: Means of changes in beliefs and effort across treatments, for those who played low effort ($\in \{110, 111, ..., 134\}$) in the first round. (Number of independent observations for each sample at the top of the bar).

between the exogenous and endogenous sanction treatments. We now try to falsify Conjecture 3*a*). We test the null hypothesis that the distribution of change in effort is identical in the endogenous and exogenous sanction settings. Since we are interested in any difference between the distributions we use the WMW test. The results in Table 4.4 show that we cannot reject the null hypothesis of identical distributions in the exogenous and endogenous treatments, both for effort and beliefs.

	WMW p-values dEnS vs. dExS
Effort	0.29
Belief	0.97

TABLE 4.4: p-values of the Wilcoxon Mann-Whitney rank sumtest of the exogenous and endogenous treatments for those who played low effort ($\in \{110, 111, ..., 134\}$) in the first round. * Denotes significance at 10%, ** denotes significance at 5%, *** denotes significance at 1%.

The problem is that the sample sizes are small, so we can only provide limited evidence of similarity.[7] Therefore, we will now show that we can make similar claims about the effectiveness of sanctions, regardless of the way they were introduced. We compare first and second round efforts and beliefs between the exogenous and the endogenous treatments, both for sanction and no sanction. We report results in Table 4.5.

	Stochastic Difference	
	ExS1 vs. ExS2	EnS1 vs. EnS2
Effort	1***	1***
Belief	0.5**	0.8***

TABLE 4.5: Estimates of stochastic difference between round 1 and round 2 of treatments ExS and EnS, for those who played low effort ($\in \{110, 111, ..., 134\}$). * Denotes significance at 10%, ** denotes significance at 5%, *** Denotes significance at 1%..

We find very similar estimates of stochastic difference in both sanction treatments. We feel confident therefore to draw the following conclusion:

Summary 4.3. *For subjects that made low efforts in the first round we find no significant evidence that endogenous and exogenous sanctions have different effects on either efforts or beliefs.*

We will now test conjecture 3*b*). In Figure 4.4 we report average changes in efforts and beliefs across treatments for subjects who played high efforts ($\in [135, 165]$) in the first round. Eyeballing the figure, it seems like the exogenous sanctions are more effective than the endogenous ones for those who played high effort. The results based on stochastic differences, reported in Table 4.6, confirm this. We observe significant evidence that exogenous sanctions are more effective in raising effort than endogenous sanctions. There is marginal significant evidence that beliefs tend to change more under exogenous sanctions. One wonders whether endogenous sanctions have any effect at all. To find out we test if there is a difference between the endogenous sanction

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

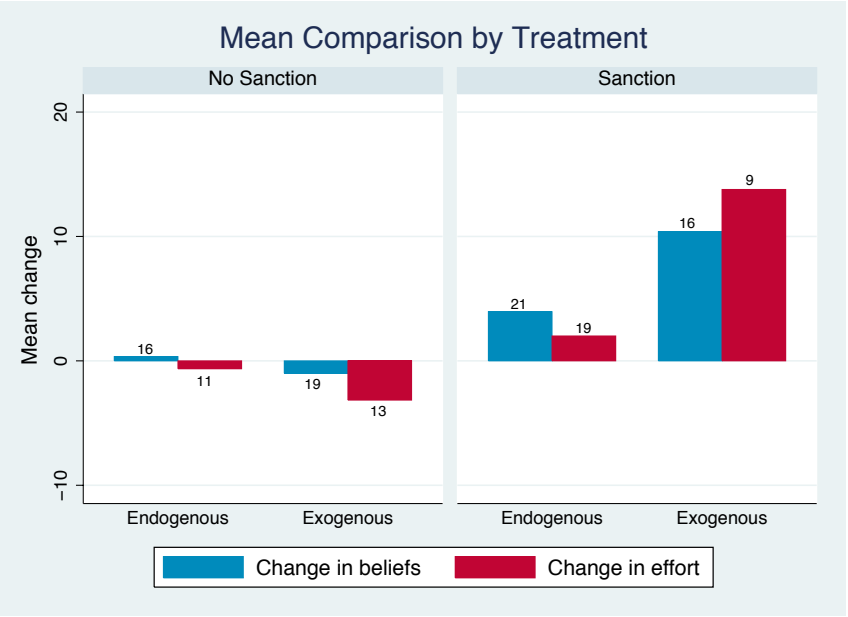

FIGURE 4.4: Means of changes in beliefs and effort across treatments, for those who played high effort ($\in \{135, 136, ..., 165\}$) in the first round. (Number of independent observations for each sample at the top of the bar).

	Stochastic Difference dEnS vs. dExS
Effort	0.66**
Beliefs	0.39*

TABLE 4.6: Estimates of stochastic difference between the exogenous and endogenous sanction treatments for those who played high effort ($\in \{135, 136, ..., 165\}$) in the first round. * Denotes significance at 10%, ** denotes significance at 5%, *** denotes significance at 1%.

	WMW p-value dExNS vs. dEnS	Stochastic Difference dExNS vs. dExS
Effort	0.35	0.78***
Belief	0.49	0.48**

TABLE 4.7: Comparison of the baseline (ExNS) treatment and the sanction treatments for those who played high effort ($\in \{135, 136, ..., 165\}$) in the first round. * Denotes significance at 10%, ** denotes significance at 5%, *** denotes significance at 1%.

treatment and the baseline treatment (ExNS). In the first column of Table 4.7 we report the p-values of the WMW test for this comparison.

We find that endogeneity dampens the increase in efforts and beliefs. In fact, it dampens it so

[7] Using statistical hypothesis testing we can show at most that the differences are not too large, since formally it is impossible to obtain significant evidence that the effect of endogenous and exogenous sanctions is equal. However, the larger the sample size, the more powerful the test, and the more confident we are that the effect, if it exists, is small.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

much that the effect of endogenous sanctions cannot be distinguished from not mentioning and introducing sanctions at all. However, the sample sizes are small, so it is possible that we would not be able to reject the null hypothesis of equal distributions, even if the actual difference is quite large. To counter this criticism, the second column of Table 4.7 shows the comparison with the baseline treatment with the *exogenous* sanction. It is clear that for similar sample sizes we get very significant results of the effectiveness of exogenous sanction.

Summary 4.4. *For subjects who played high effort in the first round, endogenous sanctions are less effective in raising efforts and beliefs than exogenous sanctions. In fact, the effect of endogenous sanctions cannot be distinguished from the effect of not introducing a sanction at all.*

4.5.5 Belief intervals

Before we move to the conclusions, we investigate the results pertaining to the width of the belief interval $U - L$. One of the reasons we asked the participants to specify an interval rather than a point belief was that we are interested in the impact of sanctions on uncertainty about the behavior of the other player, for which the size of the interval $U-L$ is a proxy (Schlag and Van der Weele 2009). Figure 4.5 shows the changes in the width of the belief interval for those who chose the lower belief interval in $\{110, 111, ..., 165\}$ in the first round). As Figure 4.5 shows, uncertainty

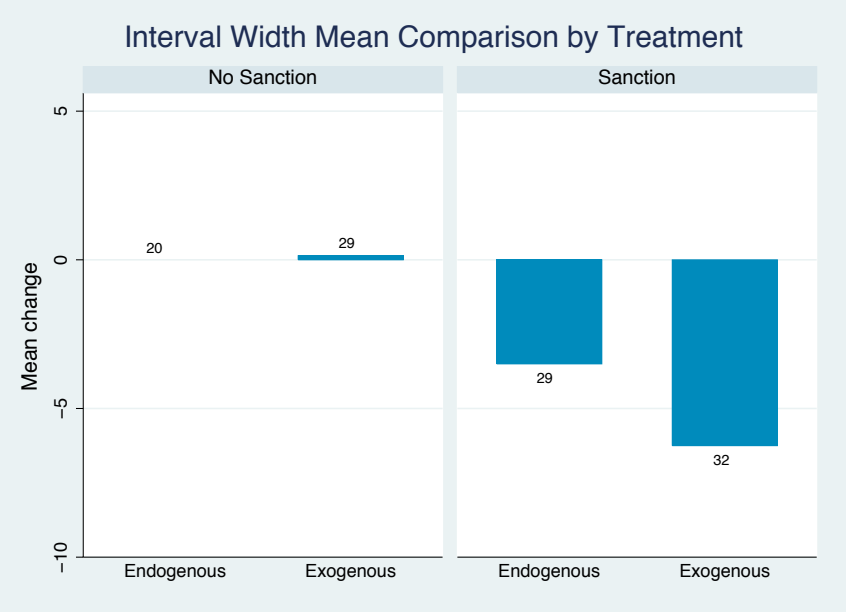

FIGURE 4.5: Means of change in the width of the interval across treatments, for those who chose the lower belief interval in the first round in ($\in \{110, 111, ..., 165\}$) in the first round (number of independent observations for each sample at the top of the bar).

did not change between rounds in both no-sanction treatments, while uncertainty went down in

both sanction treatments. We can confirm this result with statistical analysis. Table 4.8 presents the estimates of stochastic difference between the first and the second round interval width in all treatments. In both no sanctions cases a test of stochastic inequality cannot reject the null

	Stochastic Difference			
	EnNS1 vs. EnNS2	ExNS1 vs. ExNS1	EnS2 vs. EnS2	ExS1 vs. ExS2
Interval Width	0.067	0.0	-0.31^{*}	-0.46^{**}

TABLE 4.8: Estimates of stochastic difference between the round 1 and round 2, for those who chose the lower belief interval in the first round in ($\in \{110, 111, ..., 165\}$) in the first round. * Denotes significance at 10%, ** denotes significance at 5%, *** Denotes significance at 1%..

hypothesis that the distributions in the two rounds are equal at the 10% level. By contrast, we find that there is significant evidence that the interval decreases under exogenous sanctions and marginally significant evidence that the interval decreases under endogenous sanctions. This reinforces our conclusion that sanctions facilitate coordination partly by reducing uncertainty about the behavior of others.

If sanctions were to have a signaling effect, we would expect for those subjects who chose high effort ($\in \{135, 136, ..., 165\}$) in the first round, that the reduction in uncertainty is smaller under endogenous sanctions than under exogenous sanctions. Testing the direction of the effect with stochastic inequality, we find that the tendency of the decrease in uncertainty is in fact significant at 1% in the exogenous sanction treatment, while under endogenous sanctions it is no longer significant. Moreover, the estimates for stochastic inequality do not reveal a significant difference between EnS and either ExNS or EnNS. This indicates that endogenous sanctions do not reduce uncertainty for those who played high effort in the first round relative to the no-sanction treatments. It thus seems that sanctions reduce uncertainty in general, except for endogenous sanction applied to those who played high effort. This is congruent with our signaling explanation. However, when we directly compare the change in the interval width between both sanction treatments for those who played high effort, we cannot reject the null hypothesis of no difference.

Summary 4.5. *Uncertainty about the choice of the other player, as measured by the width of the belief interval, declines in the sanction treatments. There is no evidence of a change in the no-sanction treatments. For those who played high effort in the first round, the reduction in uncertainty only occurs under exogenous sanctions.*

4.6 Discussion and conclusion

The results of our experiment allow us to conclusively answer our two questions. Over the whole sample, exogenous sanctions clearly have a positive effect on effort levels and beliefs

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

about others' effort level. However, the way in which sanctions are introduced matters. This manifests itself in the fact that for people who played relatively high effort in the first round, the difference between the effect of an endogenously and an exogenously introduced sanction is significant. In fact, the endogenously introduced sanctions cannot be distinguished from the treatment without (exogenous) sanctions.

We think that the most plausible rationale for this result is the idea that underlies our hypotheses. The endogenous introduction of sanctions gives subjects a signal that the other group member did not 'cooperate', in the sense that she selected low effort. This tends to make people more pessimistic about the effort played by their companion in the group and less willing to move up in effort themselves. For those who played high effort initially this pessimism is reflected in the fact that beliefs and effort do not significantly increase under endogenous sanctions. We also found that uncertainty, as measured by the width of the belief interval, does not go down under endogenous sanctions as it does under exogenous sanctions. A signaling effect also explains why the difference between the sanction treatments does not occur for people that play low effort in the first round. For them this signaling effect is less pronounced, because they may think that the sanction was aimed at them rather than at the other player in the group.

Our results discredit a naive view of deterrence in which it is only the economic incentives that matter for behavior. The literature on crowding and intrinsic motivation had already established that sanctions may have adverse effects in some situations. We have identified another reason why sanctions may be ineffective. The result supplies a motivation why 'mild law' may not work. In contrast to Tyran and Feld (2006), we provide evidence that the *endogenous* introduction of sanctions rather than the exogenous one may be the cause of problems. In Tyran and Feld, a voting procedure for the introduction of a mild sanction gives people the opportunity to send a public signal that they are willing to cooperate. This in turn leads to increased cooperation. In our experiment, the introduction is under the discretion of a third player who has observed past play of the game. This setup reflects more closely the arrangements of a society where people make the laws through representatives, rather than directly. In this case a sanction sends exactly the opposite signal: sanctions are apparently necessary to keep people from deviating from the efficient outcome. The results show not only that such an effect can exist when the information conditions are right, but also that it is potentially quite substantial. Our study thus suggests that mild law may not be the best instrument in this case, because it does not compensate for this signaling effect by providing adequate incentives for efficient behavior.

In our experiment we observed the fact that the signaling effect was not present for low effort players, because the groups were so small that the sanction was likely to reflect their own behavior. However, in real life, relevant communities consist of many more than two people. This means that even people who play low effort may interpret the sanction as a signal, because it is unlikely that a sanction is introduced on the basis of the behavior of one person. Assuming some

external validity of the experiment, one can conclude that a sanctioning authority needs to attain a careful balance between correcting the behavior of deviants or pessimists and maintaining the optimistic beliefs of cooperators. The results of this study have implications for both public policies and manager-employees relationships in firms. As pointed out by Brandts and Cooper (2006), coordination failure can cause corporations and other organizations to become trapped in unsatisfactory situations both for managers and employee.

How to attain such a balance is an interesting further research question that goes beyond the aim of this paper. One possibility is to try to avoid the issue altogether by implementing harsh laws making undesired action very costly. Such a deterrent law would presumably override the signaling effect. However, such laws and their enforcement may be costly to implement in the real world, since they require at least some probability of detection for undesired activity and potentially costly sanctioning activities. Another possibility to investigate is whether appropriate framing of the introduction of a law can mitigate the signaling effect. In the tradition of experimental economics, this paper has tried to use neutral framing, replacing "effort" with "a number", and "sanction" with "subtraction". In real life however, a policy maker could attempt to surround the introduction of sanctions by soothing or stimulating messages. For example, one may say the actual number of people who deviate from the efficient strategy is small, or express the expectation that they will conform to the sanction. However, it is theoretically unclear why such cheap talk would be effective. The experiments by Brandts and Cooper (2008) and Van Huyck *et al.* (1992) incorporate the possibility of a principal to send written messages and suggestions to the agents. These studies could be combined with the asymmetric information structure in this paper in order to study this issue.

Last but not least, we wish to push forward the use of exact tests that "let the data speak" and do not add distributional assumptions. One approach in the experimental literature on crowding out has been to use the Wilcoxon-Mann-Whitney test to uncover differences in distributions, and to complement this test by looking at the descriptive statistics to make statements about the direction of the effect. A more popular approach throughout the experimental literature has been to implicitly use the WMW test as test for comparing means, without mentioning the condition needed for its validity, namely that all moments of the distributions except the first have to be the same. A contribution of our paper is the use of new tests (that are exact but do not impose additional distributional assumptions) that allow us to test directly for a negative impact of sanctions. We think these tests are an important addition to the toolbox of economists working with small data sets.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Chapter 5

How procedures can improve voluntary compliance

5.1 Introduction

The previous chapters investigated the use of sanctions in the presence of social interactions. In this chapter we will move away from the focus on sanctions and look at an alternative policy instrument to induce compliance: participatory procedures.

A large literature in social psychology, sociology and political science is devoted to the phenomenon of participatory decision-making. One of the main findings of this literature is that procedures that allow participation by employees or citizens, increase cooperation and compliance with decisions. This effect occurs independently of whether the actual outcome of the procedure is favorable to the agent. A second, more recent finding is that people pay more attention to procedures if they are uncertain about key aspects of their environment, e.g. if there is the threat of layoffs in their company, or uncertainty about the character of the authority.

These facts have led social scientists to construct a variety of theories that aim to explain why people value procedures. In psychology, the 'group value' (Lind and Tyler, 1988, Tyler and Lind, 1992) links procedures to the identity of the agents. Procedures that exclude an agent from the decision-making process will weaken the identification of the agent with the authority or the wider community and lead to less cooperative behavior. To explain the second fact, Lind and van den Bos (2002) propose that people use 'cognitive shortcuts' to substitute information about the nature of procedures for information about their environment that they are lacking. In economics, Frey *et al.* (2004) have used the stylized facts above to argue that people have preferences over different procedures which should be incorporated in economic models of institutions.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

This chapter abstracts from postulating preferences over procedures, cognitive shortcuts or the identity of agents. Instead, I proceed in two steps to explain how participation can lead to more cooperation even under adverse outcomes. First, I propose a stylized definition of participatory procedures. I define decision-making procedures as stochastic processes, in which the degree of participation is reflected in the ex-ante probability p that the agent (rather than the authority) gets her preferred outcome. The larger this probability, the higher the degree of participation. Such a definition interprets participatory procedures as institutions that endow the agent with 'bargaining power' or 'stochastic control rights'.

