Fehr, Ernst; Williams, Tony

Working Paper

Social Norms, Endogenous Sorting and the Culture of Cooperation

IZA Discussion Papers, No. 11457

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Fehr, Ernst; Williams, Tony (2018) : Social Norms, Endogenous Sorting and the Culture of Cooperation, IZA Discussion Papers, No. 11457, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/180475

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11457

Social Norms, Endogenous Sorting and the Culture of Cooperation

Ernst Fehr
Tony Williams

APRIL 2018

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11457

Social Norms, Endogenous Sorting and the Culture of Cooperation

Ernst Fehr
University of Zurich and IZA

Tony Williams
TSG Interactive Services Limited

APRIL 2018

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Social Norms, Endogenous Sorting and the Culture of Cooperation*

Throughout human history, informal sanctions played a key role in the enforcement of social norms and the provision of public goods. However, a considerable body of evidence suggests that informal peer sanctions often cause large efficiency costs. This raises the question whether alternative (peer) sanctioning systems exist that avoid these costs and will be preferred by the people. Here, we show that welfare-enhancing peer sanctioning without much need for costly punishment emerges quickly if we introduce two relevant features of social life into the experiment: (i) subjects can migrate across groups with different sanctioning institutions and (ii) they have the chance to achieve consensus about normatively appropriate behavior. The exogenous removal of the norm consensus opportunity reduces the efficiency of peer punishment and renders centralized sanctioning by an elected judge the dominant institution. However, if given the choice, subjects universally reject peer sanctioning without a norm consensus opportunity – an institution that has hitherto dominated research in this field – in favor of peer sanctioning with a norm consensus opportunity or an equally efficient institution with centralized punishment by an elected judge. Migration opportunities and normative consensus building are key to the quick emergence of an efficient culture of universal cooperation because the more prosocial subjects populate the two efficient institutions first, elect prosocial judges (if institutionally possible), and immediately establish a social norm of high cooperation. This norm appears to guide subjects' cooperation and punishment choices, including the virtually complete removal of antisocial punishment when judges make the sanctioning decision.

JEL Classification: D02, D03, D72, H41

Keywords: cooperation, punishment, endogenous institutions, public goods

Corresponding author:
Ernst Fehr
Department of Economics and Laboratory for Social and Neural Systems Research
University of Zurich
Blumlisalpstrasse 10
CH-8006 Zurich
Switzerland

E-mail: ernst.fehr@econ.uzh.ch

* We thank Björn Bartling, Sam Bowles, Rob Boyd, Donja Darai, Joe Henrich, Holger Herz, Louis Putterman, Laura Gee, Pete Richerson, Frederic Schneider, Christian Thöni and Roberto Weber for useful comments on earlier drafts of the paper. We also thank participants at the 2013 Economic Science Association World Meeting, the 2015 Conference on Social Dilemmas, the 2016 North American Meeting, and participants at seminars in Antwerp, Lausanne, and Zurich for useful comments. Williams especially thanks Dan Burghart for helpful discussions and Tim Salmon for comments, discussions, and a great deal of encouragement. Williams gratefully acknowledges funding from the Swiss National Science Foundation under SNF Doctoral Program (ProDoc) Grant PDFMP1-123113/1. The views expressed herein are those of the authors and do not necessarily reflect the views of TSG Interactive Services Limited.

All known human societies – from simple hunter-gatherers societies to modern nation states – enforce rules and norms through sanctions (Boehm, 2001; Boehm et al., 1993; Knauft et al., 1991; Sober & Wilson, 1998; Wiessner, 2005). Throughout human evolution, peer sanctions, in particular, have played a role in the provision of important public goods that ranged from the establishment of social insurance through food sharing among hunter-gatherers (Kaplan & Gurven, 2005) and cooperation during warfare between neighboring groups in pre-state societies (Mathew & Boyd, 2011) to the organization of collective action in industrial conflicts in capitalist societies (Francis, 1985; Roethlisberger & Dickson, 1939), from the provision of effort under team production in capitalist firms (Nalbantian & Schotter, 1997; Rehder, 1990)[1] to the voluntary recruitment of millions of soldiers for national defense in World War I (Simkins, 1988)[2].

However, circumstantial evidence and a sizeable experimental literature on the effects of decentralized sanctions in public goods indicates that peer sanctioning causes high collateral damage for extended periods of time. These costs frequently even exceed or are as large as the gains from cooperation. For example, in the 10-round experiment of Gächter, Renner, & Sefton (2008) the cost of punishment was considerably higher than the gains from cooperation; subjects would have thus been better off in each of the 10 periods in an environment that ruled out the sanctioning of free-riders.[3] Peer sanctions are also associated with antisocial (Hermann, Thöni, & Gächter, 2008) and perverse punishment (Ertan, Page, and Putterman 2009) which further questions their suitability.[4]

In view of the high cost of peer-sanctioning, the question arises whether (i) there are peer-sanctioning systems that avoid these costs and (ii) whether more centralized punishment institutions are superior to peer sanctioning. In particular, if an institution based on peer sanctions is very costly and inefficient, it seems likely that people would try to leave these groups if possible. They may have an incentive to join groups with potentially more efficient sanctioning institutions that, e. g., allocate the

[1] The repertoire of peer sanctions is very broad. It ranges from a "raised eyebrow" in case of inappropriate behavior to ridicule, from physical punishment and harassment to ostracism and social exclusion. For example, those who failed to share food with other hunter-gatherers without a good reason were ostracized and, in extreme cases, excluded from the sharing network (Kaplan & Gurven, 2005). Deserters and cowards in raids against neighboring groups were chastised, tied to a tree and beaten by age-mates (Mathew & Boyd, 2011). Rate busters in capitalist firms who undermined the going piece rates by working too hard faced harassments from their co-workers (Roethlisberger & Dickson, 1939) and strike breakers in industrial conflicts were often subjected to extreme social pressure and isolation. (Francis, 1985) describes this with regard to the strikes of the British Miners in 1984-85: "*To isolate those who supported the 'scab union', cinemas and shops were boycotted, there were expulsions from football teams, bands and choirs and 'scabs' witnessed their own 'death' in communities which no longer accepted them*". The peer sanctioning of free-riding workers under team production in Japanese-owned auto factories in the US is described in (Rehder, 1990) who reports that "*the entire team suffers when one person is absent, and the returning team members can receive both formal sanctions and informal group pressures upon his or her return. The system is designed to function that way and it works very well*".

[2] At the beginning of World War I the British Army relied entirely on the voluntary recruitment of soldiers. Between August 1914 (when Britain declared war on Germany) and September 1914 roughly 479 000 volunteers were recruited and until December 1915 roughly 2.5 million men joined the British Army voluntarily. Those who did not join faced the contempt of their community members who attached big red patches to the free-riders' front doors at night, so that everybody could see that the person living there was a dodger. However, even in those countries – such as Germany – that had conscript armies and thus forced a subset of young men into military service, large numbers of young men voluntarily joined the army.

[3] In Fehr and Gächter (2000), the total of cost punishment was on average as large as the benefits from increased cooperation. Thus, the peer punishment opportunity did not generate any overall welfare gains. In Dreber et al. (2008), Rand et al. (2009) and in O'Gorman, Henrich, and van Vugt (2009), the peer punishment institution even led to a substantial overall welfare loss.

[4] We define antisocial punishment as the sanctioning of above-average contributors to the public good. Others (e.g., Ertan, Page, and Putterman (2009)) denoted this as "perverse punishment".

authority to punish to an elected central authority that refrains from antisocial punishment. In fact, ethnographic evidence indicates that early human groups were already characterized by high mobility and frequent migration in and out of existing groups (Boehm et al., 1996; Fehr & Henrich, 2003; Hill et al., 2011; Kaplan & Gurven, 2005; Mathew & Boyd, 2011; Wiessner, 2005). Migration across groups thus captures an important component of social life in the early evolution of humans and throughout human history.

The results in this paper show that under plausible conditions peer sanctioning is a very effective way of *quickly* establishing a culture of cooperation that *enhances its members' welfare*. In this culture, free-riding is extremely low *from the very beginning without much need for punishment*. In those rare cases in which free-riding occurs, subjects face a punishment probability of nearly 100%, which explains the quick and nearly universal compliance with a high cooperation norm. In contrast to previous evidence, subjects in our peer sanctioning institution are quickly better off than they would have been in an institution without the sanctioning of free-riders. We also show that an institution with an elected central punishment authority performs equally well in terms of extremely high cooperation rates from the start and an even more rapid welfare improvement relative to an institution without the sanctioning of free-riders. However, even when subjects can freely move to the centralized punishment institution, peer sanctioning remains popular and attracts a substantial and stable share of the experimental population.

How is it possible for peer sanctioning and centralized sanctioning by an elected authority to become so efficient so quickly? We introduced two elements into our experiment that were also already present in pre-state societies but capture also key features of many current societies. First, the opportunity to form a social consensus about normatively appropriate behaviors and second, the introduction of the possibility to migrate across groups. The reason for the first element is that throughout human evolution – from hunter-gatherer societies to modern nation states – cooperation and punishment in the context of public goods is associated with strong normative content and often relies on a widely shared normative consensus (Ostrom, 2000). For example, corporal punishment among the Turkuna – a politically decentralized, nomadic pastoral society in East Africa – for those who deserted and behaved cowardly during warfare requires a community consensus based on an assessment of the "badness" of the coward's behavior (Mathew & Boyd, 2011).

We introduced the opportunity to achieve a normative consensus within a group by asking each subject the following question privately: "How many points do you think each participant should contribute to the project?"[5] Subjects only needed to insert a single number to answer this question; every group member was then informed about the average number announced in the group. The subjects thus announced a normative request by answering the question above, and the average normative request could potentially become a social norm that coordinates and guides group members' contribution and punishment decisions. Recent research has shown (Fehr & Schurtenberger, 2017), however, that the normative request we implemented is completely futile for sustaining cooperation per se, i.e., in the absence of a punishment opportunity. This ineffectiveness of the normative request in the absence of punishment is a desirable feature because we are not interested in general cooperation-enhancing mechanisms in this study. Rather, in view of the widespread role of peer punishment in social life, we want to know whether peer punishment (and other punishment institutions) may work well when we allow for a normative consensus opportunity.[6]

[5] In the experimental instructions, we used the word "project" to describe the public good.

[6] This is also one reason why we did not allow for face-to-face communication among the subjects. In addition, face-to-face communication among subjects introduces potential confounds such as sympathy or antipathy among the

We introduced migration opportunities in our main experimental treatment by giving subjects the choice between different institutions at the beginning of each of 20 periods. This is similar to Gürerk, Irlenbusch, and Rockenbach (2006) but in contrast to their experiment our set of available institutions includes those with an opportunity to form a normative consensus and an institution with a centralized punishment authority. Within an institution, subjects can contribute to, and benefit from, a public good that only benefits members of the institution. For simplicity, and to have a stark contrast between individual and collective interest, it is in an individual's rational self-interest to contribute nothing to the public good when he or she faces no sanctions for free-riding, but group welfare is maximized if everybody contributes the whole endowment to the public good. One of the available institutions is characterized by the absence of any explicit opportunity for sanctioning individual free-riders ("*no punishment*"). The second institution provides an opportunity for each group member to sanction any other group member after having observed each of their contributions to the public good. We denote this institution as "*uncoordinated peer punishment*" because it does not offer any explicit possibility for group members' to express their normative views on the appropriate contribution level, and these views therefore could not coordinate subjects' contribution behavior. This institution has dominated the experimental economics literature on punishment for almost two decades (e.g., Chaudhuri, 2011; Hermann et al., 2008; Masclet, Noussair, Tucker, & Villeval, 2003). In our third institution, subjects could announce their normative requests so that these requests could, at least in principle, guide and coordinate their contribution behavior. This institution is therefore denoted as "*coordinated peer punishment*". The fourth and final institution ("*coordinated central punishment*") maintains the normative request but allows for the delegation of punishment to a single (central) authority while also socializing the cost of punishment. This type of institution exists in both small-scale societies (e.g. village elders and tribal chiefs, with collective punishment by the group) and large-scale societies (e.g., judges, police, courts, and prisons funded by taxes). In the experiment, the selection of the central authority is based on group members' views about who should be in that position. For simplicity these views are turned into the group's social choice via majority vote. While elections based on the majority rule may not occur explicitly in small scale societies, the superior authority of single individuals in these societies is often based on a group consensus (Henrich, Chudek, & Boyd, 2015), and the simplest way to implement a consensus in our experiment is through voting.

As mentioned above, both coordinated peer punishment and centralized punishment function very well and establish extremely high cooperation levels *from the beginning with little need for sanctions*. After a short adjustment phase, subjects thus predominantly choose these two institutions, while the other two institutions - no punishment and uncoordinated peer punishment - become depopulated. In fact, uncoordinated peer punishment was basically never chosen from the very beginning. This means that the previous literature on the welfare effects of peer punishment – including the early paper by Fehr and Gächter (2000) – may have led to a misleading view by focusing on an institution that subjects not only universally reject in our setting, but that arguably also lacks an important externally valid feature of many environments.[7]

The centralized punishment institution completely removes the inefficiencies of uncoordinated peer punishment and already leads to payoff levels that are substantially greater than in "no punishment" in

subjects, idiosyncratic communication styles or reputational concerns that go beyond the experiment because subjects are no longer anonymous to each other.

[7] Because uncoordinated peer punishment is basically inexistent, we use the term peer punishment for the coordinated peer punishment institution in the following, when it is clear what is meant.

the first period. Furthermore, centralized punishment removes antisocial punishment. The high efficiency of this institution is based on the two key facts. First, many prosocial individuals (i.e., those with prosocial other-regarding preferences) enter this institution at the very beginning, leading to high normative contribution requests and the election of a prosocial central authority who refrains from antisocial punishment and who is capable of enforcing high cooperation levels with relatively low levels of punishment. Rather than being merely cheap talk, the high normative requests are associated with high actual contributions, suggesting that individuals seem to be guided by the average contribution request. Subjects thus quickly establish a strong cooperative culture in the centralized punishment institution. The second reason for the high efficiency of centralized punishment is due to its intrinsically beneficial properties - even in the absence of endogenous sorting of subjects, this institution converges towards high cooperation with comparably little punishment; without endogenous sorting, however, it takes considerably more time to establish a culture of universal cooperation.

Coordinated peer punishment shares many of the good properties of centralized punishment. Many prosocial individuals immediately enter this institution; they establish very high normative requests followed by equally high contributions. However, this institution requires more actual sanctions during a short initial phase, and some antisocial punishment still persists. Nevertheless, payoffs are already larger than in "no punishment" in period 1, but the difference is only marginally significant ($p = 0.066$). This is in stark contrast to the uncoordinated punishment institution in, e.g., Fehr & Gächter (2000) and Gächter et al. (2008).

An interesting feature of our experiment is that despite the quick achievement of very high efficiency levels under centralized punishment, the coordinated peer punishment institution attracts a stable and substantial share of subjects: after a few initial periods migration rates become very low and roughly 45% of the subjects prefer coordinated peer punishment while roughly 55% enter the centralized punishment institution. However, when we take away the opportunity to announce normative requests in a control treatment[8], cooperation and efficiency levels under peer punishment become significantly lower for an extended period of time. As a consequence, peer punishment becomes much less "competitive" and the long-run share of subjects in peer versus centralized punishment is only 30%:70%. Thus, it appears that the opportunity to form a normative consensus is of particular importance for a well-functioning peer punishment system.

Taken together, our results show that efficient punishment institutions emerge endogenously through a competitive process in an environment in which people can "vote with their feet." Normative consensus building and prosocial individuals play a key role in this process. Prosocial individuals quickly migrate to coordinated peer or coordinated centralized punishment institutions and establish a cooperative culture that is characterized by the general normative request to cooperate at very high levels. This request is credibly enforced by peers or the centralized judge and considerably shortens the length of time that it takes to render an institution efficient. In contrast to uncoordinated peer punishment, which typically incurs large initial costs, the combination of endogenous sorting of prosocial individuals with the possibility of coordinating group behavior through normative requests very quickly makes both peer punishment and centralized punishment the superior institutions.

