Caste and Punishment

The Legacy of Caste Culture in Norm Enforcement*

Karla Hoff
The World Bank
khoff@worldbank.org

Mayuresh Kshetramade
Affinnova Inc, Waltham, MA
mayurvk@yahoo.com

Ernst Fehr
Laboratory for Social and Neural Systems Research
Department of Economics, University of Zurich
ernst.fehr@econ.uzh.ch

Well-functioning groups enforce social norms that restrain opportunism. Here we study how the assignment to the top or bottom of an extreme social hierarchy – the caste system – affects individuals' willingness to punish violations of a cooperation norm altruistically. We find that individuals at the bottom of the hierarchy exhibit a much lower willingness to punish norm violations that hurt members of their own caste. We can rule out self-selection into castes and control for wealth, education and political experience. Our results thus plausibly identify the impact of caste status on individuals' willingness to punish norm violations. The lower willingness to punish suggests that the low castes may be less able to enforce contracts and ensure their property rights, which disadvantages them with respect to trading opportunities and production incentives.

"When a Dalit argued with an upper caste farmer [over discrimination towards Dalits, formerly Untouchables]…, the upper caste villagers attacked 80 Dalit families in retaliation. When the same Dalit man then went to the police to report the incident, a social boycott was imposed on all of the Dalits from [his village]; they were thrown out of their village and denied every opportunity to earn their livelihood."

Tejeshwi Pratima, "Dalits thrown out of their village for raising their voice against discrimination," June 29, 2006 [1]

Every society requires restraints on opportunism. The conventional simplifying assumption in economics is that government provides these and that individuals obey them because it is in their self-interest to obey the law. More recently, scholars have questioned the usefulness of this convention, arguing that social norms are a key source of restraints and that without them even formal rules would be unenforceable.[2] Underneath the level of behavior that most of economics is concerned with are the *social capabilities* to constrain opportunism, underpinning even modern societies with well-developed legal institutions. Economic historians such as Deirdre McCloskey (2006) and Joel Mokyr (2009) have emphasized that such capabilities may play a large role in the enforcement of contracts and property rights and, thus, in economic development and growth.[3] As social norms are part of cultural traditions that may inhibit or enhance trade and production by affecting honesty, trust, trustworthiness, and cooperation, these arguments bring cultural factors into the focus of economists' attention.[4]

---

[1] Cited in New York University Center for Human Rights and Global Justice and Human Rights Watch, 2007, p. 60. (hereafter, NYU). The incident took place in June 2006,

[2] See, *e.g.,* Hayek (1973, ch. 2), Greif (1993), Platteau (1994), Weingast (1997), Lindbeck, Nyborg and Weibull (1999), Basu (2000), and Ostrom (2000).

[3] Mokyr argues that the unusual strength of Britain's social capability to punish dishonest behavior in business helped elevate Britain to the leading position in the Industrial Revolution. In 18th century Britain, "opportunistic behavior was made so taboo that in only a few cases was it necessary to use the formal institutions to punish deviants" (p. 384). "Entrepreneurial success was based less on multi-talented geniuses than on successful cooperation between individuals who had good reason to think they could trust one another." In this secure environment, "Boulton found his Watt, Clegg his Murdoch, Marshall his Murray, Muspratt his Gamble, and Cooke his Wheatstone" (p. 386).

[4] See, *e.g.*, Gintis 1972, 2008; Bowles 1998; Bowles and Gintis 1998; Henrich et al, 2001; Guiso, Sapienza and Zingales 2004, 2006; Banerjee and Iyer 2005; Fisman and Miguel 2006; Herrmann, Thöni and Gächter 2008; Tabellini 2008, Algan and Cahuc (2009), and Nunn and Wantchekon (forthcoming).

Social norms are enforced by informal sanctions that are often imposed by individuals even when sanctioning is costly and yields no material benefits to the punisher (Fehr and Gächter 2000; Fehr and Fischbacher 2004). In laboratory and real world settings, the efficiency and volume of trade in markets[5] and the ability of communities to undertake collective action[6] can critically depend on the altruistic willingness of individuals to punish those who do not keep their formal or informal obligations.

In this paper, we investigate how the social structure of a society affects people's willingness to sanction norm violations. For this purpose, we study how the assignment to the top or bottom of an extreme hierarchy—the Hindu caste system– affects individuals' willingness to punish violations of a cooperation norm. The Hindu caste system is a set of discrete communities that are ranked on the basis of "natural superiority" and in which notions of purity and pollution are embedded (Gupta, 1991, p. 2). For thousands of years, there were stark inequalities of social, economic and political rights between the castes at the top of the hierarchy (hereafter, the "high castes") and those at the extreme bottom (hereafter, the "low castes"). The high castes had basic freedoms, the power to extract forced labor from low castes, and high ritual status. The low castes could not sell their labor and goods in markets; were barred from schools, temples, and courthouses; and were relentlessly stigmatized through the practice of Untouchability (Shah *et al.* 2006). The low castes provided forced labor to high-caste individuals on demand, and the demand for cheap labor was a factor in the persistence of Untouchability (Bayley 1999). Members of these castes are today called Dalits, a non-pejorative term that literally means "oppressed" or "ground down."

---

[5] See Fehr, Gächter and Kirchsteiger (1997), who demonstrate that the willingness to punish shirkers strongly increases the gains from trade. Logan and Shah (2009) demonstrate in an illegal market—the male sex market—the power of informal policing to permit honest providers to signal their type. The enforcement in this market is decentralized and is altruistic in the sense that it yields no benefit to the individual who punishes. One enraged client for 10 years policed the web to warn others about a specific provider who had robbed him (personal communication of Trevon Logan)

[6] See Ostrom 2000, Miguel and Gugerty 2005, Rustagi, Engel, and Kosfeld 2010.

A priori, the assignment to extreme positions in a social hierarchy could affect informal norm enforcement in a variety of ways. The hierarchy could lead to conflict and hostility between groups, which would cause dysfunctional punishment in inter-group interactions. Alternatively, the everyday practice of power could turn the objects of repression into its subjects, which would lead them to tolerate violations by members of the dominant group but not by others. A third possibility is that a history of repression diminishes the repressed group's capability for altruistic third-party punishment.

We provide evidence in favor of the third hypothesis. In doing so, we extend Sen's notion of capabilities to include the capability to punish altruistically. Our evidence takes us considerably beyond correlation and suggests instead that people from low castes have different capabilities to engage in third-party punishment *because* they are low caste and not as a result of factors correlated with caste. Two features of the caste system are key to identifying this effect:

1. *Heritability of caste and rigidities of caste ranking at the extreme ends of the caste hierarchy.* An individual's caste is determined by the accident of birth, and individual mobility across castes is basically not possible in an individual's lifetime. In rural India, castes are endogamous. There are strong norms against cross-caste marriages. Marriages between high and low caste persons are particularly harshly punished and sometimes lead to "public lynching of couples or their relatives, murder (of the bride, groom or their relatives), rape, public beatings and other sanctions" (NYU 2007, p. 11). Although caste boundaries and caste rank can change over long time periods, the status of the *specific* high-ranked castes (Brahmin and Thakur) and low-ranked castes (Chamar and Pasi) from which we draw our experimental subjects goes back millennia (Gupta 2000). The absence of across-caste mobility for these groups rules out selection bias and enables us to study the impact of caste status on individuals' willingness to punish norm violations. The high

castes in our sample thus constitute a meaningful control group for the low castes in our sample.

2. *Overlapping wealth distributions.* Despite the vast difference in social status between high and low castes, there is today considerable individual variation in wealth, consumption, and education within the high and low castes.[7] We have not only poor low-caste subjects, but also many poor high-caste subjects in our sample. By controlling for individual differences in wealth and education, we can rule out that differences in punishment behavior across castes are caused by those individual differences.

We implemented a third-party punishment experiment with subjects from high and low castes. We use this experiment to study the strength of informal punishment of norm violators. The essence of such an experiment is that one player, whom we call player B, can obey or violate a social norm in an interaction with another player, player A. Then the third party, player C, learns what B did and has the opportunity to sanction him. In order to elicit C's preferences to sanction the norm violation, the punishment is costly for C. In our experiment, as our findings will clearly indicate, the social norm is that B reciprocates a cooperative choice of A in a sequential social dilemma game.

We report here data from our study of 205 triples composed of adult males in over 100 villages in one of the poorest states of India, Uttar Pradesh. We implemented four treatments, called HHH, HLH, LLL, and LHL. The first letter in each treatment indicates the caste status – high (H) or low (L) – of player A, the potential injured party; the second letter indicates the caste status of player B, the potential norm violator; and the third letter indicates the caste status of

---

[7] With the abolition of the Zamindari (landlord) system in Uttar Pradesh in 1952, many Dalit agricultural laborers acquired ownership rights of the lands that they had been cultivating.

player C, the third party punisher. The subjects in the experiment were informed about the caste status of the matched players in an unobtrusive way (see Section 1.3).

This design enables us to test for double standards in punishment and for differences between high-and low-caste individuals in the willingness to punish. Since our experiment takes place against a backdrop of a decades-long effort at political mobilization of the low castes (Narayan 2010), it might be that hostility and conflict between high and low castes would lead to particularly high punishment levels in triples that include both high-and low-caste players. For example, in an HLH treatment— in which the victim of a norm violation is high caste, the norm violator is low caste, and the third party is high caste— hostility might induce the third party to punish the low-caste norm violator much more harshly than if all three players had come from a high caste. That is, punishment in HLH would be higher than in HHH. We call this the *caste conflict hypothesis.* It also predicts that punishment in an LHL treatment, in which the victim is low caste, the norm violator is high caste, and the punisher is low caste, would be greater than in an LLL treatment. We express this by

$$\text{pun}^{\text{HLH}} > \text{pun}^{\text{HHH}} \quad \text{and} \quad \text{pun}^{\text{LHL}} > \text{pun}^{\text{LLL}},$$

where pun^{XYZ} denotes the mean punishment for defection in treatment XYZ.

A contrary, but also plausible, hypothesis is that a history of fierce retaliation by the high caste when low-caste individuals refused to submit to the prevailing social hierarchy[8] would lead the latter to tolerate norm violations by the former. That is,

$$\text{pun}^{\text{LHL}} < \text{pun}^{\text{LLL}}.$$

We call this the *caste submission hypothesis*.