Armed with this definition, I formulate a simple signaling model between two players, an authority and an agent. There is asymmetric information about the type of the authority. She can either be selfish and care only about her own payoff, or she can be benevolent and take the payoffs of the agent into account to some extent. The two players have a conflict of interest whether to implement a project A, favored by the agent, or B, favored by the authority. The model has two stages, a decision-making stage and a cooperation or execution stage. In the decision-making stage the authority can decide on the degree of participation p of the procedure, where p is simply the ex-ante probability that the project will be A. Nature then determines the outcome of the procedure according to the chosen p. In the execution stage, the authority and the agent simultaneously choose a costly effort level. The effort levels are complements in determining the payoffs of the project. For given effort levels, project A provides higher payoffs to the agent whereas project B provides higher payoffs to the authority.

I then show formally that in this game there exists a unique separating equilibrium in which the degree of participation allowed by the authority is a credible signals of the latter's type. In such a separating equilibrium, procedures affect compliance for two reasons. First, a participatory procedure is more likely to yield the project that the agent prefers. Since this project gives him higher returns, the agent is more motivated to exert effort. I call this the *outcome effect*, which does not depend on the signaling role of procedures. Second, participatory procedures increase cooperation *even* if they result in a decision to carry out the inferior project (from the agent's point of view). This *procedural effect* arises because a fair procedure reveals the benevolent intentions of the authorities. An authority that reveals herself to be benevolent will be expected to exert higher effort in the cooperation phase, because she will internalize some of the benefits of her effort to the agent. Since efforts are complements, participatory procedures induce the agent to raise his effort level independently of the outcome of the procedure. I show that the separating equilibrium exists for arbitrarily small levels of benevolence, as long as the conflict of interest between the agent and the authority if high enough, so that signaling is sufficiently costly.

The predictions of this model are in line with the evidence on participatory procedures. Most importantly, it replicates the fact, mentioned above, that both favorable outcomes and procedures

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

per se raise cooperation. The model also predicts that uncertainty about the trustworthiness of the authority makes the nature of procedures more important. Furthermore, perceived trustworthiness of authorities is indeed a major factor in compliance decisions by agents (Tyler and DeGoey 1996). Finally, survey studies show that job satisfaction is positively related to the participatory decision-making. The model predicts this, since cooperation, and hence utility is higher under participatory procedures.

The definition of participation used in this chapter is closely related to that of 'control' used in the economic literature on delegation. However, the focus of the present study is different. In the delegation literature, the central trade-off is between the loss of control of the authority and the amount of information available at the decision-making level. In this study I focus on the information that is transmitted by the act of sharing power itself. The chapter also relates to the literature on gift giving (Carmichael and MacLeod 1997). In essence, this chapter interprets participatory procedures as costly gifts, in which the authority accepts a probability of losing his favorite project to signal his trustworthiness and increase cooperation by the other party.

The chapter proceeds as follows. The next section gives a more elaborate account of research on participatory decision-making in social psychology, and provides more detailed evidence for the stylized facts mentioned in this introduction. Section 3 introduces the stylized definition of participation used in this chapter. Section 4 presents the model and Section 5 the main results, which are discussed in Section 6.

5.2 Literature

The literature on (participatory) procedures in social psychology often goes under the term 'procedural fairness' or 'procedural justice'[1]. This stems from the well-documented tendency of people to attribute a subjective label of 'fairness' to procedures only if these allow sufficient possibility for participation of the agent (see Lind and Tyler (1988) and Tyler (2004) for surveys). The terminology used to indicate participation is somewhat diffuse. Many studies use the term "voice", which can refer to direct decision-making control or to the mere possibility by agents to present evidence or arguments for their position. Some studies explicitly refer to the former as 'decision control' and the latter as 'process control'.

[1]This literature is huge by all measures: a metastudy by Cohen-Charash and Spector (2001) counts more than 400 empirical studies into the effect of organizational procedures alone. I cannot do more here than give a representative flavor of the results.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Why do participatory procedures matter? Early explanations for the question why agents value participatory procedures focused on instrumental reasons. On the basis of experiments in dispute resolution, Thibaut and Walker (1975) argued that participation is important because it allows agents to secure better outcomes for themselves.

However, research done in the 1980s showed that although the outcome of a decision matters, participation generates increased compliance with the rules and cooperation with the authorities regardless of the outcome of the procedure to the agent. A well-known study is Tyler (1990), who reports the result of a large panel survey in which people are interviewed before and after they had interactions with the Chicago court and police system. In telephone interviews, people were asked (among other things) about several aspects of the procedures used by the authorities, including their possibilities to express their opinions and influence the outcomes of the decisions. They were also asked to evaluate the authorities and their attitudes towards compliance with the law. Tyler finds that trust in the authorities and positive attitudes towards compliance depend strongly on possibilities for participation in decision-making, regardless of the outcome of the procedure (e.g. the decision in the court case).

Participatory procedures also induce a positive evaluation of authorities (see Lind and Tyler (1988) for a survey) which in turn fosters cooperation. Tyler and DeGoey (1996) survey evidence about trust in institutions in different areas, such as the family and the workplace, and even national institutions as the police, congress and the supreme court. They show that trust in authorities consistently increases feelings of obligation to organizational rules and laws. Feld and Frey (2001) find that if there is a relationship based on trust between the taxpayer and the administration, tax evasion is lower.

In the area of tax evasion Pommerehne and Weck Hannemann (1996) and Frey (1997) show people are willing to cooperate more under participatory procedures. Controlling for demographic variables, income and the size of deterrence variables, they show that Swiss cantons that implement more direct democracy measures (referenda, town-hall meetings etc.) have lower rates of tax evasion. Smith (1992) produced similar evidence for the United States. In a laboratory experiment, Alm *et al.* (1993) find that the level of voluntary compliance increases if tax payers are able to vote on the type of public good that is provided (i.e. the charity towards which the contributions are directed).

In the realm of organizational decision-making, Cohen-Charash and Spector (2001) conduct a meta-analysis of the psychological literature that studies the effects of measures of 'organizational justice'. They use data from 190 experimental and survey studies (both in the laboratory and field), comprising a total of 64,757 participants. They find that the variable "voice", a blanket expression for diverse forms of participation in decision-making procedures, correlates

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

significantly[2] and strongly (.52) with measures of procedural justice that were used in these studies. In turn, procedural justice correlates significantly with measures of 'work performance' (.45), 'compliance with decisions' (.14), 'job satisfaction' (.43), 'organizational citizenship behavior' (.23), 'organizational commitment' (.50), 'trust in the supervisor' (.59), 'trust in the organization' (.43), 'counterproductive work behaviors' (-0.28) and 'turnover intentions' (-0.22). These correlations are often stronger than the corresponding correlations for 'distributional fairness'.

The fact that participatory procedures *per se* induce people to be more cooperative, has led researchers to argue that people value participation for other reasons than being able to influence the outcome. For example, the premise of the *relational model* (Tyler and Lind 1992) and the related *group value model* (Lind and Tyler 1988), is that people are anxious to belong to social groups and communities. Participation is important because it conveys to the individual that she is a full-fledged member of the community, which increases self-esteem, identification with the group and the motivation to contribute. As a result of such theories, Frey *et al.* (2004) argue that economic modellers should take into account preferences that are specified over procedures rather than outcomes.

When do participatory procedures matter? Based on a survey of several studies, Lind and van den Bos (2002) argue that the details of procedures matter most when people are uncertain about key elements in their environment. Van den Bos *et al.* (1998) test the hypothesis that voice is especially important when people are insecure about the character of authorities. In an experiment, a third party distributed lottery tickets between two people that had concluded an experimental task. Between treatments, the experimenters varied the information supplied to the subjects about the trustworthiness of the third party. They find that the satisfaction with the allocation of the tickets depends on whether the subjects were able to communicate their preferences to the decision maker. However, this is only the case when people were not informed about the trustworthiness of the authority. The authors conclude that when people do not know the trustworthiness of the authority, they rely more heavily on the participatory aspects of the procedure at the time of evaluating the final outcome. In a field study, Van den Bos *et al.* (2000) interviewed parents about the the quality of their children's daycare centres. They found that parents who indicated to be more unsure about the trustworthiness of the centre's organization, were more influenced by the quality of the centre's procedures in their final evaluation of its reliability.

This and similar evidence has led to new theories in social psychology, most notably 'uncertainty management theory' (Lind and van den Bos, 2002). This theory holds that procedures help people cope with uncertainties that come up in their lives. People use 'cognitive shortcuts'

[2] I report the correlations only for field studies, which were the most numerous in the sample. The correlations for laboratory studies were similar. 'Significantly' refers to the fact that the 95% confidence intervals do not contain 0.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

to substitute information that they lack about their environment with information about the perceived 'fairness' of procedures, of which participation is an important element. According to Lind and van den Bos (2002, p. 196), fair procedures thus allow people to "maintain positive affect, feel favourable towards the organization, and engage in the sort of pro-organizational behavior (e.g. accepting supervisor's orders, obeying company policies, going "above and beyond" the call of duty) that have long been known to be linked to fair process and fair outcomes [...]. These pro-organizational attitudes and behaviors are "safe" because fairness reduces the anxiety about being excluded or exploited, anxieties that might otherwise become very worrisome in uncertain contexts".

Economic literature on delegation and information. The definition of participation that I will propose is formally close to that of control rights in the economic literature on delegation. This literature investigates the trade-off between a loss of control from delegation and the beneficial effects of increased information at the decision-making level. In Aghion and Tirole (1997), delegation gives the agent incentives to gather more information and take better decisions. However, she may also use her freedom to carry out sub-optimal projects (from the authorities' point of view). In Aghion *et al.* (2002, 2004) the principal learns the type of the agent by delegating control to her and observing her behavior. In contrast, the present paper asks how control can be used to transfer information to the agent and how this affect cooperation.

Closest to the present paper, at least in it's formal setup, is Dessein (2005). Dessein considers a model between an entrepreneur and an investor. Ex-ante, the entrepreneur has more more information about the viability of the project. After the contract is signed, new, public information about the project arrives. If the information is bad, restructuring the project is optimal for the investor, but not for the entrepreneur, who receives private benefits from carrying out the original project. Dessein (2005) shows that the entrepreneur can signal his initial private information by contractually giving away control over the restructuring decision. The reason is that the 'good' investor knows that future information is likely to be positive, and hence the control will not actually be exercised. The current paper offers a more general interpretation of 'stochastic control rights'. It also makes less specific assumptions about the timing and the information structure of the game.

5.3 An operational account of participatory procedures.

In this section I will propose a definition of participation that ranks procedures with respect to the degree of control or influence that they delegate to the agent. In the real world there is

a plethora of institutions that embody some form of participation. As mentioned above, psychologist have distinguished between 'decision control' (voting, vetoing, etc.) and the weaker 'process control' (the expression of arguments). There is extensive evidence that both forms of control matter (see Lind and Tyler (1988) for a survey), but having some influence on the outcome seems to be a necessary condition to increase evaluations of the procedures and subsequent compliance. In an experiment, Lind *et al.* (1990) find that 'instrumental participation' (participation that allows an agent to have an influence on the outcome) has a bigger effect on the positive evaluation of procedures than 'non-instrumental participation' (e.g. the mere opportunity to express opinion). Tyler (1987) shows that when people have the impression that their arguments are not taken seriously, the beneficial effect of non-instrumental participation disappears. Thus, the beneficial effects of participation are due in large part (although not exclusively) to its link with control.

In the literature on deliberative democracy, Arnstein (1969) has ranked various forms of participatory procedures according to the degree of control over decisions in her well-known 'ladder of participation', which is reproduced in Figure 5.1. The sports of the ladder represent several

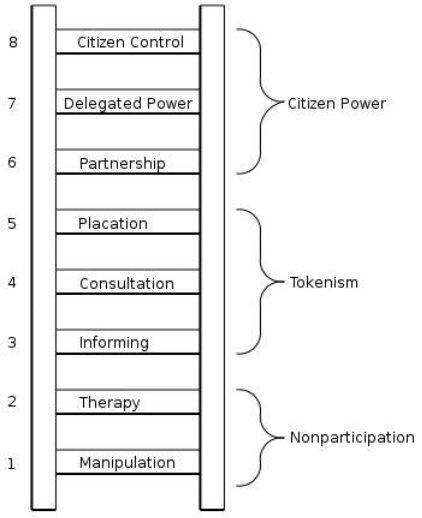

FIGURE 5.1: Arnstein's (1969) ladder of participation. (Reproduced with permission of the Taylor and Francis Group.)

different stages of participation, ranging from non-participation, via tokenism (i.e. procedures designed as window dressing for participation), to direct citizen control. The picture suggests that in any relationship between an authority and agents (i.e. employees or citizens), there is a continuum of participatory procedures that can be ranked according to the influence that they allow.

These considerations lead me to propose the following stylized definition of participatory procedures

Definition 5.1 (Participation)**.** *A procedure implements 'participation of degree p' if the ex-ante probability that the agent obtains his preferred outcome through the procedure is p.*

According to this definition, procedures are stochastic processes, the outcome of which is uncertain ex-ante. However, one *can* establish ex-ante how likely the agent is to get a preferred outcome in a given procedure. This probability is a consequence of the degree to which the procedure invites participation by the agent. Alternatively one can interpret the probability p as the bargaining power of the agent embedded in the procedure, or the degree to which control is delegated. The fact that participation is not a binary phenomenon is reflected in the assumption that $p \in [0, 1]$.

Consider the following concrete examples from an organizational context. An employer who puts up a suggestion-box on the wall would implement a p close to zero: no employee expects to exert great influence through such an institution. On the other end of the spectrum, having a representation of the employees amongst the senior management of the company and endowing them with significant bargaining (or even veto) powers would implement a p close to 1. Another example of a high p would be a government that commits itself to the outcome of a public referendum.

The notion of procedures proposed here abstracts completely from all aspects that relate to the process of decision-making itself. For example, it does not take into account that agents may have preferences to express their opinion or the beneficial effects of self-determination (Lind *et al.* 1990). I also abstract from the possibility that people change their preferences and that "the force of the better argument" (Habermas, 1990, p.158-9) may resolve conflicts of interest. In the current setup, procedures do not resolve conflicts of interests, they are merely institutions for managing them.

5.4 The model

In this section I present a model of costly signaling. In the model, there are two different kind of players: an authority (she) and a subordinate (he). The players are indicated by $t \in \{a, s\}$, where a stands for 'authority' and s for 'subordinate' or agent. There are two different types or natures $n \in \{A, B\}$ of the project, over which preferences of the players diverge. Players decide first which project is chosen, and subsequently cooperate on the chosen project. The crux of the model is that the authority can signal a concern for the agents' welfare by giving up decision-making power over a the type of project. Such delegation of decision-making power generates a motive to contribute for the subordinate because he is less afraid that he will be exploited in the subsequent cooperation stage.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Timing. The timing of the model is as follows:

1. Nature determines the preferences of the authority. With probability 1/2 the authority is 'selfish' and with probability 1/2 she is 'benevolent' (as explained below).
2. *Procedural stage.*
   (a) The authority decides on the degree of participation $p \in [0, 1]$ of the procedure.
   (b) Nature decides the outcome of the procedure. The agent gets his preferred project (A) with probability p, and the authority gets her preferred project (B) with probability $1 - p$.
3. *Cooperation stage.* After having observed both p and the outcome of the project, the agent and the authority simultaneously decide their level of cooperation $e_t(n, p) \geq 0$.

Payoffs from the project. The output G of the project is determined by the effort of the two players as follows:

$$G = e_a^{1-\alpha} e_s^{\alpha}, \tag{5.1}$$

where $\alpha \in [0, 1]$. Apart from the effort levels, each player's payoff depends on the type of project. The agent prefers project A whereas the authority prefers project B. The payoffs from the project are:

$$\pi_t(n, e_a, e_s) = \begin{cases} gG & \text{if project } n \text{ is the preferred project of player } t \\ G & \text{if project } n \text{ is not the preferred project of player } t, \end{cases} \tag{5.2}$$

where $g > 1$. Thus, g is a measure of the conflict of interest between the two players. The larger is g, the larger the difference in payoffs between the two projects, and the larger is the conflict of interest. It is not necessary for the results that the authority is certain that the agent prefers project A. What matters is that the agent with some probability prefers project A.