Our study is related to several lines of research. First, it is connected to the experimental literature on the role of sanctions in the provision of public goods and the enforcement of social norms. Influential

[8] In this control treatment normative requests were not possible in both the peer punishment and the centralized punishment institution. All other features of these institutions remained unchanged.

and insightful previous work has examined the sanctioning rules that subjects in predetermined groups select through democratic voting or other collective choice mechanisms (e. g., Yamagishi, 1986; Kosfeld, Okada, & Riedl, 2009; Ertan, Page & Putterman 2009; Sutter, Haigner, & Kocher, 2010; Putterman, Tyran & Kamei 2011; Andreoni & Gee, 2012; Markussen, Putterman & Tyran 2014). A key difference between this literature and our paper is that we introduce the opportunity to form a normative consensus and show that this has important consequences for the relative efficiency of peer versus centralized sanctioning systems. Another relevant feature of much of this literature is that once the sanctioning rule is determined through a collective choice procedure (e.g. voting), the sanctions themselves and/or the constraints on feasible sanctions are automatically executed, which presupposes a reliable institutional structure that captures important components of modern democratic states.[9] This contrasts with our peer punishment institution where subjects unconstrained by formal rules execute sanctions. Subjects have, in particular, the complete freedom to determine their individual distribution of sanctions and to engage in antisocial punishment – an important potential impediment to the efficiency of sanctioning systems in uncoordinated peer punishment. In our experiment, the punisher *may voluntarily* limit sanctioning due to his own normative views or the group members' informal normative requests, but explicit formal rules that are automatically enforced do not constrain him in any way. We therefore believe that one key insight from our paper is that despite the complete absence of formal constraints on individual sanctions, peer punishment with a normative consensus opportunity turns out to be very effective in establishing *a welfare-improving* culture of cooperation quickly.

Second, our work is related to the literature that has examined the role of competition – through individual migration opportunities – between institutions and groups that organize the provision of public goods in different ways. Gürerk, Irlenbusch, and Rockenbach (2006 and 2014) pioneered this literature and showed that uncoordinated peer punishment slowly becomes more popular and more efficient than an institution that rules out any punishment of free-riders. However, during this slow process of acquiring a competitive advantage over the no-punishment institution, *uncoordinated* peer punishment produces the well-known collateral efficiency costs typically associated with this institution. In Gürerk, Irlenbusch and Rockenbach (2014), for example, uncoordinated peer punishment acquires a stable positive welfare advantage over "no punishment" only after period 10, and it takes almost 20 periods to achieve reliable efficiency levels of roughly 80% and more. This contrasts with the quick success of *coordinated* peer punishment in our experiment, which already achieves a higher group welfare than "no-punishment" in period 1, and generates efficiency levels of close to 80% and more from period 4 onwards. This may be one reason why the subjects in our experiment universally reject uncoordinated peer punishment in favor of coordinated sanctioning institutions from the very beginning.

Like Gürerk, Irlenbusch and Rockenbach (2014), Nicklisch, Grechenig, and Thoeni (2016) only include uncoordinated peer punishment in the set of competing institutions; like us, they also allow for a central sanctioning authority but do not enable the formation of a normative consensus in this institution. In addition, they exogenously and permanently assign one subject to be the central authority which

[9] For example, in Kosfeld, Okada, and Riedl (2009), subjects can form a (sub)coalition where every member of the coalition is automatically punished if cooperation is less than complete. In Ertan, Page, and Putterman (2009), subjects can rule out antisocial punishment at a voting stage and an exogenous mechanism then fully enforces that rule. In Sutter, Haigner, and Kocher (2010), subjects can vote whether they prefer an institutional environment with mutual reward or mutual sanctioning opportunities, and this collective choice is then automatically enforced. In Andreoni and Gee (2012), subjects can pay to hire a mechanism that automatically punishes the largest free-rider – a proposal that resembles the mechanism proposed by Yamagishi (1986). Finally, in Markussen, Putterman, and Tyran (2014), subjects vote on whether to introduce a regime with informal peer sanctioning opportunities or one with formal sanctions.

sharply contrast with our centralized punishment institution where members elect the central authority each period, and the authority also bears part of the cost of punishment. In addition, individuals' normative requests may have guided our central authority – a feature that is absent in their design. These institutional differences give rise to important differences in the functioning of centralized punishment: In their experiment, a considerable share of the central authorities engage in the antisocial punishment of subjects with above average contributions, making centralized punishment even substantially less popular than *uncoordinated* peer punishment.[10] This contrasts with our results, where subjects selected very prosocial individuals as central authorities. These individuals avoid antisocial punishment altogether and quickly achieve very high cooperation and efficiency levels through prosocial punishment patterns. This result suggests that the efficiency of centralized punishment depends not only on centralization per se, but also on how the central authorities are selected and who gets selected.

Third, we believe that our results also inform the literature on the evolution of human cooperation. The rapid success of our two punishment institution appears to rely on the introduction of two key features that are already present in pre-state societies – the possibility to form a normative consensus and the opportunity to migrate across groups, which is associated with a predominantly prosocial migration to the punishment institutions with a norm consensus opportunity. These results therefore lend support to evolutionary models that allow for biased migration (Boyd & Richerson, 2009), for consensus building (Boyd, Gintis, & Bowles, 2010) and that emphasize the evolution of a norm-psychology (Chudek & Henrich, 2011). More generally, the fact that prosocial individuals appear to prefer particular institutions suggests that nonrandom interactions and assortative matching may have been important forces in the evolution of human cooperation (Eshel & Cavalli-Sforza, 1982).

Fourth, by showing (i) that migration patterns may depend on social preferences and (ii) the causal impact of norm formation opportunities on migration and cooperation, our paper emphasizes the importance of cultural aspects – such as the evolution of informal social norms – and may thus be useful for the economic literature that examines the impact of culture and social norms on economic outcomes (e.g., Benabou & Tirole 2011; Bowles, 1998; Giuliano & Nunn, 2016; Guiso, Sapienza, & Zingales, 2006; Lowes, Nunn, Robinson, & Weigel, 2017; Nunn, 2014; Tabellini, 2008).

Finally, our findings may also be interesting for those who study the emergence and the effectiveness of centralized punishment institutions from a theoretical perspective (e.g., Acemoglu & Wolitzky, 2016), for the literature on the foundations of the enforcement of laws and norms (e.g., Dixit, 2007; Greif, 1989, 1993), and the literature that discusses the relative efficiency of formal versus informal enforcement of norms and contracts (e.g., Benabou & Tirole 2011; Fafchamps, 1996; Johnson, McMillan, & Woodruff, 2002; Kranton, 1996; McMillan & Woodruff, 1999).

The remainder of the paper is organized as follows. Section 1 presents our experimental design in detail. Section 2 presents our results. Section 3 concludes the paper and discusses open questions and possible avenues for future research.

[10] This holds true for the treatment where subjects have perfect information about other group members' cooperation levels – a condition that also prevailed in our experiments. However, the random assignment of central authorities in their imperfect information treatments also led to a large share of central authorities who punished subjects with cooperative signals.

1 Experimental design and procedures

The experiment consists of three parts. Parts 1 and 2 are conducted in the lab during the same session. Part 3 is conducted online after subjects leave the lab and provides an independent measure of subjects' social preferences. Subjects are initially assigned to a large group of size $N \in \{9,11,12\}$ that stays fixed for Parts 1 and 2 of the experiment. We attempted to have 12 members in all groups, but occasionally used smaller groups due to subjects registering for the study but failing to show up at the lab. Subjects are randomly assigned a unique identification number from the set {1,2,. . . ,N} which also stays fixed for Parts 1 and 2 of the experiment.

Part 1 consists of five periods of a typical public goods game without punishment. Part 2 consists of 20 periods of a public goods game in which subjects can endogenously form subgroups by sorting themselves into 4 different institutions. Part 1 serves the purpose of familiarizing subjects with the public goods game and avoid that they are overwhelmed by the considerable complexity of Part 2 before they have learned the basic rules of a public goods game. In addition, individuals' cooperation rates in Part 1 provide a measure of their cooperativeness and prosociality in the absence of pecuniary incentives for cooperation, which proves useful in interpreting our results. Part 3 is conducted online and measures social preferences using the Social Value Orientation scale of (Murphy, Ackermann, & Handgraaf, 2011). The experiment timeline is summarized in Figure 1.

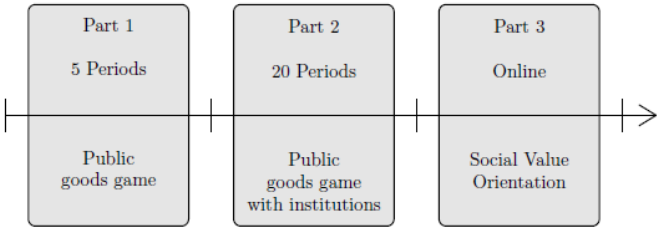


Figure 1: Experiment timeline.

1.1 Part 1: Public goods game

Subjects begin by playing five rounds of a typical public goods game without punishment. At the beginning of each period, subjects receive an endowment of points, e = 20, and can contribute any amount $g_h \in \{0, 1, \dots, 20\}$ to a group project. Each point contributed to the group project is multiplied by m and shared equally among all N group members. Each point not contributed to the project goes into a private account and can be kept by the subject. Thus, per-period earnings of subject h are given by

$$\pi_h = e - g_h + \left(\frac{m}{N}\right) \sum_{k=1}^{N} g_k$$

As mentioned above, we have N = 12 in most cases, but occasionally we had N= 9 or N = 11. The group benefit from a contribution, m, is set to m = 1.5. The parameter values for Part 1 are summarized in Table 1. They imply that the marginal benefit from a contribution to the group project for the contributor, m/N, ranged from 0.125 when N = 12 to 0.167 when N = 9. These marginal benefits provide a considerable incentive for free-riding but they are partly counterbalanced by the relatively large group size; previous evidence indicates that an increase in group size N that leaves the marginal benefit m/N unchanged (i.e., is accompanied by an increase in m) tends to increase cooperation (Isaac, Walker, & Williams, 1994). In each period and after making private contribution decisions, group members are informed of every group

member's contribution. In addition, they are informed about their earnings from the private account and from the group project.

Parameter	Value	Meaning
e	20	Endowment
m	1.5	Multiplier for contribution to public good
N	9, 11, or 12	Group size

Table 1: Parameter values used in Part 1.

1.2 Part 2: Public goods game with endogenous punishment institutions

Subjects remain in the same group and retain the same identification number in Part 2 of the experiment. Part 2 lasts for 20 periods and provides subjects with the opportunity to form institutions with other group members. At the beginning of each period, subjects individually select into one of four institutions and interact only with other group members who select the same institution in that period. Migration between institutions is costless, and subjects can select any of the institutions at the start of each period. The institutions, which we describe in more detail in Section 1.2.1 below, are (i) No Punishment, (ii) Uncoordinated Peer Punishment, (iii) Coordinated Peer Punishment, and (iv) Coordinated Central Punishment. By "Coordinated," we mean that each group member could make a normative request regarding how much each group member should contribute. In principle, this provides an opportunity to coordinate on a social contribution norm that might guide subjects' contributions. Note, however, that it was not possible to state normative requests regarding punishment behavior, and the group members were completely free to neglect both their own and others' normative requests.

Each period contains a contribution stage which is identical to the situation in Part 1, except that contributions to the group project only affect group members who select the same institution. In addition, a punishment stage is added which provides an additional endowment in each period. In the event that only one subject selects a particular institution in the period, both the contribution stage and punishment stage endowments go directly to the private account, and the subject is not able to contribute to a group account. This design feature was included because the idea of a public good necessarily involves more than one person benefitting from contributions.

We next describe the four institutions in greater detail. During each period, subjects make several decisions, and their choice opportunities depend partly on the selected institution. We then describe the information provided to subjects at each decision point. Finally, we describe the material payoffs resulting from subjects' choices.

1.2.1 Institutions

Subjects begin each period by selecting which institution they want to choose in the current period. For the remainder of the period, subjects only interact with other group members who have also chosen the

same institution. The sequence of decisions made during each period is summarized in Figure 2.

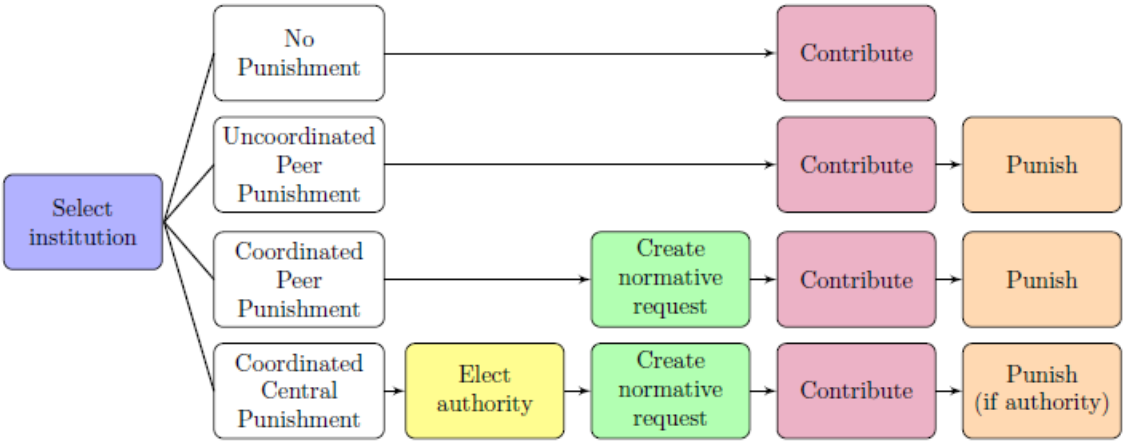


Figure 2: Sequence of decisions in Part 2.

No Punishment. No Punishment is identical to Part 1, except that (i) contributions to the group project only affect members of the group who adopt the No Punishment institution and (ii) subjects receive a second endowment in the punishment stage which is added directly to their earnings.

Uncoordinated Peer Punishment. In Uncoordinated Peer Punishment, subjects make the same contribution decision as in No Punishment. However, during the punishment stage, subjects can assign deduction points to other institution members which reduce the recipient's earnings at a cost to the person assigning punishment. The cost structure for punishment is described in Section 1.2.3 and defined in equation (4) below.

Coordinated Peer Punishment. The contribution and punishment stages in Coordinated Peer Punishment are identical to Uncoordinated Peer Punishment. However, prior to the contribution stage, each institution member makes a normative request by privately answering the question, *"How many points do you think each participant should contribute to the project?"* The average of these requests is reported on each institution member's computer screen during the contribution stage. The normative request is, however, non-binding and it is public knowledge that each institution member is informed about the average normative request. The cost structure for punishment is described in Section 1.2.3 and defined in equation (4) below.

Coordinated Central Punishment. In Coordinated Central Punishment, one member of the institution is elected to assign all of the punishment for the group, and all institution members share the total cost of punishment equally. At the start of the period, subjects in Coordinated Central Punishment select a single institution member – the central authority – by majority vote who has the exclusive right to assign the punishment; the central authority is the person who receives the most votes, and ties are broken randomly. After casting their votes, subjects determine their normative requests in the same manner as in Coordinated Peer Punishment. Then, subjects enter the contribution stage, where they are informed of the normative request and make the same contribution decision as in all other institutions. Finally, during the punishment stage, only the central authority can assign deduction points to institution members, and these deduction points reduce the recipient's earnings at a cost which is shared equally by all members of the institution.

The cost structure for punishment is described in Section 1.2.3 and defined in equation (5) below.

The role of normative requests. The rationale for the introduction of a normative request is that social norms are in reality likely to be pervasive in public goods contexts, but the typical public goods experiments without communication do not enable people to express their normative views on group members' contribution decisions (Ostrom, 2000). However, letting group members speak to each other face-to-face before taking actions in the experiment introduces other potential confounds such as sympathy or antipathy among the subjects, idiosyncratic differences in communication styles, reputational concerns that go beyond the experiment because subjects are no longer anonymous to each other, etc. Some researchers have therefore introduced the possibility of communicating purely numerical information regarding contribution intentions or contribution promises among the group members (Bochet, Page, & Putterman, 2006; Bochet & Putterman, 2009; Wilson & Sell, 1997).

