Culture and the Efficiency of Coordination: Experiments with High- and Low-Caste Men in Rural India*

Benjamin A. Brooks† Karla Hoff‡ Priyanka Pandey§

February 6, 2015

Abstract

We study the relationship between culture and the efficiency of coordination. In a field experiment in India, men from high and low castes repeatedly played a simple coordination game with efficient and inefficient equilibria. Compared to their low-caste counterparts, the high-caste men coordinated far less efficiently. We trace the divergence in outcomes to how individuals respond to the low payoff that results when a player attempts efficient coordination but his partner does not. After this event, high-caste men are significantly less likely than low-caste men to continue trying for efficiency. This difference can be explained by the culture of honor among the high castes, which may lead them to see this low payoff as an insult rather than an accident and to respond in a manner that impedes efficient coordination.

Keywords: Culture, coordination, stag hunt, learning.

JEL codes: C72, O12, O17, Z1.

*We greatly benefitted from discussions at an early stage of this project with Paul Attewell, Gary Charness, Ernst Fehr, Rocco Macchiavello, Matthew Rabin, and Jennifer Stellar. We thank seminar participants at Indiana University, Princeton University, the World Bank, the University of California–San Diego, the University of California–Santa Barbara, the University of Chicago, University of Chicago Booth School of Business, the University of Memphis, and Warwick University. We also thank Cristina Bicchieri for detailed comments on our draft. Finally, thanks to Sonal Vats for excellent research assistance and the World Bank Research Support Budget for financial support.

†Becker Friedman Institute, University of Chicago, babrooks@uchicago.edu

‡The World Bank, khoff@worldbank.org

§The World Bank, ppandey@worldbank.org

1 Introduction

Much of what we value in society depends on coordination: for example, fiat money, the rule of law, language, and standardization. The conventions that societies coordinate on can vary widely, with some being more efficient than others. Ample evidence suggests that inefficient conventions can impede economic growth (Matsuyama, 1996; Ray, 2000; Hoff, 2001). Thus, it is important to understand why inefficient conventions emerge and persist. Communities differ from one another in many characteristics, but a prominent distinguishing feature of a society is its culture. By culture, we mean the broadly shared set of beliefs and mental models used by a group of individuals to understand the world around them, and the norms and rules of thumb that are used to guide decision making.[1] Conventions do not emerge arbitrarily, but rather through a process of decentralized coordination. Culture may influence this process by affecting how individuals interpret and respond to others' actions.

In this paper, we report evidence that cultural differences can have a large impact on how individuals form a convention. To our knowledge, this is the first experimental evaluation of the effect of culture on the efficiency of coordination.[2] We conducted a field experiment in rural Uttar Pradesh, India, in which men from different levels of the Indian caste hierarchy were paired anonymously to solve a simple coordination problem. This part of the world is notable in that individuals from different casteslive side-by-side, speak the same language, and overlap in the distribution of wealth and education, but are distinguished by culture. By comparing behavior across these two populations, we can identify the relationship between culture on behavior without the concern that the variation is due to translation or differences in socioeconomic characteristics Andersen et al. (cf 2008).

The participants in our experiment were men from the extreme top and bottom ranks of the Hindu caste system. These subjects repeatedly played the simple coordination game known as the Stag Hunt. This game has two Nash equilibria, only one of which is Pareto efficient. The efficient action, however, entails some risk, in that trying for efficiency when the other player does not results in a low reward that we term the *loser's payoff*. In contrast, the inefficient action results in a the same reward, regardless of the other player's action. The Stag Hunt experiment allows us to investigate how the different cultures at the top and bottom of the caste hierarchy affect the players' ability to coordinate on the efficient outcome.

[1] This definition follows much recent work on culture and its effect on economic outcomes (e.g., DiMaggio, 1997; Alesina, Giuliano and Nunn, 2013; Algan and Cahuc, 2013). This definition differs from other common usages of the word culture, such as to describe the collective artistic and intellectual accomplishments of a society.

[2] An important and recent paper of Jackson and Xing (2014) also addresses how culture affects coordination outcomes. They show that nationality (Chinese versus American) is strongly related to the distribution of payoffs and *utilitarian* efficiency, whereas we show that culture can have a large effect on *Pareto* efficiency.

A priori, there are many reasonable conjectures for how the subjects' caste statuses would correlate with the outcome of coordination. One might predict that coordination would be more efficient between individuals with similar cultures, since social proximity might lead to more trust or to greater concern for one's partner's welfare. Alternatively, one might predict that non-cultural traits would shape the outcome of the experiment: high-caste men generally have more education than low-caste men, and might have an easier time understanding and "solving" the game.

The striking result of the experiment was that low-caste individuals coordinated far more efficiently than their high-caste counterparts. Low-caste men playing with other low-caste men achieved the highest rate of efficient coordination, followed by low-caste men playing with high-caste men, with high-caste men playing with other high-caste men having the least efficient outcomes. For example, 73 percent of low-low pairs played the efficient equilibrium in the final round of the partnership, compared to 50 percent of low-high, and only 32 percent of high-high pairs. Caste status is correlated with many other individual characteristics, and one might suspect that it is covariates of caste rather than caste culture itself that is driving the divergence in outcomes. However, we observe measures of the subjects' wealth and education, and once we control for caste status in regression analysis, these covariates are statistically insignificant.