The agent. There is one agent (occasionally referred to as a subordinate), with the following von Neumann-Morgenstern utility function:

$$u_s(n, e_a, e_s) = \pi_s(n, e_a, e_s) - \frac{1}{2} e_s^2. \tag{5.3}$$

Here, the second term in the utility function captures the loss from the costly effort to the authority.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

The authority. The authority has the following von Neumann-Morgenstern preferences

$$u_a(n, e_a, e_s) = \pi_a(n, e_a, e_s) + a\pi_s(n, e_a, e_s) - \frac{1}{2}e_a^2. \tag{5.4}$$

The parameter a is drawn by nature from $\{0, \theta\}$. If $a = 0$ we call the decision maker 'authoritarian' or 'selfish'. If $a = \theta \in (0, 1)$ we call the decision maker 'benevolent' or 'altruistic'. If the decision maker is benevolent, she has preferences over the payoffs π_s of the agent from the project[3]. Everything else equal, the decision maker prefers higher payoffs of the agent from the project. The size of θ determines the strength of the benevolent motives. Since $\theta < 1$ there is always a conflict of interest between the authority and the agent even if the authority is benevolent, because she always values her own payoffs more.

In summary, the game consists of a simultaneous move game, preceded by a procedural stage that opens up the opportunity for the authority to signal. Note that neither player has a preference over procedures p; they care only about the payoffs from the project.

Applications. The model can be applied to an authority that relies on cooperation from subordinates or agents but has imperfect sanctioning possibilities. Applications to workplace situations are perhaps most salient. In this case the manager or supervisor is the authority and the employee the agent. To the extent that work effort is non-contractible, management needs to rely on the voluntary cooperation of it's employees. In many cases the output of the company will depend on the ability of management and the workers to cooperate constructively. That is, both the effort levels of the management and the employees are necessary for a good result, so the complementarity between effort levels assumed in the model arises naturally.

Another application is between two partners in a joint venture. Whereas one partner may have the decision-making power about the nature of the venture, its success depends on the effort and contributions of both partners.

Finally one can think of organizations without real sanctioning power that need to rely on the voluntary cooperation of their members. This is true for many volunteer organizations, but also of large international decision making bodies such as the UN or the IMF. In this case the authority may consist of a subset of the members while the agents are the remaining members. Complementarity arises because cooperation between all the members is necessary for effective policy making.

[3]The fact that π_s is the same function as in the specification of the agent's preferences is for notational simplicity. What is important is that both functions represent the same preferences. The model results go through similarly if we posit altruistic preferences over u_s rather than π_s. The present specification is however computationally easier.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

5.5 Participatory procedures as a signal

In this section we look for a perfect Bayesian equilibrium of the game. As is customary, we mandate that this equilibrium satisfies the Cho and Kreps (1987) Intuitive Criterion.

Let $e_t(n,p)$ denote the effort level of player t when procedure p resulted in outcome n, and denote equilibrium values by *. We solve the game backwards. In the last round of the game, both players simultaneously choose an effort. Because the payoffs are concave and the costs of effort are convex, the optimal effort exists and is bounded for each player. The first important observation is that

$$e^*_\theta(n,p) \geq e^*_0(n,p), \tag{5.5}$$

i.e. for any given n and p, the benevolent authority will always exert a higher effort than the selfish authority. The reason is that the benevolent authority internalizes a part of the payoffs of the agent, and therefore her marginal utility of effort is higher than that of the selfish authority.

The efforts of the two players are complements, because a higher effort of the other player raises the marginal utility of the other player. This leads to the second observation: it follows from (5.5) that the effort of the agent will depend on his beliefs about the type of the authority. If he believes the authority is benevolent, he will be more motivated to exert high effort.

We now move to the procedural stage. Before characterizing the optimal p, we formulate a useful lemma, that will lay the basis for the existence of a separating equilibrium. We assume that authority chooses p, knowing that her own effort in the second stage is a best response against the effort of the agent, i.e. $e_a(n,p) = e^*_a\left(e_s(n,p)\right)$, which is a function of $e_s(n,p)$ only. With some suppression of notation we can now express the expected utility of the authority as follows:

$$U_a\left(p, e_s(n,p)\right) = p * u_a\left(A, e_s(A,p)\right) + (1-p) * u_a\left(B, e_s(B,p)\right).$$

It is possible to show that this expected utility function satisfies the following single crossing property

Lemma 5.1 (Single crossing property)**.** *For any $p' > p$,*

$$U_0\left(p', e_s(n,p')\right) \geq U_0\left(p, e_s(n,p)\right) \Rightarrow U_\theta\left(p', e_s(n,p')\right) > U_\theta\left(p, e_s(n,p)\right). \tag{5.6}$$

Lemma 5.1 says that whenever the selfish authority is indifferent between a pair $\{p', e_s(n,p')\}$ and $\{p, e_s(n,p)\}$, the benevolent authority will strictly prefer the pair with the higher p. We call this a single-crossing condition because in effect it ensures that the indifference curves of

the different types cross at most once. The proof of this claim is intuitive: raising p is less costly for the benevolent authority than for the selfish authority. Whereas the loss of her favorite project is a pure loss to the latter, the benevolent authority internalizes some of the gain to the agent. It should be no surprise that a useful single crossing property in this game should be in *expected* utility. The signaling variable of the authority is p, the probability with which her non-preferred outcome A occurs. Thus, although the signaling variable does not directly influence the authority's outcome in any given state, it changes her expected outcome through modifying the probability of each state. Before we move on to the equilibrium results, I first define a distinction that is central to the model.

Outcome and procedural effects. A central point of the paper is to distinguish between two effects of procedures on the cooperation level of the agent. The first is the effect of decision outcomes, the second is the effect of procedures. To avoid confusion, I provide formal definitions of both.

Definition 5.2. *An 'outcome effect' exists for a given level of participation $\bar{p}$ if $e_s(A, \bar{p}) \neq e_s(B, \bar{p})$. The outcome effect is positive if $e_s(A, \bar{p}) > e_s(B, \bar{p})$.*

Thus, an 'outcome effect' exists when for a given p, the outcome of the procedure changes the cooperation level of the agent. A positive output effect means that contributions rise when the project is the one favoured by the agent (A). Note that the term 'outcome' refers here to the outcome of the decision-making procedure (i.e. A or B), not to the utility level of the agent at the end of the game.

Definition 5.3. *A 'procedural effect' exists for a given outcome $n \in \{A, B\}$ and some values p and p' where $p' \neq p$, if $e_s(n, p') \neq e_s(n, p)$. The procedural effect is positive if $p' > p \Rightarrow e_s(n, p') > e_s(n, p)$.*

Definition 5.3 states that the contribution level of an agent does not only depend on the outcome itself, but also on the degree of participation of the procedure by which the outcome was established. A positive procedural effect means that contributions rise with a more participatory procedure. Note that the effect cannot exist in a pooling equilibrium, because it requires that there are at least two equilibrium values of p.

With these definitions in hand, we can derive the existence and characteristics of a separating equilibrium in our game.

Proposition 5.1 (Separating equilibrium)**.** *If and only if $\theta \leq \frac{1}{g}\left(g^{\frac{4-2\alpha}{1-\alpha}} - 1\right)$ then there exists a unique separating equilibrium in pure strategies in which*

a) the selfish authority chooses $p_0^* = 0$,

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

b) the benevolent authority chooses

$$\begin{array}{cc} 0 < p^*_\theta(g,\theta) < 1 & \text{if } \theta < \bar{\theta}(g) \\ 1 & \text{if } \theta \geq \bar{\theta}(g), \end{array} \tag{5.7}$$

c) both the outcome and the procedural effects are positive, i.e.

$$e^*_s(B, p^*_0) < e^*_s(B, p^*_\theta) < e^*_s(A, p^*_\theta) \tag{5.8}$$

Proposition 1 says that the procedure is a signal of the type of the authority. Observing $p^*_\theta > 0$ means that the authority is benevolent, whereas observing $p^* = 0$ means that the authority is selfish. The intuition behind the existence of the separating equilibrium is the following. Consider the choice between p^*_0 and p^*_θ. From the point of view of the authority, two things are relevant. On the one hand, choosing p^*_θ leads to higher equilibrium contributions, as one can see from Proposition 1*c*). This increases utility to the authority for a given outcome A or B. On the other hand, choosing a higher p increases the probability that the authority will end up with the wrong project. The existence of the separating equilibrium comes from the fact that delegating control is more costly for the selfish authority. The equilibrium exists only if θ is not too high relative to g. The intuition behind this condition is that a high θ increases the effort of the agent in the second round, and thus it becomes more profitable to the selfish type mimic the signal. Thus, if g is low relative to θ, the selfish type is willing to mimick even the strongest signal (setting $p = 1$) in which case the separating equilibrium collapses.

Proposition 5.1*b*) tell us that the procedure chosen by the benevolent authority depends on θ and g. If θ is high relative to g (but not so high that it violates the equilibrium condition), then the benevolent authority prefers to implement project A. Note that the conflict of interest is still there: for given effort levels, the authority would prefer to carry out project B. However, when θ is high, the increased cooperation of the agent makes implementing A so attractive, that it outweighs the loss to the authority of her favorite project. As a consequence, the authority will set $p^*_\theta = 1$.

On the other hand, if θ is low relative to g, the increased cooperation by the agent does not compensate the loss of project B. Then, the benevolent authority sets p^*_θ at the lowest level such that the incentive compatibility constraint of the low type is satisfied. This incentive compatibility constrained is graphed in Figure 5.2. If the conflict of interest increases, inducing participation becomes more costly. Thus, a lower p will be sufficient to deter the low type from copying the signal. Furthermore, if θ increases, the curve in Figure 5.2 shifts upwards. The reason is that the higher is θ, the higher is the level of effort of the agent under participation, because he anticipates a higher effort level of the authority. Therefore, it becomes more attractive for the selfish authority to mimic the benevolent authority, and the benevolent authority

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

needs to set a higher p^*_θ in order to signal her type in equilibrium. Note that this implies that there exist equilibria even for arbitrarily small levels of benevolence. All that is required for an equilibrium to exist is that the conflict of interest is large enough relative to the level of benevolence. Naturally, when benevolence is low, the level of p^*_θ will also be low.

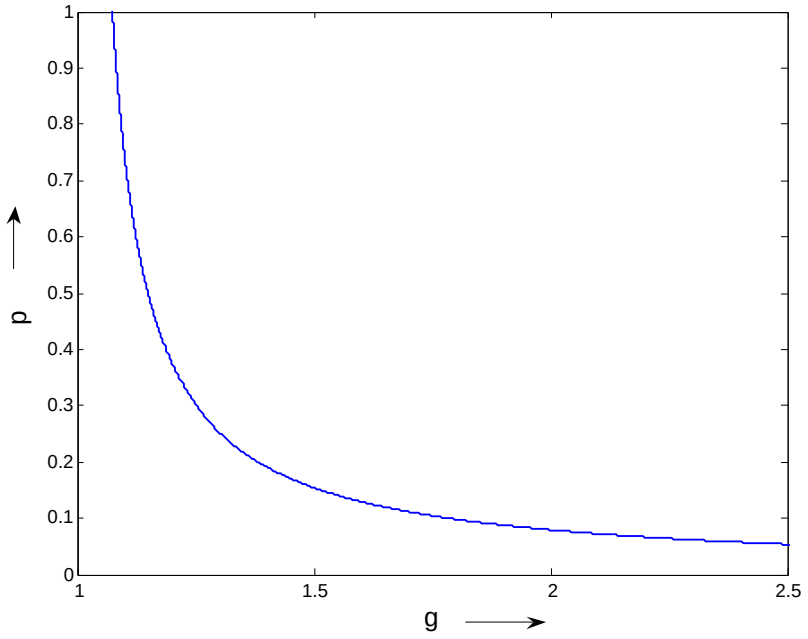

FIGURE 5.2: The equilibrium p^*_θ as a function of the conflict of interest g.

Proposition 5.1*c*) tells us that both the output and the procedural effect are positive. This means the model can account for the main stylized facts observed in the procedural fairness literature: the positive impact on contribution levels of both favorable decision outcomes and participatory procedures. Consider someone who observes $e^*_s(B, p^*_0) < e^*_s(B, p^*_\theta)$. The observer will note that for the same (adverse) result B of the procedure, the agent selects a higher effort if the procedure allowed more participation. Not observing the type of the authority and ignoring the underlying strategic considerations of the game, the observer may be tempted to conclude that the agent in the game has preferences over procedures. In the context of the model however, the increased cooperation can be explained by conventional preferences.

The uniqueness of the equilibrium results from applying the intuitive criterion to rule out pooling equilibria on low participation levels. In any pooling equilibrium, deviations to higher levels of p by the low type are dominated by the equilibrium strategy. As a result, the only off-equilibrium beliefs that are admitted by the intuitive criterion are $\mu(\theta) = 1$, which causes the high type to deviate from the (candidate) pooling equilibrium.

5.6 Discussion

In this section we discuss some implications of the model, relate it to the literature discussed in section 2 and contrast the predictions of the signaling explanation with explanations by other models.

Procedures and cooperation. In the second round of the game the agents play a game that resembles the Prisoners' dilemma. If both players are selfish, the effort levels in the Nash equilibrium are inefficient. That is, a Pareto improvement could be obtained if both agents would exert higher effort. However, this is not an equilibrium, because the players do not take into account the positive externality of their effort on the other player. To some extent, the benevolent authority *does* take this externality into account and this improves efficiency. Participatory procedures serve to improve efficiency further by making the benevolence common knowledge, which in turn increases the effort of the agent and the authority. Thus, utility of the agent is higher under participatory procedures, regardless of the outcome of the procedure. This feature means that the model can explain the robust fact that participation correlates with higher rates of job satisfaction (Cohen-Charash and Spector 2001).

The analysis of signaling equilibria also fits well with the idea of 'token participation', introduced by Arnstein (1969). Token participation refers to participatory procedures that are designed as windowdressing to give people the idea that they have influence, whereas in fact they have little. An example of this would be to invite agents to a decision meeting but to ignore their remarks. The model predicts that only sufficiently high levels of real influence will serve as a signal of benevolence. Indeed, Thibaut *et al.* (1974) provide evidence that if people feel that participation is fake, and that they cannot actually exert any influence, they provide especially low evaluations of the procedure.

In essence, this paper interprets participatory procedures as costly gifts, intended to improve cooperation in (repeated) dilemma games (Carmichael and MacLeod 1997). Of course there are other, more direct ways to give gifts, but in cases where decisions about disputed options have to be made, participation is likely to be a highly salient way to do so.

Relation to social psychology. The signaling model does have important overlaps with some of the models in the social psychology literature that were mentioned in the introduction. It is related closely to the 'fairness heuristic' and 'uncertainty management' theories (Lind and Van den Bos 2002). In these theories, the agent uses information implicit in the fairness of procedures to form judgements about his environment, in this case the trustworthiness of the authority. However, in contrast to the psychology literature, the present model outlines a clear reason for such judgements (the existence of a separating equilibrium) and a straightforward

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

mechanism (Bayesian updating) by which it takes place. The paper is also related to the relational model (Lind and Tyler 1992). Like in the relational model, the agent takes the fairness of procedures as a signal of his importance to the authority.

In contrast to most of the psychology literature the present paper gives an account of procedures in purely instrumental terms. The psychological models just mentioned are conceptually richer, because they allow for other factors such as the identity of the agent or preferences to express opinions. This makes such theories more versatile, but also more complicated. The use of instrumental models can help clarify where richer conceptual models are needed.

Testing instrumental versus intrinsic models of participation. This paper argues for a conception in which preferences for procedural fairness are *instrumental* or extrinsic. This contrasts to the approach of Frey *et al.* (2004) who argue that people derive utility directly from fair institutions, i.e. the preferences are *intrinsic*. The signaling model explains the stylized fact that fairness matters more in the presence of uncertainty (about trustworthiness). An intrinsic approach cannot explain this. On the other hand, a signaling approach does not explain empirical results that pure expression without any potential to change outcomes matters for the evaluation of procedures and authorities. Lind *et al.* (1990) show that such preferences play a role, but also that they are quantitatively smaller than the effects of decision control. Tyler (1987) shows that people react more positively to procedures if they believe that the authority seriously considered their arguments prior to decision-making.

The current model also predicts that participatory or 'fair' procedures mandated by a third party will be less effective than when these procedures are voluntarily introduced. Mandatory participation may raise cooperation levels by providing better outcomes but will not send a signal of benevolence, because they do not reveal the character of the authority. By contrast, an approach that posits (only) intrinsic preferences for fair procedures would predict that mandatory and voluntary institutions are equally succesful in raising cooperation.

Feldman and Tyler (2008) have attempted to test whether mandating voice procedures is an effective way to increase compliance. They interviewed employees in Israeli firms and asked the reactions of the employees to two different (and fictitious) introductions of participatory procedures: voluntarily by the employer or mandated by the government. The results are not fully conclusive. They find that the mandated introduction raises willingness to comply by more, but only for those whose actual employer had no participatory procedures in place. For those who enjoyed such procedures in their real working environment, the effect of the voluntarily introduced procedure was bigger. This evidence may reflect that employees who did not enjoy voice procedures in reality, were skeptical towards the voluntary provision and have more faith in mandated procedures. On the other hand, those that already had a good view of their employer may have seen the voluntary provision as further evidence of the trustworthiness of

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

the employer. However, the fact that this study relies on fictitious scenarios and self-reported compliance under such fictitious scenarios, makes further research desirable.