These studies show that purely numerical communication is generally unable to establish cooperation and occasionally performs worse than an environment without communication. These findings suggest that giving merely the opportunity to express a normative request may just enable cheap talk without behavioral consequences. A recent study (Fehr & Schurtenberger, 2017) strongly supports this conclusion. The study compares a voluntary contribution treatment (i.e., one with no punishment) that allows for individual normative requests with an otherwise identical treatment condition *without* such requests. The results show that the normative request is completely ineffective in influencing behavior in this environment[11].

If the normative request is merely cheap talk, why did we then nevertheless introduce it in our setting? The reason is that the normative request may help subjects coordinate more easily on a particular contribution norm when peer punishment and central punishment are possible. There is, of course, no guarantee that this norm involves high cooperation, but it is possible. In addition, a high average normative request may delegitimize low contributions which, in turn, may enhance the credibility of the punishment threat. Our design also enables us to measure whether subjects prefer, ceteris paribus, a normative request because we allow for peer punishment institutions with and without normative requests. And finally, the existence of the normative request allows us to measure whether subjects indeed live up on average to the group's request.

1.2.2 Information

The information provided during each period is summarized in Figure 3. At the beginning of each period and before subjects join an institution, all subjects are informed of the number of group members who joined each institution in the previous period and the average earnings for each institution in the previous period. After subjects join an institution, they learn each current institution member's contribution in the previous period. Following this, the Coordinated Central Punishment institution elects the central

[11] The fact that the normative request does not affect behavior in the absence of punishment options is also the reason why we did not include an institution with "No Punishment but a normative request". Because our experiment is already relatively complicated, the inclusion of such an institution would have made Part 2 of our experiment more complex without providing much insight. To keep the design as simple as possible, we also did not include a Central Punishment Institution without normative requests. In addition, in naturally occurring environments where a central authority has the power to punish, this power is typically deeply embedded in the normative structure of a society. We kept the possibility of Uncoordinated Peer Punishment because this institution has played a major role in research over the last 15 years. The inclusion of this institution and its relative performance also allows us to assess whether the absence of normative requests in peer punishment constitutes a serious omission.

authority[12], and the Coordinated Peer Punishment and Coordinated Central Punishment institution members determine their normative requests; the average of these requests per institution (rounded to the nearest integer) is then transmitted back to the institution members.

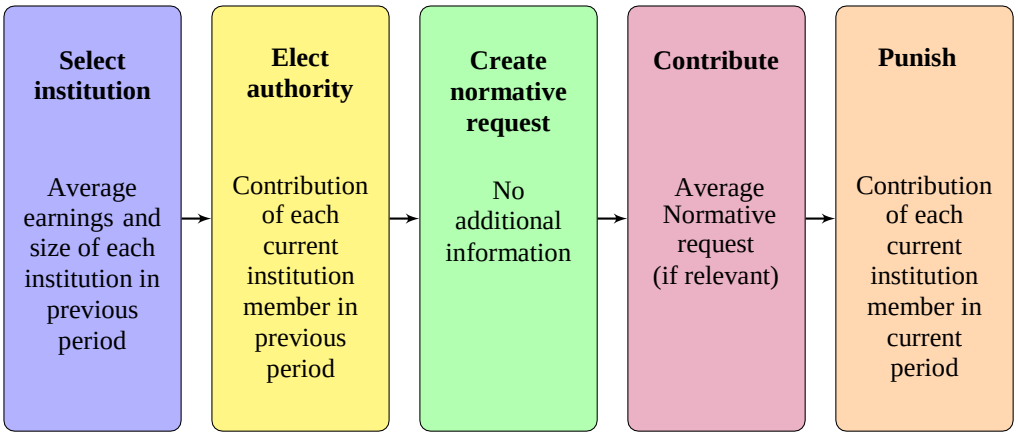


Figure 3: Sequence of information in Part 2.

Finally, at the punishment stage, members of all institutions observe the contribution of each institution member in the current period. Information about particular acts of punishment was exclusively available to the punisher and the punished subject but not to the whole membership of an institution.

With the exception of the information feedback on the average normative request, which is a distinguishing feature of Coordinated Peer Punishment and Central Punishment, all subjects receive the information described in Figure 3 – even if no decision has yet been made at that point. For example, all subjects in a matching group are informed about each current institution member's contribution in the previous period. We intentionally chose this feature to rule out the desire for increased information as a confounding explanation for institution selection. All institutions share the characteristic that subjects receive more information about their own institution than they do about the others. This captures the notion that we know more about the people we interact with, but only have limited information about those with whom we do not interact.

1.2.3. Material payoffs

Earnings in each period are the sum of the contribution stage earnings and the punishment stage earnings. Contributions stage earnings for individual *h* in institution *i*, π_{hi}^{1}, are given by

$$\pi_{hi}^{1} = e^{1} - g_{hi} + \left(\frac{m}{s_i}\right)\sum_{k=1}^{s_i} g_{ki}, \qquad (2)$$

where e^1 is the contribution stage endowment, g_{hi} is the individual's contribution to the institution project, *m* is the total pecuniary gain per contribution to the institution's project for all institution members together, and s_i is the endogenously determined institution size (number of institutions members). Punishment stage

[12] In the first period of Part 2 (period 6 overall), subjects are informed of the average contributions of each current institution member during Part 1.

earnings for individual *h* in institution *i*, π_{hi}^2, are given by

$$\pi_{hi}^2 = e^2 - c_{hi}(d) - r\sum_{k=1}^{s_i} d_{khi}, \quad (3)$$

where e^2 is the punishment stage endowment, $c_{hi}(d)$ is the cost of punishing other institution members for individual *h*, *r* is the reduction in earnings for each deduction point received from other institution members, and d_{khi} is the number of deduction points that institution member *k* assigns to individual *h* in institution *i*. Therefore, the total payoff reduction individual *h* receives because others punish him is given by $r\sum_{k=1}^{s_i} d_{khi}$. The cost *h* bears from assigning deduction points, denoted by d_{hki}, to other members *k* in his institution *i* is given by

$$c_{hi}(d) = \sum_{k=1}^{s_i} d_{hki}, \quad (4)$$

if *h* is a member of one of the two peer punishment institutions, and

$$c_{hi}(d) = \frac{\sum_{k=1}^{s_i} d_{Aki}}{s_i} \quad (5)$$

if *h* is a member of the Central Punishment Institution where the cost of all punishments assigned by the central authority *A* are given by $\sum_{k=1}^{s_i} d_{Aki}$. These costs are shared equally by all s_i institution members. Thus, the total cost of assigning a deduction point is 1, but the cost for each institution member is determined according to equation (4) in institutions utilizing peer punishment and equation (5) in the central punishment institution. In the No Punishment institution c_{hi}, d_{khi} and d_{Aki} are always equal to zero for all *h* and *k*.

Finally, we imposed a bankruptcy condition so that earnings in a single period could not be negative, but subjects still had to pay for assigning deduction points to others even if the tokens could not reduce the recipient's earnings any further.[13] Therefore, per-period earnings are given by

$$\pi_{hi} = \max\{\pi_{hi}^1 + \pi_{hi}^2, 0\}. \quad (6)$$

Table 2 summarizes the parameter values used in the experiment. We set e^1 = e^2 = 20, *m* = 1.5, and *r* = 3.

Parameter	Value	Meaning
e^1	20	Endowment for contribution stage
m	1.5	total pecuniary gain per contribution to the public good
s_i	Endogenous	Institution size
e^2	20	Endowment for punishment stage
r	3	Reduction in earnings from receiving one unit of punishment
$c_{hi}(d)$	See equations (4) and (5)	Cost of assigned punishment to others

Table 2: Parameter values used in Part 2.

[13] This happened in less than 1% of all cases and was thus rarely a binding constraint.

1.2.4 Control treatments

Under endogenous institution selection, institutions may be successful in establishing and maintaining cooperation for two primary reasons. First, cooperative individuals may select into the same institutions, and these institutions therefore perform better. Second, the institution may create incentives that induce cooperative behavior, regardless of whether the individuals joining the institution are generally cooperative. It is, of course, also possible that these two effects are mutually reinforcing. Without a control treatment, our design with endogenous selection does not allow us to disentangle the "self-selection effect" from the "pure institutions effect". We therefore conducted a control treatment with exogenous assignment to the different institutions. For each matching group in our endogenous selection sessions, we create a matching group of the same size and with identical migration patterns. This means that for each session in the endogenous selection treatment, and each individual in that session with a given migration path, there is matched session in the exogenous assignment treatment with matched individuals that face exactly the same migration paths.[14] This also means that for each period and each session of the endogenous selection treatment there is matched session in the exogenous assignment treatment in which the distribution of subjects across institutions is completely identical. The exogenous assignment treatment allows us to examine the effects of institutions independently of self-selection. In addition, by comparing outcomes under endogenous selection and exogenous assignment, we can identify the impact of endogenous selection.

In addition to the control treatment with exogenous assignment to institutions we conducted a further control treatment in which we maintained the possibility for voluntary migration but removed the opportunity to achieve consensus about normatively appropriate behavior. Therefore, in this treatment there were only three institutions – No Punishment, Uncoordinated Peer Punishment and Uncoordinated Central Punishment. We introduced this control treatment primarily to study how the absence of a norm consensus opportunity affects subjects' selection into the different institutions. It is, for example, a plausible conjecture that the absence of normative coordination is more detrimental for peer punishment and, hence, induces subjects to predominantly enter into the uncoordinated centralized punishment institution.

1.3 Part 3: Social Value Orientation

Part 3 of the experiment consists of the Social Value Orientation (SVO) measure, which consists of six allocation decisions between oneself and one other anonymous individual (Murphy et al., 2011). The allocation decisions constitute modified versions of the dictator game in which the relative price of giving varies (Andreoni & Miller, 2002; Forsythe, Horowitz, Savin, & Sefton, 1994). The SVO measure provides a numeric score which can be used to measure each individual's prosociality. Higher SVO scores indicate greater prosociality. Full details of the SVO measure are provided in Online Appendix A.2. This part of the experiment was conducted online either two weeks before or two weeks after subjects completed Parts 1 and 2 in the lab. Subjects knew that one of their allocation decisions would be selected and implemented, and that they would be the recipient of a different anonymous person's allocation decision; payments were mailed to subjects.

[14] More formally, for every session s in the endogenous sorting treatment we created a matched session s' in the exogenous assignment treatment such that for every subject j in s there exists a subject j' in s' with the exact same migration path, implying that the allocation of subjects across institutions at time t, and the previous migration experiences of subjects in every institution at time t, are identical across treatments.

1.4 Experimental procedures

Sessions for Part 1 and Part 2 were conducted in a computer lab at the University of Zurich in December 2012 and April 2013. Part 3 was conducted online using Qualtrics (www.qualtrics.com). Experimental instructions are provided in the Online Appendix. Subjects were mostly students from the University of Zurich and the Swiss Federal Institute of Technology (ETH-Zurich). Recruitment was conducted using ORSEE, and we did not invite students who listed economics or psychology as their major (Greiner, 2015). Experiments were programmed in z-Tree (Fischbacher, 2007). Points were used as the experimental currency and converted to Swiss Francs CHF at the end of the study; subjects were informed of the exchange rate in the instructions. In Parts 1 and 2, conducted during the same lab session, the exchange rate is 1 point = CHF 0.05. In Part 3, conducted online, the exchange rate is 1 point = CHF 0.10. Average earnings were CHF 62.29 for the lab session (consisting of both Parts 1 and 2), including a show-up fee of CHF 10. Earnings from Part 3 were CHF 15-20. Lab sessions lasted 2.5-3 hours on average, and the online portion of the experiment took 10-20 minutes.

Overall, 256 subjects participated in the lab sessions; 128 subjects participated in the endogenous selection treatment, and another 128 subjects participated in the exogenous assignment treatment. Each treatment consisted of eleven groups in total. Nine of these groups had twelve members. One group of nine members and one group of eleven members were used in each treatment because some invited subjects did not come to the lab. Partner matching was used in Parts 1 and 2, so each group remained fixed for the entire lab session.

For subjects in the endogenous treatments (where we predict assortment effects), 84 of 128 subjects (66%) completed Part 3 of the experiment online.[15] We do not find obvious evidence for selection bias among respondents, i.e., subjects who completed Part 3 (average age = 22.0, percent female = 0.37, average earnings in Parts 1 and 2 = CHF 63.40) are similar in observable characteristics to those who did not complete Part 3 (average age = 22.3, percent female = 0.34, average earnings in Parts 1 and 2 = CHF 62.67).

2 Results

2.1 Institution selection and institution performance

A key question of our study relates to individuals' preferences for the different available institutions. We are, in particular, interested in (i) how frequently Coordinated Peer and Central Punishment are chosen relative to Uncoordinated Peer Punishment and (ii) how quickly the punishment institutions dominate in terms of their popularity among subjects. We summarize our findings with regard to these questions in

Result 1 (Institution Selection):

- **(a)** In period 1, subjects are initially evenly distributed between the No Punishment (NP), Coordinated Peer Punishment (PP), and Coordinated Centralized Punishment (CP) institutions, while Uncoordinated Peer Punishment is basically ignored from the beginning.
- **(b)** After only a few periods, however, the punishment institutions with norm coordination attract more than 80% of the subjects and thereafter the share of subjects attracted by NP becomes negligible.

[15] In the exogenous assignment treatment 96 of 128 subjects (75%) completed Part 3.

Evidence for Result 1 is provided by Figure 4 and Table A1 in the Appendix, which are based on the pooled data from all matching groups and displays how the distribution of subjects between the different institutions evolves over time. These data show that, from the very beginning, the share of subjects selecting Uncoordinated Peer Punishment is negligible. In contrast, 37% of the subjects sort into NP, 36% into PP, while the remaining subjects prefer CP in period 1. By period 2, however, roughly 75% of the subjects already prefer the punishment institutions with normative requests, PP and CP. Following this, more than 80% always enter these institutions, while the percentage of subjects entering NP becomes negligible. It is interesting to compare these results with those in Gürerk et al. (2014), where normative requests were ruled out; there it took 15 periods until 80% of the subjects entered the peer punishment institution.

Once almost all subjects had selected themselves into the two punishment institutions with normative requests, CP attracted on average roughly 55% of the subjects while the other 45% selected into PP. These average numbers hide however predominant selection into one of the two institutions at the level of the individual matching group (Figure A2 in the Appendix A2). In fact, in four of the eleven matching groups literally all subjects selected into the CP institution and in one matching group the large majority preferred CP; in six matching groups the overwhelming majority of the subjects (i.e., two thirds or more) selected into the PP institution. In addition, after the short initial phase where subjects migrated out of NP, migration rates quickly became rather low – typically none or only one subject per matching group changed the institution in later periods.

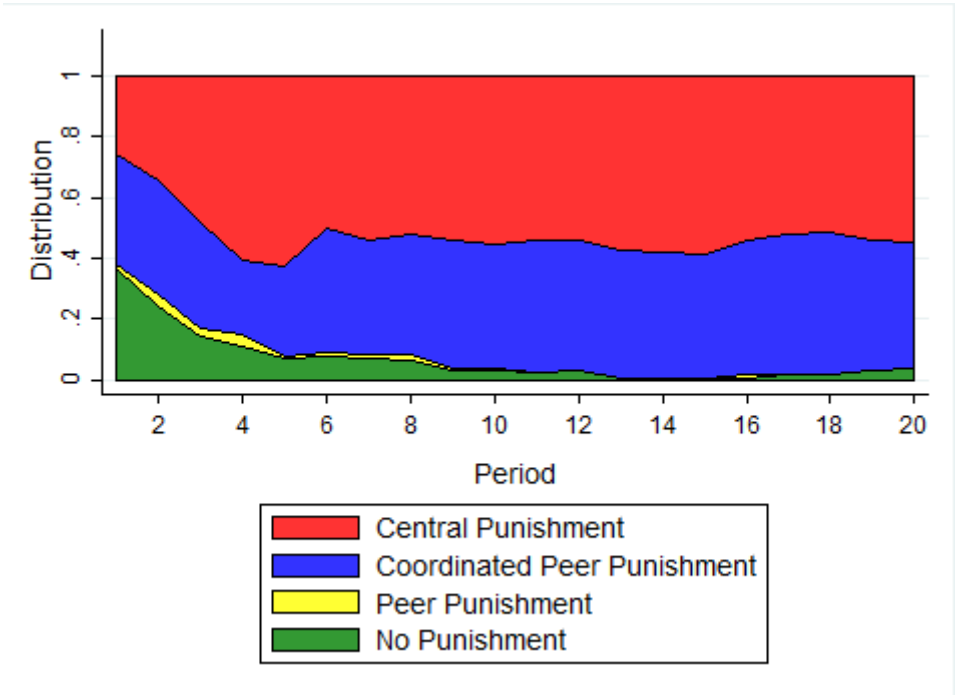


Figure 4: Distribution of subjects across institutions.