This caste gap in pair outcomes is the consequence of a caste gap in individuals' decisions about what action to take. Even in the first period of play, there are differences in behavior between high- and low-caste subjects, with 16 percent more low-caste than high-caste players initially trying for the efficient outcome. However, this difference is modest in comparison to the disparity in behavior that arose later in the experiment in response to the history of play. A natural descriptive statistic to consider is how likely an individual is to attempt efficient coordination conditional on the previous period's outcome. For example, one might suppose that after a round of efficient coordination, players would be likely to try again for efficient coordination in the subsequent round. This is indeed the case: for every pair of caste statuses, over 85 percent of individuals try again for efficiency after achieving efficient coordination. In fact, we find that there is only one outcome after which the caste difference in behavior is statistically significant, which is when a player attempts efficient coordination while his partner does not and therefore obtains the loser's payoff. After this event, 71 percent of low-caste individuals attempted efficient coordination, compared to only 42 percent of high-caste players. This difference is even more stark if one compares high-high and low-low pairs, for which 68 percent of low-caste subjects try again for efficiency versus only 32 percent of high-caste subjects, a difference of 36 percentage points! Caste differences after other one-period histories are much smaller and statistically insignificant. Regression analysis further

confirms that this is the outcome after which the largest difference between high- and low-caste behavior occurs, even after controlling for other individual characteristics.

Why do high- and low-caste individuals behave differently after obtaining the loser's payoff? To address this question, we conducted a survey on how individuals would respond to everyday situations that produce for them an outcome similar to the loser's payoff. The loser's payoff has the feature that the player receives a low payoff because of his partner's action, but the partner's *intention* is not obvious. Should the partner's behavior be interpreted as a malicious offense worthy of punishment, or as an accident to be forgiven? In a survey, we presented high- and low-caste men with different vignettes in which a person A takes some action that adversely affects a person B, though A's intentions are ambiguous. For each vignette, the survey respondents were asked: is B's action justified? In every such scenario, high-caste men were much more likely than low-caste men to think that the severe punishment was justified. As an example, one scenario has A marry B's daughter, even though the latter is from a slightly higher-status caste. Subsequently, B encounters A in the village, they argue, and B physically assaults A. For this vignette, 70 percent of high-caste respondents thought that B was justified in assaulting A, versus only 22 percent of low-caste men. The survey also included a control vignette in which it is obvious that A intended B harm. In this case, a high proportion of both high- and low-caste respondents approved of B's behavior, indicating that it is only when intentions are ambiguous that high- and low-caste culture lead to different behavior. We also asked the respondents to explain the reasons behind their response. The comments for the marriage scenario are especially telling. Approximately half of those who thought that B was justified explained that the punishment was necessary to preserve *izzat*, which is the Hindu word for honor.

The survey results are consistent with differences between high- and low-caste culture as described in the anthropology and sociology literatures. In particular, the high castes of north India possess a version of what has been called the "culture of honor", in which individuals are highly concerned with protecting and advancing their honor in the eyes of their peers. Honor is diminished by being in a subordinate position to others, and honor is enhanced by aggressively defending oneself against perceived insults or offenses. In contrast, the concern for honor is less characteristic of low castes, who have traditionally been assigned a subordinate role in the north Indian social structure. This evidence supports hypothesis that the culture of high-caste men leads them to respond differently than low-caste men to miscoordination. A player who receives the loser's payoff has a lower reward than his partner, which can be construed as a subordinate position. Under this interpretation, the appropriate response may be to play the inefficient action in order to punish the partner and preserve honor.

A subtlety of this story is that the culture of honor has a direct effect on behavior only at the miscoordination stage, i.e., when a player receives the loser's payoff. If players could deterministically play the efficient action forever, miscoordination would not occur, no one would be offended, and the culture of honor would have no effect. However, if players have to converge to efficient coordination along a path of behavior that involves miscoordination, then the culture of honor could divert the players to the inefficient outcome. We illustrate this mechanism within the standard learning model of logistic fictitious play, which posits that subjects play a noisy best response to the historical distribution of their partner's actions. For the Stag Hunt, the distribution of actions converges over time to a steady state that approximates one of the two Nash equilibria. We add to this basic framework the feature that preferences depend on how others' intentions are construed: if a player feels insulted by the loser's payoff, then he prefers that his partner receive a low payoff in the subsequent round. We find that if this preference for punishment is sufficiently strong, then efficient coordination is unsustainable in the long run. The logic is straightforward. Even starting from a regime in which players expect a high degree of efficient coordination, mistakes will inevitably occur due to the noise in behavior. This causes players to become insulted and the subsequent punishment lowers the empirical frequency of attempts at efficient coordination. This in turn leads even non-insulted players to go for efficiency less often, which increases the likelihood of miscoordination, leading to more insults, and an unraveling of the efficient equilibrium ensues.

Thus, the culture of honor among high-caste men can explain the caste gaps in (a) behavior after the loser's payoff, (b) the rate of coordination on the efficient outcome, and (c) responses to the survey on appropriate responses to injury. An alternative hypothesis that would explain these findings is that high-caste individuals are less trusting of their partners than are individuals from low castes, which makes high-caste men more likely than low-caste men to play the riskless but inefficient action. We implemented a second experiment in which high- and low-caste subjects played the Trust Game: one player, the principal, is given an endowment and can choose whether or not to give it to another player, the agent. If the principal hands over his endowment, it grows by a large amount. Subsequently, the agent has to choose whether or not to return a share of the surplus. The principal will hand over the endowment only if he "trusts" the agent, in the sense of assigning a high probability to the event that the agent returns enough of the surplus to make the investment worthwhile. We implemented this experiment with all combinations of caste statuses in the principal and agent roles. We find no statistically significant caste difference in the rate at which principals share with the agents. In fact, more principals share when both principal and agent are high caste than when both are low caste.