Finally, it will be difficult to disentangle the effects of signaling and reciprocity. In the model, agent are always selfish, while the authority may be benevolent. The reason for this modeling choice is to show that we need not infer reciprocal attitutes in agents if they react positively to fair procedures. Complementarity in levels of cooperation suffices for the result. However, it may be more realistic to assume that the agent is *reciprocal*, in the sense that he will care about the authority's payoff if he thinks the authority cares about his. If this is the case, the result in this paper will obtain more easily, because by signaling his type the authority does not only induce trust, but also reciprocity.

5.7 Conclusion

In this paper I model participation in decision-making procedures as the degree of (stochastic) influence that people have on the decision-making process. I then showed that if an authority can commit to the outcomes of such procedures, they can be used as a signaling device. A procedure with ample participation possibilities indicates that the authority is of a good type that will not exploit the agent. Thus, participation increases trust and cooperation, even if the actual outcome of the procedure is not beneficial to the agent. This matches the most important stylized fact in the literature on procedural fairness, namely that participatory procedures increase compliance, regardless of their outcome. The model can also explain why participation is especially important when there is uncertainty about the type of the authorities, and why participation increases job satisfaction.

Decision procedures and procedural fairness are complex phenomena and I do not claim to have delivered anything more than a first step to an economic understanding of it. Nevertheless, I believe that taking a formal approach to decision-making procedures allows some important conceptual clarifications, and among other things, will help to understand when a less reductionist approach is necessary.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Chapter 6

Epilogue: from deterrence to participation?

> "The purpose of getting power is to be able to give it away."
>
> Aneurin Bevan (1897-1960).

When the allied forces landed in Sicily on the 10th of July 1943, they split the task of conquering the island. The combined British-Canadian forces would advance north along the eastern coast, while the Americans would march west towards Palermo. On the paper, the former task was by far the easier. The Italian-German force in the east was outnumbered five to one and ill-equiped. Nevertheless, they put up a staunch fight, employing clever tricks, such as making up for their lack of ammunition by using firecrackers to divert enemy fire. It took the British and Canadians 5 weeks and some thousands of casualties to arrive in Messina on the Northern shore.

Their American Seventh Army on the other hand had to conquer the mountainous inland of the island where about 60,000 Italian-German defense forces were concentrated, amongst which a German tank division. Moreover, these forces had taken up strategic positions in the difficult mountainous terrain that had been proven to be an almost unconquerable hideout. Nevertheless, the Americans covered the 100 mile distance from Agrigento to Palermo in a remarkable four days without meeting any noticeable resistance.

What caused this enormous difference between the two forces? A popular account is that the Americans employed the services of the Sicilian-born American gangster 'Lucky' Luciano, who was also a patriotic American. His contacts within the Sicilian Mafia, and especially with the powerful Sicilian boss Don Calò have been said to be instrumental in guaranteeing the American Army a free passage across the island. Although the actual influence of the Mafia

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

on the outcome of the invasion is debated[1] there is little doubt that the Allies cooperated with the Mafia in conquering the island (Newark, 2007). Through the contacts of Luciano the Mafia allegedly gathered intelligence for the Allies and sabotaged Axis war efforts. After the invasion, the organization helped suppress dockworkers' strikes and communist gatherings. On their part, immediately after they reached Palermo, the Allies appointed Don Calò mayor of his village as well as Honorary Colonel of the US Army. Luciano's 30 to 50 year sentence was converted and he was paroled in 1946. Moreover, the Allies appointed Mafiosi to important positions all over Sicily. On the basis of recently declassified documents Newark (2007) writes that the whole allied change of command, all the way up to Roosevelt and Eisenhower condoned the cooperation, and that the Allies on some occasions even armed the Mafia.

This tale underscores several points that have been made in this thesis. First, it is a stark example of the fact that effective rule flows not merely from force, but from the informal forces in society. While the British and Canadians were fighting hard in the east, the American army walked south to north across Sicily's mainland without encountering any resistance.

Second, the example drives home the message in chapter 2, namely that knowledge of the social landscape in a society is indispensable for any governing authority. In the Sicilian case, the Allies were well prepared. They had made sure that 15% of the invading force consisted of Sicilians who had migrated to the US, so that they would be more likely to be greeted as liberators. They had primed their connections with the Mafia via Lucky Luciano, facilitating the invasion and subsequent occupation. One can contrast these preparations with the conquest of Iraq in 2003, where the Allies did not have a good idea of the sociological conditions in the country, and were not able to respond adequately to the ethnic tensions.

Third, it shows the importance of the strategic devolution of power. After the invasion of Sicily, the first thing the American's did was to enlist the informal groups in society to secure order. Through the quick devolution of power to local underground power structures (the Mafia) they immediately established a form of effective local rule, and filled the power vacuum that existed after the fall of fascism. Again, this contrasts sharply with the events in Iraq. There, instead of delegating power, the Americans immediately *dissolved* the most important power structures (the Sunni dominated army and Ba'th party). This augmented the post-Saddam power vacuum so that it could not be filled even by the supremely effective American force. Setting up alternative power structures proved slow and costly: although a token Iraqi governing council was quickly established, by the time general elections were held in December 2005 (the earlier elections for provisional government had been boycotted by the Sunni's) the country was already engaged in severe sectarian violence. It is telling that the fragile order in

[1] The account here is based largely on Norman Lewis' (1964), who argues in his book the *The Honored Society* that the Mafia captured the Italian commander of the defense forces and caused the Italian troops to desert. However, Newark (2007) plays down the influence of the Mafia on the outcome of the invasion and argues in *Mafia Allies* that the Italians simply gave up the fight.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Iraq today relies on the US paying and arming their former enemies in the Sunni insurgency (see Dawisha (2009) for a political history of Iraq before and after the invasion).

These examples vividly illustrate that governing without civil society is impossible. Cooperation with organizations like the Sunni insurgents and the Mafia was vital to establish social order in the short run. Although Sicily and Iraq are war time examples, they carry lessons for peace-time policy making. If social order stems from the informal structures of civil society, and not just from the threat of force by the state, social control cannot simply be imposed by strong incentives. Instead, to rule effectively the state has to share it's powers and devolve responsibilities to individuals and groups in civil society. This point is made well by criminologist David Garland. In his book *The Culture of Control* he has studied crime-fighting policies in Britain and the United States over the last decades. In his conclusion he states that

> "The lesson of the twentieth century experience is that the nation state cannot any longer hope to govern by means of sovereign commands issued to obedient subjects, and this is true whether the concern is to deliver welfare, to secure economic prosperity, or to maintain 'law and order'. In the complex, differentiated world of late modernity, effective, legitimate government must devolve power and share the work of social control with local organizations and communities."
>
> D. Garland (2001, p. 205).

Thus, as was argued in chapter 2, economists need a radical overhaul from their traditional Hobbesian way of thinking about policy making. Instead, what is needed is better understanding of the possibilities of governing through participation and devolution of power.

The importance of this can again be understood by references to the invasion of Sicily. Although the cooperation with the Mafia brought short-term gains, the long-term effects are not at all pretty. In the first weeks after the allied invasion the Mafia re-established its grip over Sicily that had been weakened under the fascist regime. Today, Sicilian society is still bearing the consequences of this Allied trade-off. It does not take a wizard to predict that installing a criminal organization in the driver's seat is not conducive to long-run prosperity. However, as Garland's quote demonstrates, effective government will have to find some way to devolve power in order to govern effectively. Economists understand relatively little of these processes. Some elements may fit relatively easily into an economics framework. For example, in the Sicilian story, the collaboration between the Allies and the Mafia seems to have been a simple quid pro quo exchange. The Mafia supplied help in securing the Allies war objectives, whereas the Allies helped, or at least did not obstruct, the Mafia's post-war power grab. These kind of deals and their associated complexities (issues of commitment, reputation etc.) are well-understood by economists.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

However, this does by no means exhaust the possibilities to use delegation for effective rule. One of the most important issues connected to participatory governance that has escaped rational choice theorists is that of *legitimacy*. Many studies have shown that participation in decision-making extends 'legitimacy' to the authority (Tyler, 2004). In turn, this legitimacy is one of the main currencies by which effective rule is exercised: People tend to cooperate more with authorities that they view to as legitimate and tend to ignore the commands of those that they do not (see also the references in Chapter 5).

This suggests a completely different style of governing than the command and control framework. Indeed, participation is in some sense the opposite of deterrence, in that it gives *less* control to the authority. Tyler (2008) sketches an optimistic picture of how participatory procedures and legitimacy can all but replace deterrent strategies. I am not similarly optimistic, but I do think that legitimacy is an under-researched topic in economics. How legitimacy is established and how it can be used to generate voluntary compliance are important questions that rational choice theorists have ignored.

By contrast, social psychologists have done much work on this topic. Chapter 4 outlined the main psychological theories about the benefits of participatory government. These revolve around the material benefits that an individual can secure by being able to influence decisions (Thibaut and Walker, 1975), identification with the group (Lind and Tyler, 1988) and the resolution of uncertainty about the environment (van den Bos and Lind, 2002).

I think that an interdisciplinary approach to this subject would benefit both social psychology and economics. Formal definition of the concepts of 'legitimacy' itself, but also of related concepts such as 'respect', and 'authority', could bring intellectual clarity to this field. Chapter 5 shows an example of how economic models can deliver such clarity. I think the model presented in this chapter considerably sharpens the concepts that are implicit in both the 'fairness heuristic' and 'uncertainty management' theories developed by van den Bos and Lind (2002).

On the other hand, economists and rational choice theorists can also learn from psychologists. This is especially the case with respect to the aspects of compliance and decisions making that have to do with identity. Lind and Tyler (1988) stress the concept of group identity in their 'group value' theory. The idea is that by being able to participate in decision-making and express their opinions, agents identify with the group, and to some degree internalize the group benefit as their own. There are many subtle issues here that on first glance seem to defy economic modeling techniques. Nevertheless, there is progress. For example, Bénabou and Tirole (2007) provide a very interesting model of identity formation, that can explain a wide variety of behaviors and institutions that were previously unintelligible to standard economic theory. Another area of progress is on the concept of *esteem*. The group value theory says that participation in decision-making in the group is important because it confers the esteem of the group on the individual. The economics of esteem-based incentives is a topic that currently

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

receives a lot of attention in economics (Ellingsen and Johannesen 2007, Brennan and Pettit 2004).

In short, an integration of the conceptual frameworks of economics and social psychology may enrich these two disciplines. Moreover, by recognizing that less control is sometimes more, we may end up with a more effective and humane way to exercise social control.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Appendix A

Proofs

Proofs of Chapter 2

Proof of Proposition 2.1. A separating equilibrium exists if the threshold type is indifferent between complying or not. It is easy to see that a single-crossing condition holds, such that if the threshold type is indifferent, all types higher than the threshold type will prefer not to comply, and all types lower than the threshold type will prefer to comply.

Denote the threshold type by θ^*. The utility of the threshold type is:

$$u(d) = \theta^* - h + a(s)E[\theta \mid \theta^*, d].$$

Requiring indifference between complying and not complying yields the following:

$$\begin{aligned} u(0) &= u(1) \\ \theta^* - h + a(s)E[\theta \mid \theta^*, d=0] &= \theta^* - h + a(s)E[\theta \mid \theta^*, d=1] \\ \delta(\theta^*) &= \frac{h - \theta^*}{a(s)} \end{aligned} \tag{A.1}$$

where $\delta(\theta^*) = E[\theta \mid \theta^*, d=0] - E[\theta \mid \theta^*, d=1]$, the difference in respect for compliers and deviators.

We know that $\delta(\theta^*)$ is continuous and defined on $[0,1]$. Thus, if and only if

$$\begin{aligned} &\delta(0) \leq \frac{h}{a(s)} \text{ and } \delta(1) \geq \frac{h-1}{a(s)} \text{ , or} \\ &\delta(0) \geq \frac{h}{a(s)} \text{ and } \delta(1) \leq \frac{h-1}{a(s)}, \end{aligned} \tag{A.2}$$

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

we know that $\frac{h-\theta^*}{a(s)}$ crosses $\delta(\theta^*)$ at least once, guaranteeing the existence of at least one separating equilibrium.

We know from the definition of $\delta(\theta^*)$ that $\delta(0) = E[\theta]$ and $\delta(1) = 1 - E[\theta]$. Using this it is easy to derive that $min\left\{a(s)E[\theta], 1 + a(s)\left(1 - E[\theta]\right)\right\} \leq h < max\left\{a(s)E[\theta], 1 + a(s)\left(1 - E[\theta]\right)\right\}$ is a sufficient condition for (A.2). (The fact that the second inequality is strict has to do with the tiebreaking rule that an indifferent type complies. For a separating equilibrium one needs to a positive fraction of agents who do not comply.) □

Proof of Proposition 2.2. In the proof of Proposition 1 we derived the equilibrium condition for the threshold type: $h = \theta^* + a(s)\delta(\theta^*)$. If we take the total derivative of this expression with respect to h and θ^* we get:

$$\begin{aligned} dh &= d\theta^* + a(s)\frac{d\delta(\theta^*)}{d\theta^*}d\theta^* \\ \frac{d\theta^*}{dh} &= \frac{1}{1 + a(s)\frac{d\delta(\theta^*)}{d\theta^*}} \end{aligned} \tag{A.3}$$

It follows that $\frac{d\theta^*}{dh} < 0 \Leftrightarrow \frac{d\delta(\theta^*)}{d\theta^*} < -\frac{1}{a(s)}$. □

Proofs of Chapter 3

Proof of Proposition 3.1. Proof of 1. We work backwards through the game, and start by characterizing the agents reaction functions. We know from (3.2) and (3.4) that both types have a 'threshold sanction': for lower sanctions than this threshold they defect, for higher sanctions they cooperate. Low types cooperate when the sanction is higher than 1, and defect otherwise. From (3.4) we know that high types cooperate when $g \geq 1 - (1 - \theta)p(m > \overline{m})$, and defect otherwise. In the symmetric information when $\geq \overline{m}$ it is sufficient that

$$g \geq \theta \tag{A.4}$$

The reaction functions imply that when $g < 1$, all egoists defect and the conditional cooperators face a coordination game between themselves. If all other high types defect it is best for a high type to also defect. If all other high types cooperate, it is a best response for the high types to cooperate (at least when $\omega \geq \overline{m}$). Suppose high types coordinate on defection. In this case the government can set $g < 1$ resulting in $m = 0$, or it can set $g = 1$ resulting in $m = 1$. From the objective function of the government it is straightforward to verify that when $\alpha < 1$, the latter strategy dominates the former.

Proof of 2. Above we derived the reaction functions of the citizens. We known that (A.4) holds with equality in equilibrium, so that $g_2^* = \theta$, because the government always sets the lowest

possible sanctions to induce cooperation. The government will set low sanctions iff:

$$\begin{aligned} W(\omega, g_2^*) &\geq W(1, g_1^*) \\ \omega - \alpha g_2^* &\geq 1 - \alpha \\ \omega &\geq 1 - \alpha(1 - g_2^*) \end{aligned} \tag{A.5}$$

In equilibrium, this 'incentive compatibility constraint' holds with equality for the lowest government type that sets low sanctions, and with inequality for all higher types. Since the government will always set the lowest possible sanctions in equilibrium, i.e. $g_2^* = \theta$, the threshold government type is given by $1 - \alpha(1 - \theta)$.

Naturally, government will set high sanctions if $\omega < \overline{m}$. Thus, we have

$$\omega^* = \begin{cases} 1 - \alpha(1 - \theta) & \text{if } \overline{m} < 1 - \alpha(1 - \theta) \\ \overline{m} & \text{if } \overline{m} \geq 1 - \alpha(1 - \theta) \end{cases} \tag{A.6}$$

□

Proof of Proposition 3.2. Identical to that of Proposition 3.1, Part 1. □

Proof of Lemma 3.1. From the reaction functions derived above, we see that contribution by the low types implies contribution by the high types. Thus, there are at most three different equilibrium action profiles for the citizens in the economy: one where both types contribute, one where only the high types contribute, and one where nobody contributes. This means that in equilibrium there are at most three different levels of sanctions g. If there were more, two such levels induce the same strategic reactions by the agents. This cannot be an equilibrium since the government would always deviate to the lower and cheaper sanction that induces a given reaction. This means that the three sanction levels that are candidates to feature in equilibrium are the ones that most cheaply induce the three possible citizens' strategy profiles described above. From the citizens' reaction functions, we see that setting $g = 0$ and $g = 1$ is the cheapest way of inducing respectively no cooperation and full cooperation. Since we assumed that $\alpha < 1$, we see from the welfare function that setting $g = 1$ and inducing full cooperation always yields a higher payoff to the government than setting $g = 0$ and leaving everybody to defect.