Result 1a is a powerful indication that in the presence of punishment institutions with an opportunity for normative requests subjects strongly reject a punishment institution that does not provide this opportunity. This suggests that previous research on peer punishment that rules out normative requests could have missed an important aspect of informal peer punishment because choices regarding contributions to public

goods often seem to be deeply embedded in an existing normative consensus. However, achieving such a consensus in the absence of explicit normative requests may be difficult or may at least take a lot of time, during which normative mis-coordination occurs. Because Uncoordinated Peer Punishment is basically absent from the beginning, we reserve the term peer punishment for Coordinated Peer Punishment for the rest of the paper (unless we indicate otherwise).

The strong dominance of coordinated peer sanctioning and centralized sanctioning raises the question of why these two institutions prevail. A plausible hypothesis is that they prevail because they are successful in quickly establishing high cooperation and efficiency levels. We find evidence for both conjectures:

Result 2 (Cooperation and the Efficiency of Punishment Institutions):

(a) Under both Coordinated Peer Punishment and Central Punishment, very high cooperation levels close to 100% of the endowment are quickly obtained.

(b) In terms of efficiency gains, central punishment already outperforms the no-punishment institution from period one onwards, and coordinated peer punishment achieves the same result within a few periods.

Figure 5a and 5b illustrate Result 2. Figure 5a shows the average contributions over time in the NP, the PP and the CP. The unit of observation is the matching group, so that the average contribution for an institution is first taken at the matching group level, and these results are then averaged across matching groups.[16] Figure 5a shows that the average contribution (as a percentage of the endowment) in PP and CP is already about 90% in the first period of Part 2, and soon reaches close to 100%. In contrast, average contributions in NP quickly decline to low levels.[17]

This rapid achievement of maximum cooperation levels contrasts sharply with the results observed under non-coordinated peer punishment. For example, in Fehr and Gächter (2000), subjects' cooperation level in period 1 is only about 60% of the endowment, and, it takes 20 periods until cooperation levels reach 90% of the endowment in Gächter et al. (2008). Likewise, stable cooperation levels of 90% or more are only reached after 10 periods in Gürerk et al. (2014).[18]

High cooperation levels do not necessarily imply, however, that efficiency gains have been optimally exploited.[19] If high cooperation is only achieved through high punishment, efficiency gains may even be negative because punishment is costly for the punished and the punisher. In Fehr & Gächter (2000), for

[16] When we show average cooperation and average efficiency across institutions we always proceed in this way. In addition, our non-parametric statistical tests always use the average outcome of an institution in the matching groups as the unit of observation, and when we run regressions, standard errors are clustered at the matching group level.

[17] We do not show cooperation and efficiency levels in NP after period 6 because the membership of this institution during these periods, if there is any, consists typically of N = 1 but cooperation and gains from cooperation require at least N=2.

[18] When comparing these experiments one has to keep in mind the potentially different MPCR's across studies because a higher MPCR typically leads to higher cooperation. However, the quick achievement of very high cooperation levels in our study relative to Gächter et al. (2008) and Fehr and Gächter (2000) cannot be attributed to a higher MPCR. In our study we typically had quickly 5 members in PP and in CP. This yields an MPRC of m/N = 1.5/5 = 0.3 which is *smaller* compared to the MPCR of 0.5 and 0.4 in Gächter, Renner and Sefton (2008) in Fehr and Gächter (2000), respectively.

[19] We define efficiency as follows. In the absence of any cooperation and punishment zero percent of the potential cooperation gains are reaped, i.e., efficiency is zero. If there is full cooperation and no punishment 100% of all potential cooperation gains are acquired, i.e., efficiency is 100%. Negative efficiency gains accrue if the gains achieved through cooperation are outweighed by the cost of punishment.

example, the efficiency levels were negative during the first few periods, i.e., subjects were worse off during these periods compared to a situation with zero cooperation and zero punishment, and the average efficiency under peer punishment was not higher during the whole 10 periods of the experiment than in the absence of a punishment opportunity (Mann-Whitney U, p = 0.45). The situation was even worse in Gächter et al. (2008) because efficiency gains in their 10-period experiment were lower in the presence of a punishment opportunity *in each period*, and the overall efficiency of this institution was therefore lower compared to the no punishment treatment (Mann-Whitney U, p =0.033). In fact, the efficiency gains of peer punishment were even negative most of the time and, this institution therefore resulted on average in *negative* efficiency gains of -5.1%.

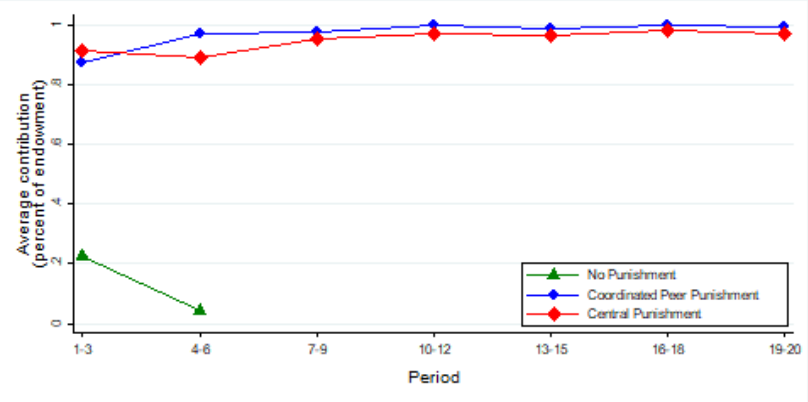


Figure 5a. Average contributions over time

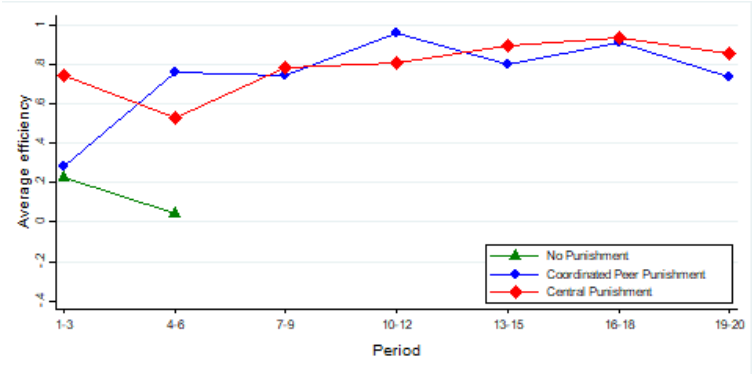


Figure 5b. Average efficiency over time

Figure 5b reveals that the efficiency properties of our two punishment institutions with normative requests are radically different because Central Punishment already attains roughly 80% of the available efficiency gains in the first period, and Peer Punishment *always* shows, i.e., from period 1 onwards, higher efficiency than NP. Statistically, efficiency is significantly higher in CP relative to NP (Mann-Whitney U, p=0.001) in the first period, and the efficiency difference between PP and NP is marginally significant (Mann-Whitney U, p=0.066). However, the efficiency of PP, as those in CP, is always substantially and significantly higher than in NP from period 4 onwards (Mann-Whitney U tests, all $p < 0.05$).

2.2 Institution selection and performance in the absence of a normative consensus opportunity

Why do coordinated peer punishment and centralized punishment perform so well in terms of quickly establishing and maintaining cooperation and achieving high efficiency levels? And, of equal importance, what are the reasons that the typical efficiency costs of peer punishment are almost completely absent under PP? One reason could be that the normative consensus building opportunity enables subjects to avoid the collateral damages typically associated with uncoordinated peer punishment. Another reason could be that prosocial individuals are the first to migrate into coordinated PP and CP and quickly establish a beneficial social norm of full cooperation so that those who join later can be smoothly integrated into a "culture of cooperation." To study these conjectures in more detail, we examine what happens if we remove the normative consensus building opportunity in this section, and in the next section we study the impact of removing voluntary migration opportunities.

When we remove the normative request, coordinated peer punishment is no longer a feasible institution. For this reason there are only three different institutions available in the absence of a normative request: NP, uncoordinated PP and CP without normative request. We are particularly interested in whether the existence of a normative consensus building opportunity changes the relative popularity of centralized versus peer punishment and of NP versus peer punishment. The following result summarizes our findings:

Result 3 (Role of Normative Request):

(a) When we remove the normative consensus building opportunity, the peer punishment institution becomes much less popular and a clear majority of subjects initially prefer to be in NP while the popularity of CP is initially not affected. However, over time, CP becomes by far the most popular institution and NP slowly becomes de-populated almost completely.

(b) The strong reduction in the popularity of peer punishment is associated with a considerable reduction in the cooperation and efficiency level of this institution that prevails for an extended period of time. In contrast, the efficiency of CP is only marginally affected by the removal of normative consensus building.

Support for Result 3a is provided by Figures 6a, 6b and Table A1b in Appendix 1, which provides numerical information on the selection of subjects across institutions. Figures 6a and 6b indicate that the removal of the normative request reduces the share of subjects that select into peer punishment by roughly 20 percentage points. This reduction in PPs popularity is present from the very beginning. The figures also show that initially it is primarily the NP institution that benefits from subjects' reluctance to enter uncoordinated PP. However, over time this advantage of NP vanishes completely and the institution becomes basically de-populated. This process of de-population of NP is associated with an increase in the popularity of CP that attracts – from period 7 onwards – roughly 15 percentage points more subjects in the absence of normative requests relative to a situation with normative requests. In the absence of normative requests, roughly 70 percent of the subjects selected themselves into CP and only about 30 percent entered the PP institution in the last five periods.

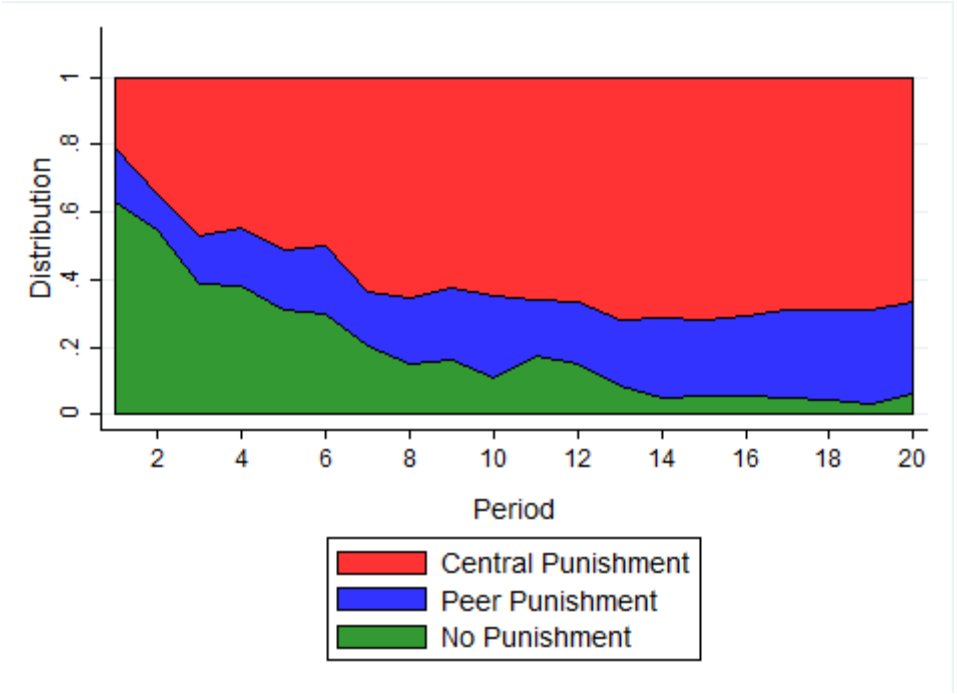


Figure 6a: Distribution of subjects across institutions in the absence of normative requests.

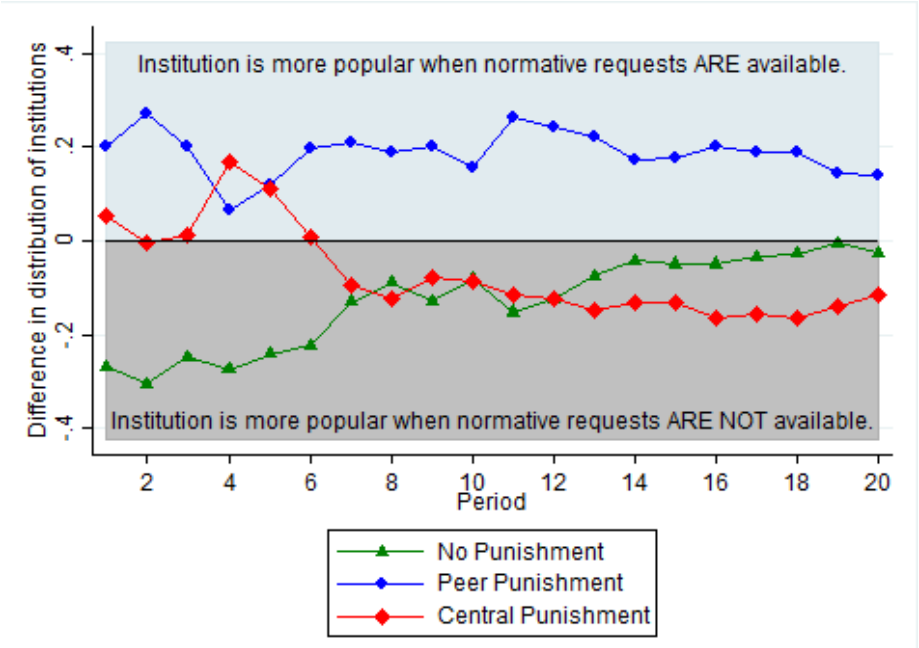


Figure 6b: The impact of normative consensus building on institutional preferences.

The figure shows the relative share of subjects in each institution in the presence of normative requests minus the relative share of subjects in the absence of normative requests.

These results indicate that the opportunity to form a normative consensus has a striking causal impact on subjects' willingness to enter peer punishment versus no punishment or centralized punishment. In the longer run the absence of normative consensus building opportunities causes a substantial shift in favor of

centralized punishment which raises the question about the mechanisms underlying this shift. It turns out that if we take away normative requests, cooperation and efficiency of peer punishment is strongly hampered for extended periods of time (see Figures A3a and A3b in Appendix A3). In particular, during the first 12 periods, cooperation (Mann Whitney test: $p = 0.005$) and efficiency (Mann Whitney test: $p = 0.020$) are significantly lower under PP without normative requests compared to PP with normative requests. Interestingly, however, the cooperation and efficiency of centralized punishment is affected by the removal of normative requests at most during the first three periods (MW test for cooperation: $p = 0.094$; MW-test for efficiency: $p = 0.020$; see also Figures A3c and A3d in Appendix A3) but not beyond (all $p > 0.11$ for later periods). Thus, it appears that the possibility of forming a normative consensus is particularly important for an institution that relies on peer punishment while our centralized punishment institution with democratically elected judges can compensate in some way for the absence of normative consensus building opportunities.