---

[8] Even in contemporary rural India events like the following are still reported: "When a Dalit … refused to sell *bidis* [hand-rolled cigarettes] on credit to the nephew of an upper caste village chief, the upper caste family retaliated by forcibly piercing his nostril, drawing a string through his nose, parading him around the village, and tying him to a cattle post" (cited in NYU, 2007, p. 60, from *Indian Express* (Bombay), April 28, 1998).

The third hypothesis that our design enables us to examine is that, controlling for the caste of the norm violator, high-caste compared to low-caste individuals punish more severely norm violations that hurt members of their community. Sen (2000) and Rao and Walton (2004) argue that *inequality of agency* is a consequence of social exclusion. Repression and social exclusion might induce fatalism, undermining the self-confidence required to punish violators of cooperation norms; and restraints on social and economic life might reduce affiliation with members of one's community, undermining the motivation to punish norm violators who hurt other members. We call this the *caste culture* hypothesis. It predicts that among the members of the historically repressed group, there is a lower willingness to punish those who violate the cooperation norm of the group. Controlling for individual differences in education and wealth, this hypothesis implies

$$pun^H > pun^L$$

where pun^H denotes the mean punishment of a norm violation when the punisher is high caste (*i.e.* punishment in HHH and HLH), while pun^L denotes the mean punishment when the punisher is low caste (*i.e.* punishment in LLL and LHL).

Our results unambiguously refute the caste conflict and the caste submission hypotheses. Instead we find, in line with the caste culture hypothesis, that low-caste compared to high-caste individuals punish norm violations less often and less severely. This result is robust to controls for wealth, education, and participation in village government. In fact, the effect of individual differences in wealth is very small, always insignificant, and sometimes not in the expected direction.

Further, we show that the low castes' lower propensity to punish has nothing to do with differences in the underlying social norm. We measure the underlying social norm in our

experiment by player B's expectations about punishment. Regardless of caste status, the vast majority of subjects in the role of player B expected that they would *not* be punished for cooperation in the social dilemma game but *would* receive high punishment for defection – a clear indication that cooperation was considered the normatively right thing to do in this game. Thus, low and high castes have the same cooperation norm in the social dilemma game.

Why then do low-caste individuals punish norm violations less severely than high-caste individuals in our experiment? A factor known to influence altruistic third-party punishment is in-group affiliation (Bernard, Fehr and Fischbacher, 2006; Goette, Huffman, and Meier 2006; Chen and Li 2009 show a similar result for second-party punishment). In the experiment discussed above, we always ensured that the potential victim of the norm violation (player A) and the potential punisher (player C) belonged to the same specific caste[9], while the potential norm violator (player B) belonged to a different specific caste. In this setting, if a player C who is Brahmin has a strong concern for the victim of the norm violation from his own specific caste, he will be more willing to punish than would a player C who is Chamar, who may not care much for the victims from his own specific caste. To test this hypothesis, we conducted a second experiment in which we ensured that this kind of in-group concern could not affect the punishment pattern. If the willingness to punish is affected by differences in in-group concerns between high and low castes, then the caste gap in punishment should be reduced in the second experiment. In fact, we observe that the caste gap in punishment vanishes in the second experiment: when in-group concern for the victim cannot play a role, the high castes punish at about the same level as the low castes. This result provides support to the hypothesis that high-

---

[9] In the context of this argument, it is very important to recognize that the caste system consists of discrete endogamous communities (*jati,* translated as caste in English). An individual belongs to a *specific* caste, such as the Brahmins, the Thakurs, the Chamars or the Pasis. The specific castes constitute a large part of an individual's social network and social life. Therefore, the *specific* castes represent the relevant in-group. For example, for a Brahmin the relevant in-group are the other Brahmins, and for a Pasi the other Pasis constitute the relevant in-group.

caste compared to low-caste individuals exhibit a greater willingness to sanction violations of cooperation norms that hurt members of their own specific caste.

This difference may have far-reaching implications. First, it means that the high castes are better able to enforce contracts and ensure their property rights, which advantages them with respect to trading opportunities and production incentives. Although it has been studied much less than *second*-party punishment, *third*-party punishment can be a much greater restraint on norm violations. If only those whose economic payoff is directly affected by the norm violation (the so-called "second parties") are willing to enforce the norm, then a person could violate it with impunity as long as he was stronger than his victims. In contrast, the number of third-party punishers is potentially as large as the community itself. If third parties are willing to punish unfairness, then the sheer number of potential punishers can constrain even powerful individuals, as Mokyr suggests was generally the case in England under the Industrial Revolution, and which he argues helped to make the revolution possible.

Second, as noted above, experimental and field evidence indicates the importance of informal punishment for voluntary contributions to public goods and voluntary participation in collective action. If high castes are more able than low castes to sanction free riders, they are in a better position to provide public goods and to organize collective action. This advantage may be one reason why Untouchability continues to be practiced in almost 80 percent of Indian villages despite the constitutional abolition of Untouchability (Shah *et al*. 2006). The quote at the beginning of the introduction nicely illustrates the superiority of the high caste in organizing collective action for the purpose of sustaining their caste status and power differences with

collective force.[10] When a single Dalit argued with a high-caste farmer over discrimination, the high-caste villagers *collectively* attacked 80 Dalit families and imposed a boycott on Dalits from the village. In contrast, the Dalits were unable to respond with collective action.

Finally, the lower willingness in low-caste communities to punish those who hurt members of their community may undermine the power of the state to direct resources to the low caste—for "state initiatives do not operate in a social vacuum," but rather require social pressures to sustain them (Drèze, Lanjouw and Sharma, 1998).[11]

1. The Experimental Design

We are interested in how the assignment to castes with different social status affects the willingness to punish violations of a conditional cooperation norm. For this purpose, we developed a simple experimental game in which certain behaviors are likely to constitute a normative obligation that, if violated, would be punished by an impartial observer. We describe the game and then the different treatment conditions.

1.1. *The Experimental Game*

Figure 1 depicts the game between three individuals, A, B, and C. Each individual plays the game in his home village with anonymous players from other villages. Players A and B interact in a sequential exchange game. They each have an endowment of 50 rupees, which is a

---

[10]A British official in 1947 wrote that attempts by low-caste individuals to exercise the right to use public wells, a right granted to them by law under British rule, were commonly met with "social boycotts" – collective punishments imposed by high castes that might refuse to trade with an entire low-caste community, or might destroy their crops and dwellings. The official concluded that "No legislative or administrative action can restore to the depressed class people the right to use public wells" (cited in Galanter, 1972, p. 234). In a social boycott in 1998, high-caste individuals who gave employment to low-caste individuals were fined by the village council (Human Rights Watch, 1999, p. 30).

[11] The required social pressures from the low castes were absent during the 1957–1993 period spanned by the five surveys of Palanpur, in the state of Uttar Pradesh. After reviewing all public services in the village in this period, Drèze, Lanjouw and Sharma (p. 220) conclude that predatory actors had derailed essentially *every* public initiative to help disadvantaged groups: "With few exceptions, Jatabs [the main low caste in Palanpur] have remained outside the scope of constructive government intervention"; in one particularly egregious case, the intended low-caste beneficiaries became "victims of merciless extortion" by the high-caste managers of a cooperative lending society.
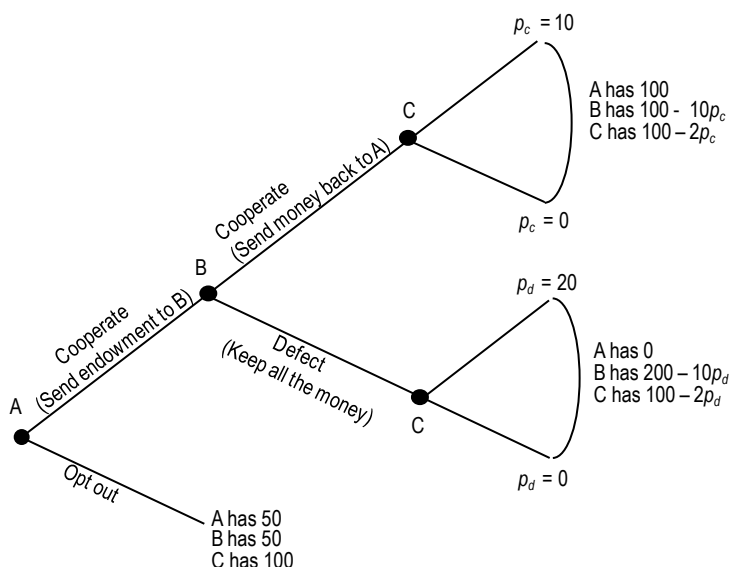
considerable amount of money compared to the daily wage of an unskilled agricultural worker of about 50-75 rupees. Player A has to choose between two actions: he can "send" his total endowment to B, in which case the experimenter triples its value so that B has altogether 200 rupees;[12] or he can opt out, in which case A keeps his endowment and the game ends. If A sends his endowment to B, then B has to make a binary choice: to keep everything for himself or to send 100 rupees back to A. We allowed players A and B only binary choices because we wanted "keeping the money" by player B to be an unambiguous norm violation. We expected a widely shared understanding that if A sends money to B, then B should send money back to A; *i.e.* that there is a social norm of conditional cooperation.

Player C is an uninvolved outside party who can punish B at a cost to himself. His endowment is 100 rupees. For each two-rupee coin that player C spends on punishment, Player B is docked a 10-rupee note. We asked C to make a choice for the case where B keeps all the money, and also for the case where B sends money back to A. Player C makes this choice before he learns B's decision. He indicates his choice by moving coins on a game board in private.

The conventional assumption that individuals are purely self-interested implies a unique subgame perfect Nash equilibrium in which the third party never punishes, the second party keeps all the money if it is offered to him, and so the first party opts out. Thus the equilibrium would be (Opt out, Keep all the money, Don't punish). In this outcome, A avoids being a sucker and does not enter into a relationship with B, and C keeps clear of punishing C. Players A and B each then earn 50 rupees, whereas if they cooperated they would each earn 100 rupees.

---

[12] 200 rupees is of the order of one week's per capita gross state product in Uttar Pradesh (based on the official estimate of per capita annual gross state product in 2004-05, the latest year for which such data are available).

Figure 1. Sequential Exchange Game with Third Party Punishment

$p_c = 10$

C

A has 100
B has 100 - $10p_c$
C has $100 - 2p_c$

$p_c = 0$

Cooperate
(Send money back to A)

B

$p_d = 20$

Cooperate
(Send endowment to B)

Defect
(Keep all the money)

C

A has 0
B has $200 - 10p_d$
C has $100 - 2p_d$

$p_d = 0$

A

Opt out

A has 50
B has 50
C has 100

Note: p_c and p_d, respectively, indicate the number of two-rupee coins that Player C spends to punish Player B conditional on B's cooperation or defection (norm violation).