We also consider the possibility that low- and high-caste cultures have the same rules for understanding others' intentions and for how to punish, but that low-caste individuals simply feel that they are unable to respond to transgressions like those depicted in the vignettes. We conducted a survey involving 30 high-caste and 32 low-caste men to test for caste differences in self-efficacy. We found no evidence that low-caste men have lower locus of control than high-caste men. Thus, heterogeneity in trust and differences in self-efficacy are unlikely to explain the relationship between caste status and coordination. Based on the Trust Game experiment and the surveys on culture, we maintain that the culture of honor is the most likely explanation.

Our work is related to voluminous literatures in economics, psychology, sociology, and anthropology that address culture and its effects on economic outcomes. The topic has recently experienced a resurgence of interest within economics. This literature expresses a diverse array of viewpoints on the definition of culture on the mechanisms through which culture operates. Greif (1994) compares institutions across societies that derive from either Medieval Latin or Muslim culture, and argues that a cultural legacy of collectivism versus individualism affects which kinds of institutions emerge. Lazear (1999) views culture as being tied to language, and examines how economic incentives affect the assimilation of immigrants into American culture through learning English. Algan and Cahuc (2013) survey work on the relationship between trust and economic growth, with the qualities of trustfulness and trustworthiness being one aspect of culture. Alesina, Giuliano and Nunn (2013) study how a society's historical production technology can affect today's labor force participation of women through persistent norms and beliefs about gender roles. See also Fisman and Miguel (2007), Fehr, Hoff and Kshetramade (2008), Tabellini (2008), Greif and Tabellini (2010), Henrich and Ensminger (2011), and Fernández (2013).

Our work studies how culture affects the way individuals solve coordination problems. A particularly close reference on this agenda is Jackson and Xing (2014), who study how cultural differences can affect the *distributional* consequences of coordination: they show that Chinese subjects are more likely to play an equilibrium with unequal payoffs, while American subjects play equilibria with egalitarian payoffs. In light of their result, it should not be surprising that culture can also have a large effect on the Pareto efficiency of the outcome of coordination. Economists have identified several generic impediments to efficient coordination: large numbers of individuals (Schelling, 1978; Van Huyck, Battalio and Beil, 1990), pessimistic beliefs (Merton, 1968; Hoff and Stiglitz, 2002; Sapienza, Zingales and Guiso, 2006; Tabellini, 2008), and social distance (Chen and Chen, 2011). We argue that culture is an additional factor that can help or hinder individuals' ability to coordinate on efficient outcomes.

Culture affects coordination through at least two channels. First, culture is a source of *mental frameworks* for interpreting sensory input, including feedback from strategic interactions. For example, is a given action an accident, an insult, or something else? Second, culture provides *rules of thumb* that guide behavior. What is the culturally appropriate response to a perceived insult? We argue that these elements of culture influence what kind of convention will emerge through the way individuals respond to miscoordination. A mental framework that leads a player to interpret miscoordination as an insult and a rule of thumb that prescribes punishment as the response will together impede the individual's ability to coordinate with his partner on an efficient outcome.

Rules of thumb, in the sense we are using the term, are closely related to the notion of culture as a "toolkit" in Swidler (1986) and Henrich and Ensminger (2011), and also work by Nunn and colleagues (e.g., Alesina, Giuliano and Nunn, 2013). Mental frameworks, referred to as schemas in the psychology literature, have long been a subject of interest. Kahneman (2011) describes a classic experiment that illustrates how mental frameworks shape the way humans process information: the psychologists Heider and Simmel showed experimental subjects an animated film in which geometric shapes floated around and "interacted" with each other. The subjects subsequently interpreted the film by assigning decidedly human motives and emotions to triangles and circles. The point here is that even though emotion was not inherent in the movement of the geometric shapes, the subjects used mental frameworks to interpret and assign meaning to what they saw. The notion of mental frameworks is closely related to the concept of *frames* in economics and psychology. A frame is an aspect of the way a situation is presented to the individual, whereas a mental framework is brought by the individual to the situation. For both concepts, the key idea is that events and actions do not always speak for themselves, but instead may depend on framing for their meaning (Goffman, 1974). Moreover, changing the frame can lead to a large difference in behavior, even though there is no change in the possible actions or material outcomes (Andreoni, 1995; Ariely, 2009).

Many mental frameworks and rules of thumb are not universal, but are specific to a given region or culture (Zerubavel, 1999; Heine, 2012). A prime example is the culture of honor, which is broadly characteristic of the US South but not of the US North. This difference is the subject of a set of celebrated studies by Cohen, Nisbett, and colleagues. Nisbett and Cohen (1996) show that southerners respond more positively than northerners to job applicants who have committed a violent crime that was in defense of honor, whereas there was no difference for other kinds of crime. Cohen et al. (1999) expose northern and southern men to the same provocative behavior, but find significantly more aggressive responses from the southern subjects.