2.3 Endogenous sorting and the culture of cooperation

After we have examined the causal role of normative requests for the prevalence and the efficiency of different institutions we study in this section the role of endogenous sorting in our main set-up as described in Figure 2. We ask, in particular, about the role of endogenous sorting for the success of CP and PP with normative requests? In principle, these two institutions could have intrinsically beneficial properties that lead to high cooperation and efficiency even in the absence of voluntary migration. Yet, it could also be the case that voluntary migration has itself distinct welfare relevant effects. To study the intrinsic beneficial properties of PP and CP with normative requests, we conducted experiments where we ruled out endogenous selection into institutions, and assigned subjects randomly to the migration paths the individuals in the endogenous selection institution chose. The comparison of outcomes in the presence and absence of endogenous selection enables us to state

Result 4 (Role of Endogenous Sorting):

- **(a)** Under endogenous sorting into sanctioning institutions, the average cooperation and efficiency levels under Coordinated Peer Punishment and under Central Punishment are higher than under exogenous assignment during the initial periods, but are indistinguishable after period 12.
- **(b)** Under exogenous assignment, the two punishment institutions initially do not outperform NP with regard to efficiency, and peer punishment is even significantly worse.

Evidence for Result 4a is provided in Figures 6a – 6d. Figures 6a and 6c show that cooperation levels are already higher under endogenous selection at the beginning and throughout the first 12 periods. This pattern is also supported by Mann-Whitney U tests. The null hypothesis of equal contributions can already be rejected for Coordinated Peer Punishment ($p = 0.021$) and for Central Punishment ($p = 0.012$) for the very first period.

A similar picture emerges with regard to efficiency; in both Coordinated Peer Punishment and in Centralized Punishment. Endogenous selection is already significantly better in period 1 (MW-tests; PP: $p = 0.016$; CP: $p = 0.035$)

The lower cooperation and efficiency levels in the exogenous assignment treatment also mean that the two punishment institutions initially do not outperform the No Punishment treatment in terms of efficiency

under exogenous assignment: PP is in this regard equally as good as NP during the first three periods (MW-test; p=0.14), and also during periods 4-6 (MW-test; p=0.13). CP is also equally as good as NP during the first three periods (MW-test; p = 0.18) but already significantly better in periods 4-6 (MW-test; p = 0.049).

Result 4 means that the two punishment institutions have intrinsically beneficial properties because efficiency and cooperation are eventually very high, even under exogenous assignment.[20] However, the opportunity to select into institutions endogenously has important cooperation- and efficiency-enhancing influences for many periods, and it apparently helps quickly establish high cooperation and efficiency levels. These results raise the question what exactly makes endogenous selection such a powerful device. Our next result provides the answer to this question.

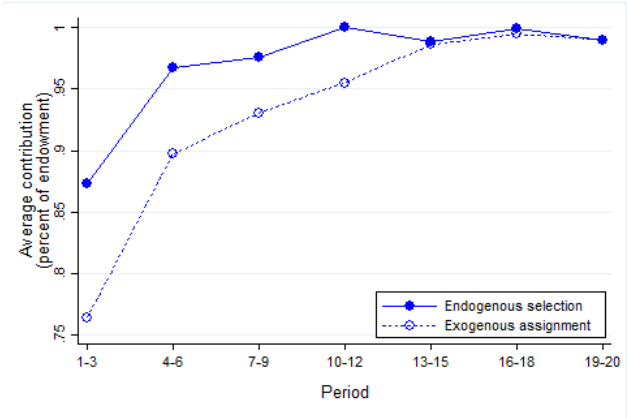


Figure 7a. Contributions over time in Coordinated Peer Punishment

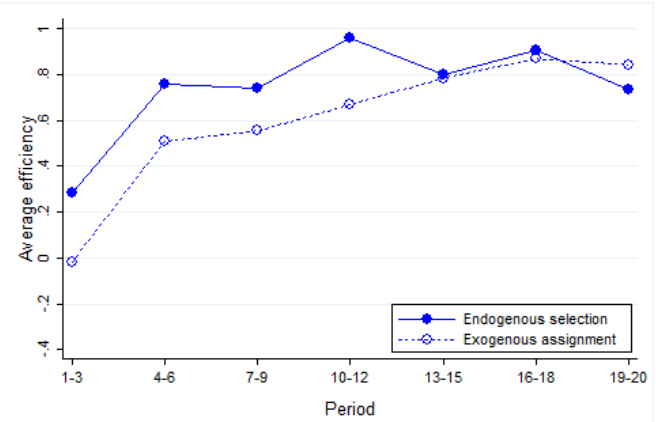


Figure 7b. Efficiency over time in Coordinated Peer Punishment

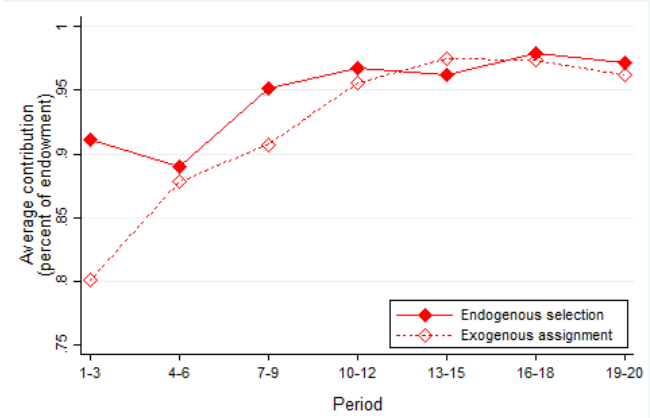


Figure 7c. Contributions over time in Centralized Punishment

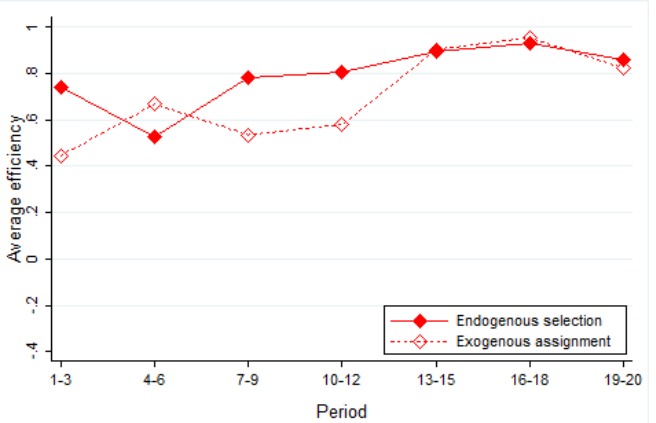


Figure 7d. Efficiency over time in Centralized Punishment

[20] This finding is in stark contrast with the results of (Gürerk et al., 2014), where cooperation under exogenous assignment never moves beyond 60% of the endowment under peer punishment while in our setting – where normative requests are possible – cooperation under Coordinated Peer Punishment reaches 100%.

Result 5 (Culture of Cooperation):

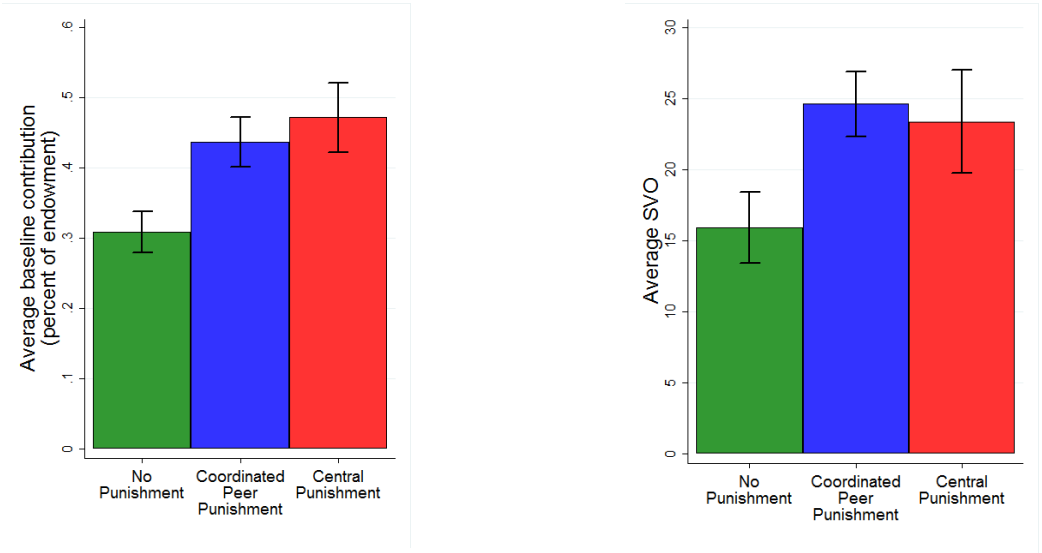
(a) Under endogenous sorting, prosocial individuals are quick to migrate into Coordinated Peer Punishment and Central Punishment institutions and establish a culture of high cooperation.

(b) The culture of cooperation is characterized by a strong consensus from the very beginning that subjects should make high contributions and widespread obedience with this social norm.

(c) Both punishment institutions successfully integrate migrants into their cooperative culture by immediately inducing them to make high contributions to the public good.

(d) In the centralized punishment institution with endogenous sorting, this culture of cooperation also includes the selection of very prosocial individuals as authorities, whereas the chosen authorities are much less prosocial under exogenous assignment.

Figure 8 provides evidence for Result 5a. The figure uses two different measures of prosociality. In Figure 8a, we use subjects' average cooperation level during the baseline public goods game in Part 1 of the experiment. This measure makes sense because complete free-riding was in subjects' self-interest in Part 1 where no sanctions for free-riding were possible. The average cooperation level in Part 1 is therefore a good proxy for a subject's prosociality. As a robustness check for our findings, we also use subjects' Social Value Orientations (SVOs), that we solicited in Part 3, as a measure of prosociality.

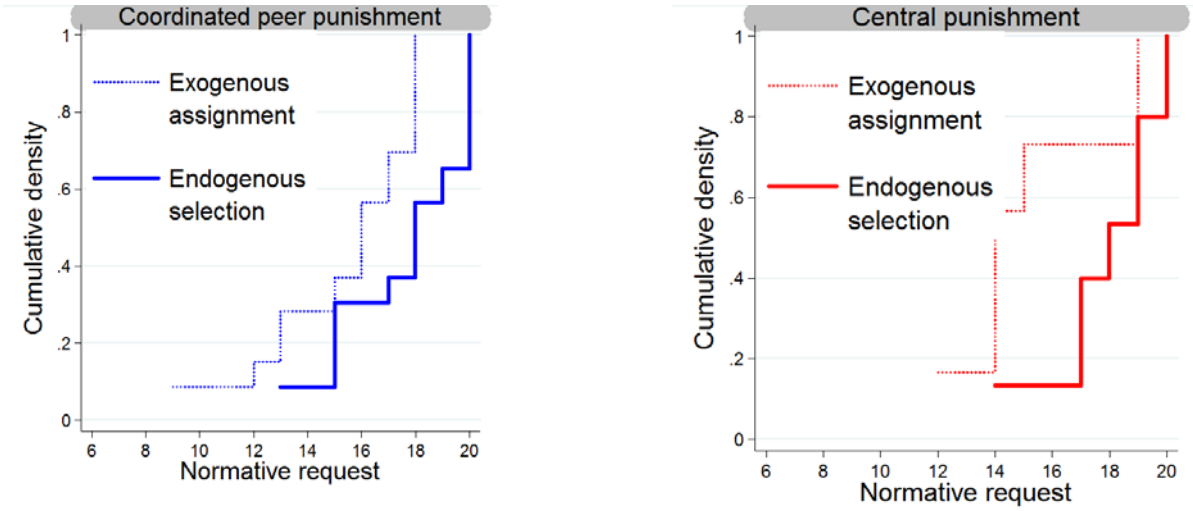


(a) average baseline contribution **(b) Social Value Orientation**

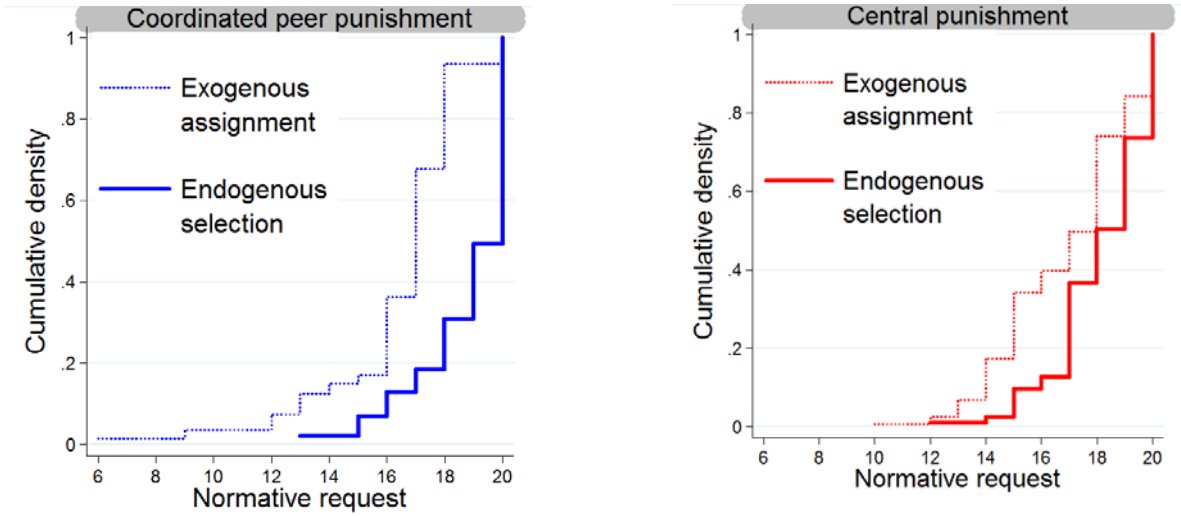
Figure 8: Assortment in first period of Part 1 when subjects could endogenously select into institutions. Error bars represent standard error of the mean.

Figure 8a shows that higher average contributors in Part 1 are more likely to adopt coordinated peer punishment and centralized punishment institutions in the first period of endogenous selection. These differences are also statistically significant (MW-tests based on individuals' average contributions in Part 1 as the unit of analysis; PP > NP, p=0.012; CP > NP, p=0.015). Figure 8b and the associated test shows a very similar result if we use each subject's SVO as a measure of prosociality (MW-test; PP and CP > NP,

p=0.021).[21] Result 5b states that subjects in the endogenous selection treatment quickly reached a consensus that everybody should make very high contributions. Support for this result is given in Figure 9 below.



(a) Period 1.



(b) Periods 1-5.

Figure 9: Cumulative density functions of groups' average normative requests in (a) the first period and (b) the first five periods. In all cases, the normative request under endogenous selection first-order stochastically dominates the normative request under exogenous assignment. Density functions are based on individual normative requests in each period.

[21] Since not all subjects completed Part 3 of the experiment on-line, we have fewer SVO observations. We therefore pool the PP and CP institutions for the statistical test. See also Section 1.4 for lack of evidence for bias in completing Part 3. For those who completed Part 3, 30 subjects selected NP in the first period of Part 2, 33 selected PP, and 19 selected CP – a distribution that is very similar to the overall distribution in period 1 (see Figure 4). In addition, the baseline contributions of those who select into the PP and the CP in period 1, and their SVO scores, are very similar (Figure 7a and 7b), so that it appears justified to pool the SVO scores across the two punishment institutions for the statistical test.

Figure 9 shows that individuals in the coordinated peer punishment and centralized punishment institutions make higher normative requests than members of the same institution under exogenous assignment in the early periods. Figure 8 displays the cumulative density functions for groups' average normative requests[22] in both institutions under endogenous selection and exogenous assignment for (a) the first period and (b) the first five periods. In all cases, the normative request under endogenous selection first-order stochastically dominates the normative request under exogenous assignment.

In the first five periods of the endogenous selection treatment, more than 80% of the subjects in PP had a normative request of 18 or more, while only about 30% of the subjects had similarly high requests under exogenous assignment. Likewise, almost 90% of subjects in the CP institution requested a contribution level of 17 or more when they self-selected into this institution, while only 60% had similarly high requests under exogenous assignment. These results show that the vast majority of subjects requested very high contribution levels right from the beginning under endogenous selection, indicating a strong social consensus about what is normatively expected from everybody in the group, but a substantial number of subjects were satisfied with lower requests under exogenous assignment.