Therefore, defection by all agents in the economy can not be an equilibrium outcome. We are left with at most two possible equilibrium outcomes: one where both types contribute, one where only the high types contribute. As a consequence, there are at most two sanction levels, one associated with each outcome. □

Proof of Lemma 3.2. We prove the lemma by showing the following:

1. A government that observes $\omega = 0$ sets $g = 1$ in equilibrium.

2. For a government that observes $\omega = 1$, the upper bound on the equilibrium sanction is $max\left\{\theta, 1 - \frac{1}{\alpha}(1 - \overline{m})\right\} < 1$

Proof of 1. In a state of society $w = 0$ where everybody is egoistic, setting $g < 1$ will lead everyone to defect which cannot be optimal for the government.

Proof of 2. The proof is based on the application of the 'intuitive criterion' (Cho and Kreps 1987), a refinement of Bayesian Nash equilibrium. An equilibrium fails the intuitive criterion (IC) if it requires off-equilibrium beliefs that place positive probability on types for whom deviation payoffs are dominated by equilibrium payoffs. The idea is that it is 'unreasonable' to believe that such types would have deviated. Denote by $\Omega(g')$ the set of government types who will deviate to an off equilibrium sanction g'. We call beliefs with full density inside $\Omega(g')$ 'IC-admissible'. Then $[0, 1]/\Omega(g')$ is the set of types who would never deviate to a sanction g'. Beliefs with density in this set are 'non IC-admissible'.

We make two observations that restrict the set of deviations that we need to consider. First, off-equilibrium sanctions are attractive deviations for a given government type ω if they induce at least as many contributions as in equilibrium for a lower sanction level. We need not consider deviations to $g = 1$ because by Lemma 1 and the first part of this proof, these are always on the equilibrium path. For a deviation to a sanction level $g < 1$, the contribution level will be either ω or 0. A deviation cannot be profitable if contributions are 0. Thus, we focus on deviations to sanctions g' that induce a contribution level of ω. From (A.4) we know that if $g' < \theta$, sanctions will never (for any beliefs) induce cooperation from high types, and so we look only at deviations to sanctions $\theta \leq g' < 1$.
Second, we can restrict our attention to deviations by the government type $\omega = 1$. In this case the whole population consists of high types, and a contribution level of ω equals the maximum contribution level. Therefore, if this type does not deviate, other types will not do so either.

In sum, a pooling equilibrium exists if for $\omega = 1$ and for all $g' \neq g*$ there exist off-equilibrium beliefs that:

1. are 'IC-admissible', and

2. lead to zero contributions, thus making deviations unprofitable.

Consider deviations from the pooling equilibrium in which the government plays $g = 1$ and everyone contributes. If the government $\omega = 1$ deviates to a lower sanction $g' < 1$, it will generate full contributions if the off-equilibrium sanction induces cooperation from the high types.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

The set $\Omega(g')$ of government types that will deviate under such circumstances is determined by comparing the government's utility in equilibrium to that of a deviation:

$$\begin{aligned} EW(\omega, g') &\geq EW(\omega, g = 1) \\ \omega - \alpha g' &\geq 1 - \alpha \\ \omega &\geq 1 - \alpha(1 - g') \end{aligned} \tag{A.7}$$

Thus, we have $\Omega(g') = [1 - \alpha(1 - g'), 1]$. The best case for a pooling equilibrium is made when off-equilibrium beliefs are as low as possible. The most negative off-equilibrium beliefs that are admissible by the IC are a degenerate distribution with full density on $1 - \alpha(1 - g')$. These beliefs will lead to zero contributions if $\overline{m} > 1 - \alpha(1 - g')$. Solving for g' yields

$$g' < 1 - \frac{1}{\alpha}(1 - \overline{m}) \tag{A.8}$$

Thus, if an off-equilibrium sanction $\theta \leq g' < 1$ satisfies $g' \geq 1 - \frac{1}{\alpha}(1 - \overline{m})$, we can find IC admissible beliefs that induce positive contribution levels.

It is easy to see that we can always find sanctions $max\left\{\theta, 1 - \frac{1}{\alpha}(1 - \overline{m})\right\} \leq g' < 1$, for which there are no IC-admissible beliefs that are sufficiently low to induce a zero cooperation level. Thus, for the type $\omega = 1$ there is a profitable deviation to a sanction that is slightly lower than 1. Thus, a pooling equilibrium on $g = 1$ cannot exist. □

Proof of Proposition 3.3. Note that for various reasons we cannot use a standard single crossing property condition (sanctions are equally costly for each government type, and the decision variable of the agents is binary). The proof proceeds in four steps. First, we characterize the citizens' posterior belief about the distribution of types in the economy. Agents base their beliefs on the government's policy and their own type. We derive only the posterior beliefs of conditional cooperators (high types) under a sanction $g < 1$, because this is the only case in which beliefs matter for the choice of action[1].

Conditional on $g_2 < 1$ and $\Theta = \theta$ we compute from Bayes' rule the posterior belief distribution $\mu(\omega)$ that a given distribution ω has been chosen by nature. The common prior is that each distribution is equally likely to be chosen by nature. Obviously $\mu\left(\Omega = \omega < \omega^* \mid g = g_2, \Theta = \theta\right) = 0$, because the agent knows that a low sanction is played only if $\omega \geq \omega^*$. The posterior for Ω

[1] Concerns of space lead me to omit the full characterization of posterior beliefs of agents. These are available on request.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

$= \omega \geq \omega^*$ is:

$$\begin{aligned}\mu\left(\Omega = \omega \geq \omega^* \mid \Theta = \theta, g = g_2\right) &= \frac{P\left(\Omega = \omega \cup \Theta = \theta \cup \omega \geq \omega^*\right)}{P\left(\Theta = \theta \cup \omega \geq \omega^*\right)} \\ &= \frac{\frac{\omega}{1-\omega^*}}{\int\limits_{\omega^*}^{1} \frac{\omega}{1-\omega^*} d\omega} \\ &= \frac{2\omega}{1-(\omega^*)^2}\end{aligned}$$

Hence,

$$\mu\left(\Omega = \omega \mid \Theta = \theta, g = g_2\right) = \begin{cases} 0 & if\ \omega < \omega^* \\ \frac{2\omega}{1-(\omega^*)^2} & if\ \omega \geq \omega^* \end{cases} \tag{A.9}$$

Second, we determine the best response of the citizens in the economy to any government policy given their posterior beliefs and their type. Both types will cooperate under $g_1 = 1$. We know that the best response of a low type is to defect whenever $g < 1$. Remains to analyze the case of a high type who observes g_2. From (3.4) we know that best response of a high type is to cooperate if and only if $P(m > \overline{m}) \geq \frac{1-g}{1-\theta}$.

To get the best response of the citizens, we have to compute the equilibrium value $P^*(m > \overline{m} \mid g_2^*)$ from the equilibrium beliefs. If $\overline{m} \leq \omega^*$, it is straightforward that $P^*(m > \overline{m} \mid g_2^*) = 1$. Substituting this in (3.4) yields the equilibrium condition for the cooperation of high types

$$g_2^* \geq \theta \tag{A.10}$$

If $\overline{m} > \omega^*$ the equilibrium beliefs are given by the following equation:

$$\begin{aligned}P^*(m > \overline{m}) &= \int\limits_{\overline{m}}^{1} \frac{2\omega}{1-(\omega^*)^2} d\omega \\ &= \frac{1-\overline{m}^2}{1-(\omega^*)^2}\end{aligned}$$

Substituting this in (3.4) yields the equilibrium condition for the cooperation of high types:

$$g_2^* \geq \frac{\overline{m}^2 - (\omega^*)^2 + \theta(1-\overline{m}^2)}{1-(\omega^*)^2} \tag{A.11}$$

Third, the best response of the government types is described by the incentive compatibility constraint (A.5) derived above, that gives the threshold type that is indifferent between the high and the low sanction. We now know the reaction functions of all the players, depending on the parameters.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

The fourth step is deriving the equilibrium conditions on the parameter values, starting with the equilibrium sanction. We need to consider both the case when $\overline{m} > \omega$ and the complement.

Case 1: $\overline{m} \leq \omega^*$.

In this case, equilibrium beliefs $P(m > \overline{m}) = 1$, and so from (3.4) it follows that $g_2^* \geq \theta$ is sufficient for cooperation of the high types. From the ICC of the government (A.5) it follows that $\omega^* \geq 1 - \alpha(1-\theta)$. If these two conditions hold with equality it is easy to check that they constitute an equilibrium. Deviations to $g_2 > \theta$ are never profitable and deviations to $g_2 < \theta$ lead to $m = 0$.
Now suppose that $g_2^* > \theta$. Consider a deviation to $g' = \theta$. The intuitive criterion specifies (see proof of Lemma 3.2) that the lowest reasonable off equilibrium beliefs are $1 - \alpha(1-g')$. A deviation to $g' = \theta$ is thus profitable as long as $\overline{m} \leq 1 - \alpha(1-\theta)$.

Case 2: $\overline{m} > \omega^*$.

In this case $P(m > \overline{m}) = \frac{1-\overline{m}^2}{1-(\omega^*)^2}$ and $g_2^* \geq \frac{\overline{m}^2 - (\omega^*)^2 + \theta(1-\overline{m}^2)}{1-(\omega^*)^2}$. From the government's ICC one can derive that the (lower bound of the) equilibrium threshold $\underline{\omega}$ is given implicitly by:

$$(1-\underline{\omega})\left(1-\underline{\omega}^2\right) \leq \alpha(1-\theta)(1-\overline{m}^2) \tag{A.12}$$

Suppose that $g_2^* = \frac{\overline{m}^2 - (\omega^*)^2 + \theta(1-\overline{m}^2)}{1-(\omega^*)^2}$, so $\omega^* = \underline{\omega}$ and is given by (A.12) with equality. It is clear that deviations to $g' > g_2^*$ are never profitable. Deviations to $\theta \leq g' < g_2^*$ are unprofitable as long as $\overline{m} > 1 - \alpha(1-g')$ (see proof of Lemma 3.2). We know from the government's ICC that $\underline{\omega} > 1 - \alpha(1-g')$. Thus we have that $\overline{m} > \underline{\omega} > 1 - \alpha(1-g')$. This means we can always find off-equilibrium beliefs that make a deviation unprofitable and the equilibrium exists.

Now consider as an equilibrium sanction $g_2^* > \frac{\overline{m}^2 - (\omega^*)^2 + \theta(1-\overline{m}^2)}{1-(\omega^*)^2}$, and thus (by the ICC), $\omega^* > \underline{\omega}$. It is clear that deviations to $g' > g_2^*$ are never profitable. Deviations to $\theta \leq g' < g_2^*$ can be ruled out by reasonable off-equilibrium beliefs by the same reasoning as above. Thus this equilibrium exists.

Summarizing, we have

$$\omega^* \begin{cases} = 1 - \alpha(1-\theta) & \text{if } \overline{m} < 1 - \alpha(1-\theta) \\ \in [\underline{\omega}, \overline{m}] & \text{if } \overline{m} \geq 1 - \alpha(1-\theta) \end{cases}$$

where $\underline{\omega}$ is given in (A.12).

Proof of 2. Comparing $\overline{\omega}$ just derived, with ω^* in Proposition 3.1, the proof is immediate. □

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Proofs of Chapter 5

Proof of Lemma 5.1. A sufficient condition for the result is that the indifference curves of the two types cross only once in the identified region. This will guarantee that the change in e_s to compensate the authority for a change in p is larger for the selfish authority than for the benevolent authority. Let $\{p, e_s(n,p,a)\}$ be the indifference curve of type a, i.e. $U_a(p, e_s(n,p,a)) \equiv \overline{U}_a$. We assume that the effort of the authority is a best response to the effort of the agent, i.e. $e_a(n,p) = e_a^*(e_s(n,p,a))$, which is a function of $e_s(n,p,a)$ only. Taking the derivative of $U_a(p, e_s(n,p,a))$ w.r.t. p yields

$$
\begin{aligned}
0 &= \frac{\partial U_a(p, e_s(n,p,a))}{\partial p} + \frac{\partial U_a(p, e_s(n,p,a))}{\partial e_s(n,p,a)} \frac{\partial e_s(n,p,a))}{\partial p} \\
0 &= u_a\left(A, e_s(A,p,a)\right) - u_a\left(B, e_s(B,p,a)\right) \\
&\quad + p\frac{\partial u_a\left(A, e_s(A,p,a)\right)}{\partial e_s(A,p,a)} \frac{\partial e_s(A,p,a)}{\partial p} + (1-p)\frac{\partial u_a\left(B, e_s(B,p,a)\right)}{\partial e_s(B,p,a)} \frac{\partial e_s(B,p,a)}{\partial p}
\end{aligned}
$$

We want to show that for $n \in \{A, B\}$:

$$
\left.\frac{\partial e_s(n,p,0)}{\partial p}\right|_{U_0=\bar{U}_0} > \left.\frac{\partial e_s(n,p,\theta)}{\partial p}\right|_{U_\theta=\bar{U}_\theta} \tag{A.13}
$$

Let us first investigate the case in which the authority is compensated for the change in p by a change in $e_s(A,p,a)$, i.e. $\partial e_s(B,p,a)/\partial p = 0$. Then we can write:

$$
\frac{\partial e_s(A,p,a)}{\partial p} = \frac{u_a\left(B, e_s(B,p,a)\right) - u_a\left(A, e_s(A,p,a)\right)}{p\frac{\partial u_a(A, e_s(A,p,a))}{\partial e_s(A,p,a)}}. \tag{A.14}
$$

We can simplify the denominator as follows:

$$
\begin{aligned}
\frac{\partial u_a\left(A, e_s(A,p,a)\right)}{\partial e_s(A,p)} &= \frac{\partial \pi_a}{\partial e_s(A,p,a)} + \frac{\partial \pi_a}{\partial e_a^*(A,p)} \frac{\partial e_a^*(A,p)}{\partial e_s(A,p,a)} - e_a^*(A,p)\frac{\partial e_a^*(A,p)}{\partial e_s(A,p,a)} \\
&= \frac{\partial \pi_a}{\partial e_s(A,p,a)},
\end{aligned}
$$

where in the second line we used the envelope theorem to set $\frac{\partial e_a^*(A,p)}{\partial e_s(A,p,a)} = 0$. We can now write (A.14) as:

$$
\frac{\partial e_s(A,p,a)}{\partial p} = \frac{u_a\left(B, e_s(B,p,a)\right) - u_a\left(A, e_s(A,p,a)\right)}{p\frac{\partial \pi_a}{\partial e_s(A,p,a)}} \tag{A.15}
$$

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

It remains to derive the components of (A.15) and show that (A.13) holds. First consider the denominator. To ease notation we denote $e_s(n,p,a)$ by e_{sn}. We find

$$\frac{\partial \pi_0\left(A, e_{sA}\right)}{\partial e_{sA}} = (1-\alpha)\left(e_{sA}\right)^{-\alpha}\left(e_{aA}\right)^{\alpha}, \text{ and} \tag{A.16}$$

$$\frac{\partial \pi_\theta\left(A, e_{sA}\right)}{\partial e_{sA}} = (1+\theta g)(1-\alpha)\left(e_{sA}\right)^{-\alpha}\left(e_{aA}\right)^{\alpha} \tag{A.17}$$

Now we specify the numerator of (A.15). Denote $\Delta_a = u_a\left(B, e_{sB}(p,a)\right) - u_a\left(A, e_{sA}(p,a)\right)$. We have

$$\Delta_0 = \left[g\left(e_{sB}\right)^{1-\alpha}\left(e_{0B}\right)^{\alpha} - \frac{\left(e_{0B}\right)^2}{2}\right] - \left[\left(e_{sA}\right)^{1-\alpha}\left(e_{0A}\right)^{\alpha} - \frac{\left(e_{0B}\right)^2}{2}\right] \tag{A.18}$$

$$\Delta_\theta = \left[(g+\theta)\left(e_{sB}\right)^{1-\alpha}\left(e_{\theta B}\right)^{\alpha} - \frac{\left(e_{\theta B}\right)^2}{2}\right] - \left[(1+\theta g)\left(e_{sA}\right)^{1-\alpha}\left(e_{\theta A}\right)^{\alpha} - \frac{\left(e_{\theta B}\right)^2}{2}\right] \tag{A.19}$$

The best responses $e_a^*(n,p)$ in the simultaneous move game are readily calculated by taking first order conditions from the authorities' payoffs:

$$e_{0A}^* = \left[\alpha\left(e_{sA}\right)^{1-\alpha}\right]^{\frac{1}{2-\alpha}} \tag{A.20}$$

$$e_{0B}^* = \left[g\alpha\left(e_{sB}\right)^{1-\alpha}\right]^{\frac{1}{2-\alpha}} \tag{A.21}$$

$$e_{\theta A}^* = \left[(1+\theta g)\alpha\left(e_{sA}\right)^{1-\alpha}\right]^{\frac{1}{2-\alpha}} \tag{A.22}$$

$$e_{\theta B}^* = \left[(g+\theta)\alpha\left(e_{sB}\right)^{1-\alpha}\right]^{\frac{1}{2-\alpha}} \tag{A.23}$$

We insert these best responses of the authority into the expressions for the denominator (A.16), (A.17) and the numerator (A.18), (A.19). After some tedious algebra we can define $Z = \left[\alpha^{\frac{\alpha}{2-\alpha}}\left(e_{sB}\right)^{\frac{(1-\alpha)^2}{2-\alpha}} - \alpha^{\frac{2}{2-\alpha}}\left(e_{sB}\right)^{\frac{2(1-\alpha)}{2-\alpha}}\right]$ and $Q = -e_{sA} + \alpha^{\frac{2}{2-\alpha}}\left(e_{sA}\right)^{\frac{2(1-\alpha)}{2-\alpha}}$ and write:

$$\begin{aligned}
\frac{\partial e_s(A,p,0)}{\partial p} &\geq \frac{\partial e_s(A,p,\theta)}{\partial p} \\
g^{\frac{2}{2-\alpha}} Z + Q &\geq \left(\frac{g+\theta}{1+\theta g}\right)^{\frac{2}{2-\alpha}} Z + Q \\
g &\geq \frac{g+\theta}{1+\theta g} \\
g^2 &\geq 1,
\end{aligned}$$

which is always satisfied.