The high castes of north India have been characterized as possessing a similar culture of honor. Hitchcock (1958) describes a typical member of the martial high castes in Uttar Pradesh in the 1950s as "brave, mettlesome, and very quick to perceive and resent an insult. It is part of his code that a slight to his prestige should be avenged" (p. 12). Similarly, Drèze and Gazdar (1997) write that for Thakurs, one of the highest castes, "honor lies pimarily in not doing anything that would put them in a position of subordination or moral debt" (pp. 32-33). The anthropologist David Mandelbaum (1993) writes that although it is hard earned, "[honor] has to be continually reaffirmed in practice, reinforced in action, defended against challenge, and rewon and advanced in competition" (p. 23). The sociologist Steve Derné (1992, pp. 277-9) reports on the mechanisms that perpetuate the culture of honor: men "see their family honor as important for their success... Men who dishonor themselves jeopardize marriage prospects for themselves, their children, and their brothers and sisters." Concern with honor is prevalent among high castes but not low castes. Since social norms assign low-caste individuals inferior institutional roles, they do not have either the need or the opportunity to develop the culture of honor that is typical of the high castes (Khare, 1984).

To summarize, culture influences the mental frameworks and rules of thumb that guide individual behavior in strategic settings. In Uttar Pradesh, the concern for honor is part of the culture of the high castes and not of the low castes, which could lead individuals from these groups to behave differently in the same strategic situation.

Our paper also relates to the study of fairness in behavioral economics, such as the concept of fairness equilibrium (Rabin, 1993), in that we consider how the perception of others' intentions is incorporated into preferences. A difference is that we study how such perceptions will influence how individuals learn to play a game over repeated interaction, whereas work on fairness equilibrium considers predictions in static settings. And finally, we contribute to the experimental and theoretical literatures on learning in games (Van Huyck, Battalio and Beil, 1990, 1991; Van Huyck, Cook and Battalio, 1997; Battalio, Samuelson and Van Huyck, 2001; Fudenberg and Levine, 1998).

We wish to emphasize that while our field experiment studies coordination in an artefactual setting,[3] there is no shortage of real world problems that are the result of or are exacerbated by inefficient coordination. Economic development has long been viewed as a process of coordinating on efficient institutions (Matsuyama, 1996; Ray, 2000; Hoff, 2001), and inefficient coordination and the resulting adverse consequences for development have been documented by many researchers, including North (1990, 2006), Greif (1994), Hoff and

[3]We borrow this terminology from Harrison and List (2004) to refer to field experiments in which subjects are put in strategic settings that are not part of their everyday life.

Stiglitz (2002), and Guiso, Sapienza and Zingales (2008). Uttar Pradesh in particular has a long history of institutional failures that have stymied efforts at economic development. Traditionally, basic services such as street maintenance and sanitation were provided by those in the lower ranks of the caste hierarchy via a system of customary obligations. That system eroded following Indian independence and the emancipation of the laboring classes, but as of fifty years later Drèze and Gazdar (1997) write that "the challenge of creating [institutions]... to address those needs has been largely unmet." Drèze and Sharma (1998) describe numerous examples of inefficient conventions in their case study of the village of Palanpur, such as a planting season that starts much too late and the lack of effective drainage systems. While we would not claim that inefficient coordination is the primary culprit behind underdevelopment in UP, our results suggest that improving the efficiency of coordination is one way to promote economic growth.

The rest of the paper is structured as follows. We describe the design of our Stag Hunt experiment in Section 2. Section 3 presents our analysis of the outcome of the experiment and documents the caste differences in behavior described above. Section 4 describes our survey on attitudes towards punishment, which provide concrete evidence of cultural differences betweeh high and low castes. Section 5 develops a mechanism by which the culture of honor can impede efficient coordination within the learning model of logistic fictitious play. Section 6 discusses our field experiment on trust and survey on self-efficacy, and Section 7 concludes. Appendices contain the formal analysis of the model of Section 5 and supporting materials from the experiments.

2 The Stag Hunt experiment

In September 2005, we conducted a field experiment in the district of Lucknow in the Indian state of Uttar Pradesh. We recruited 122 male subjects, half of whom were drawn from "General Castes" at the top of the caste hierarchy, with the remaining half drawn from the "Scheduled Castes" at the bottom.[4][5] Throughout the paper, we will write H to denote a subject from a high-status caste and L to denote a subject from a low-status caste.

The subjects were organized into pairs and repeatedly played the coordination game known as the Stag Hunt, which is depicted in Figure 1. The name Stag Hunt comes from

[4]General Castes and Scheduled Castes are the official terms used by the Indian government. Each group comprises about one-fifth of the population of the state of Uttar Pradesh. Scheduled Castes were formerly known as "untouchables".

[5]The exact caste composition of the high-status sample was: Brahmin (55 percent), Thakur (44), and Lala (1). For the low-status sample, the distribution was: Chamar (56), Rawat (38), Dhanuk (3), and Pasi (2). Two percent of the low-status sample identified themselves by the generic term for low caste, *harijan*.