Requesting a high contribution is one thing, but obeying these requests may well be another. We therefore next examine whether the average normative requests were actually obeyed as claimed in Result 5b. For this, we restrict attention to the first five periods of Part 2 because contributions tend towards full contributions quickly and eliminate any variation in the data afterwards. To provide support for Result 4b, we regress subjects' contributions levels on a constant and on the average normative request in a group. In addition, we use a restricted model in which we set the coefficient on the constant equal to zero (i.e. $\beta_0 = 0$). The restricted model appears justified because we observe relatively few low normative requests and cooperation levels so that the estimate of the constant is based on only few observations in the relevant range. In fact, we cannot reject the restricted model based on a likelihood ratio test in 3 of the 4 conditions. The restricted model can only be rejected in favor of the unrestricted in Central Punishment under endogenous selection. Our regression results are reported in Table 3 below.

The remarkable fact in Table 3 is that in the restricted model the coefficients on the normative request are always very close to one. They range from 0.968 to 1.001, indicating that groups' average normative requests are clear coordination devices leading to contributions that are not significantly different from the request. Even if we allow the constant to deviate from zero, the slope on the normative request is again close to one – except for the condition under Central Punishment with endogenous selection[23] – and the constant term is always insignificant. Thus, overall a high obedience with the average normative request prevailed in both punishment institutions and under both treatment conditions, but the normative requests were significantly higher, and less dispersed, under endogenous selection.

[22] The cumulative distributions functions of individuals' normative requests are shown in appendix A4.

[23] However, even in this condition, the predicted contributions for commonly observed values of the normative request (i.e., for requests from 16-20) almost perfectly match the normative request. For example, with a normative request of 16, the predicted contribution is 16.669 (= 5.693 + 16*0.686). With a normative request of 20, the predicted contribution is 19.413 (=5.693 + 20*0.686).

	Endogenous Selection		Exogenous Assignment	
	Peer Punishment	Central Punishment	Peer Punishment	Central Punishment
Panel A – Restricted model (constant constrained to zero)				
Normative request	0.981	0.997	1.010	0.968
	(0.017)***	(0.019)***	(0.020)***	(0.014)***
Panel B – Unrestricted model				
Normative request	1.078	0.686	1.144	1.073
	(0.170)***	(0.210)	(0.109)***	(0.066)***
Constant	-1.827	5.693	-2.253	-1.803
	(3.394)	(3.683)	(1.895)	(1.314)
Wald test, $\chi^2(1)$	0.290	2.390	1.414	1.882
	(0.590)	(0.122)	(0.234)	(0.170)
N	201	292	201	292
Number of clusters	11	11	11	11
Bootstrap samples	9,999	9,999	9,999	9,999

Table 3. The role of normative requests as a coordination device for contributions in the first five periods of Part 2. OLS regressions with robust standard errors clustered by matching group. Bias-corrected accelerated (BCA) standard errors using 9,999 bootstrapped samples are given in parentheses. Wald test is unable to reject the null hypothesis that the constant term is zero; values of chi-squared are reported with p-values for the Wald test given in parentheses. *p<0.10, **p<0.05, ***p<0.01.

Because the prosocial subjects tend to be among the first to enter the punishment institutions, the question arises how these institutions manage to integrate the more selfish individuals into their cooperative culture. Previous studies have shown that the migration of selfish individuals into cooperative groups may undermine cooperation (Ehrhart & Keser, 1999). However, Result 5c states that both punishment institutions successfully integrated migrants into their cooperative culture and immediately induced them to make high contributions. Figure 10 below shows that this was indeed the case. The figure displays subjects' contribution types *after* entering a punishment institution as a function of contribution type in the period immediately *before* entering the institution.

The figure shows that under endogenous selection 95% of the migrants into the punishment institution were low or moderate contributors before they entered; 87% of the migrants under exogenous assignment were previously low and moderate contributors. However, approximately 70% of the subjects

who were previously low contributors under endogenous selection (far left in Figure 10) immediately become high contributors upon entering a punishment institution. The figure also shows that the corresponding number is much lower (roughly 40%) under exogenous assignment. Taken together, these findings show that the big majority of previous low contributors is quickly integrated under endogenous selection into the cooperative culture prevailing in our punishment institutions.

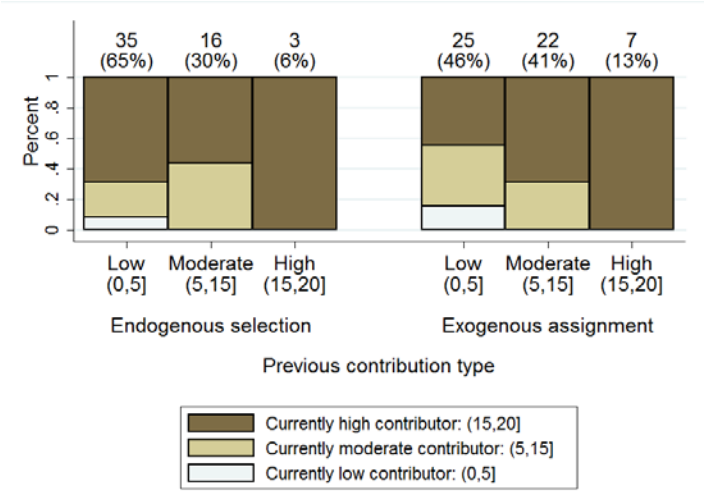
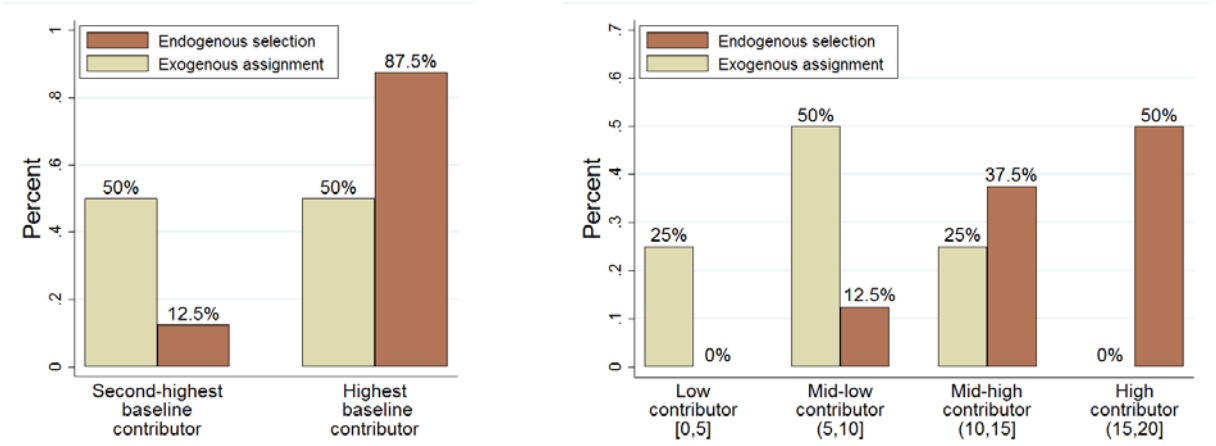


Figure 10: Contribution behavior immediately before and after entering a punishment institution. Horizontal axis indicates the contribution type in the No Punishment institution in the last period before joining a punishment institution. Bar height indicates relative frequency of contribution type upon entering a punishment institution conditional on type before migrating. Number above bars indicates the overall number of subjects from that treatment who fall into the category. The number in parentheses indicates the relative share of these subjects in the treatment population.

The quick establishment of a cooperative culture under endogenous sorting is further confirmed by the selection of prosocial authorities (Results 4d) in Central Punishment. This result is displayed in Figure 11, which shows the prosociality characteristics of the authorities selected in the first period of Part 2. In the overwhelming majority of cases (87%), the authority to punish was delegated to the highest baseline (Part 1) contributor when endogenous selection was possible, while this occurs only 50% of the time under exogenous assignment (Figure 11a).

Furthermore, Figure 10b illustrates that the selected authorities' baseline cooperation levels are very different between treatments. Under endogenous selection, 50% of the initial authorities contributed on average at least 16 points to the group project during Part 1, and 87.5% contributed at least 11 points. In contrast, 75% of the initial authorities in the exogenous assignment treatment contributed on average half of their endowment (10 points) or less. Subjects under endogenous selection thus not only select the highest contributor most of the time (Figure 10a), but the chosen authorities' baseline cooperation level is much higher compared to that of the elected authorities under exogenous assignment (Figure 11b). This means that endogenous selection ensured from the first period onwards that prosocial subjects were selected for the authority position in Central Punishment.



(a) Contribution rank of the authority (b) Contribution type of the authority

Figure 11: Baseline contribution behavior in Part 1 for subjects elected as central authority in first period of Part 2.

Taken together, these results suggest that subjects selecting punishment institutions under endogenous selection quickly established a cooperative culture. These institutions attracted a high share of prosocial individuals who quickly established high normative requests; these requests successfully coordinated the whole group to high contribution levels so that little punishment may have been necessary to enforce the high and widely agreed upon contribution norm. The immediate selection of very prosocial authorities further supports the cooperative culture in Central Punishment. In contrast, exogenous assignment of subjects to institutions puts sand into the gears of cooperation. It prevents the self-selection of prosocial individuals, leads to a lower prosociality of the selected authorities, and causes substantial adjustment costs during the initial phases because subjects demand lower contributions and cooperate less, which then requires higher punishment costs to establish cooperation.

2.3 Punishment in a culture of cooperation

What role did actual punishment of free-riders play in the quick establishment and compliance with a culture of cooperation? To what extent do the two punishment institutions differ with regard to the sanctioning of free-riders? Are these institutions capable of removing or diminishing the problem of antisocial punishment, i.e., the sanctioning of those who contribute more than the group average to the public good? And how does endogenous selection affect the punishment behavior? Our next result provides answers to these questions.

Result 6 (Prosocial and antisocial Punishment):

(a) Regardless of the sorting mechanism, the sanctioning probability for free-riders, i.e. those who contribute below the group's average, is generally very high. However, it is higher under Peer Punishment compared to Central Punishment, and those who are sanctioned generally receive considerably harsher punishments in Peer Punishment.

(b) Regardless of the sorting mechanism, we observe a non-negligible frequency of antisocial punishment under peer punishment while, under Central Punishment, anti-social punishment is almost completely eliminated.

(c) Free-riders generally face a lower probability of being punished in the exogenous assignment

treatment.

Empirical support for Result 6a is provided by Figure 12, Figure 13, and Table 4. Figure 12 shows that free-riders who deviate more than 2 units from the group average are punished in 100% of the cases under Peer Punishment, while the punishment probability is somewhat lower under Central Punishment, but in most cases still substantially above 50%. Table 4 shows that across all levels of free-riding, the probability of punishment is 96% (69%) in Peer Punishment under endogenous selection (exogenous assignment) and 65% (41%) in Central Punishment under endogenous selection (exogenous assignment). The difference across punishment institutions is largely driven by the fact that there are more small deviations (between 0 and 2 units below the group average) under Central Punishment that are punished with a considerably smaller probability than under Peer Punishment. The strength of punishment for the punished is shown in Figure 13, which indicates that Peer Punishment imposes considerably stronger sanctions on free-riders. This result is further corroborated by econometric analyses in appendix A5.

	Endogenous Selection		**Exogenous Assignment**	
Contribution level	Peer Punishment	Central Punishment	Peer Punishment	Central Punishment
Below average	95.7	64.7	68.6	41.1
Average or above	8.1	2.6	8.6	2.9

Table 4: Probability of punishment for below-average and above average contributors based on data from all 20 periods. Under Peer Punishment this is the probability that at least one institution member punishes, under Central Punishment it is the probability of punishment by the judge.

Table 4 and figures 12 and 13 also provide insights into the prevalence and the strength of antisocial punishment across institutions as described in Result 6b. In line with the literature, we define antisocial punishment as the sanctioning of those who contribute above the group average. Table 4 shows that 8-9% of the positive deviations from the group average were sanctioned in Peer Punishment, and Figure 12 indicates that the probability of antisocial punishment even increases for stronger positive deviations from the group average.[24] This pattern contrasts sharply with the observations in Central Punishment, where Figure 12 illustrates that antisocial punishment is basically absent. Only 2-3% of the positive deviations were sanctioned in Central Punishment (Table 4 and Figure 12).

Finally, Table 4 and Figure 12 also provides support for Result 6c by showing that the probability of punishment is generally lower in the exogenous assignment treatment. This lower probability is largely driven by a much higher prevalence of small deviations (between 0 and 2 units below the group's average contribution) in the exogenous assignment treatment. Under Peer Punishment, for example, the number of small deviations is more than five times higher in the exogenous treatment (110 versus 21 cases). Likewise, under Central Punishment the number of small deviations is 245 under exogenous assignment but only 92 under endogenous selection. Both in PP and CP these small deviations are punished with much higher frequency under endogenous selection (Figure 12).

[24] The regression analyses in Appendix A5 show that under peer punishment with endogenous sorting this leads to a significant increase in the average payoff reduction for higher positive deviations from the group average.

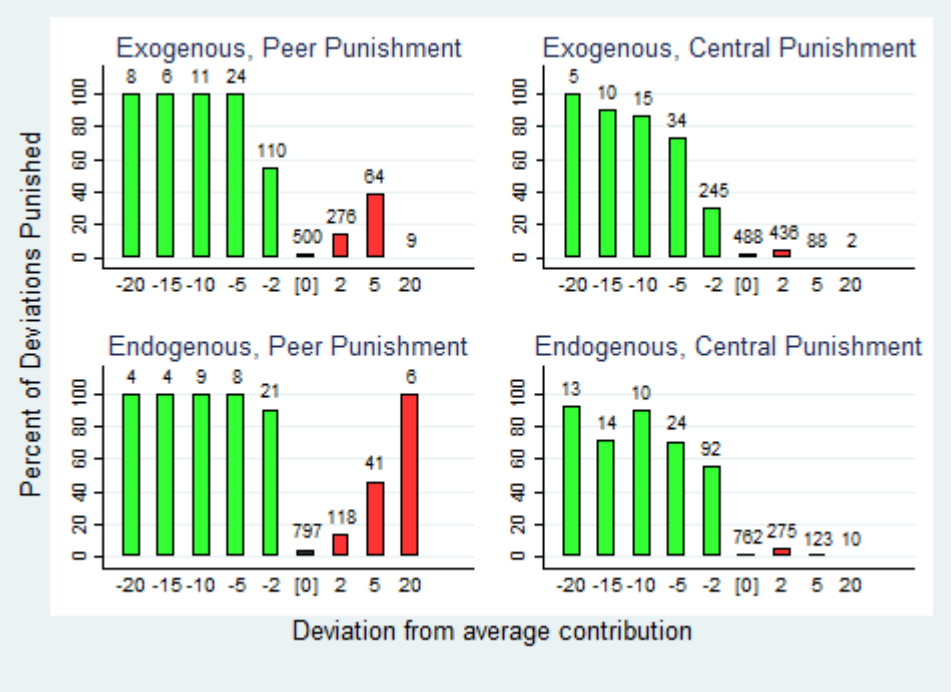


Figure 12: Probability of punishment for given deviations from the group's average contribution. The number above each bar indicates the absolute number of observations for that bar. The number of punishment cases for each bar can be computed by the probability of punishment times the number of observations for that bar.