An analogue argument exists if compensation is in $e_s(B,p,a)$, but this is omitted here for reasons of space. Thus, we showed that for a small increase in p, the high type needs to be compensated

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

with a smaller increase in $e_s(p)$ than the low type. This means the indifference curves cross only once. □

Proof of Proposition 5.1. The outline of the proof is as follows. We first derive the optimal levels of effort of the authority and the agent. Then we show the conditions under which exists a $\tilde{p} > 0$ with $\mu(a = \theta \mid \tilde{p}) = 1$, such that the selfish authority is indifferent between setting $\tilde{p}$ and $p = 0$ (the equilibrium level in the separating equilibrium). We then show that the benevolent authority always prefers to set $\tilde{p}$ to any other p. Finally, we prove uniqueness by ruling out pooling equilibria. To ease notation we denote $e_t(n, p)$ by e_{tn}.

Equilibrium effort levels. Optimal effort levels are derived by taking derivatives of π_t for the agent and both types of authorities, yielding the reaction functions. Equilibrium levels are then computed by solving the system of reaction functions for effort levels. Note that effort levels will depend on the information of the agent. It is straightforward to compute the equilibrium effort levels under complete information (as is the case in a separating equilibrium). The optimal effort level of the agent is:

$$e^*_{sA} = [g(1-\alpha)]^{\frac{2-\alpha}{2}} [\alpha(1 + 1_\theta \theta g)]^{\frac{\alpha}{2}} \tag{A.24}$$

$$e^*_{sB} = [1-\alpha]^{\frac{2-\alpha}{2}} [\alpha(g + 1_\theta \theta)]^{\frac{\alpha}{2}} \tag{A.25}$$

Where $1_\theta = 1$ if the authority is a high type and 0 otherwise. The equilibrium level of the authority is:

$$e^*_{aA} = [g(1-\alpha)]^{\frac{1+\alpha}{2}} [\alpha(1 + 1_\theta \theta g)]^{\frac{1-\alpha}{2}} \tag{A.26}$$

$$e^*_{aB} = [1-\alpha]^{\frac{1+\alpha}{2}} [\alpha(g + 1_\theta \theta)]^{\frac{1-\alpha}{2}} \tag{A.27}$$

Note that when the we calculate the deviations to off-equilibrium actions, the beliefs of the agent may not be correct. Specifically, when the agent (mistakenly) thinks the authority is a high type, these are the optimal efforts of the selfish authority:

$$e^*_{0A} = [g(1-\alpha)]^{\frac{1+\alpha}{2}} [\alpha]^{\frac{1-\alpha}{2}} [1 + \theta g]^{\frac{\alpha(1-\alpha)}{2(2-\alpha)}} \tag{A.28}$$

$$e^*_{0B} = [1-\alpha]^{\frac{1+\alpha}{2}} [\alpha g]^{\frac{1-\alpha}{2}} [g + \theta]^{\frac{\alpha(1-\alpha)}{2(2-\alpha)}} \tag{A.29}$$

Indifference of the selfish type. Assume there is a $\bar{p} > 0$ with associated beliefs $\mu(a = \theta \mid \bar{p}) = 1$ and denote by $\bar{e}_{sn}$ the optimal effort level of the agent under such maximally optimistic beliefs (found in A.24 and A.25 when $1_\theta = 1$). Similarly, assume that $\mu(a = 0 \mid p = 0) = 1$ and denote by $\underline{e}_{sn}$ the optimal effort level of the agent under such maximally pessimistic beliefs (found in A.24 and A.25 when $1_\theta = 0$). Then, it is easy to see from the optimal effort levels derived above that $\underline{e}_{sn} < \bar{e}_{sn}$. Thus, we know that $u_0(A, \underline{e}_{sA}, e^*_{aA}) < u_0(A, \bar{e}_{sA}, e^*_{aA})$. Then, by the continuity

of U_0 in p, if $u_0\left(B, \underline{e}_{sB}, e^*_{aB}\right) \geq u_0\left(A, \bar{e}_{sA}, e^*_{aA}\right)$, there exists a $\tilde{p}$ such that the selfish type is indifferent between $\{\tilde{p}, \bar{e}_{sn}\}$ and $\{0, \underline{e}_{sn}\}$. The condition for this to be satisfied is

$$\begin{aligned}
u_0\left(B, \underline{e}_{sB}, e^*_{aB}\right) &\geq u_0\left(A, \bar{e}_{sA}, e^*_{aA}\right) \\
g\tilde{G}\left(\underline{e}^*_{sB}, e^*_{0B}\right) - \frac{1}{2}\left(e^*_{0B}\right)^2 &\geq \tilde{G}\left(\bar{e}_{sA}, e^*_{0A}\right) - \frac{1}{2}\left(e^*_{0A}\right)^2 \\
g\left[1-\frac{\alpha}{2}\right][1-\alpha]^{1-\alpha}[\alpha g]^{\alpha} &\geq \left[1-\frac{\alpha}{2}\right][g(1-\alpha)]^{1-\alpha}[\alpha]^{\alpha}[1+\theta g]^{\frac{\alpha(1-\alpha)}{2-\alpha}} \\
g^{2\alpha} &\geq [1+\theta g]^{\frac{\alpha(1-\alpha)}{2-\alpha}} \\
\theta &\leq \frac{1}{g}\left(g^{\frac{4-2\alpha}{1-\alpha}} - 1\right)
\end{aligned} \tag{A.30}$$

Furthermore, because $u_0\left(A, \bar{e}_{aA}, e^*_{aA}\right) < u_0\left(B, \bar{e}_{aB}, e^*_{aA}\right)$ and U_0 is linear in p, it follows that the selfish type always prefers $p = 0$ to any $\bar{p} \in [\tilde{p}, 1]$. (We will follow the tiebreaking rule that an indifferent authority sets $p = 0$).

Strategy of the benevolent type. If the selfish type is indifferent between $\{\tilde{p}, \bar{e}_{sn}\}$ and $\{0, \underline{e}_{sn}\}$, then by Lemma 5.1 we know that the high type prefers to set $\tilde{p}$ to $p = 0$. Therefore $\tilde{p}$ is a canditate for a separating equilibrium. What about deviations to other levels of p? Consider first deviations to $\bar{p} \in (0, \tilde{p})$. Suppose off equilibrium beliefs are such that $\mu\left(a = \theta \mid \bar{p}\right) = 0$. These beliefs do not violate the Intuitive Criterion, because we know from (A.30) that the selfish type would always mimic the benevolent type if she were to set $p \in (0, \tilde{p})$ and beliefs are such that $\mu\left(a = \theta \mid \bar{p}\right) = 1$. Then, because U_θ is linear in p, $U_\theta(\bar{p}) < U_\theta(p = 0) < U_\theta(\tilde{p})$. Thus, such deviations are not profitable.

Now consider deviations to $\bar{p} \in (\tilde{p}, 1]$. Because a low type will never set p in this area, the intuitive criterion tell us that the only reasonable off-equilibrium beliefs are $\mu = (a = \theta \mid \bar{p} \geq \tilde{p}) = 1$. Thus, deviations are profitable if:

$$\begin{aligned}
u_0\left(A, \bar{e}_{aA}, e^*_{aA}\right) &\geq u_\theta\left(B, \bar{e}_{sB}, e^*_{aB}\right) \\
G\left(\bar{e}_{sA}, e^*_{\theta A}\right) - \frac{1}{2}\left(e^*_{\theta A}\right)^2 &\geq gG\left(\bar{e}_{sB}, e^*_{\theta B}\right) - \frac{1}{2}\left(e^*_{\theta B}\right)^2 \\
\left[(1+g\theta)\left(1-\frac{\alpha}{2}\right)\right][1-\alpha]^{1-\alpha}[\alpha(1+\theta g)]^{\alpha} &\geq \left[(g+\theta)\left(1-\frac{\alpha}{2}\right)\right][g(1-\alpha)]^{1-\alpha}[\alpha(g+\theta)]^{\alpha} \\
g^{1-\alpha}(1+g\theta)^{1+\alpha} &\geq (g+\theta)^{1+\alpha} \\
\theta &\geq \frac{g^{\frac{1-\alpha}{1+\alpha}} - g}{1 - g^{\frac{2}{2-\alpha}}} = \bar{\theta}
\end{aligned} \tag{A.31}$$

Suppose first that (A.31) holds. In that case, the high type is better off under project A then under project B. Thus, she will set $p^*_\theta = 1$.

Now suppose that (A.31) does not hold. In this case the high type prefers to implement project B. Because U_θ is linear in p, the high type will never deviate to $\bar{p} \in (\tilde{p}, 1]$. It follows that there

is a separating equilibrium in which the low type sets $p = 0$ and the high type sets $p^*_\theta = \tilde{p}$, where $\tilde{p}$ is such that

$$u_0\left(B, \underline{e}_{sB}, e^*_{aB}\right) = \tilde{p} * u_0\left(A, \bar{e}_{sA}, e^*_{aA}\right) + (1-\tilde{p}) * u_0\left(B, \bar{e}_{sB}, e^*_{aB}\right) \text{ or}$$
$$\tilde{p} = \frac{u_0\left(B, \bar{e}_{sB}, e^*_{aB}\right) - u_0\left(B, \underline{e}_{sB}, e^*_{aB}\right)}{u_0\left(A, \bar{e}_{sA}, e^*_{aA}\right) - u_0\left(B, \bar{e}_{sB}, e^*_{aB}\right)}$$

Substituting in the appropriate expressions yields

$$\tilde{p} = \frac{1 - \left(1 + \frac{\theta}{g}\right)^{\frac{\alpha(1-\alpha)}{2-\alpha}}}{\frac{1}{g^2}\left(1 + g\theta\right)^{\frac{\alpha(1-\alpha)}{2-\alpha}} - \left(1 + \frac{\theta}{g}\right)^{\frac{\alpha(1-\alpha)}{2-\alpha}}} > 0. \tag{A.32}$$

It remains to show that both these high type strategies can occur in equilibrium. We have derived two constraints on the parameteres θ and g: (A.30) is a sufficient condition for the existence of an equilibrium, whereas (A.31) gives the optimal strategy of the high type. We show that there is a $\bar{g} > 1$ such that (A.30) implies (A.31) and only if $g \leq \bar{g}$. It is easy to derive that both functions are monotonic and that

$$\lim_{g \downarrow 1} \frac{g^{\frac{1-\alpha}{1+\alpha}} - g}{1 - g^{\frac{2}{2-\alpha}}} = \alpha \qquad \text{(by l'Hôpitals Rule)}$$
$$\lim_{g \to \infty} \frac{g^{\frac{1-\alpha}{1+\alpha}} - g}{1 - g^{\frac{2}{2-\alpha}}} = 0 \qquad \text{and}$$
$$\lim_{g \downarrow 1} \frac{1}{g}\left(g^{\frac{4-2\alpha}{1-\alpha}} - 1\right) = 0$$
$$\lim_{g \to \infty} \frac{1}{g}\left(g^{\frac{4-2\alpha}{1-\alpha}} - 1\right) = \infty$$

So, (A.31) crosses (A.30) only once and from above.

Uniqueness. Consider now a candidate pooling equilibrium on some level $0 \leq p_{pool} < p^*_\theta$ and some level of effort of the agent $e_{sn}(p_{pool}) = e_{sn}\left(E[a]\right)$. This would indeed be an equilibrium if it were supported by sufficiently low off-equilibrium beliefs. However, we can rule out such beliefs for deviations to high levels of p by applying the intuitive criterion.

We know that in this pooling equilibrium the low type has a higher utility than in the separating equilibrium above (or she would deviate to $p = 0$). This implies that if (A.30) holds, there exists a $\hat{p} < \tilde{p}$ such that the low type is indifferent between $\{p_{pool}, e_{sn}(E[a])\}$ and $\{\hat{p}, \bar{e}_{sn}\}$, where, as before, $\bar{e}_{sn}$ is the optimal effort level of the agent under maximally optimistic beliefs. Because the expected utility of the low type is continuously decreasing in p, deviating to any $p > \hat{p}$ is dominated for the low type, and as a consequence the intuitive criterion prescribes that $\mu(a = \theta \mid p \geq \hat{p}) = 1$. By Lemma 1 we know that if the low type is indifferent between

$\{p_{pool}, e_{sn}(E)\}$ and $\{\hat{p}, \bar{e}_{sn}\}$, the high type prefers to set $\{\hat{p}, \hat{e}_{sn}\}$, and therefore prefers to deviate. This implies that the candidate pooling equilibrium is not an equilibrium. □

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Appendix B

Stochastic difference and inequality

Given two random variables Y_1 and Y_2, $\delta(Y_1, Y_2) = \Pr(Y_2 > Y_1) - \Pr(Y_2 < Y_1)$ is called the *stochastic difference* of Y_1 verses Y_2. The stochastic difference can be estimated by computing the sample analogues. Consider first the case of matched pairs where data is given by joint observations of Y_1 and Y_2. The estimate is calculated by ignoring all pairs in which $Y_1 = Y_2$ and then taking the difference between the empirical frequency of pairs with $Y_2 > Y_1$ and of pairs in which $Y_2 < Y_1$. Now consider the case in which there are two independent samples, one associated to each variable. Here one can estimate δ by considering the frequency of $Y_2 > Y_1$ among all possible pairs and subtracting from this the frequency in which $Y_2 < Y_1$ among all these pairs. The resulting estimates are unbiased.

If $\delta(Y_1, Y_2) > 0$ then one says that Y_2 tends to yield larger outcomes than Y_1. We wish to identify significant evidence that Y_2 tends to yield larger outcomes than Y_1. So we wish to test the null hypothesis $H_0 : \delta(Y_1, Y_2) \leq 0$ against the alternative hypothesis $H_1 : \delta(Y_1, Y_2) > 0$ for a given specified level α. This is called a test of *stochastic inequality* (Cliff, 1993, Brunner and Munzel, 2000).

Assume that data has the form of *matched pairs* as given by n independent observations of (Y_1, Y_2). Then this test reduces to a sign test. One uses a binomial test to test whether the probability that $Y_2 > Y_1$ conditional on $Y_2 \neq Y_1$ is $\leq 1/2$.

Now assume instead that data is given by two independent samples of Y_1 and of Y_2. Let n_i be the number of observations of Y_i, $i = 1, 2$. We present an exact test of these hypotheses due to Schlag (2008).

Randomly match one observation of each sample to generate $\min\{n_1, n_2\}$ matched pairs. Then determine a rejection probability based on the randomized version of the sign test with size $0.2 \cdot \alpha$. The combination of the matching and the probabilistic recommendation yields an exact randomized test with size $0.2 \cdot \alpha$. We proceed as follows to derive an exact

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

nonrandomized test that has level α. Reject the null hypothesis if the rejection probability of the above randomized test is above 0.2. Note that the factor used to reduce the size of the randomized test is equal to the threshold used to translate the randomized recommendation into a deterministic recommendation.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Appendix C

Experiment instructions

I report instructions for the endogenous sanction treatment.

Originally in Italian

Instructions for the first round

Introduction

Welcome! You are going to take part in an experimental study of decision making. Please follow these instructions carefully. You will be paid according to your performance. At the end of the experiment we will tell you how much you earned.

Once everyone is seated we will formally start the experiment by reading the instructions. After this reading you will have the opportunity to ask us questions about the procedure. However at no time may you communicate with any of the other participants of your session. Please also refrain from talking to others about your experience until tomorrow in order not to influence others taking part in our experiment. Please turn off your mobiles in case they are still switched on. We hope you have fun.

Matching and assignment to a role

The computer will assign you by chance (i.e. at random) to a group consisting of three participants. You will not know the identity of the other two in your group and they will not know your identity. The computer will also assign a role to each in this group. Two of this group (from now on: player 1 and player 2) will have to take a decision as described below, the third (from now on: player 3) will be inactive but still will earn some money.