	Contribute 6 (*Stag*)	Contribute 2 (*Hare*)
Contribute 6 (*Stag*)	(10,10)	(3,7)
Contribute 2 (*Hare*)	(7,3)	(7,7)

Figure 1: Payoffs of the period game (in rupees)

a parable told by Rousseau: Each of several hunters has a choice between hunting a stag and hunting a hare. Regardless of what others do, pursuing the hare is riskless in the sense that it will always result in a meager meal. If everyone pursues the stag, the stag hunt will be successful, and each hunter will obtain a rich meal. However, if some hunters pursue the hare, the stag hunt will be unsuccessful, and a player who attempts to hunt the stag will not eat at all. The Stag Hunt has long been viewed as a metaphor for the problem of coordinating on efficient conventions (cf. Skyrms, 2003).

We implemented the Stag Hunt as follows. Each subject was given an initial endowment of 6 rupees (equivalent to 15 US cents). They were then asked to choose between contributing either 2 or 6 rupees to a common pool. If a player contributed 2, he received back 3 rupees, regardless of the choice of the other player, for a total payoff of 7. On the other hand, if a player contributed the entire endowment of 6, he received a payoff of 10 if the other player also contributed 6, but received only 3 if the partner contributed 2. In this case, the player's payoff of 3 rupees was lower than the partner's 7. We refer to this outcome as the *loser's payoff*. The device of presenting the players with an endowment was done to frame the loser's payoff as a loss in the sense of Kahneman and Tversky (1979), since it represents a loss of 50% of the initial endowment of 6 rupees.

For the remainder of the paper, we eschew the labeling of actions as "Contribute 6" and "Contribute 2" in favor of the more colorful *Stag* and *Hare*, respectively. However, we wish to emphasize that the experiment was explained to the subjects using neutral language: invest 6 or 2 tokens in the common pool.

Each subject played the Stag Hunt for five periods with a partner of the same status and for five periods with a partner of the opposite status, for a total of ten periods of play. Thus, we have three groups: low caste playing with low (LL), high playing with high (HH), and low playing with high (LH). We used a counterbalanced design to avoid confounding the caste status of the pair and the order of play: In Cohort I, subjects were organized into LL and HH pairs for periods 1-5, followed by LH for periods 6-10. In Cohort II, the order

was reversed: *LH* for periods 1-5, and *LL* and *HH* for periods 6-10. Subjects were randomly assigned to a cohort. We will refer to a set of five periods played with a fixed partner as a "pairing".

Before the game began, all subjects were given an orientation from the same individual who explained the overall structure of the game. Subjects were given a plastic "game box" with which to visualize the game. The box was separated into bins, in which plastic tokens represented the player's and partner's contributions and payoffs for each period. Each subject was assigned a monitor who stayed with the subject until the experiment was over, and the subject was not allowed to talk to anyone besides the monitor for the duration of the game. However, the monitor never saw the decisions of the subject or his partner, so that the monitor could not indicate a judgment of approval or disapproval. At the beginning of each pairing, the monitor told the subject the caste status of his partner but did not convey any other information about the partner's identity. After the tenth period, players received their payoffs in private and in cash, with an exchange rate of one rupee per token. Each experimental session lasted about four hours. Mean earnings were 77 rupees, which was 77 percent of the maximum of 100 rupees and approximately 1.5 times the daily unskilled wage.

The subjects were drawn from seven different villages within the block of Bakshi Ka Tulab in the Lucknow district. Within each village, we drew about 12 subjects, of which an equal number were from high castes and low castes. Subjects were recruited using systematic sampling. As an example, if 30 low-caste households lived in the village, every fifth household was chosen, starting from the first that was encountered. Every selected household volunteered one adult male to participate in the game. Low- and high-caste subjects were kept in two separate locations within their own neighborhoods, and partners with the same caste status were kept in separate areas of the experiment site. Cellular service was unreliable, so runners were used to communicate to each player their partner's actions as the experiment progressed.

The experiment was designed to answer two questions. First, *how does a community's culture affect the ability to establish an efficient convention?* Any number of cultural differences between high- and low-caste individuals—in trust, social preferences, mental frameworks, and rules of thumb—could lead to differences in coordination outcomes.

Second, *does social distance affect the ability to coordinate?* By social distance, we mean the perceived status difference between individuals. A decrease in social distance leads, in general, to an increase in empathy, which could lead to the selection of more efficient outcomes (Bernhard et al., 2006; Charness and Gneezy, 2008; Chen and Li, 2009). In the next section, we will explore the outcome of the Stag Hunt experiment for answers to these questions.

3 Results of the Stag Hunt

3.1 Pair outcomes

A striking pattern that emerged from the experiment is that pairs with low-caste players achieved more efficient outcomes than pairs with high-caste players. Over the course of the entire experiment, LL pairs spent 59 percent of periods in efficient coordination (*Stag*,*Stag*), versus 49 percent for LH, and only 34 percent for HH. These numbers are even more stark for periods 5 and 10, after the subjects have had ample time to learn a convention. In these final periods of a pairing, the proportion of pairs playing (*Stag*,*Stag*) was 73 percent for LL, 50 percent for LH, and only 32 percent for HH. All of these differences are statistically significant at the 5 percent level.[6]

These outcomes are graphically depicted in Figure 2. The top panel shows the outcome distribution over all periods of a pairing, and the bottom panel shows results for final periods. HH pairs spend much more time in inefficient coordination than do LH or LL. In particular, HH pairs spent 29 percent of all periods and 35 percent of final periods in inefficient coordination. The comparable numbers for LH and LL pairs pooled are 11 percent overall and 9 percent in the final period. We reject the hypothesis that LL and LH pairs were in inefficient coordination at the same frequencies as HH at the 5 percent level of significance, both for all periods pooled and for final periods. The difference in behavior had a substantial impact on earnings over the course of the experiment. HH pairs generated an average surplus per player of 73 rupees, while LL pairs' average surplus was 82 rupees, approximately 12 percent higher.