The sequential exchange problem is akin to A having a good that B values more – that is the tripling of the money. By entering into an exchange, A gives B the opportunity to keep all the resources – a risk that typically arises in exchanges with separation between the *quid* and the *quo* over time or space and imperfect contractibility. In developing countries a large proportion of all exchange is probably characterized by such features. If B sends back nothing and C punishes him, the punishment by C mimics a disinterested third party's sanctioning a norm violator – for example, reproaching or bullying him or gossiping about him, which entails some cost or risk to the punisher but a larger cost to the individual punished. The deterrent to defection by a purely self-interested player is $10(p_d - p_c)$, where p_d denotes what Player C spends to punish defection and p_c denotes his spending to punish cooperation.

The game instructions avoid value-laden words, such as "cooperate," "defect," and "punish."[13] Instead they use neutral terms such as "send the money," "keep the money," and "impose a loss."

1.2. *Treatment Conditions*

Before we describe the treatments in detail, it is important to distinguish between two meanings of the term "caste." First, belonging to a caste means that the individual belongs to a specific endogamous social grouping consisting of thousands of families – such as belonging to the caste of Brahmins or the caste of Chamars. Each such social grouping is associated with a traditional set of occupations and culture. Within a village, individuals of the same caste are generally clustered in neighborhoods. Networks organized around the specific endogamous castes provide mutual insurance, which contributes to the very low rate of migration from villages in India (Munshi and Rosenzweig 2007).

The second meaning of belonging to a caste is that the individual is assigned the *social status* of the caste. For example, both Brahmins and Thakurs are castes at the high end of the caste hierarchy, but they constitute nevertheless clearly distinct social groups. Chamars and Pasis are castes in the lowest stratum of the caste hierarchy, but constitute clearly distinct endogamous groups. In the following, we use the terms "caste divide," "high caste," and "low caste" to indicate the status dimension of caste affiliation. We reserve the term "caste" without a modifier to mean the specific endogamous social group.

Because we were interested in the effect of assignment to an extreme position in a social hierarchy on the willingness to punish norm violations altruistically, we implemented four treatments that varied the composition in a triple of individuals with high (H) and low (L) caste

---

[13]The instructions are in the Web Appendix.

status. In the "single status" treatment, all three players are either high caste or low caste (treatments HHH and LLL). In the "mixed status" treatment, there is a deep status divide among players: in HLH only player B is low caste, and in LHL only player B is high caste. The number of triples by treatment was 62 in HHH, 61 in HLH, 41 in LLL, and 41 in LHL.

Because caste affiliation is associated with a certain status *and* with membership in a specific social group, we developed an experimental design that controls for the in-group-out-group relationships among the players. If we had not done this, we would confound the effect of the caste divide between players with the effect of in-group favoritism or out-group hostility. To see this, notice that in the single status treatments, player B could, in principle, be a member of the *same* specific caste as A and C, whereas in the mixed status treatments, player B would necessarily be of a *different* specific caste than A and C. In this case, it would be impossible to know whether any treatment difference between, say, HHH and HLH, was due to the caste divide or to in-group affiliation or to both. To avoid this confound, we formed triples in which player B – the potential norm violator – was always from a different *specific* caste than players A and C. Figure 2 illustrates this by giving examples of the triples we used.

Figure 2. Examples of Interacting Players

13
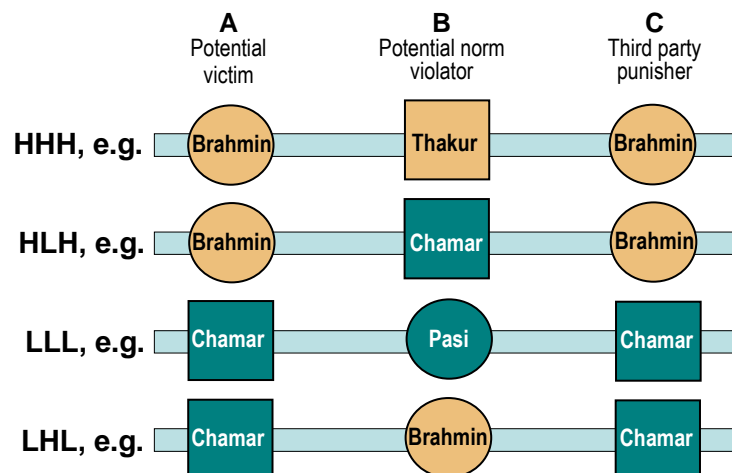
We drew our subjects from Brahmins and Thakurs, who are high caste, and Chamars and Pasis, who are low caste.[14] Figure 2 provides examples of interacting players in the game. In both the HHH and the HLH treatments, the potential victim of a norm violation (player A) and the third party punisher (player C) are from the same specific caste – in this example, Brahmin, while the potential norm violator (player B) comes from a different specific caste – in this example, Thakur or Chamar; player B is therefore an out-group member relative to players A and C. Figure 2 shows that this feature – A and C belong to the same specific caste and B belongs to a different specific caste – holds across all four treatments.[15]

1.3. *Procedures*

We recruited adult male subjects for each role (A, B, and C) from three non-overlapping sets of villages in central Uttar Pradesh: players A came from 61 villages, B from 53 villages, and C from 48 villages, randomly chosen from the hundreds of villages within 2.5 hours' drive from the town of Unnao (comprising three subdistricts of Unnao district). Informants in each village told us the neighborhoods in which the different castes in the village lived. In public places in a village, the recruiters asked individuals if they were interested in participating in an experiment about decision making that would last two hours and in which they would earn some money. We generally recruited five subjects for a single treatment (and never more than six) from a given village, and no two subjects from the same household. No subject participated in more than one treatment.

---

[14] The number of observations where C is Brahmin is 63, where C is Thakur is 60, where C is Chamar is 39, and where C is Pasi is 43.
[15] We varied this feature in a follow-up experiment, described in Section IV, which was designed to examine the psychological mechanisms behind the observed caste differences in the willingness to punish norm violators.

The average age of players was 35 years (with standard deviation 8.0) for high caste players and 34 years (with standard deviation 7.6) for low caste players. The age range was between 24 and 50 years.

To ensure that subjects understood the instructions, the rules of the game were explained to them at great length. A few subjects who did not pass a basic test of comprehension did not go on to participate in the game. Individual sessions were held inside a Qualis car. Each subject made his decision in private while the experimenter waited outside the car. After Player C had indicated his decision by moving coins on a game board, the experimenter reentered the car, informed C of A's and B's decisions, and paid C.

Implementing this experiment in rural India raised two ethical concerns. First, the players should never learn the identity of those with whom they interact in the game. Second, our concern with caste relationships, a politically fraught issue, should not be salient. To address these concerns, we recruited subjects for each role (A, B, and C, respectively) from three distant sets of villages. We carefully thought about how to communicate the caste affiliation of the participants to the players. One possibility was to use names, since names generally convey both a person's individual identity and his specific caste. The use of names was thus an unobtrusive way of indicating caste. We checked explicitly in a pre-experiment that individuals were generally willing to reveal their last names to our team of recruiters, and also verified the ability of individuals to recognize caste membership from names. Last names vary in the degree to which they identify a person's specific caste. In order to accurately convey the caste of the subjects, we used fictitious last names, each of which was a clear marker of the player's true caste.[16] This is a minor deception of the subjects that meets APA guidelines

---

[16] The last names were Bajpayee and Shukla for Brahmin, Chauhan and Thakur for Thakur, Goutham and Kureel for Chamar, and Rawat and Pasi for Pasi.

(, section 8.07) and is of a degree common in other social science experiments. On the basis of our pretest of the recognizability of names, we find that the vast majority of members of the high castes and the low castes recognized the specific castes of all names used in the experiment.

During the individual sessions with each subject, the experimenter conveyed information about his partners by saying, *e.g.* in the case of Player C, "You are playing the game with two other people. You are person C. [NAME], who is from another village, is person A; and [NAME], who is from another village, is person B." The advantage of using names is that we can convey information about caste but still maintain effective anonymity among the players, since thousands of people with the same last name live in the state and each of the players in a given triple came from a different and distant village. The experimenter emphasized that no player would ever know the villages to which the two players with whom he was matched belonged. The experimenter also never used the word "caste" in his interactions with the subjects before or during the game. In post-play interviews, we also asked questions about a subject's beliefs about the other players' actions, about the reasons for his own actions, and about his wealth and other individual characteristics.

2. Results

Our data give us two measures of punishment for defection: p_d measures the *absolute punishment* of defectors (*i.e.* norm violators), and p_d - p_c measures the extent to which defectors are more strongly punished than cooperators. We denote p_d - p_c as *relative punishment*. Throughout, we measure punishment in units of two-rupee coins spent. Recalling that player B loses one 10-

rupee note for each coin spent by player C, it follows that $2p_j$ is the rupee spending on punishment and $10p_j$ is the rupee punishment imposed, where $j = d$ or c.

Although our data also provide us with a measure of the sanctions imposed on cooperators, p_c, in this paper we focus on the role of caste affiliation in the sanctioning of defectors.[17] Before we present the results, we present evidence that cooperation by player B is indeed viewed as the normatively right thing to do.

2.1. *Testing the existence of a conditional cooperation norm*

We measure whether there is a social norm of conditional cooperation by player B's beliefs about punishment. After player B had chosen his action, we asked him how much punishment he expected in the case of cooperation and how much punishment he expected in the case of defection.[18] If there is a widely shared belief in the existence of a normative obligation to reciprocate cooperation, the beliefs of player B should reflect this obligation. In Figure 3 we show player B's beliefs about punishment. The figure shows little variation across treatments and between castes. On average, in each treatment and for each caste, player B expected that C would spend nearly 10 coins to punish B if he defected, and would spend almost nothing on punishment if B cooperated. Thus, in each treatment the null hypothesis of equal expected punishment across cooperation and defection can be unambiguously rejected (Mann-Whitney test, $P < 0.01$ in all cases), indicating a conditional cooperation norm in each treatment. Moreover, the differences between expected punishment of defection and cooperation are almost identical across treatments

---

[17] We find non-negligible levels of punishment for cooperation but no significant differences across treatments (see Fehr, Hoff and Kshetramade 2008). Although the first result may seem surprising, recent evidence, including evidence from cross-cultural studies (Herrmann, Thöni and Gächter 2008; Nikiforakis 2008 and Cinyabuguma, Page and Putterman 2006) indicates that punishment of cooperators is frequent.
[18] We asked these questions in a neutral language, *i.e.* we did not us the terms cooperation, defection, reciprocation or punishment, as we discussed in Section I.