Decisions and Earnings

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

During the experiment any choice will lead to some earnings expressed in tokens. Total earnings at the end of the experiment are determined by the sum of all earnings and will then be converted into money at the exchange rate of

1 token = 7.5 Eurocents (or equivalently: 100 tokens=7,5 Euro)

It will not be possible to have negative earnings at the end.

Player 1 and Player 2

Players 1 and 2 will simultaneously each be asked to make two decisions: to choose a number and to make a guess about which number the other player chooses. Both decisions have to be entered into a decision screen that is described in more detail below. Neither player will observe the decisions of the other player.

Choosing a Number

Both player 1 and player 2 have to choose a number. This number can be any number between and including 110 and 170 (fractions or decimals not allowed).

The earnings in tokens of either player 1 or player 2 from choosing a number are determined as follows. A player receives the lower of the two numbers chosen by player 1 and player 2 minus 85% of their own number.

This has the following implications:

- Assume players 1 and 2 chose the same number. Then a player will receive his/her own number (since both numbers are equal, this is also the lowest number) minus 85% of his/her own number.

- Assume that players 1 and 2 chose different numbers. Then, the player who chose the lower number could have increased his earnings by choosing a slightly higher number. However, the player who chose the higher number could have increased his earnings by choosing a slightly lower number.

The following mathematical representation will not be read out loud.

Suppose (among players 1 and 2) that one of them chooses the number Y and the other chooses the number Z.

If $Y = Z$ then the player who chose Y receives $Y - 0.85\times$ Y.

If $Y < Z$ then the player who chose Y receives $Y - 0.85\times$ Y.

If $Y > Z$ then player who chose Y receives $Z - 0.85\times$ Y.

In addition, players 1 and 2 first receive a fixed amount of 35 tokens.

Guessing the other's choice

In addition to specifying a number, both player 1 and player 2 are asked to make a guess about the number chosen by the other player. The guess is made by specifying a range (given by its lower bound L and its upper bound U) in which the other player's choice is believed to belong.

The earnings in tokens of either player 1 or player 2 from making this guess are determined as follows. A wrong guess (the actual number chosen by the other player falls outside the specified range) yields nothing. A correct guess (the actual number chosen by the other player lies within the specified range) yields 15% of the difference between 60 and the width of the range U-L. Therefore the smaller the specified range, the higher the earnings if the guess is correct. However, a smaller range also increases the risk that the guess is not correct, in which case no tokens are earned.

The following mathematical representation will not be read out loud:

If the number Z chosen by the other player lies in the range (it is greater than or equal to L and less than or equal to U) then the player who has chosen L and U gets $0.15 \times (60 - (U - L))$ tokens if this number Z does not lie within the range then the player who has chosen L and U gets nothing.

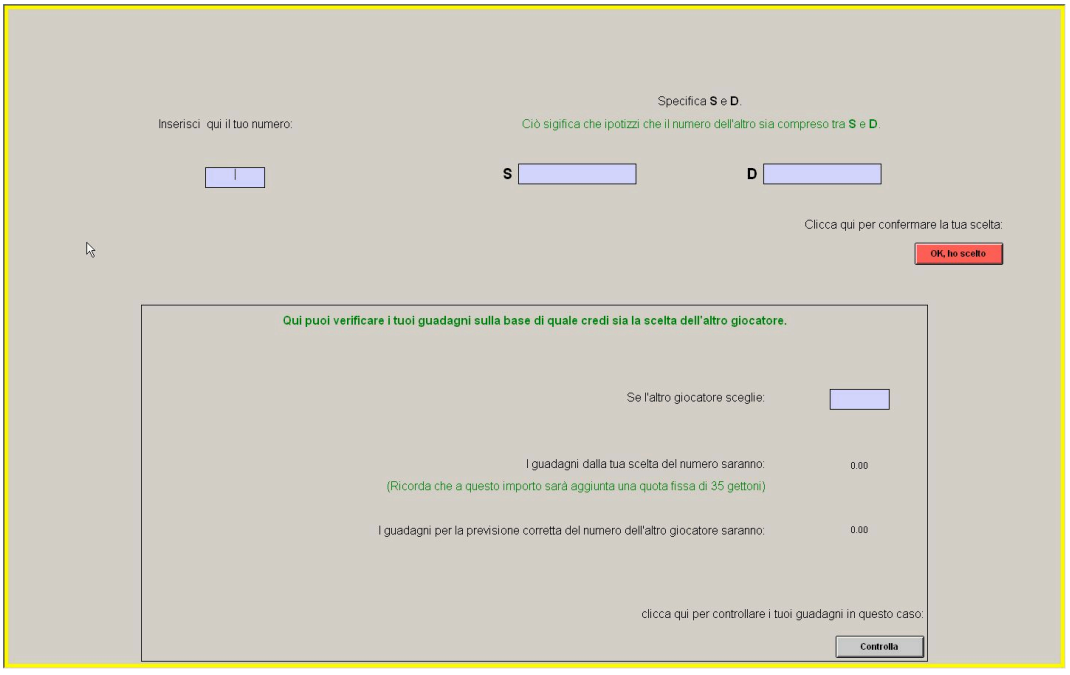

FIGURE C.1: Input screen in the first round.

Player 3

Player 3 does not make any decision during the experiment and earns an amount of tokens equal to 25% of the smaller of the two numbers chosen by players 1 and 2.

A more mathematical representation of this statement will not be read out loud:

Tokens earned by player three = 0.25× (smaller of the two numbers chosen by player 1 and player 2)

Tutorial

Before the experiment starts, so before roles are assigned, all participants have the possibility to practice and to get used to the structure of the game. To this end, you will participate in a tutorial round, where you will see the decision screen as described above. You will have 5 minutes to enter as many different values as you like for both your own number and your guess, and the other player's hypothetical number. You can then use the check button to see what your earnings from these numbers and your guess would be. You are encouraged to verify the calculation behind the earnings of both the number choice and the guess. The values entered in this tutorial have no influence on your earnings and will not be recorded. After 5 minutes the tutorial will stop and the experiment will start.

Final Remarks

During the experiment, you are not permitted to speak or communicate with the other participants. If you have a question while the experiment is going on, please raise your hand and one of the experimenters will come and answer it.

At this time, do you have any questions about the instructions or procedures? If you have a question, please raise your hands and one of the experimenters will come to your seat to answer it.

Instructions for the second round

Introduction

Now we run a second and final experiment. Earnings will be added to your previous earnings. After this new experiment everything is over and your total payment will be calculated.

This new experiment is very similar to the previous one up to some changes we highlight.

Matching and roles

All participants are matched with the same people as before and keep the roles they had before.

Decisions and Earnings

IN CONTRAST to the previous experiment, player 3 now also makes a decision.

Player 3

At the start of the experiment, before player 1 and 2 make any decisions, player 3 observes the numbers chosen by players 1 and 2 in the previous experiment. After having observed these numbers, player 3 makes a decision that determines how earnings of players 1 and 2 are calculated in this new experiment. The outcome of this decision is observed by players 1 and 2 before they make their choices. Player 3 has the following two choices:

a) NOT CHANGE: To choose "not change" means that the earnings of all players are as in the previous experiment. In particular, player 3 earns 25% of the smaller of the two numbers chosen by players 1 and 2.

b) CHANGE: To choose "change" means that earnings in tokens of all players are changed as follows. Players 1 and 2 receive the lower of the two numbers chosen minus 85% of their own number minus 50% of the difference between 170 and the player's own chosen number. . That is, relative to the previous experiment, there is an extra amount subtracted to your earnings that is larger the smaller your number is. Player 3 earns 25% of the smaller of the two numbers chosen by players 1 and 2 minus 4. The terms that are new as compared to the previous experiment have been underlined.

Mathematical illustration not to be read out loud:

Suppose player 3 chooses "change" and (among players 1 and 2) that one of them chooses the number Y and the other chooses the number Z.

If $Y = Z$ then the player who chose Yreceives $Y - 0.85 \times Y - 0.5 \times (170 - Y)$.

and player 3 receives $0.25 \times Y - 4$.

If $Y < Z$ then the player who chose Y receives $Y - 0.85 \times Y - 0.5 \times (170 - Y)$.

and player 3 receives $0.25 \times Y - 4$.

If $Y > Z$ then player who chose Y gets $Z - 0.85 \times Y - 0.5 \times (170 - Y)$

and player 3 receives $0.25 \times Z - 4$.

Regardless of the choice of player 3, player 1 and 2 also receive a fixed amount of 35 tokens.

Player 1 and Player 2

As in the previous experiment, players 1 and 2 make two decisions: choose a number and make a guess by specifying a range. Earnings from making the guess are as in the previous experiment, earnings from choosing a number are specified above.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Final Remarks

If you have any questions then please ask them now.

Please do not log off the computer when the experiment is over.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Bibliography

Aghion, Philippe, Mathias Dewatripont, and Patrick Rey. 2002. "On Partial Contracting." *European Economic Review*, 46, pp. 745-53.

Aghion, Philippe, Mathias Dewatripont, and Patrick Rey. 2004. "Transferable Control." *Journal of the European Economic Association*, 1:2, pp. 115-38.

Aghion, Philippe and Jean Tirole. 1997. "Formal and Real Authority in Organizations." *Journal of Political Economy*, 105:1, pp. 1-29.

Alm, James. 1998. "Tax Compliance and Administration," in *Handbook on Taxation*, Hildreth, B. W. and J.A. Richardson eds. CRC Press.

Alm, James, Betty Jackson, and Michael McKee. 1993. "Fiscal exchange, collective decision institutions, and tax compliance." *Journal of Economic Behavior and Orgnaization*, 22, pp. 285-303. Arbor.

Ariely, Dan, Anat Bracha, and Stephen Meier. 2009. "Doing well or doing good? Image motivation and monetary incentives in behaving procsocially," *American Economic Review*, Forthcoming.

Ayres, Ian and Steven D. Levitt. 1998. "Measuring Positive Externalities from Unobservable Victim Precaution: An Empirical Analysis of Lojack," *Quarterly Journal of Economics*, 113:1, pp. 43-77.

Allingham, Michael and Agnar Sandmo. 1972. "Income Tax Evasion: A Theoretical Analysis," *Journal of Public Economics*, 1, pp. 323-38.

Andreoni, James, Brian Errard, and Jonathan Feinstein. 1998. "Tax Compliance," *Journal of Economic Literature*, XXXVI, pp. 818-60.

Bacharach, Michael, Gerardo Guerra, and Daniel J. Zizzo. 2007. "The Self-Fulfilling Property of Trust: An Experimental Study," *Theory and Decision*, 63:4, pp. 349-88.

Bar-Gill, Oren and Chaim Fershtman. 2005. "Public Policy with Endogenous Preferences," *Journal of Public Economic Theory*, 7:5, pp. 841-57.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Beccaria, Cesare. 1992(1770). *An Essay on Crimes and Punishments.* Brookline Village: International Pocket Library.

Becker, Gary S. 1968. "Crime and Punishment: An Economic Approach," *Journal of Political Economy*, 76:2, pp. 169-217.

Bénabou, Roland and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation," *Review of Economic Studies*, 70, pp. 489-520.

Bénabou, Roland and Jean Tirole. 2005. "Incentives and Prosocial Behavior," *American Economic Review*, 96:5, pp. 1652-78.

Bénabou, Roland and Jean Tirole. 2007. "Identity, Dignity and Taboos: Beliefs as Assets," IZA discussion paper 2583.

Black, Dan A. and Daniel S. Nagin. 1998. "Do Right-to-Carry Laws Deter Violent Crime?" *Journal of Legal Studies*, 27, pp. 209-19.

Blumstein, Alfred and Allan J. Beck. 1999. "Population Growth in U.S. Prisons, 1980-1996, in *Prisons: Crime and Justice- A Review of Research. Volume 26.* Tonry, M. and J. Petersilia eds. Chicago: University of Chicago Press, pp. 17-61.

Bohnet, Iris and Robert D. Cooter. 2003. "Expressive Law: Framing or Equilibrium Selection?" KSG Working Paper 03-046.

Bohnet, Iris, Bruno S. Frey, and Steffen Huck. 2001. "More Order with Less Law: On Contract Enforcement, Trust, and Crowding," *American Political Science Review*, 95:1, pp. 131-44.

Bolton, Gary E. and Axel Ockenfels. 2000. "ERC: a theory of equity, reciprocity and competition," *American Economic Review* 90, pp. 166-93.

Bowles, Samuel. 1998. "Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions," *Journal of Economic Literature*, XXXVI, pp. 75-111.

Bowles, Samuel. 2008. "Policies designed for self-interested citizens may undermine the moral sentiments: evidence from economic experiments," *Science*, p. 320.

Bowles, Samuel and Sung-Ha Hwang. 2008. "Social preferences and public economics: mechanism design when social preferences depend on incentives," *Journal of Public Economics*, 92:8-9, 1811-20.

Brandts, Jordi and David J. Cooper. 2006. "It's What You Say Not What You Pay: An Experimental Study of Manager Employee Relationships in Overcoming Coordination Failure," *Journal of the European Economic Association*, 5:6, pp. 1223-68.

Brandts, Jordi and Cooper, David. 2006. "A Change Would Do You Good: An Experimental Study on How to Overcome Coordination Failure in Organizations", *American Economic Review*, 96:3, pp. 669-93.

Brennan, Geoffrey and Philip Pettit. 2004. *The Economy of Esteem. An Essay on Civil and Political Society.* New York: Oxford University Press.

Brunner, Edgar and Ullrich Munzel. 2000. "The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation," *Biometrical Journal*, 42, pp. 17-25.

Cameron, Samuel. 1988. "The Economics of Crime Deterrence: A Survey of Theory and Evidence," *Kyklos*, 41:2, pp. 301-23.

Cardenas, Juan Camilo, John K. Stranlund, and Cleve E. Willis. 2000. "Local Environmental Control and Institutional Crowding-out," *World Development*, 28:10, pp. 1719-33.

Carmichael, H. Lorne and Bentley W. MacLeod. 1997. "Gift Giving and the Evolution of Cooperation." *International Economic Review*, 38:3, pp. 485-509.

Chaudhuri, Ananish, and Laura Bangun. 2007. "Credible Assignments in the Minimum Effort Coordination Game," Mimeo.

Cho, In-Koo and David M. Kreps. 2007. "Signaling games and stable equilibria," *Quarterly Journal of Economics*, 102:2, pp. 179-221.

Cliff, Normann. 1993. "Dominance Statistics: Ordinal Analyses to Answer Ordinal Questions" *Psychological Bulletin* 114, pp. 494-509.

Cohen-Charash, Yochi and Paul E. Spector. 2001. "The role of justice in organizations: a meta-analysis." *Organizational Behavior and Human Decision Processes*, 86, pp. 278-321.

Coleman, Stephen. 1997. "Income tax compliance: A unique experiment in Minnesota," *Government Finance Review*, 13, pp. 11-15.

Corman, Hope and Naci H. Mocan. 2000. "A Time-Series Analysis of Crime, Deterrence, and Drug Abuse in New York City," *American Economic Review*, 90:3, pp. 584-604.

Cornwell, Cristopher and William N. Trumbull. 2000. "Estimating the Economic Model of Crime with Panel Data," in *Readings in Urban Economics: Issues and Public Policy*, Robert Wassmer (ed.). Blackwell Publishing. pp. 382-92.

Cooter, Robert 1998. "Expressive Law and Economics," *Journal of Legal Studies*, 27, pp. 585-607.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Dawisha, Adeed. 2009. *Iraq: A Political History from Independence to Occupation.* New Jersey: Princeton University Press.

Deci, Edward L. 1975. *Intrinsic Motivation.* New York: Plenum Press.

Dessein, Wouter. 2005. "Information and Control in Ventures and Alliances." *Journal of Finance*, 60:5, pp. 2513-50.

Devetag, Giovanna and Ortmann, Andreas. 2007. "When and Why? A Critical Survey on Coordination Failure in the Laboratory," *Experimental Economics*, 10:30, 331-344.

Dezhbakhsh, Hashem and Paul H. Rubin. 1998. "Lives Saved or Lives Lost? The Effects of Concealed-Handgun Laws on Crime." *American Economic Review, Papers and Proceedings*, 88:2, pp. 468-74.

Dills, Angela K., Jeffrey A. Miron, and Garrett Summers. 2008. "What Do Economists Know About Crime," NBER Working Paper 13759. Cambridge.

Donohue, J.J. and Wolfers, J. 2006. "Uses and Abuses of Empirical Evidence in the Death Penalty Debate," *Stanford Law Review*, 58, pp. 791-846.

Drago, Francesco, Roberto Galbiati and Pietro Vertova. forthcoming. "The Deterrent Effects of Prison: Evidence from a Natural Experiment," *Journal of Political Economy.*

Dubin Jeffrey A. and Louis L. Wilde. 1988. "An Empirical Analysis of Federal Income Tax Auditing and Compliance", *National Tax Journal*, 41:1, 61-74.

Dur, Robert. 2006. "Status-Seeking in Criminal Subcultures and the Double Dividend of Zero-Tolerance," CES-Ifo Working Paper 1762.