Thus, efficiency of coordination is strongly related to caste status. Of course, caste status is itself correlated with many other individual characteristics, such as education and wealth, and it is important to consider whether these covariates can explain the difference in pair outcomes. We collected data in post-play interviews on land ownership, education, and housing.[7][8] Figure 3 shows the distribution of these covariates by caste status. Within our sample, H subjects own more land, are more likely to have a high school education, and are more likely to live in a house that is not made out of mud. However, there is still substantial within-group variation, and the distributions of these characteristics overlap. For example,

[6] All standard errors for the pair analysis are clustered by village.

[7] Houses were classified as either brick (*pucca*), impermanent materials like sticks and mud, or a mixture of the two.

[8] Using data from the 1997-98 Survey of Living Conditions in Uttar Pradesh, we find that land ownership, housing, wealth, and education have a large and significant impact on adult per capita consumption. Together, these variables explain a substantial part (between 30 and 40 percent) of the variation in consumption for both high- and low-caste individuals (Hoff, Kshetramade and Fehr, 2011).

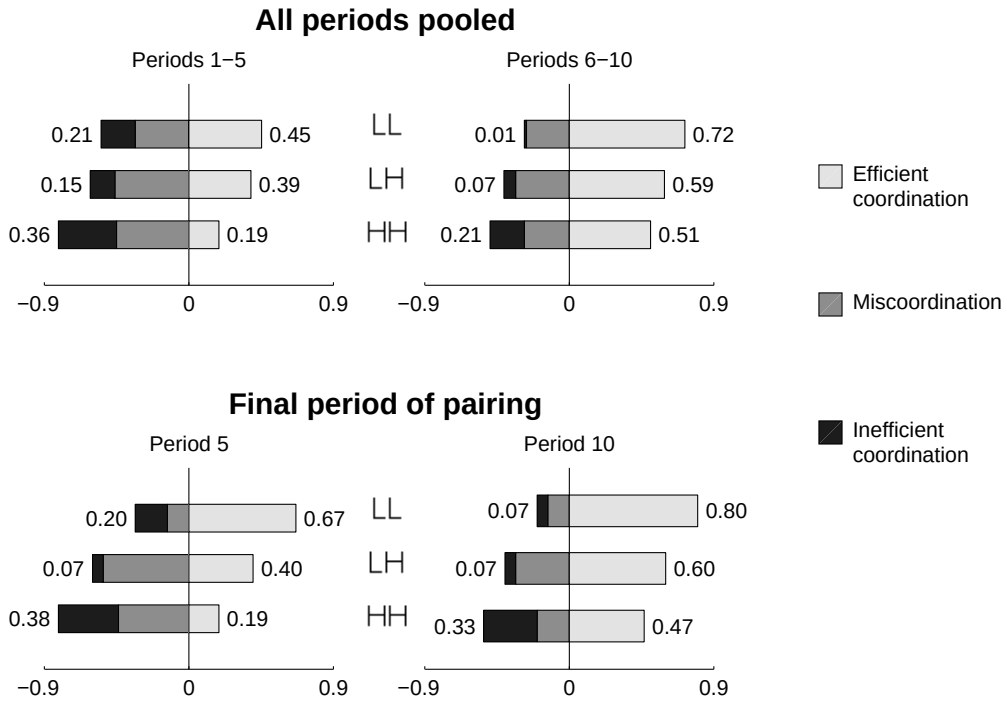

Figure 2: Outcomes for fixed pairs in the Stag Hunt. Black bars are proportional to the number of pairs with outcome (*Hare*,*Hare*), gray bars are proportional to the number of pairs with outcomes (*Hare*,*Stag*) or (*Stag*,*Hare*), and light gray bars are proportional to the number of pairs with outcome (*Stag*,*Stag*).

approximately 15 percent of the low-caste sample completed high school, compared to 18 of high-castes; 63 percent of low-caste subjects live in a mud house, compared to 19 of high-caste.

Table 1 reports probit regressions of whether or not the outcome in a given period was efficient as a function of the caste statuses of the players, mean land holdings, and the percentages of the players who have high school educations and live in non-mud houses. Column 1 pools all periods, and column 2 restricts the sample to final periods. The omitted case is that both players are from low-status castes. We report average marginal effects, so the interpretation of the coefficient on *HH* in column 1 is that on average, *HH* pairs were 25 percent less likely to achieve the efficient outcome than *LL* pairs. For both samples, the coefficient on *HH* is significantly different from zero at the one-percent level, and for the final period sample, we also reject that the coefficient on *LH* is zero. In both regressions, the coefficients on land, education, and house type are jointly insignificant at all conventional levels ($p > 0.5$).