(Kruskall-Wallis test, $P = 0.994$), suggesting that the same conditional cooperation norm applies across treatments and castes.

The actual decisions of B to cooperate provide further evidence of a strong conditional cooperation norm across treatments and castes. The fraction of B who are high caste who cooperated by treatment are 85% (HHH) and 88% (LHL). The comparable figures for B who are low caste are 77% (LLL) and 87% (HLH). A probit regression controlling for individual characteristics (not shown) indicates that that there is no significant difference between HHH and LHL, nor between LLL and HLH ($p > .40$). There is also no significant difference in the probability of cooperation between players B who are high caste and those who are low caste ($z = 1.24$, $p = .21$).

Figure 3. Player B's Beliefs about the Absolute Punishment of Cooperation and Defection
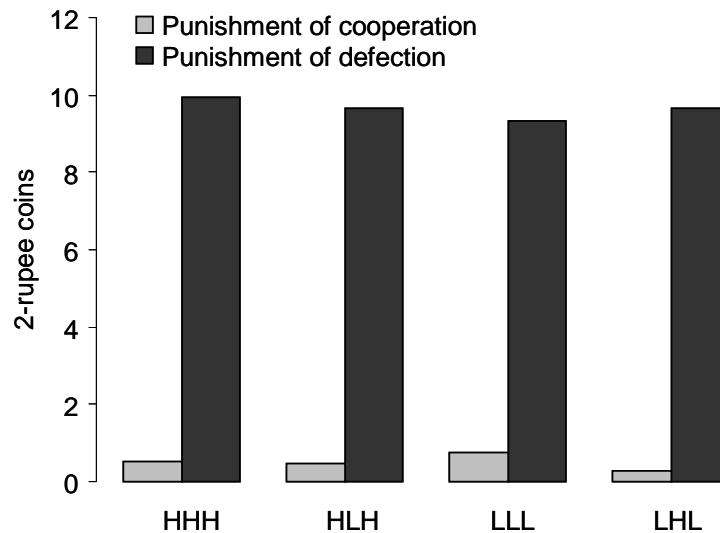
2.2. *Testing the caste conflict and caste submission hypotheses*

The *caste conflict* hypothesis suggests that we should observe a difference between mean punishments in the single status treatments and the mixed status treatments. With regard to punishment by *high* caste members, this means that we should observe a *higher* level in HLH

than in HHH. *Contra* this hypothesis, Figure 4 shows that mean relative punishment in HLH is not higher than in HHH; and the difference between conditions is not significant according to a Mann-Whitney test ($P = 0.18$). This result is reinforced if we examine absolute punishment levels. Mean punishment in HLH (6.80 two-rupee coins) is again not higher than in HHH (8.45 two-rupee coins), and the difference is not significant (Mann-Whitney test, $P = 0.09$).

The caste conflict hypothesis also predicts a higher level of punishment in mixed status treatments by *low* caste members, that is, punishment should be higher in LHL than in LLL. *Contra* this hypothesis, there is no significant difference between mean relative punishment in these two treatments (Mann-Whitney test, $P = 0.85$). The same picture emerges if we examine absolute punishment: the mean in LHL (6.15) is not significantly different than in LLL (5.37; Mann-Whitney test, $P = 0.91$).

The caste submission hypothesis predicts that punishment in LLL is higher than in LHL. Figure 4 and the results discussed above show that this hypothesis is not borne out.

Figure 4. Relative Punishment of Norm Violators
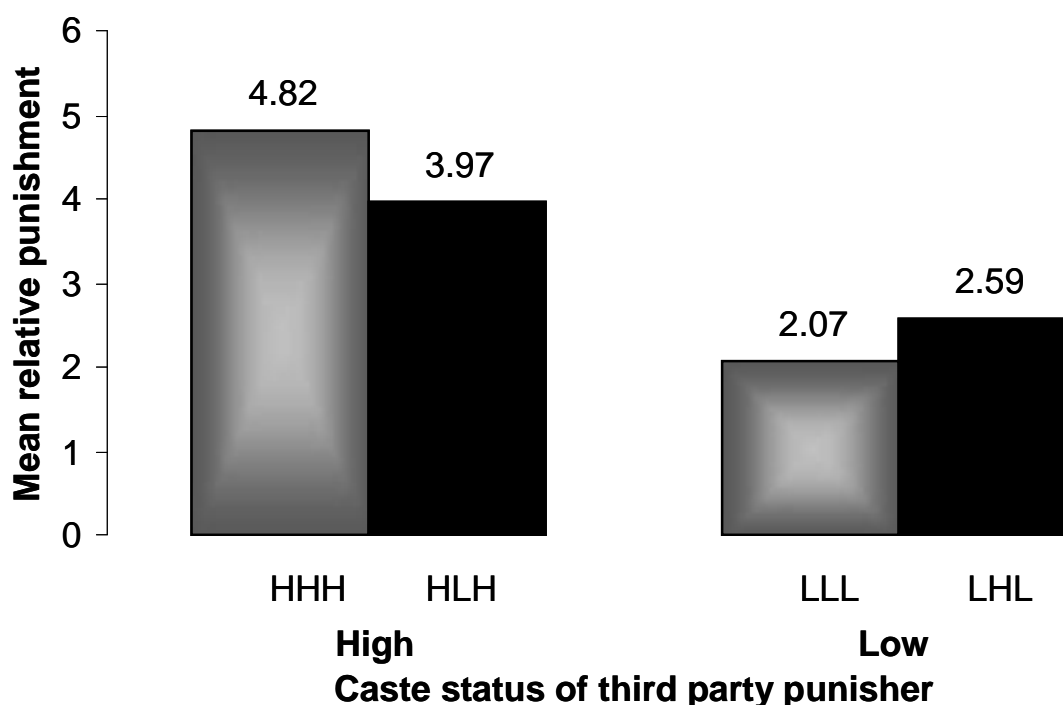
In order to check the robustness of these results, we conducted OLS and tobit regressions that included a number of factors that capture social and economic differences among individuals in the role of player C.[19] For each individual, we control for land owned, education, and house quality. In rural Uttar Pradesh, the site of our experiments, land owned and house quality – whether an individual lives in a mud house or a pure brick house – are major indicators of wealth; in Appendix Table A1, we show that these variables are also important predictors of per capita consumption. In addition, the regressions control for whether the individual cultivates his own land, which might affect his attitudes towards the norm that we are investigating, and

---

[19]Tobit analyses (available on request) support all results reported here. Tobit deals with censored data better than OLS. The censoring problem may play a role in the case of absolute punishment for defection, which ranges from 0 to 20, but much less so in the case of relative punishment for defection, which ranges from -10 to 20.

whether he has participated in village government.[20] In regressions (1), (2), (4) and (5) of Table 1, we show the results of OLS regressions that are based on the following model:

$$pun = \alpha + \beta \cdot (\text{player C is high caste}) + \gamma \cdot (\text{treatment is mixed status}) +$$

$$\delta \cdot (\text{player C is high caste} \cdot \text{treatment is mixed status}) + \mu \cdot Z + error$$

where "pun" denotes absolute or relative punishment, and Z is a vector of variables measuring individual characteristics of player C.

The omitted category in regressions (1), (2), (4) and (5) is a low caste punisher who is in the single status treatment. Thus, the constant α measures the punishment level in LLL. The next three coefficients are measures of the caste and treatment effects when we control for individual characteristics: β measures the difference between a high and low caste player's punishment in the single status treatments, *i.e.*, the difference between HHH and LLL, and thus, $\alpha + \beta$ indicates the punishment level in HHH. The coefficient for the mixed status treatment, γ, measures the difference between the LLL and LHL treatments, implying that $\alpha + \gamma$ represents a measure of punishment in LHL. Finally, punishment in the HLH treatment is measured by $\alpha + \beta + \gamma + \delta$.

With respect to the *high* caste, the caste conflict hypothesis is that $pun^{HLH} > pun^{HHH}$, which implies that $\gamma + \delta > 0$, which we assess with an *F*-test in Table 1. In all four regressions the *F*-test indicates that $\gamma + \delta$ is not significantly different from zero. For example, in regression models (2) and (5), which control for the above-mentioned socioeconomic characteristics, the *P*-values for the *F*-statistics are 0.343 and 0.108, respectively.

---

[20] The regressor "Has political experience" is a dummy variable equal to one if the respondent has ever been a village government chief ("Pradhan"), vice-Pradhan, or member of the village government council.

With respect to the *low* caste, the caste conflict hypothesis is that $pun^{LHL} > pun^{LLL}$, which implies that $\gamma > 0$, while the caste submission hypothesis implies that $\gamma < 0$. In all the regressions, the coefficient for the mixed status treatment (γ) is not significantly different from zero ($P > 0.4$ in all cases). Thus, after controlling for important socioeconomic characteristics such as land ownership, education, and house quality, we refute the caste conflict and caste submission hypotheses.

2.3. *Testing the caste culture hypothesis*

The *caste culture* hypothesis predicts a higher level of punishment when the punisher is a high caste member compared to when he is a low caste member. Figure 4 provides preliminary support for this hypothesis. Mean *relative* punishment imposed by high caste individuals is roughly 90 percent higher than mean relative punishment by low caste individuals (4.40 compared to 2.33). A similar picture emerges for mean *absolute* punishment: Members of high castes punish 32 percent more than low caste members (7.63 compared to 5.76). These differences are significant according to Mann-Whitney tests at $P \leq 0.01$.

To what extent is this systematic difference in the willingness to punish a norm violation a result of differences in wealth and education across high and low castes? We collected information in post-play questions on land ownership, housing wealth, and education, which are all known to be important predictors of per capita consumption.[21] We note that although people belonging to high castes are on average wealthier, there is a substantial overlap across castes: many people who belong to a high caste are nevertheless very poor. Likewise, a significant number of low caste people have managed to acquire more wealth than poor high caste people.

---

[21]We show this in Table A1 in the appendix, where we use data from the 1997-98 Survey of Living Conditions in Uttar Pradesh. The table indicates that land ownership, housing wealth, and education have a large and significant impact on adult per capita consumption. Together, they explain between 30 and 40 percent of the variation in consumption for both high and low caste individuals.