Eide, Erling. 2000. "Economics of Criminal Behavior," in *Encyclopedia of Law and Economics*, Boudewijn Bouckaert and Gerrit de Geest eds. Cheltenham: Edward Elgar, pp. 345-89.

Ellingsen, Tore and Magnus Johannesson. 2007. "Paying Respect," *Journal of Economic Perspectives*, 21:4, pp. 153-49.

Ellingsen, Tore and Magnus Johannesson. 2008. "Pride and Prejudice: The Human Side of Incentive Theory," *American Economic Review*, 98:3, pp. 990-1008.

Fajnzylber, Pablo, Daniel Lederman and Norman Loayza. 2002. "What causes violent crime?" *European Economic Journal*, 46, pp. 1323-57.

Falk, Armin and Michael Kosfeld. 2006. "The Hidden Cost of Control," *American Economic Review*, 96, pp. 1611-30.

Fehr, Ernst and Armin Falk. 2002. "Psychological foundations of incentives." *European Economic Review*, 46, pp. 687-724.

Fehr, Ernst and Simon Gächter. 2000. "Fairness and Retaliation: The Economics of Reciprocity," *Journal of Economic Perspectives*, 14:3, pp. 159-81.

Fehr, Ernst and Klaus Schmidt. 1999. "A theory of fairness, competition and cooperation," *Quarterly Journal of Economics*, 114(3), pp. 817-68.

Feld, Lars P. and Bruno S. Frey. 2001. "Trust breeds trust: how taxpayers are treated." *Economics of Governance*, 3, pp. 87-99.

Feldman, Yuval and Tom R. Tyler. 2008. "Mandated justice: The potential promise and pitfalls of mandating procedural justice." 3rd Annual COnference on Empirical Legal Studies Papers.

Fischbacher, Urs and Simon Gächter. 2006. "Heterogeneous Social Preferences and the Dynamics of Free Riding in Public Goods," IZA Working Paper 2011.

Fischbacher, Urs. 2007. "z-Tree: Zurich Toolbox for Ready-made Economic experiments," *Experimental Economics*, 10:2, 171-78.

Forsythe, Robert, Horowitz, Joel L., N. E. Savin, and Martin Sefton. 1994. "Fairness in Simple Bargaining Experiments," *Games Economic Behavior*, 6, pp. 347-69.

Franzoni, Luigi G. 1999. "Tax Evasion and Tax Compliance," in *Encyclopaedia of Law and Economics*, B. Bouckaert and G. de Geest eds. Cheltenham: Edward Elgar.

Frey, Bruno S. 1997a. "A Constitution for Knaves Crowds out Civic Virtues," *The Economic Journal*, 107, pp. 1043-53.

Frey, Bruno S. 1997b. *Not Just for The Money. An Economic Theory of Personal Motivation.* Cheltenham, UK: Edward Elgar.

Frey, Bruno S., Matthias Benz, and Alois Stutzer. 2004. "Introducing procedural utility: not only what, but also how matters." *Journal of Institutional and Theoretical Economics*, 160, 377402.

Frey, Bruno S. and Reto Jegen. 2001. "Motivation Crowding Theory," *Journal of Economic Surveys*, 15:5, pp. 589-611.

Frey, Bruno S. and Stephan Meier. 2004. "Social Comparisons and Pro-social Behavior: Testing 'Conditional Cooperation' in a Field Experiment," *American Economic Review*, 94:5, pp. 1717-22.

Frey, Bruno S. and Felix Oberholzer-Gee, 1997. "The cost of price incentives: an empirical analysis of motivation crowding-out," *American Economic Review*, 87:3, pp. 746–55.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

Frey, Bruno S. and Benno Torgler. 2007. "Tax Morale and Conditional Cooperation," *Journal of Comparative Economics*, 35: 136-59.

Frohlich, Norman and Joe A. Oppenheimer. 1995. "The Incompatibility of Incentive Compatible Devices and Ethical Behavior: Some Experimental Results and Insights." *Public Choice Studies* 25, pp. 24-51.

Garland, David. 2001. *The Culture of Control. Crime and Social Order in Contemporary Society.* Oxford: Oxford University Press.

Gächter, Simon. 2006. "Conditional cooperation: Behavioral regularities from the lab and the field and their policy implications," CeDEx Discussion Paper, 2006-03.

Gächter, Simon, Esther Kessler, and Manfred Königstein. 2007. "Performance Incentives and the Dynamics of Voluntary Cooperation," mimeo.

Gächter, Simon and Elke Renner. 2006. "The Effects of (Incentivized) Belief Elicitation in Public Good Experiments," CeDex discussion paper 2006-16.

Garoupa, Nuno. 1997. "The Theory of Optimal Law Enforcement," *Journal of Economic Surveys*, 11:3, pp. 267-95.

Glaeser, Edward L. and Bruce Sacerdote. 1999. "Why Is There More Crime in Cities?" *Journal of Political Economy*, 107:6, pp. 225-59.

Glaeser, Edward L., Bruce Sacerdote, and Jose A. Scheinkman. 1996. "Crime and Social Interactions," *Quarterly Journal of Economics*, 111:2, pp. 507-48.

Gneezy, Uri and Aldo Rustichini. 2000. "A Fine is a Price," *Journal of Legal Studies*, 29, pp. 1-17.

Goeree, Jacob and Charles A. Holt. 2005. "An Experimental Study of Costly Coordination," *Games and Economic Behavior*, 51, pp. 349-64

Goeree, Jacob and Charles A. Holt. (2001). "Then Little Treasures of Game Theory and Ten Intuitive Contradictions," *American Economic Review*, 91:5, pp. 1402-22.

Gonzalez-Navarro, Marco. 2008. "Deterrence and Displacement of Auto Theft," CEPS Working Paper 177.

Guillen, Pablo, Schwieren, Christiane and Gianandrea Staffiero. 2006. "Why feed the Leviathan?" *Public Choice*, 130, pp. 115-28.

Güth, Werner and Axel Ockenfels. 2005. "The coevolution of morality and legal institutions: and indirect evolutionary approach," *Journal of Institutional Economics*, 1:2, pp. 155-74.

Habermas, Jürgen. 1990. *Moral Consciousness and Communicative Action.* Cambridge: MIT Press.

Hardin, Russell. 1991. "Trusting Persons, Trusting Institutions," in *The Strategy of Choice.* Richard J. Zeckhauser ed. Cambridge, MA: MIT Press, pp. 185-209.

Helland, Eric, and Alexander Tabarrok . 2007. "Does Three Strikes Deter?: A Nonparametric Estimation," *Journal of Human Resources*, XLII, pp309-330.

Hörish, Hannah and Christina Strassmair. 2008. "An experimental test of the deterrence hypothesis." University of Munich Working Paper 2008-4.

Huck, Steffen. 1998. "Trust, Treason, and Trials: An Example of How the Evolution of Preferences Can Be Driven by Legal Institutions," *Journal of Law, Economics and Organisation*, 14:1, pp. 44-60.

Kahan, Dan M. 1997. "Social Influence, Social Meaning, and Deterrence," *Virginia Law Review*, 83:2, pp. 349-95.

Kahan, Dan M. 2005. "The Logic of Reciprocity: Trust, Collective Action, and Law," in *Moral Sentiments and Material Interests.* Herbert Gintis, Samuel Bowles, Robert Boyd and Ernst Fehr eds. Cambridge, Massachusetts: MIT Press, pp. 339-78.

Kessler, Daniel P. and Steven D. Levitt. 1997. "Using Sentence Enhancements to Dinstinguish between Deterrence and Incapacitation," *Journal of Law and Economics*, 17:1, pp. 343-63.

Lappi-Seppälä, Tapio. 2000. "The Fall of the Finnish Prison Population." *Journal of Scandinavian Studies in Criminology and Crime Prevention*, 1, pp. 27-40.

Lappi-Seppälä, Tapio. 2001. "Sentencing and punishment in Finland: the decline of the repressive ideal", in *Punishment and penal systems in western countries.* Tonry, M. and R. Frase eds. New York: Oxford University Press.

Lee, David S. and Justin McCrary. 2005. "Crime, Punishment, and Myopia, NBER working paper 11491.

Levitt, S. D. 1997. "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime," *American Economic Review*, 87:3, pp. 270-90.

Levitt, S. D. 2002. "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime: Reply," *American Economic Review*, 92, pp. 1244-50.

Levitt, Steven D. and Thomas J. Miles. 2007. "Empirical Study of Criminal Punishment," in *The Handbook of Law and Economics*, A. M. Polinsky and S. Shavell eds. North Holland, pp. 455-95.

Lewis, Norman. 1964. *The Honored Society: the Sicilian Mafia Observed.* Collins.

Lind, Allan E., Ruth Kanfer and Christopher P. Earley. 1990. "Voice control and procedural justice: Instrumental and noninstrumental concerns in fairness judgements." *Journal of Personality and Social Psychology*, 59, pp. 952-59.

Lind, Allan E. and Tom R. Tyler. 1988. *The Social Psychology of Procedural Justice.* New York and London, Plenum Press.

Lind, Allan E. and Kees van den Bos. 2002. "When fairness works: Toward a general theory of uncertainty management." *Research in Organizational Behavior*, 24, pp. 181-223.

Lott, John R. and David B. Mustard. 1997. "Crime, Deterrence, and Right-to-Carry Concealed Handguns," *Journal of Legal Studies*, 26, pp. 1-68.

Mansour, Abdala, Nicolas Marceau, and Steeve Mongrain. 2006. "Gangs and Crime Deterrence," *Journal of Law, Economics and Organisation*, 22:2, pp. 315-39.

McAdams, Richard H. 1997. "The Origin, Development, and Regulation of Norms," *Michigan Law Review*, 96, pp. 338-433.

McAdams, Richard H. 2000a. "A Focal Point Theory of Expressive Law," *Virginia Law Review*, 86:1649.

McAdams, Richard H. 2000b. "Signaling Discount Rates: Law, Norms, and Economic Methodology," Yale Law Review, 110:625-689.

McAdams, Richard H. and Eric B. Rasmusen. 2007. "Norms in Law and Economics," in *The Handbook of Law and Economics*, A. M. Polinsky and S. Shavell eds. North Holland, pp. 1573-628.

McAdams, Richard H. and Janice Nadler. 2005. "Testing the Focal Point Theory of Legal Compliance: Expressive Influence in an Experimental Hawk/Dove Game," *Journal of Empirical Legal Studies*, 2, pp. 87-123.

McCarthy, Bill. 2002. "New Economics of Sociological Criminology," *Annual Review of Sociology*, 28, pp. 417-42.

McCrary, Justin. 2002. "Using Electoral Cycles in Police Hiring to Estimate the Efffect of Police on Crime: Comment," *American Economic Review*, 92:4, pp. 1236-43.

Moody, Carlisle and Thomas B. Marvell. 1996. "Police Levels, Crime Rates, and Specification Problems, *Criminology*, 24, 606-646.

Newark, Tim. 2007. *Mafia Allies. The True Story of Americas Secret Alliance with the Mob in World War II.* Saint Paul (MN): Zenith Press.

van der Weele, Joël (2009), On Sanctions and Signals: How Formal and Informal Mechanisms Produce Compliance
European University Institute DOI: 10.2870/12176

OECD. 2007. Factbook 2007. Paris: OECD.

Ostman, Axel. 1998. "External control may destroy the commons," *Rationality and Society* 10:1, pp. 103–22.

Pettit, Philip. 1995. "The Cunning of Trust," *Philosophy and Public Affairs*, 24:3, pp. 202-25.

Pew Center on the States. 2008. "One in 100: Behind Bars in the US 2008". Washington. Available from http://www.pewcenteronthestates.org/report_detail.aspx?id=35904.

Pommerehne, Werner W., and Hannelore Weck-Hannemann. 1996. "Tax rates, tax administration and income tax evasion in Switzerland". *Public Choice*, 88, pp. 161-70.

Polinsky, Mitchell A. and Steven Shavell. 2000. "The Economic Theory of Public Enforcement of Law," *Journal of Economic Literature*, 38:1, pp. 45-76.

Polinsky, Mitchell A. and Steven Shavell. 2007. "The Theory of Public Enforcement of Law," in *Handbook of Law and Economics*. Mitchell A. Polinsky and Steven Shavell eds. Elsevier, pp. 403-54.

Nussbaum, Martha C. 1997. "Flawed Foundations: The Philosophical Critique of (a Particular Type of) Economics." *University of Chicago Law Review*, 64:4, pp. 1197-214.

Posner, Eric A. 2000. *Law and Social Norms*. Harvard University Press: Cambridge.

Robinson, P.H., and J.M. Darley. 1997. "The utility of desert," *Northwestern University Law Review*, 91, pp. 453-499.

Robinson, P.H., and J.M. Darley. 2004. "Does Criminal Law Deter? A Behavioral Science Investigation," *Oxford Journal of Legal Studies*, 24:2, pp. 173-205.

Schlag, Karl H. 2008. "A New Method for Constructing Exact Tests without Making any Assumptions," Universitat Pompeu Fabra Working Paper 1109.

Schlag, Karl H. and Joël J. van der Weele. 2009. "An interval scoring rule," mimeo, European University Institute. Available from http://www.eui.eu/Personal/Researchers/joelvdweele/Work/Work.html.

Scholz, John T. 1998. "Trust, Taxes, and Compliance," in *Trust and Governance*. Valerie Braithwaite and Margaret Levi (eds.) New York: Russell Sage Foundation, pp. 135-65.

Sheffrin, Steven M. and Robert K. Triest. 1992. "Can Brute Deterrence Backfire? Perceptions and Attitutes in Taxpayer Compliance", in *Why People Pay Taxes*. Joel Slemrod (ed.) Ann Arbor: The University of Michigan Press.

Shinada, Mizuho and Toshio Yamagishi. 2007. "Punishing free riders: direct and indirect promotion of cooperation," *Evolution and Human Behavior*, 28, pp. 330-39.

Silverman, Dan. 2004. "Street Crime and Street Culture," *International Economic Review*, 45:3, pp. 761-86.

Sliwka, Dirk. 2007. "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes," *American Economic Review*. 97:3, pp. 999-1012.

Smith, Kent W. 1992."Reciprocity and Fairness: Positive Incentives for Tax Compliance", in *Why People Pay Taxes. Tax Compliance and Enforcement*. Joel Slemrod ed. Ann Arbor: University of Michigan Press.

Souvorov, A. 2003. "Addiction to Rewards," Mimeo GREMAQ, Toulouse.

Sunstein, Cass R. 1996. "On the Expressive Function of Law," *University of Pennsylvania Law Review*, 144, p. 2021.

Thibaut, John, Nehmia Friedland, and Laurens Walker. 1974. "Compliance with rules: Some social determinants." *Journal of Personality and Social Psychology*, 30, pp. 792-801.

Thibaut, John and Laurens Walker. 1975. *Procedural Justice : A Psychological Analyis*. New Jersey: Hillsdale.

Topalli, Volkan. 2005. "When Being Good is Bad: An Expansion of Neutralization Theory," *Criminology*, 43:3, pp. 797-835.

Tyler, Tom R. 1987. "Conditions leading to value expression effects in judgements of procedural justice: A test of four models." *Journal of Personality and Social Psychology*, 52, pp. 333-342.

Tyler, Tom R. 1990. *Why People Obey the Law*. New Haven and London: Yale University Press.

Tyler, Tom R. 2004. "Procedural Justice," in *The Blackwell Companion to Law and Society*. Austin Sarat (ed.) New York: Wiley Blackwell.

Tyler, Tom R. 2008. "Psychology and Institutional Design." *B.E. Law and Economics Review*, 4:3, pp. 801-87.

Tyler, Tom R. and Peter DeGoey. 1996. "Trust in organizational authorities: The influence of motive attributions on willingness to accept decisions", in *Trust in organizations: Frontiers of theory and research*. Roderick M. Kramer and Tom R. Tyler eds. Thousand Oaks, CA: Sage.

Tyler, Tom R. and Allan E. Lind. 1992. "A relational model of authority in groups", in *Advances in Experimental Social Psychology (Vol. 25)* M. Zanna (ed.). New York: Academic.

Tyran, Jean-Robert and Lars P. Feld, 2006. "Achieving Compliance when Legal Sanctions are Non-deterrent," *Scandinavian Journal of Economics*, 108:1, pp. 135-56.

Van den Bos, Kees, Henk A. M. Wilke and Allen E. Lind. 1998. "When do we need procedural fairness? The role of trust in authority". *Journal of Personality and Social Psychology*, 75, 1449-1458.

Van den Bos, K., Els C.M. van Schie and Suzanne E. Colenberg, 2000. "Parents Reactions to Child Daycare Organizations: The Influence of Perception of Procedures and the Role of Organizations Trustworthiness". *Social Justice Research*, 15:1, 53-63.

Van Huyck, John, Ann B. Gillette and Raymond Battalio. 1992. "Credible Assignments in Coordination Games", *Games and Economic Behavior*, 4, pp. 606-26.

Wenzel, Michael. 2004. "The Social Side of Sanctions: Personal and SocialNorms as Moderators of Deterrence," *Law and Human Behavior*, 28:5, pp. 547-67.