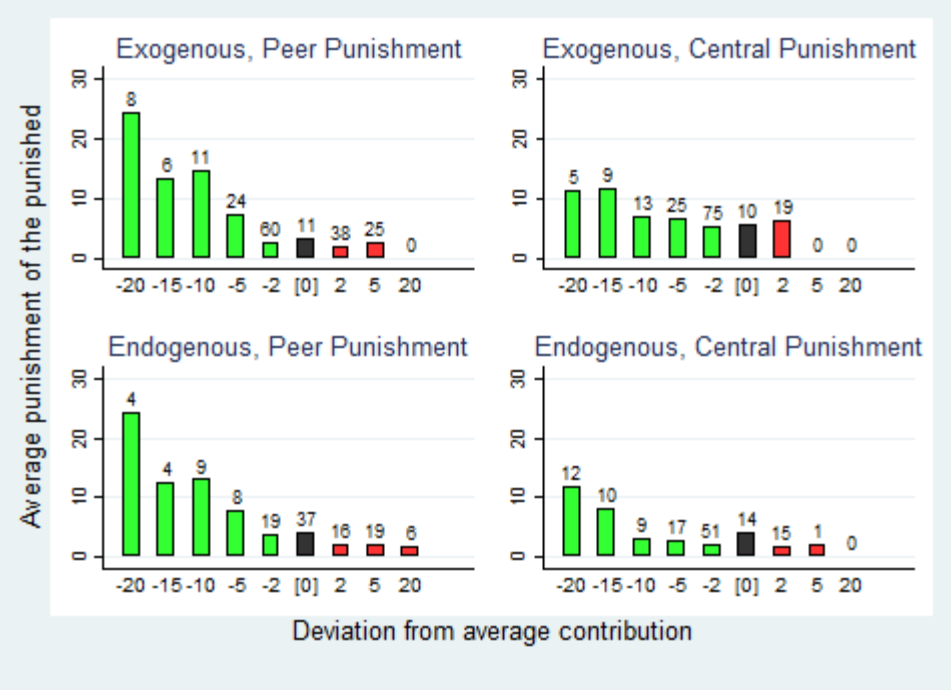


Figure 13: Average payoff reduction received by the punished for given deviations from the group's average contribution. The number above each bar indicates the absolute number of punishment cases for that bar.

Taken together, these sanctioning patterns suggest that the high probability of punishing free-riders under endogenous selection helped in quickly establishing and sustaining a culture of cooperation in this treatment, whereas the subjects under exogenous assignment not only established lower contribution norms

but minor free-riders also faced a lower probability of punishment, which led to a relatively large frequency of small negative deviations from the group average. Finally, Peer Punishment led to a harsher punishment of free-riders compared to Central Punishment, but at the same time, there is also a higher probability of anti-social punishment when peers can sanction. In fact, antisocial punishment was basically non-existent in Central Punishment, which may well be a result of the selection of prosocial authorities in this treatment.

3 Discussion and concluding remarks

The threat of peer sanctioning looms large in most human societies. The currency in which they are expressed are manifold and range from a raised eyebrow to ridicule, from verbal outrage to corporal punishment, from harassment to ostracism. A large empirical literature indicates that peer sanctioning often generates large collateral damage in the form of excessive or ill-guided punishment. The conclusions of this literature are typically based on what we have called "uncoordinated peer punishment", but in practice the threat of peer punishment often appears to be based on a wide normative consensus about what constitutes appropriate behavior. For this reason, we introduced the opportunity for normative consensus building into our experiment – an opportunity that has no behavioral effects in the absence of targeted sanctioning opportunities. In addition, we combine this with the possibility to migrate across communities with different sanctioning institutions. We allow, in particular, uncoordinated peer punishment to compete with coordinated peer punishment and with centralized punishment by an elected authority – both institutions that allow for normative consensus building. It turns out that subjects universally reject uncoordinated peer punishment in favor of the two other punishment institutions. This strong revealed preference for normative consensus building suggests that previous research – including that by one of the co-authors of this paper – failed to take a potentially important characteristic of efficient peer punishment into account.

In fact, coordinated peer punishment and centralized punishment by an elected judge in combination with the migration opportunities leads to the quick emergence of an efficient and universal culture of cooperation. Prosocial individuals migrate first into these punishment institutions and they immediately use the consensus building opportunity to establish a social norm of high cooperation that is credibly enforced by peers or the judge so that relatively little actual punishment is necessary to enforce the norm. Therefore, these two sanctioning institutions become quickly more efficient than the no punishment institution.

However, when we remove the opportunity for normative consensus building, many more subjects are initially reluctant to enter the peer punishment institutions and preferred to be in the no punishment institution. Moreover, the absence of normative consensus building also causes a reduction in the efficiency of peer punishment for an extended period of time and a permanent decrease in the popularity of peer punishment that is associated with a long run shift towards centralized punishment. These finding indicate the importance of a normative consensus for the efficiency and the success of peer punishment in attracting subjects.

In our view, these results are not only interesting for the literature on public goods and sanctioning systems, but they also may inform the literature on the evolution of human cooperation by suggesting to take biased migration and normative consensus building as potentially important mechanisms into

account (Boyd et al., 2010; Boyd & Richerson, 2009). Likewise, by providing insights into the mechanisms of social norm formation, the literature on the role of culture and social norms in economic life (e.g., Benabou & Tirole, 2011; Bowles, 1998; Giuliano & Nunn, 2013; Guiso et al., 2006; Lowes et al., 2017; Nunn, 2014; Tabellini, 2008) may benefit.

Our results raise, of course, also many new questions that could be pursued by future research. Recent papers (Buffat & Senn, 2017; Muthukrishna, Francois, Pourahmadi, & Henrich, 2017) have pointed out that those who punish under centralized punishment may be subject to the temptations of corruption, which then induces dysfunctional sanctioning patterns. Institutional competition via voting with the feet mechanisms, normative consensus building, and the election of judges may also prove to be important in combatting the corruption of leaders and judges in centralized sanctioning systems. Another literature has pointed out the potential relevance of counter-punishment opportunities (Denant-Boemont, Masclet, & Noussair, 2007; Nikiforakis, 2008) within a given period, i.e., the possibility of the sanctioned individuals to immediately strike back in the period in which the sanction occurred. Previous work on counter-punishment neither allowed for normative consensus building nor for migration across groups, but both of these institutional mechanisms possibly constrain the relevance of counter-punishment. The rationale for this conjecture is two-fold. First, the informally binding nature of the social norm could also put constraints on counter-punishment by the free-riders. Second, due to prosocial migration and the norm consensus opportunity, there is little punishment to begin with that could trigger counter-punishment.[25] Finally, it is worthwhile to point out that counter-punishment *across periods* was not ruled out in our set-up because group members knew subjects' ID numbers. It was, for example, possible to counter-punish a judge in the next period because the judge's ID number was known. But this never occurred.

Finally, it would be interesting to examine how consensus building and migration affect the relative performance of different sanctioning systems when information about others' behavior is imperfect and costly to acquire. Which normative standards of evidence emerge endogenously in such an environment, and how do various sanctioning systems differ in this regard? Studying this environment in the context of our set-up raises a whole set of exciting new questions.

[25] In Fehr and Schurtenberger (2017), normative consensus building is possible in a counter-punishment set-up. They show that the norm consensus opportunity significantly improves cooperation and efficiency under peer punishment despite the opportunity for direct counter-punishment. In addition, they show that counter-punishment is itself affected by normative concerns: the presence of a norm consensus opportunity encourages the victims of antisocial punishment (i.e., above average contributors who were punished) to strike back, which indeed constrains antisocial punishment.

Appendix A1

Table A1a
Distribution of subjects across institutions over time in the presence of normative requests

Period	No Punishment	Peer Pun	Coordinated Peer Pun	Coordinated Central Pun
1	0.37	0.02	0.36	0.26
2	0.24	0.04	0.38	0.34
3	0.14	0.03	0.34	0.48
4	0.11	0.04	0.24	0.61
5	0.07	0.01	0.30	0.63
6	0.08	0.02	0.41	0.50
7	0.07	0.02	0.38	0.54
8	0.06	0.02	0.39	0.52
9	0.03	0.01	0.42	0.54
10	0.03	0.01	0.41	0.55
11	0.02	0.00	0.44	0.54
12	0.03	0.00	0.43	0.54
13	0.01	0.00	0.42	0.57
14	0.01	0.00	0.41	0.58
15	0.01	0.00	0.41	0.59
16	0.01	0.01	0.45	0.54
17	0.02	0.00	0.46	0.52
18	0.02	0.00	0.47	0.52
19	0.03	0.00	0.43	0.54
20	0.04	0.00	0.41	0.55

Table A1b

Distribution of subjects across institutions over time in the absence of normative requests

Period	No Punishment	Peer Punishment	Centralized Punishment
1	0.63	0.16	0.21
2	0.55	0.10	0.35
3	0.39	0.14	0.47
4	0.38	0.17	0.44
5	0.31	0.17	0.51
6	0.30	0.20	0.50
7	0.20	0.16	0.64
8	0.15	0.19	0.65
9	0.16	0.22	0.63
10	0.11	0.24	0.65
11	0.17	0.17	0.64
12	0.15	0.18	0.67
13	0.08	0.19	0.72
14	0.05	0.24	0.72
15	0.06	0.22	0.72
16	0.06	0.23	0.71
17	0.05	0.26	0.69
18	0.04	0.27	0.69
19	0.03	0.28	0.69
20	0.06	0.27	0.67

Appendix A2

Distribution of subjects across institutions over time within the individual matching groups

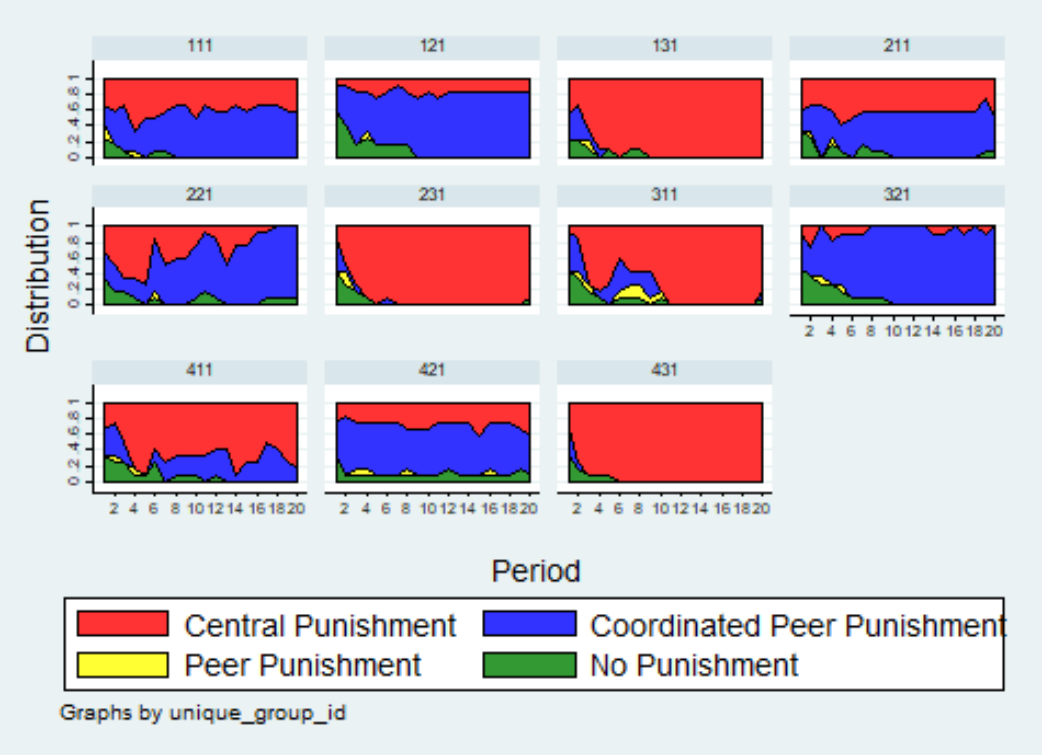


Figure A2. The figure shows how subjects selected themselves into the different institutions within the matching groups. Numbers above each graph indicate the matching group ID. Subjects predominantly preferred the centralized punishment institution in 5 out of eleven matching groups; in four of them even all subjects migrated to CP. In six of the eleven matching groups the subjects predominantly preferred coordinated peer punishment.

Appendix A3

Cooperation and efficiency with and without normative requests in PP and CP

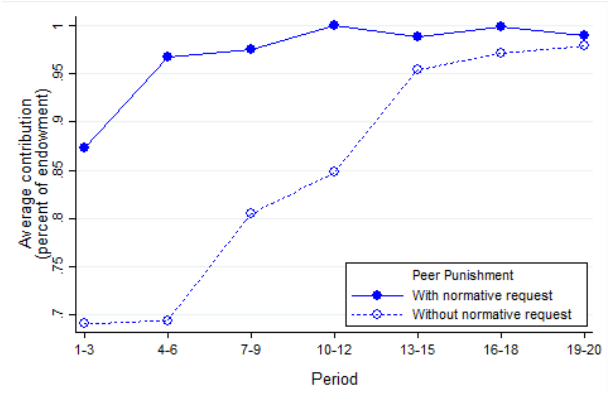


Figure A3a. Contributions over time in Peer Punishment

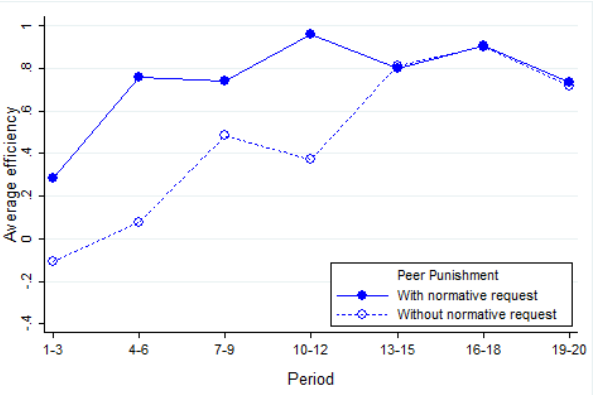


Figure A3b. Efficiency over time in Peer Punishment

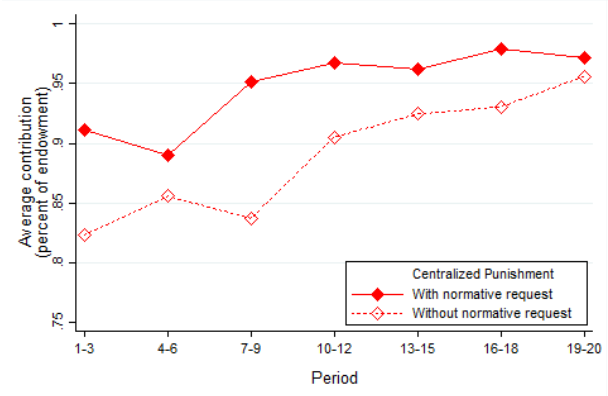


Figure A3c. Contributions over time in Centralized Punishment

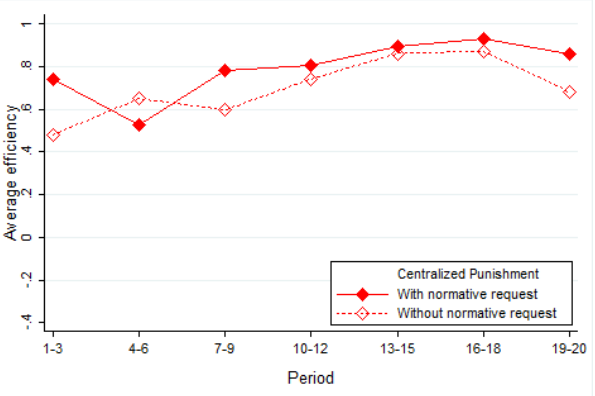
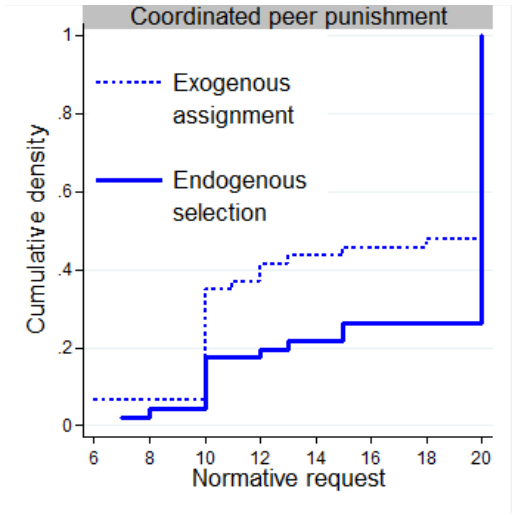


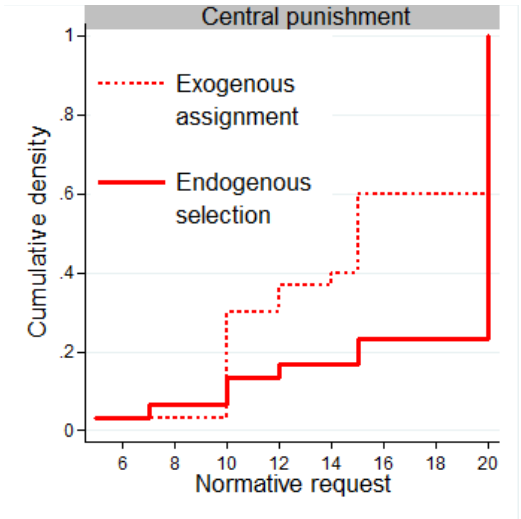
Figure A3d. Efficiency over time in Centralized Punishment

Figure A3a and A3b suggest that cooperation and efficiency is higher under PP with normative requests compared to PP without normative requests for the first 12 periods. Mann Whitney tests with average cooperation and average efficiency per matching group (for the first 12 periods) as units of observation confirm this (cooperation: $p = 0.005$; efficiency: $p = 0.020$). In contrast, in the case of CP, cooperation and efficiency are not significantly different with and without normative request when we take average observation for the first 12 periods as units of observation (cooperation: $p = 0.539$; efficiency: $p = 0.667$). Instead, under CP we find only a (weakly) significant difference during the first three periods (cooperation: $p = 0.094$; efficiency: $p = 0.020$) but not beyond these initial periods.