Therefore, we are in a position to examine the differences in punishment across castes while holding wealth constant. This shows up clearly in our sample of subjects. If we examine the distribution of castes across house quality, we find that among the 81 subjects who live in a mud house, 42% belong to a high caste and 58% belong to a low caste. If we look at those 124 subjects who live in a brick house or a mixed mud and brick house, we find that 71% belong to a high caste and 29% to a low caste. A similar pattern emerges in the case of land ownership. Among the 95 subjects who own land below the median, 36% belong to the high caste and 64% to the low caste. Among the 110 subjects who own land above the median, 83% belong to the high caste and 17% to the low caste.

In Figures 5a and 5b we show the mean punishment of defectors conditional on the caste status, house quality, and land ownership of the punisher. Figure 5a indicates that regardless of whether subjects live in a mud house or a house that is built at least partly with bricks, the high-caste subjects punish more on average than the low-caste subjects.

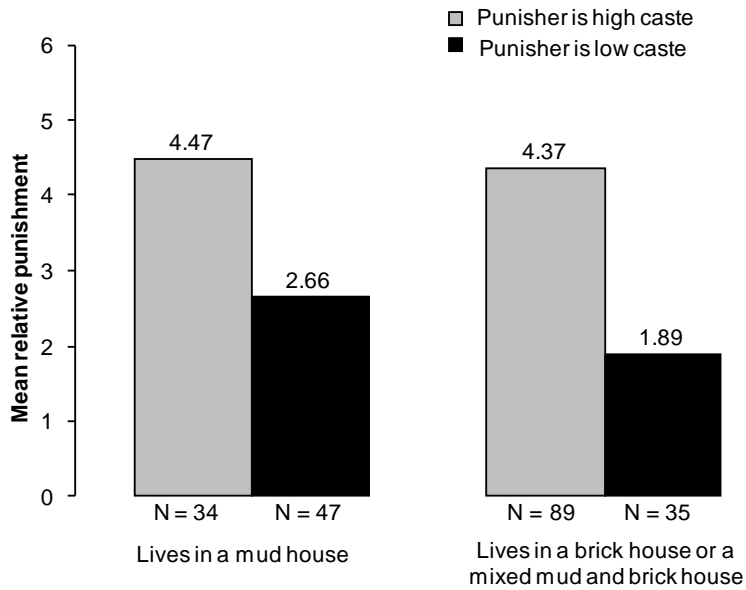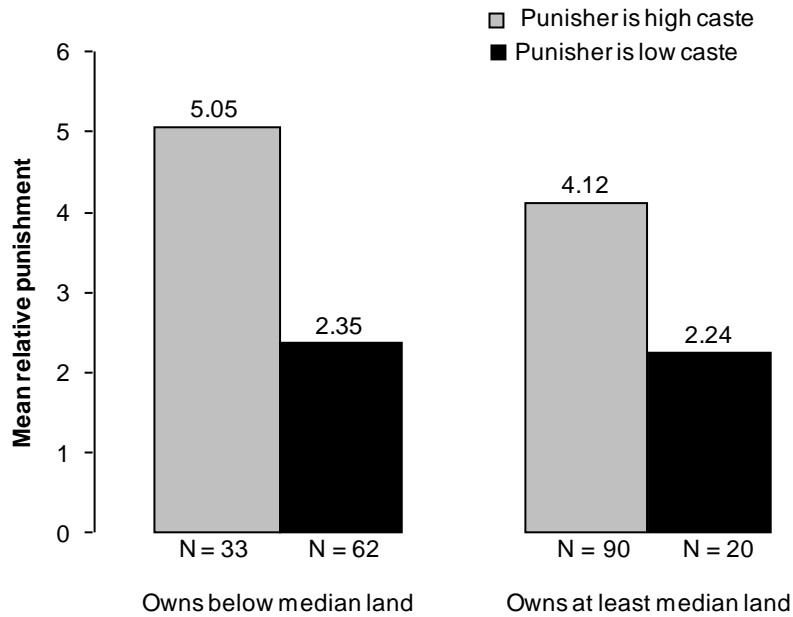**Figure 5a. Punishment by Those Who Live in Mud Houses and Those Who Do Not**

Figure 5b. Punishment by Those Who Own Below-median Land and Those Who Own

Above-median Land

Moreover, the high-caste subjects exhibit very similar mean punishment levels regardless of house quality. The same holds true for low-caste subjects: they punish on average less than

high-caste subjects but display similar mean levels regardless of house quality. Figure 5b indicates a very similar pattern with regard to land ownership. Regardless of whether subjects own land below or above the median, the high-caste subjects punish on average more than the low-caste subjects. Also, within the high caste, the mean punishment level is similar for those below and above median land ownership and, among the low caste individuals, the mean punishment level is also similar for those below and above the median.

Taken together, Figures 5a and 5b suggest that wealth does not have a big effect on punishment. This hypothesis is clearly supported by regressions (3) and (6) of Table 1, which show that the caste gap in punishment is robust to the introduction of these controls. Since these regressions do not include the mixed status dummy, the omitted category are the treatments LLL and LHL, in which only low caste players can punish. The high caste dummy measures the extent to which high caste members are *generally* more willing than low caste members to punish defectors. Both regressions confirm that high caste players spend roughly two two-rupee coins more on punishment of defection, a difference that is significant in the case of both relative punishment ($P = 0.055$) and absolute punishment ($P = 0.022$). In a dprobit regression (not shown), we find that controlling for other individual characteristics, high caste individuals are 16.6 percent more likely to punish norm violations than low caste individuals ($P = 0.023$) and that no individual characteristic other than caste has a significant effect on the probability of imposing non-zero punishment.

The regressions in Table 1 also show that land ownership has little effect on punishment and is statistically insignificant; likewise, the effect of house quality is small and insignificant. These results confirm the message conveyed by Figures 5a and 5b. Among the control variables, only completion of secondary school (10 years of schooling) has an effect on punishment that is significant at the 5 percent level in at least some of the model specifications. Controlling for the

other factors, completion of secondary school raises absolute and relative punishment by roughly 1.8 two-rupee coins, an increase of about 0.3 standard deviation units.[22] We conducted further tests (not shown here) to examine whether the education variable changes the caste gap. For this purpose we interacted "completion of high school" in regressions like those in (3) and (6) with the high caste dummy. The interaction effect is small and insignificant ($P > 0.70$ for both absolute and relative punishment), indicating that the completion of secondary school leaves the caste gap unaffected. Similarly, the interaction of political participation with the high caste dummy is not significant and does not change the qualitative results.

A final robustness check for the validity of the caste culture hypothesis is whether our results hold over subpopulations. Do the results reflect the peculiarities of a single specific caste, or do they hold for each of the four specific castes in our sample? If caste status drives the caste gap, we should observe that the punishment behavior of members of the two high castes in our sample – the Brahmins and the Thakurs – is each significantly greater than that of the two low castes in our sample – the Chamars and the Pasis. We examine this question in Table 2. Here columns (2) and (4) duplicate regressions (3) and (6) of Table 1 except that now we distinguish specific castes rather than caste status. In Table 2, Chamar, a low status caste, is the omitted category. As predicted by the caste culture hypothesis, the coefficients on Brahmin and Thakur, which are high status castes, are significant and positive, and the coefficient on Pasi, which is a low status caste, is not significant. Thus, whatever explains the caste gap in the willingness to punish defectors operates at the level of *caste status* in our sample, not at the level of *specific individual castes* (descent groups). Just as plants in spatially separated deserts have

---

[22] We find that the effect of a subject's education on his level of punishment for defection is highly nonlinear. When we measure education in years, the estimated coefficient on education is not significant, whereas in Tables 1-3, which measures education as a dummy variable that equals one if a person has completed 10 years of schooling, the estimated coefficients are nearly always significant. In India, secondary school ends at tenth grade. Among the 125 subjects without secondary school completion, 45 percent were high caste and 55 percent were low caste; among the 80 subjects who had completed high school, 84 percent were high caste and 16 percent were low caste.

independently developed similar adaptations that enable them to live in the desert, one can

conjecture that different low castes have developed similar attitudes towards altruistic

punishment that reflect the common constraints on their social life. The attitudes converge

between the two low castes in our sample and diverge from those developed by the two high

castes in our sample.

It might appear that a possible explanation for the higher punishment in HHH than in

LLL is that Brahmins and Thakurs dislike each other and therefore punish each other out of spite,

whereas Chamars and Pasis do not dislike each other and therefore have no spiteful motive to

punish.   Recalling that relative punishment equals p_d - p_c, this measure "differences out" a

propensity to punish out of spite. If, for example, punishment was motivated only by spite, then

p_d would be no greater than p_c, and thus relative punishment would be zero. But as shown in

Table 2, the same patterns of results hold whether we measure punishment in absolute or relative

terms. Further, if we estimate specification (4) of Table 2 with punishment for cooperation (*i.e.*

spiteful punishment) as the dependent variable, none of the caste variables (*i.e*. none of the

variables in rows 2-6) is significant. As we report in Fehr et al. (2008, Figure 1), we also find no

significant treatment differences when we consider across all four conditions the proportion of

the third parties who punish cooperators.

3. Interpretation and Discussion of Results

To sum up, neither differences in the recognition of names, nor differences in the perceived

obligation to reciprocate, nor income/wealth effects, nor differences in spite appear to explain the

greater third-party punishment imposed by high-caste members than by low-caste members.

What then could explain this? As noted above, a factor that influences altruistic third-party

punishment is in-group affiliation. Bernard, Fehr and Fischbacher (2006) and Goette, Huffman,
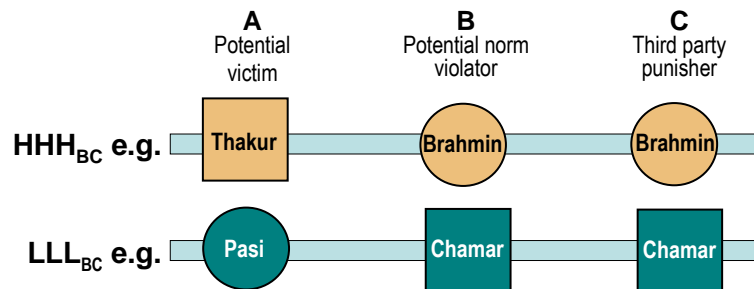
and Meier (2006) experimentally show in-group bias in punishment in two respects. If the victim of a norm violation is a member of the punisher's in-group, or if the norm violator is a member of the punisher's out-group, then punishment tends to be higher. An intuitive explanation of these results is that third parties empathize more with a victim who is an in-group member compared to an out-group member, and "give a break" to a norm violator who is an in-group member.