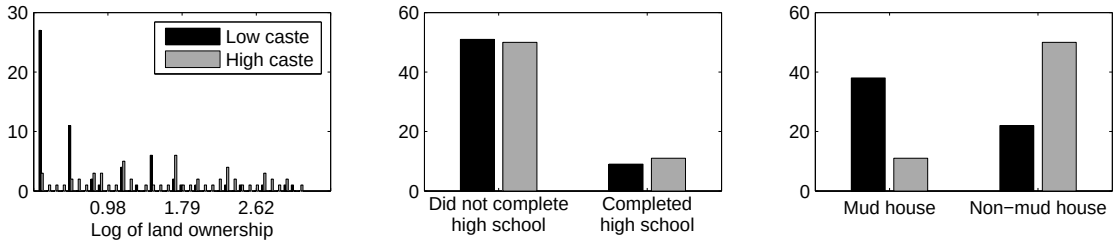

Figure 3: Distribution of covariates by caste status

We can unpack how the distribution of outcomes evolves over time for different caste statuses. Van Huyck, Battalio and Beil (1990) found that the frequency of efficient coordination in a Stag Hunt increased over the first five periods, and then did not deviate from that high level over the next five periods of a ten-period game. Did a similar pattern emerge in our experiment?

In Figure 4, we report the period-by-period outcome shares for each pair type. Over the first five periods, the proportion of *LL* pairs with the outcome (*Stag*,*Stag*) increased more or less steadily from 27 to 67 percent, a statistically significant difference ($p < 0.001$). In contrast, among *HH* and *LH* pairs, there was no statistically significant time trend to the proportion of efficient outcomes, and a plurality of both *HH* and *LH* pairs were in miscoordination or the inefficient equilibrium at period 5. For periods 6-10, the *LL* pairs consistently maintained an efficient convention, with between 67 and 80 percent of pairs playing (*Stag*,*Stag*). The *HH* pairs initially started without a single inefficient outcome, though a majority started in miscoordination. However, the fraction of pairs playing (*Hare*,*Hare*) rose over time, from zero in the initial period to 33 percent in the last two periods, while there was no time trend to the proportion playing (*Stag*,*Stag*). Evidently, the *HH* pairs learned to avoid disequilibrium by settling on an inefficient convention.[9] The time trends for *LH* pairs are more erratic. A majority achieved efficiency in period 10, though this share is not significantly different from 50 percent at the 5 percent level.

To sum up, the evidence strongly supports the conclusion that high caste status is associated with less efficient coordination. Most *LL* pairs established an efficient convention, while most *HH* pairs did not, with outcomes for *LH* pairs in between. Moreover, the summary statistics and regression results indicate that social distance was not an obstacle to efficient coordination. Mixed-status pairs achieved the efficient outcome at a higher rate than same-status high-caste pairs, and they achieved a substantially lower rate of inefficient coordination. We summarize the pair results below:

[9]We know that the *H* players who move *to* (*Hare*,*Hare*) are generally moving *from* (*Stag*,*Hare*) or (*Hare*,*Stag*), since almost 90 percent of *H* players in (*Stag*,*Stag*) in a given period of a fixed pairing continue to play *Stag* in the next period. See Table 2.

	All periods	Final period
HH	-0.268*	-0.427***
	(0.108)	(0.0819)
LH	-0.112	-0.260*
	(0.0721)	(0.101)
Land	0.00505	0.00710
	(0.0106)	(0.0148)
High school	0.0778	0.123
	(0.133)	(0.0946)
Non-mud house	0.0131	-0.00926
	(0.0768)	(0.137)
N	605	121

dfdx coefficients; se_dfdx in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1: Regression analysis for pair outcomes.

Result 1 (Caste gap in pairwise efficiency)**.** *HH pairs are less likely to coordinate on the efficient outcome than LH, who are in turn less likely to coordinate on the efficient outcome than LL.*

3.2 Individual behavior

To understand the pair outcomes, we need to unpack how individual behavior varied by caste. Table 2 shows the proportion of subjects playing *Stag*, broken down by caste status and pair treatment. In addition to reporting proportions for all periods, we also separate out the observations in initial periods of play, i.e., periods 1 and 6, and by the pair outcome in the preceding period for periods 2 to 5 and 7 to 10. It is natural to think that subjects use the history of the game in deciding which action to play. In principle, the entire history could factor into the decision, and the mapping from histories to actions could be quite complicated. Conditioning on the entire history is both impractical and unlikely to be edifying. The previous period's outcome, on the other hand, is in all likelihood the most prominent feature of the history and therefore might capture, to a first order, the relationship between history and behavior.