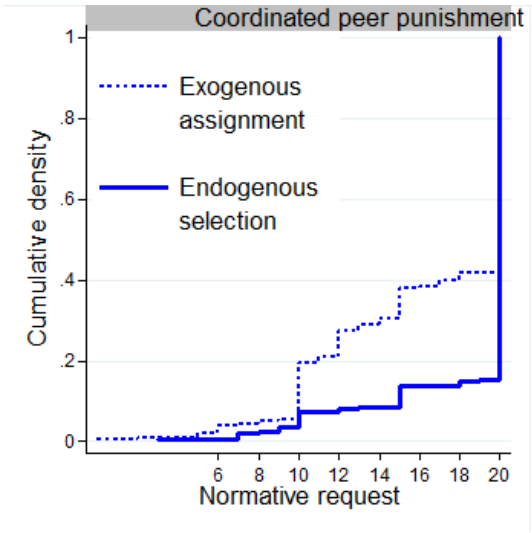
Appendix A4
Distribution of <u>individual</u> normative requests in different treatments and institutions



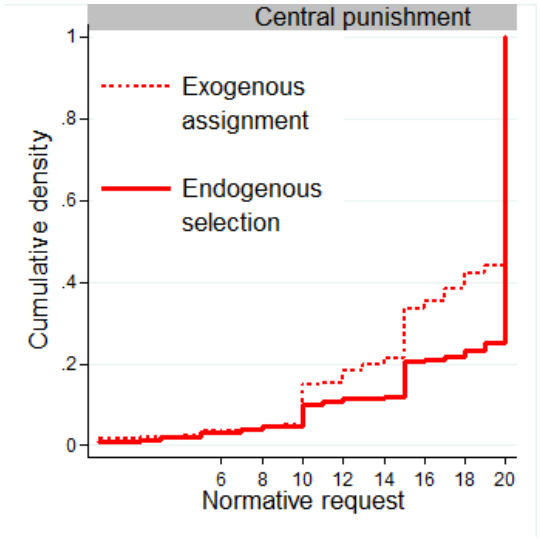
A2a. Individual Normative Requests in Coordinated Peer Punishment (period 1)



A2b. Individual Normative Requests in Coordinated Central Punishment (period 1)



A2c. Individual Normative Requests in Coordinated Peer Punishment (periods 1-5)



A2d. Individual Normative Requests in Coordinated Central Punishment (periods 1-5)

Appendix A5

Econometric Analysis of Punishment Patterns

In the table below, we use Tobit regressions to study the impact of subjects' negative and positive deviations from the group's average contribution on the received punishment. We also used negative and positive deviations from a group's average normative request as regressors, but this model performed substantially worse than the model with deviations from the average contribution.

The table shows that the coefficient for negative deviations from the average contribution is always positive – regardless of treatment and punishment institution. Thus larger negative deviations are more strongly punished. It also shows that under endogenous sorting negative deviations from the average are more harshly punished by peers than by the judge. Finally, we observe that larger positive deviations are more harshly punished under endogenous peer punishment – indicating antisocial punishment.

	Endogenous Selection		Exogenous Assignment	
	Peer Punishment	Central Punishment	Peer Punishment	Central Punishment
Own negative deviation from average contribution	2.302 (0.383)***	1.095 (0.173)***	2.027 (0.249)***	2.247 (0.406)***
Own positive deviation from average contribution	2.088 (0.500)***	-1.329 (18.094)	0.806 (0.535)	-4.718 (1.817)
Constant	-9.368 (1.777)***	-6.739 (1.052)***	-5.579 (0.943)***	-12.671 (4.485)***
Left-censored	886	1,194	825	1,167
Uncensored	122	129	183	156
N	1,008	1,323	1,008	1,323
Number of clusters	11	11	11	11
Bootstrap samples	9,999	9,999	9,999	9,999

Table A5. Punishment received based on deviation from average group contribution during the period. Tobit regressions with robust standard errors clustered by matching group. Bias-corrected accelerated (BCA) standard errors using 9,999 bootstrapped samples are given in parentheses[26]. ***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

[26] We apply bootstrapping to compute standard errors because we have a relatively small number of matching groups. Without bootstrapping inconsistent standard errors and a tendency to over-reject the null would arise (Bertrand, Duflo, & Mullainathan, 2004). We use a pairs cluster bias-corrected accelerated (BCA) bootstrap with cluster robust standard errors as proposed by Cameron, Gelbach, and Miller (2008) that performs well in their simulations. In an earlier version of the paper, we used a pairs cluster bootstrap-t procedure which performed comparably in Cameron, Gelbach, and Miller (2008)'s simulations. We had also reported average partial effects rather than Tobit coefficients; interested readers can find the average partial effects in the replication log file for this table.

References

Acemoglu, D., & Wolitzky, A. (2016). *Sustaining Cooperation: Community Enforcement vs. Specialized Enforcement*.

Andreoni, J., & Gee, L. K. (2012). Gun for hire: Delegated enforcement and peer punishment in public goods provision. *Journal of Public Economics*, *96*(11–12), 1036–1046.

Andreoni, J., & Miller, J. (2002). Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism. *Econometrica*, *70*(2), 737–753.

Benabou, R., & Tirole, J. (2011). Laws and Norms. National Bureau of Economic Research working Paper 17579.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? *Quarterly Journal of Economics*, *119*(1), 249–275.

Bochet, O., Page, T., & Putterman, L. (2006). Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior and Organization*, *60*(1), 11–26.

Bochet, O., & Putterman, L. (2009). Not Just Babble: A Voluntary Contribution Experiment with Iterative Numerical Messages. *European Economic Review*, *53*(3), 309–326.

Boehm, C. (2001). *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Harvard University Press.

Boehm, C., Antweiler, C., Eibl-Eibesfeldt, I., Kent, S., Knauft, B. M., Mithen, S., … Wilson, D. S. (1996). Emergency Decisions, Selection Mechanics, and Group Selection [and Comments and Reply]. *Current Anthropology*, *37*(5), 763–778.

Boehm, C., Barclay, H. B., Dentan, R. K., Dupre, M.-C., Hill, J. D., Kent, S., … Rayner, S. (1993). Egalitarian Behavior and Reverse Dominance Hierarchy [and Comments and Replies]. *Current Anthropology*, *34*(3), 227–254. https://doi.org/10.1086/204166

Bowles, S. (1998). Endogenous preferences: The cultural consequences of markets and other economic institutions. *Journal of Economic Literature*, *36*(1), 75–111.

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, *328*(5978), 617–20.

Boyd, R., & Richerson, P. J. (2009). Voting with your feet: Payoff biased migration and the evolution of group beneficial behavior. *Journal of Theoretical Biology*, *257*(2), 331–339.

Buffat, J., & Senn, J. (2017). *Corruption and Cooperation*.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics*, *90*(3), 414–427.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics*, *14*(1), 47–83.

Chudek, M., & Henrich, J. (2011). Culture-gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences*, *15*(5), 218–226.

Denant-Boemont, L., Masclet, D., & Noussair, C. N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory*, *33*(1), 145–167. https://doi.org/10.1007/s00199-007-0212-0

Dixit, A. K. (2007). *Lawlessness & Economics: Alternative Modes of Governance*. Princeton University Press.

Dreber, A., Rand, D. A., Fudenberg D. & Nowak M. A. (2008). Winners don't punish. *Nature* 452, 348-351.

Ehrhart, K.-M., & Keser, C. (1999). *Mobility and Cooperation: On the Run*. *CIRANO Scientific Series 99s-24*.

Ertan, A., Page, T., & Putterman, L. (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, *53*(5), 495–511.

Eshel, I., & Cavalli-Sforza, L. L. (1982). Assortment of encounters and evolution of cooperativeness. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(4), 1331–1335.

Fafchamps, M. (1996). The Enforcement of Commercial Contracts in Ghana. *World Development*, *24*(3), 427–448.
Fehr, E., & Gächter, S. (2000). Cooperation and Punishment in Public Goods Experiments. *American Economic Review*, *90*(4), 980–994.
Fehr, E., & Henrich, J. (2003). Is strong reciprocity a maladaptation? On the evolutionary foundations of human altruism. In P. Hammerstein (Ed.), *Genetic and Cultural Evolution of Cooperation*. MIT Press.
Fehr, E., & Schurtenberger, I. (2017). *The Dynamics of Norm Formation and Norm Decay*. Working Paper, Department of Economics, University of Zurich.
Fischbacher, U. (2007). Z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178.
Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in Simple Bargaining Experiments. *Games and Economic Behavior*.
Francis, H. (1985). The Law , Oral Tradition and the Mining Community. *Journal of Law and Society*, *12*(3), 267–271.
Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science*, *322*(5907), 1510.
Giuliano, P., & Nunn, N. (2013). The Transmission of Democracy: From the Village to the Nation-State. *American Economic Review*, *103*(3), 86–92.
Greif, A. (1989). Reputation and Coalitions in Medieval Trade: Evidence on the Maghribi Traders. *Journal of Economic History*, *49*(4), 857–882.
Greif, A. (1993). Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition. *American Economic Review*, *83*(3), 525–548.
Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114–125.
Guiso, L., Sapienza, P., & Zingales, L. (2006). Does Culture Affect Economic Outcomes? *Journal of Economic Perspectives*, *20*(2), 23–48.
Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2006). The Competitive Advantage of Sanctioning Insitutions. *Science*, *312*(2006), 108–11.
Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2014). On cooperation in open communities. *Journal of Public Economics*, *120*, 220–230.
Henrich, J., Chudek, M., & Boyd, R. (2015). The Big Man Mechanism: how prestige fosters cooperation and creates prosocial leaders. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1683), 20150013.
Hermann, B., Thöni, C., & Gächter, S. (2008). Antisocial Punishment Across Societies. *Science*, *319*, 1362–1367.
Hill, K. R., Walker, R. S., Božičević, M., Eder, J., Headland, T., Hewlett, B., … Wood, B. (2011). Co-Residence Patterns in Hunter-Gatherer Societies Show Unique Human Social Structure. *Science*, *331*(6022), 1286–1289.
Isaac, R. M., Walker, J. M., & Williams, A. W. (1994). Group size and the voluntary provision of public goods. Experimental evidence utilizing large groups. *Journal of Public Economics*, *54*(1), 1–36.
Johnson, S., McMillan, J., & Woodruff, C. (2002). Courts and Relational Contracts. *Journal of Law, Economics, and Organization*, *18*(1), 221–277.
Kaplan, H., & Gurven, M. (2005). The Natural History of Human Food Sharing and Cooperation: A Review and a New Multi-Individual Approach to the Negotiation of Norms. In H. Gintis, S. Bowles, R. Boyd, & E. Fehr (Eds.), *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life* (pp. 75–113). MIT Press.
Knauft, B. M., Abler, T. S., Betzig, L., Boehm, C., Dentan, R. K., Kiefer, T. M., … Rodseth, L. (1991). Violence and Sociality in Human Evolution [and Comments and Replies]. *Current Anthropology*, *32*(4), 391–428.
Kosfeld, M., Okada, A., & Riedl, A. (2009). Institution Formation in Public Goods Games. *American Economic Review*, *99*(4), 1335–1355.
Kranton, R. E. (1996). Reciprocal Exchange: A Self-Sustaining System. *American Economic Review*,

86(4), 830–851.
Lowes, S., Nunn, N., Robinson, J. A., & Weigel, J. L. (2017). The Evolution of Culture and Institutions: Evidence From the Kuba Kingdom. *Econometrica*, *85*(4), 1065–1091.
Markussen, T., Putterman, L., & Tyran, J. R. (2014). Self-organization for collective action: An experimental study of voting on sanction regimes. *Review of Economic Studies*, *81*(1), 301–324.
Masclet, D., Noussair, C., Tucker, S., & Villeval, M.-C. (2003). Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism. *American Economic Review*, *93*(1), 366–380.
Mathew, S., & Boyd, R. (2011). Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences*, *108*(28), 11375–11380.
McMillan, J., & Woodruff, C. (1999). Interfirm Relationships and Informal Credit in Vietnam. *Quarterly Journal of Economics*, *114*(4), 1285–1320.
Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. J. (2011). Measuring Social Value Orientation. *Judgment and Decision Making*, *6*(8), 771–781.
Muthukrishna, M., Francois, P., Pourahmadi, S., & Henrich, J. (2017). Corrupting cooperation and how anti-corruption strategies may backfire. *Nature Human Behaviour*, *1*(7), 138. Retrieved from http://www.nature.com/articles/s41562-017-0138
Nalbantian, H. R., & Schotter, A. (1997). Productivity Under Group Incentives : An Experimental Study, *87*(3), 314–341.
Nicklisch, A., Grechenig, K., & Thoeni, C. (2016). Information-sensitive Leviathans. *Journal of Public Economics*, *144*, 1–13.
Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics*, *92*(1–2), 91–112.
Nunn, N. (2014). *Historical Development*. (P. Aghion & S. N. Durlauf, Eds.), *Handbook of Economic Growth* (Vol. 2A). Elsevier B.V.
O'Gorman, R., Henrich, J., & van Vugt, M. (2009). Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society. Series B, Biological Sciences*, *276*(1655), 323–9.
Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives*, *14*(3), 137–158.
Rand D. G., Dreber A., Elingsen T., Fudenberg D., Nowak M. A. (2009). Positive interactions promote public cooperation. *Science* 325, Issue 5945, 1272-1275.
Rehder, R. R. (1990). Japanese transplants: After the honeymoon. *Business Horizons*, *33*(1), 87–98.
Roethlisberger, F., & Dickson, W. J. (1939). *Management and the worker: an account of a research program conducted by the Western electric company, Hawthorne works, Chicago*. Harvard University Press.
Simkins, P. (1988). *Kitchener's Army: The Raising of the New Armies, 1914-1916*. Manchester United Press.
Sober, E., & Wilson, D. S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press.
Sutter, M., Haigner, S., & Kocher, M. G. (2010). Choosing the stick or the carrot? Endogenous institutional choice in social dilemma situations. *Review of Economic Studies*, *77*(4), 1540–1566.
Tabellini, G. (2008). Institutions and Culture. *Journal of the European Economic Association*, *6*(2–3), 255–294.
Wiessner, P. (2005). Norm enforcement among the Ju/'hoansi Bushmen : A case of strong reciprocity? *Human Nature*, *16*(2), 115–145.
Wilson, R. K., & Sell, J. (1997). "Liar, Liar..." Cheap Talk and Reputation in Repeated Public Goods Settings. *Journal of Conflict Resolution*, *41*(5), 695–717.
Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, *51*(1), 110–116.