However, what the group means to an individual affects the level of in-group affiliation. Social identity theorists (Tajfel 1982, Turner *et al.* 1994) argue that group identification exists in part to provide self-esteem as the individual construes himself as associated with some group that he values. This theory predicts lower levels of identification among members of groups that are lower in prestige. A classic field study is Cialdini *et al.* (1976), which finds that students from universities with major football teams identify themselves more with their team when it wins than when it loses. Applying social identity theory to our setting, the social stigma of the low castes and their limited means to acquire an alternative group identity (*e.g.* because historically they lacked a literate class) might induce members of low castes to develop little concern for other (unknown) members of their specific caste. An individual who belongs to the Chamars – a caste at the low end of the status hierarchy – may feel ashamed of being a Chamar in view of the inferior status of his group and the humiliation he experiences in daily life. For low-caste individuals, the in-group is the category that serves as his discrediting. In contrast, the high social and ritual status of the high castes may induce high-caste individuals to show more in-group concern for unknown members of their specific caste. An individual belonging to the Thakurs, for example, a caste at the high end of the status hierarchy, derives status, prestige and material

benefits from being a Thakur. It thus seems more likely that this individual would care more about the welfare of other Thakurs.[23]

In the treatments that we have discussed, players A and C were always members of the same *specific* caste, while the potential norm violator – player B – always came from a different *specific* caste (recall Figure 2). This feature of the experimental design enabled us to examine the role of caste *status* while controlling for in-group affiliation. By varying this design, we can learn whether members of the high castes display a greater degree of in-group bias than members of low castes. We conducted a second experiment in which players B and C were members of the same specific caste, while player A belonged to a different specific caste. We implemented this condition in the single status treatment and thus denote the new triples by HHH_{BC} and LLL_{BC}. In Figure 6 we show an example. In the example, the specific caste composition for HHH_{BC} is Thakur-Brahmin-Brahmin in the roles of players A, B, and C, respectively, and for LLL_{BC} it is Pasi-Chamar-Chamar as players A, B, and C, respectively.

**Figure 6. Examples of Interacting Players under the BC Condition**

A crucial difference between the BC condition and the previous treatments is that in the BC condition, the potential victim of the norm violation (player A) is no longer a member of the punisher's specific caste. Thus, if the punisher exhibits in-group favoritism towards the victim, punishment will be lower in the BC condition than in the previous treatments. Furthermore, in the BC condition, a norm violator is a member of the punisher's in-group. Therefore, a punisher who tends to give the members of his own specific caste a break when they violate a social norm will punish less in the BC condition. Since the BC condition has players B and C of the same specific caste, it necessarily has players B and C of the same caste *status*. The only treatments in our main experiment that have players B and C of the same caste status are HHH and LLL. Therefore, to study the effect of the BC condition, we compare the new treatments with our earlier single caste treatments (LLL and HHH). We have 39 triples in HHH_{BC} and 36 triples in LLL_{BC}.

Figure 7 shows the results. The caste gap in relative punishment for defection nearly vanishes in the BC condition. While there is a large gap between HHH and LLL in the previous experiment as shown by the first two bars, which are repeated from Figure 4, the gap between HHH_{BC} and LLL_{BC} is small and insignificant (Mann-Whitney test, $P > 0.8$ for both relative and absolute punishment). The elimination of the caste gap reflects the fact that the high-caste members punish significantly less in the BC condition (Mann-Whitney test: $P \leq 0.05$ for both relative and absolute punishment), while the low-caste members do not show a difference in punishment across the LLL and LLL_{BC} conditions (Mann-Whitney test: $P = 0.99$ for relative punishment and $P = 0.356$ for absolute punishment).[24]

---

[24] We observe these differences despite the fact that the perceived social norm is the same across these two conditions. Just as we found in our main experiment (see Figure 3), we also observe in the BC condition that players B expect no or little punishment in the case of cooperation and high punishment in the case of defection. Thus, the difference between these two conditions *cannot* be attributed to differences in perceived social norms.

Figure 7. Relative Punishment When the Victim and Third Party Punisher Are Members of the Same Specific Caste, and When They Are Not
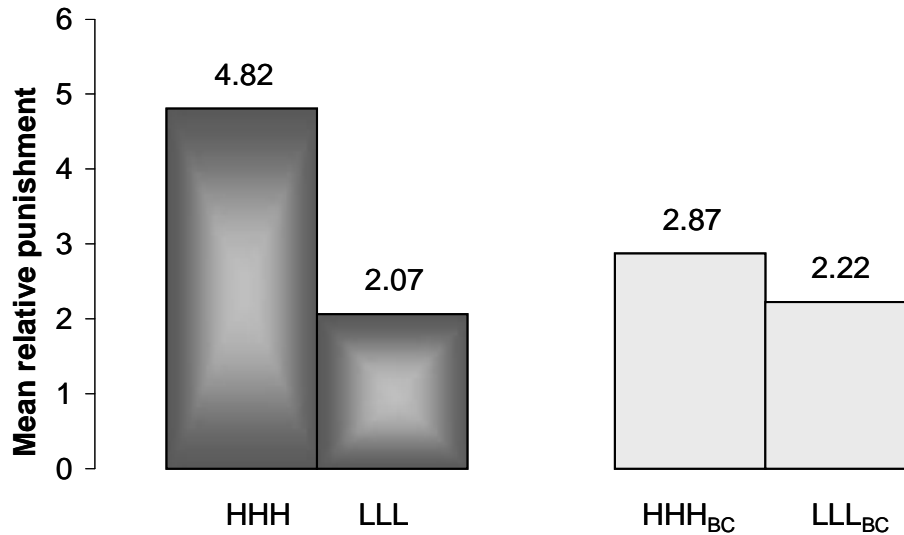
We further examine the role of in-group favoritism in Table 3, which reports the regression results for the pooled data of the HHH, HHH_{BC}, LLL and LLL_{BC} treatments. In these regressions, the LLL treatment is the omitted category. Therefore the BC-dummy measures the difference between LLL and LLL_{BC}. The table shows in row 3 that the coefficient on the BC-dummy is small and insignificant, suggesting that there is no in-group favoritism among low caste members. The difference between the HHH condition and the HHH_{BC} condition is measured by the sum of the coefficients for the BC dummy and the interaction term between the BC-dummy and the high caste dummy. The lower panel of Table 3 shows the F-test for the null hypothesis that the sum of these coefficients is zero. Controlling for the socioeconomic characteristics of the punisher, we can reject this hypothesis with high confidence ($P < 0.02$ for both relative and absolute punishment). This result is consistent with the hypothesis that high caste members' punishment decisions are driven by in-group favoritism towards their specific caste.

Finally, Table 3 provides information about whether there is a significant caste gap in the BC condition. The difference between HHH_{BC} and LLL_{BC} is measured by the sum of the coefficients for the high caste dummy and the interaction term. The *F*-test for the null hypothesis that the sum of these coefficients is equal to zero shows that the null cannot be rejected ($P > 0.6$ in all four regressions of Table 3). Thus, the punishment behavior of low- and high-caste individuals is indistinguishable in the BC treatment, suggesting that in-group favoritism among the members of the specific castes with high status plays an important role in explaining the caste gap in our main treatments.

One explanation is that the greater pride of belonging to a group with prestige and social benefits makes specific castes that are high in status, compared to specific castes at the bottom of the caste hierarchy, more willing to punish to protect members of their specific castes. A second, distinct explanation for the caste gap could be that high castes' use of punishment against other groups was much more effective historically than it was for low castes. Thus high-caste members *learned* to punish and to store the rule in its general form: "punish violators when they hurt members of my specific caste." There is persuasive evidence that cultural norms are acquired through intergenerational transmission and thus persist across generations (see an overview and new evidence in Butler et al. 2010).[25]

---

[25] With further treatments, it might be possible to distinguish between these mechanisms. This study implemented only eight of the 18 treatments possible with two status groups and two possible relationships between same status groups (insider and outsider). Because of the long travel distances over unpaved roads, a team could generally run only two sessions per day. A treatment with 35 triples entails seven 5-person sessions, hence 3.5 days. To implement 18 treatments would thus require 63 (=18*3.5) workdays, with five teams working full-time most of the days (an A-team, a B-team, a C-team, and two teams to return to the site to give payoffs to players A and B). It would be possible to reduce the number of observations per treatment if caste could be experimentally manipulated. But since that is impossible, larger samples are needed to ensure overlap in individual characteristics between high- and low-caste subjects.

4. Concluding Remarks

In this study we have shown how individuals' lifelong position at the top or bottom of an extreme social hierarchy – the Indian caste system – affects their willingness to punish norm violations. The Indian caste system is an excellent setting for studying the effects of assignment to the top or bottom of a social hierarchy because people are born into specific castes and mobility across caste status groups is basically absent, whereas the greater freedoms that low-caste individuals have enjoyed in the last 50 years, and the greater levels of landlessness among the high castes, have created a substantial overlap between high and low caste groups with respect to wealth, education, and political participation in village government. This means that we can both rule out self-selection into castes and also compare individuals who differ in caste status but do not differ with respect to our measures of wealth, education and political experience. Our results thus plausibly identify the impact of caste status on individuals' willingness to punish norm violations.

We have put forward three plausible hypotheses: the *caste conflict hypothesis*, the *caste submission hypothesis*, and the *caste culture hypothesis*. Our findings unambiguously refute the first two and support the third: compared to low-caste subjects, high-caste subjects have a considerably greater willingness to altruistically punish violations of a cooperation norm. This finding is robust to controls for wealth, education and political participation. Our results also suggest that the differences in the willingness to punish can be attributed to the caste *status* of individuals: individuals from each specific high caste in our sample – the Brahmins and the Thakurs – exhibit a significantly greater willingness to punish compared to individuals from each specific low caste, while the differences between the two specific high castes and between the two specific low castes in our sample are negligible. We can also rule out that subjects from the high caste have a different view about the prevailing social norm. Both members of high castes

33

and members of low castes believe that the third-party punisher will dock them a large amount if they defect, but not otherwise. Our findings therefore suggest that the punishment differences between high and low castes truly reflect a difference in preferences for the punishment of defection.

In a further experiment, we show that the high caste do not *generally* punish more than the low caste. Instead, they punish more when the victim of the norm violation is a member of their specific caste. In our main experiment, the injured party always belonged to the punisher's *specific* caste while the norm violator did not. In the follow-up experiment, the norm violator always belonged to the punisher's *specific* caste, while the injured party did not. If the punisher is motivated by in-group favoritism – taking revenge if the injured party is a member of his own specific caste, or giving a norm violator from his own specific caste a break – we should observe less punishment in the second experiment. We observe a substantial reduction in the severity of punishment imposed by high-caste individuals, but no significant change in the level of punishment imposed by low-caste individuals. As a result, the caste gap in punishment vanishes in the second experiment. This suggests that in-group favoritism – being more socially minded, but only towards those whom they consider part of their community – is an important driving force behind the higher castes' stronger willingness to punish norm violations altruistically.

Our results relate to a longstanding question in political economy: *How does an elite resist reforms after it loses control over the political institutions?* Acemoglu and Robinson (2006) propose a model of purely self-interested agents in which each member of the elite is large enough to internalize the benefits from resistance to reform through lobbying, bribery, intimidation, or violence, whereas each member of the repressed group is not large enough to do so. In their model, changes in *de jure* political power (*i.e.* political power allocated by political institutions) induce offsetting changes in the distribution of *de facto* political power as members

of the elite increase their contributions to collective action, whereas agents in the repressed groups do not. Our results suggest another, complementary mechanism to explain the puzzle of persistence of repressive institutions despite major *de jure* reform. In line with a large—and rapidly increasing—literature on altruistic norm enforcement, we have provided evidence that the assumption that individuals are purely selfish is inadequate to explain the observed enforcement of norms. The novelty of our findings is to provide evidence that individuals assigned to the top stratum of an extreme social hierarchy have a substantially greater willingness to altruistically enforce a cooperation norm that helps their groups than do individuals assigned to the bottom stratum of the hierarchy.[26]

This result is reminiscent of an older perspective (*e.g.* Gellner 1994) that stressed that in order to dominate a group thoroughly, the group had to be *pulverized* and *atomized*. In this view, many of the restrictions historically imposed on the low castes – such as exclusions from public celebrations and bans on marriage ceremonies and other shared rituals – make sense because they prevent the low castes from developing positive group identities that promote collective action. In a world in which everybody was completely selfish, such restrictions would make little sense, whereas if one takes into account the possibility of altruism towards one's own group, these restrictions may help the high castes to maintain their superior position.

---

[26] The example that Acemoglu and Robinson use to illustrate their argument is the persistence of a repressive economy in the southern United States after the Civil War that looked remarkably like that of the antebellum South. However, another factor that may have played a key role were the actions of Southerners who viewed dismantling white supremacy as a deep violation of a norm and who were willing to enforce the norm at a cost to themselves. As Foner (1988, pp. 431-32) reports, members of the Ku Klux Klan crossed class lines: "ordinary farmers and laborers constituted the bulk of Klan membership," which included men so poor that it was said they were "not worth the bread they eat." In one South Carolina county, Foner reports that nearly the entire white male population joined the Ku Klux Klan. This lends support to the view that material self-interest was not the only factor in the intimidation. Observers at the time noted that blacks did not have this power to act in common for self-protection (Foner, p. 436).

References

Acemoglu, Daron and Robinson, James A. "*De Facto* Political Power and Institutional Persistence." *American Economic Review, Papers and Proceedings,* 2006, *96*(2), pp. 325-330.

Algan, Yann and Pierre Cahuc, "Civic Virtue and Labor Market Institutions," *American Economic Journal: Macroeconomics* 2009, *1*(1), pp. 111–145.

Banerjee, Abhijit, and Iyer, Lakshmi. "History, Institutions and Economic Performance: The Legacy of Colonial Land Tenure Systems in India." *American Economic Review,* 2005, *95* (4), pp. 1190-1213.

Basu, Kaushik. *Prelude to political economy: A study of the social and political foundations of economics.* Oxford, UK: Oxford University Press, 2000.

Bayley, Susan. *Caste, society and politics in India from the eighteenth century to the modern age.* Cambridge, Massachusetts: Cambridge University Press, 1999.

Bernhard, Helen; Fehr, Ernst and Fischbacher, Urs. "Group Affiliation and Altruistic Norm Enforcement." *American Economic Review*, 2006, *96* (2), pp. 217-221.

Bowles, Samuel. "Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions." *Journal of Economic Literature*, 1998, *36*(1), pp. 75-111.

Bowles, Samuel and Gintis, Herbert. "The Moral Economy of Communities: Structured Populations and the Evolution of Pro-Social Norms." *Evolution and Human Behavior*, 1998, *19*(1), pp. 3-25.

Butler, Jeffrey; Giuliano, Paola; and Guiso, Luigi. "The Right Amount of Trust," manuscript, European University Institute, 2010.

Carpenter, Jeffrey P. and Matthews, Peter Hans. "Norm Enforcement: Anger, Indignation or Reciprocity?" Manuscript, Middlebury College, 2006.

Chen, Yan and Li, Sherry Xin. "Group Identity and Social Preferences," *American Economic Review,* 2009, *99,* pp. 431-457.

Cialdini, Robert; Borden, Richard; Throne, Avril; Walker, Marker; Freeman, Stephen and Sloan, Lloyd. "Basking in Reflected Glory: Three (Football) Field Studies," *Journal of Personality and Social Psychology,* 1976, *34,* pp. 366-75.

Cinyabuguma, Michael; Page, Talbot and Putterman, Louis. "Can Second-Order Punishment Deter Perverse Punishment?" *Experimental Economics,* 2006, *9*(3), pp. 265-79.

Drèze, Jean and Gazdar, Haris. "Uttar Pradesh: The Burden of Inertia," in Drèze and Amartya Sen, eds. *Indian development*. Oxford, UK: Oxford University Press, 1997.

Drèze, Jean; Lanjouw, Peter and Sharma, Naresh. "Economic development in Palanpur, 1957-93," in Lanjouw and Nicholas Stern, *Economic development in Palanpur over five decades,* Oxford, UK: Clarendon Press Oxford, 1998.

Fehr, Ernst and Gächter, Simon. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review*, 2000, *90*(4), pp. 980-94.

Fehr, Ernst; Gächter, Simon and Kirchsteiger, Georg. "Reciprocity as a Contract Enforcement Device: Experimental Evidence." *Econometrica*, 1997, *65*(4), pp. 833-60.

Fehr, Ernst and Fischbacher, Urs. 2004. "Third Party Punishment and Social Norms." *Evolution and Human Behavior,* 2004, *25*, pp. 63-87.

Fehr, Ernst; Hoff, Karla and Kshetramade, Mayuresh. 2008. "Spite and Development," *American Economic Review, Papers and Proceedings* 98(2): 494-99.

Fisman, Raymond and Miguel, Edward. "Corruption, Norms, and Legal Enforcement: Evidence from Diplomatic Parking Tickets." *Journal of Political Economy*, 2007, *115* (6), pp. 1020-1048.

Foner, Eric.[1988] *Reconstruction: America's unfinished revolution.* Reprinted: New York: Perennial Classics,1989.

Galanter, Marc. "The Abolition of Disabilities, Untouchability and the Law," in J. Michael Mahar, *ed*., *The Untouchables in Contemporary India*. Tucson: University of Arizona Press, 1972, pp. 227-314.

Gellner, Ernest. *Conditions of liberty: Civil Society and its rivals.* New York: The Penguin Press, 1994.

Gintis, Herbert. "A Radical Analysis of Welfare Economics and Individual Development." *Quarterly Journal of Economics*, 1972, *86*(4), pp. 572-99.

Gintis, Herbert. "Punishment and Cooperation." *Science*, 2008, *319*(5868), pp. 1345-46.

Goette, Lorenz; Huffman, David and Meier, Stephan. "The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups." *American Economic Review*, 2006, *96*(2), pp. 212-16.

Greif, Avner. "Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition." *American Economic Review.* 1993, *83*(3), pp. 525-548.

Guiso, Luigi; Sapienza, Paola and Zingales, Luigi. "The Role of Social Capital in Financial Development," *American Economic Review,* 2004, *94*(3), pp. 526-556.

Guiso, Luigi; Sapienza, Paola and Zingales, Luigi. "Does Culture Affect Economic Outcomes?" *Journal of Economic Perspectives*, 2006, *20*(2), pp. 23-48.

Gupta, Dipankar. 1991. *Social Stratification,* New Delhi; New York: Oxford University Press

Gupta, Dipankar. *Interrogating caste: Understanding hierarchy and difference in Indian society.* New York: Penguin Books, 2000.

Hayek, Friedrich A. *Law, legislation, and liberty: Volume 1 - Rules and order.* Chicago, IL: University of Chicago Press, 1973.

Henrich, Joseph; Boyd, Robert; Bowles, Samuel; Camerer, Colin; Fehr, Ernst; Gintis, Herbert and McElreath, Richard. "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies." *American Economic Review,Papers and Proceedings* 2001, *91*, pp. 73–78.

Herrmann, Benedikt; Thöni, Christian and Gächter, Simon. "Antisocial Punishment across Societies." *Science*, 2008, *319*(5868), pp. 1362-7.

Human Rights Watch. *Broken people: Caste violence against India's "Untouchables."* New York: Human Rights Watch, 1999.

Lindbeck, Assar; Nyberg, Sten and Weibull, Jorgen W. "Social Norms and Economic Incentives in the Welfare State." *Quarterly Journal of Econ*omics, 1999, 114, pp. 1–35.

Logan, Trevon D. and Shah, Manisha. "Face Value: Information and Signaling in an Illegal Market." NBER Working Paper 14841, 2009.

McCloskey, Deirdre N., *The Bourgeois Virtues: Ethics for an Age of Commerce.* Chicago: University of Chicago Press, 2006.

Miguel, Edward and Gugerty, Mary Kay. "Ethnic Diversity, Social Sanctions, and Public Goods in Kenya." *Journal of Public Economics,* 2005, *89* (11-12), pp. 2325-68.

Mokyr, Joel. *The enlightened economy: An economic history of Britain, 1700-1850.* Yale University Press and Penguin Press, 2008 (in press).

Munshi, Kaivan and Rosenzweig, Mark. "Why is Mobility in India so Low? Social Insurance, Inequality, and Growth." Unpublished manuscript, 2007.

Narayan, Badri, "History Produces Politics: The Nara- Maveshi Movement in Uttar Pradesh." *Economic and Political Weekly* XLV: 40 (2010), 111-119.

Nikoforakis, Nikos. "Punishment and Counter-Punishment in Public Good Games: Can We Really Govern Ourselves?" *Journal of Public Economics*, 2008, *92*, pp. 91-112.

New York University School of Law Center for Human Rights and Global Justice and Human
Rights Watch. *Hidden Apartheid: Caste Discrimination against India's "Untouchables."*
Shadow Report to the UN Committee on the Elimination of Racial Discrimination, *19* (3)
2007.

Nunn, Nathan and Leonard Wantchekon, "The Slave Trade and the Origins of Mistrust in
Africa," *American Economic Review*, forthcoming.

Oakes, Penelope, Haslam, S. Alexander, and Turner, John C. *Stereotyping and social reality.*
Oxford, England: Blackwell, 1994.

Ostrom, Elinor. "Collective Action and the Evolution of Social Norms." *Journal of Economic
Perspectives*, 2000, *14*(3), pp. 137-58.

Platteau, Jean-Philippe. "Behind the Market Stage Where Real Societies Exist—Part II: The Role
of Moral Norms." *Journal of Development Studies,* 1994, *30*(3), pp. 753-817.

Rao, Vijayendra. "Symbolic Public Goods and the Coordination of Collective Action: A
Comparison of Local Development in India and Indonesia," in Bardhan, Pranab and Isha
Ray, eds., *The Contested Commons: Conversations between Economists and
Anthropologists.* ch. 18, 2008.

Rao, Vijayendra and Walton, Michael. "Culture and Public Action: Relationality, Equality of
Agency, and Development," in Rao and Walton, eds*., Culture and public action.*
Stanford, CA: Stanford University Press, 2004, pp. 3-36.

Rustagi, D., Engel, S., and Kosfeld, M. (2010), "Conditional Cooperation and Costly Monitoring
Explain Success in Forest Commons Management," *Science* 2010, *330*, pp. 961-65.

Sen, Amartya. "Social Exclusion: Concept, Application, and Scrutiny." Social Development
Papers No. 1, Office of Environment and Social Development, Asian Development Bank,
2000.

Shah, Ghanshyam, Mander, Harsh, Thorat, Sukhdeo, Deshpande, Satish and Baviskar, Amita. *Untouchability in rural India.* New Delhi: Sage Publications, 2006.

Sobel, Joel. "Interdependent Preferences and Reciprocity." *Journal of Economic Literature,* 2005, *43*, pp. 392-436.

Tabellini, Guido. "Presidential Address - Institutions and Culture." *Journal of the European Economic Association*, 2008, *6*(2-3), pp. 255-94.

Tajfel, Henri."Social Psychology of Intergroup Relations."*Annual Review of Psychology,* 1982,*33*, pp.1-39.

Thorat, Sukhadeo. "Oppression and Denial: Dalit Discrimination in the 1990s." *Economic and Political Weekly,* 2002, pp. 572-578.

Weingast, Barry. "Political Foundations of Democracy and the Rule of Law." *American Political Science Review,* 1997, *91*(2), pp. 245-63.

Willer, Robb. "Groups Reward Individual Sacrifice: The Status Solution to the Collective Action Problem." *American Sociological Review*. 2009a, *74*, pp. 23-43.

Willer, Robb. "A Status Theory of Collective Action," in Shane R. Thye and Edward J. Lawler, eds., *Advances in group processes.* 2009b, *26*.

Table 1 — Determinants of Third-Party Punishment

	(1)	(2)	(3)	(4)	(5)	(6)
	Relative Punishment			Absolute Punishment		
Constant[a]	2.073***	4.548	4.783*	5.366***	7.258**	7.575**
	(0.789)	(2.868)	(2.831)	(0.765)	(3.061)	(3.040)
Player C is high caste, β	2.749**	2.552**	1.737*	3.086***	3.444***	2.112**
	(1.070)	(1.114)	(0.898)	(1.062)	(1.095)	(0.918)
Treatment is mixed status, γ	0.512	0.678		0.78	1.021	
	(1.284)	(1.296)		(1.341)	(1.340)	
Player C is high caste and treatment is mixed status, δ	-1.368	-1.652		-2.429	-2.703	
	(1.642)	(1.643)		(1.701)	(1.694)	
Individual characteristics of player C						
Land owned in acres (centered)		0.012	0.009		-0.184	-0.191
		(0.143)	(0.142)		(0.141)	(0.140)
Lives in a mud house[b]		0.849	0.707		0.665	0.425
		(0.913)	(0.920)		(0.894)	(0.896)
Has completed at least 10 years of schooling		1.955**	1.896**		1.819**	1.725*
		(0.892)	(0.882)		(0.906)	(0.899)
Is an owner-cultivator		-1.505*	-1.568*		-1.572*	-1.680*
		(0.852)	(0.838)		(0.872)	(0.865)
Has political experience		-1.36	-1.248		-1.159	-0.96
		(1.375)	(1.388)		(1.491)	(1.517)
R^2	0.035	0.081	0.076	0.037	0.088	0.074
F stat: $\text{pun}^{HHH} = \text{pun}^{HLH}$	0.70	0.90		2.48	2.60	
Prob ≥ F	0.404	0.343		0.117	0.108	

Notes: The table reports on player C's punishment of norm violations in treatments LLL, HHH, LHL, and HLH. Each column gives the results of an OLS regression based on 205 observations. Robust standard errors are in parentheses. Significance at the 10% level is represented by a *, at the 5% level by a **, and at the 1% level by ***.

[a] The constant is the baseline group mean with respect to which we measure changes. It is LLL in regressions (1),(2),(4), and (5) and LLL and LHL in regressions (3) and (6).

[b] The omitted category is a house constructed at least partly of brick.

Table 2 ─ Determinants of Third-Party Punishment Including Specific Caste Affiliation

	(1)	(2)	(3)	(4)
	Relative Punishment		Absolute Punishment	
Constant (baseline group mean: Player C is Chamar)	1.933*	4.124	5.263***	6.808**
	(1.068)	(2.908)	(1.090)	(3.143)
Player C is Brahmin	3.294**	3.217**	3.482**	3.945***
	(1.364)	(1.378)	(1.417)	(1.404)
Player C is Thakur	2.458*	2.988**	2.875**	3.999***
	(1.421)	(1.403)	(1.421)	(1.415)
Player C is Pasi	0.273	1.095	0.2	1.04
	(1.290)	(1.282)	(1.343)	(1.347)
Treatment is mixed status	0.506	0.653	0.776	1.002
	(1.288)	(1.297)	(1.343)	(1.338)
Player C is high caste and treatment is mixed status	-1.354	-1.621	-2.419	-2.679
	(1.647)	(1.650)	(1.705)	(1.699)
Individual characteristics of Player C				
Land owned in acres (centered)		0.015		-0.182
		(0.143)		(0.142)
Lives in a mud house[a]		0.858		0.691
		(0.923)		(0.903)
Has completed at least 10 years of schooling		2.001**		1.900**
		(0.896)		(0.935)
Is an owner-cultivator		-1.619*		-1.710**
		(0.830)		(0.860)
Has political experience		-1.408		-1.183
		(1.350)		(1.472)
R^2	0.038	0.085	0.039	0.091

Notes: The table reports on player C's punishment of norm violations by the specific caste of player C in treatments LLL, HHH, LHL, and HLH. Each column gives the results of an OLS regression based on 205 observations. Robust standard errors are in parentheses. Significance at the 10% level is represented by a *, at the 5% level by a **, and at the 1% level by ***.

[a]The omitted category is a house constructed at least partly of brick.

Table 3 ─ Determinants of Third-Party Punishment Including the Effect of In-group Affiliation

	(1)	(2)	(3)	(4)
	Relative Punishment		Absolute Punishment	
Constant (baseline group mean: LLL)	2.073***	5.401*	5.366***	7.882**
	(0.790)	(2.804)	(0.767)	(3.265)
Player C is high caste	2.749**	2.766**	3.086***	3.745***
	(1.072)	(1.167)	(1.064)	(1.104)
Players B and C have the same specific caste but different from player A's	0.149	-0.164	-0.505	-0.817
	(1.179)	(1.138)	(1.262)	(1.203)
Interaction of BC-condition and Player C is high caste	-2.1	-3.050*	-2.742	-3.857**
	(1.752)	(1.734)	(1.776)	(1.717)
Individual characteristics of Player C				
Land owned in acres (centered)		-0.251		-0.466***
		(0.193)		(0.172)
Lives in a mud house[a]		0.221		-0.242
		(0.892)		(0.887)
Has completed at least 10 years of schooling		2.823***		2.339***
		(0.922)		(0.889)
Is an owner-cultivator		-1.618*		-1.821**
		(0.844)		(0.843)
Has political experience		-1.909		-1.455
		(1.295)		(1.573)
R^2	0.043	0.125	0.071	0.161
F-stat: HHH = HHH$_{BC}$	2.270[a]	6.530[a]	6.750[a]	15.320[a]
Prob ≥ F	0.134	0.011	0.010	0.000
F-stat: LLL$_{BC}$ = HHH$_{BC}$	0.220[b]	0.040[b]	0.060[b]	0.010[b]
Prob ≥ F	0.640	0.845	0.809	0.937

Notes: The table reports on player C's punishment of norm violations in treatments LLL, HHH, LLL$_{BC}$, and HHH$_{BC}$. Each column gives the results of an OLS regression based on 178 observations. Robust standard errors are in parentheses. Significance at the 10% level is represented by a *, at the 5% level by a **, and at the 1% level by ***.

[a]The omitted category is a house constructed at least partly of brick.

Table A1 ─ Determinants of Per Capita Consumption

	Per capita adult-equivalent consumption (ln)	
	High caste	Low caste
Ln (land owned)[a]	0.282***	0.084***
	(0.039)	(0.024)
Lives in a mixed mud and brick house[b]	-0.052	0.100
	(0.111)	(0.081)
Lives in a brick house[b]	0.179*	0.238**
	(0.102)	(0.079)
Has completed at least 10 years of schooling	0.304***	0.118*
	(0.093)	(0.067)
Constant	8.001***	7.875***
	(0.082)	(0.033)
Number of observations	131	204
R^2	0.330	0.385

Notes: The table shows OLS regressions of per capita adult-equivalent consumption for the high caste and low caste households, respectively, covered in the survey of Uttar Pradesh in the *World Bank 1997-98 Survey of Living Conditions: Uttar Pradesh and Bihar*. Robust standard errors are in parentheses. Significance at the 10% level is represented by a *, at the 5% level by a **, and at the 1% level by ***.

[a] Land ownership of each household is shifted up by 1/6 acre so that landless households can be included in the regression.

[b] Omitted category is a house of walls constructed only of mud, with either thatch or tiled roof.