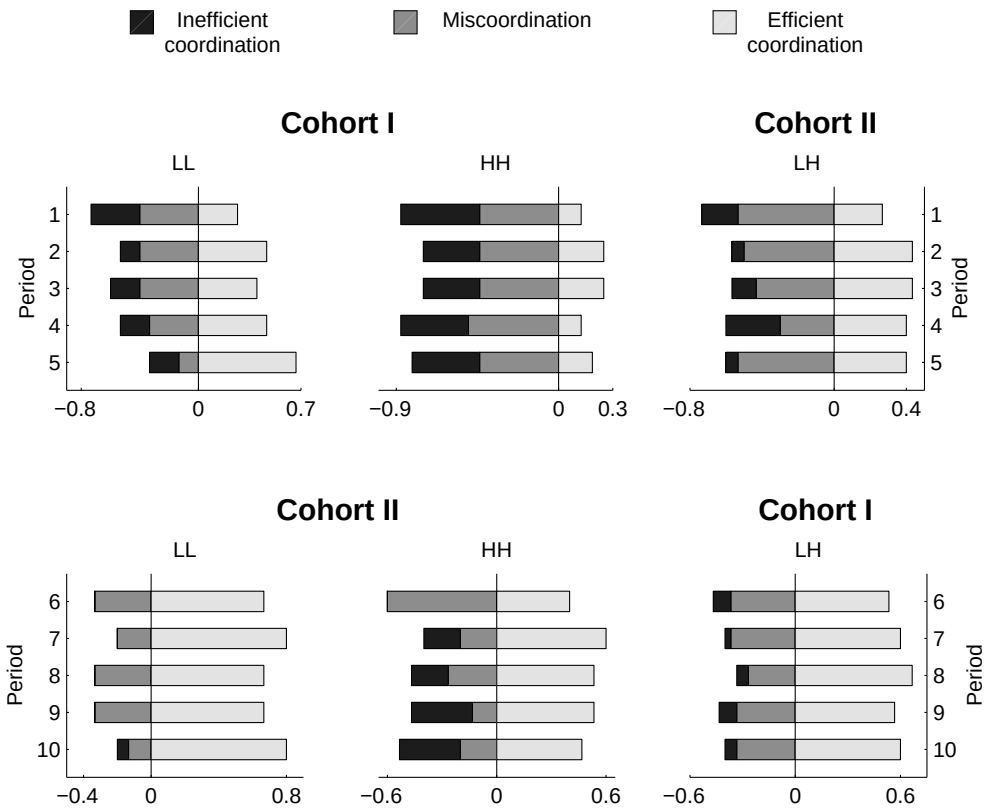

Figure 4: Period-by-period pair outcomes for the Stag Hunt.

We see in column 2 of Table 2 that even in initial periods of a pairing, before there is any history to condition on (with the current player at least), there is already a caste gap in behavior. 68 percent of low-caste subjects play *Stag* in the initial period, as opposed to only 53 percent of high-caste subjects, a difference which is statistically significant at the 5 percent level according to a two-sided t-test. This gap is more or less the same across single- and mixed-status pairs.

Columns 3-7 show the proportion who played *Stag* conditional on the previous period's outcome. In the column label (x, y), x denotes the player's own action and y is the partner's action. For instance, the last column shows that after the efficient outcome (*Stag*,*Stag*), subjects in all three pair types played *Stag* at virtually the same high rate, which was nearly 90 percent.

For other histories, however, there are substantial caste differences in behavior. The largest such difference occurs after the outcome (*Stag*,*Hare*), that is, after the subject received the loser's payoff. The proportions for this history are reported in column 5: 71 percent of low-caste players chose *Stag* in the next period but only 42 percent of high-caste

			Preceding outcome was:			
	All periods	Initial periods	(*Hare*,*Hare*)	(*Hare*,*Stag*)	(*Stag*,*Hare*)	(*Stag*,*Stag*)
L	**0.74**	**0.68**	**0.52**	**0.58**	**0.71**	**0.88**
in LL	0.74	0.65	0.50	0.66	0.68	0.86
in LH	0.74	0.70	0.53	0.47	0.72	0.89
H	**0.58**	**0.53**	**0.31**	**0.46**	**0.42**	**0.87**
in LH	0.64	0.55	0.40	0.46	0.56	0.86
in HH	0.53	0.52	0.26	0.47	0.32	0.88
$L - H$	**0.16**	**0.15**	**0.21**	**0.12**	**0.29**	0.01
$LL - HH$	0.21	0.13	0.24	0.19	0.36	-0.02
N	1,210	242	154	181	181	452

Table 2: Proportion of subjects who chose *Stag*

players, a difference of 29 percentage points! This is the only history after which the caste difference is statistically significant according to a two-sided t-test ($p < 0.05$). The gap is even more stark if one compares the single-caste pairs, in which 36 percentage points more low-caste than high-caste players tried again for efficiency.

It is important to note that there are also substantial caste gaps after the histories (*Hare*,*Stag*) and especially (*Hare*,*Hare*). In the latter case, 21 percentage points more low-caste subjects played *Stag* than high-caste subjects, and for single-caste pairs the gap was 24 percentage points. These differences no doubt contributed to the divergence in pair outcomes, though the data seem to indicate a particular salience of the outcome (*Stag*,*Hare*).

As with pair outcomes, we use regression analysis to test whether or not differences in individual behavior are explained by variation in wealth and education. Probit regressions will also allow us to better control for unobserved individual characteristics. For example, some individuals may be more trusting or have more pro-social preferences, which would be correlated with efficient outcomes. We can proxy for such characteristics using the player's

		Preceding outcome was:			
	Initial	(*Hare*,*Hare*)	(*Hare*,*Stag*)	(*Stag*,*Hare*)	(*Stag*,*Stag*)
H	-0.161*	-0.226	-0.0931	-0.363***	0.0169
	(0.0712)	(0.170)	(0.114)	(0.100)	(0.0362)
Type	0.347***	-0.118	0.194	-0.0509	0.110**
	(0.0573)	(0.105)	(0.109)	(0.0612)	(0.0452)
Land	0.00579	-0.00505	-0.00156	0.00642	-0.00138
	(0.00399)	(0.00770)	(0.00686)	(0.00437)	(0.00239)
High school	0.0643	-0.140	-0.0921	0.160	-0.0779
	(0.0870)	(0.144)	(0.137)	(0.102)	(0.0587)
Non-mud house	0.0523	0.0309	0.0506	0.0559	0.0247
	(0.0628)	(0.115)	(0.102)	(0.0855)	(0.0447)
N	242	154	181	181	452

dfdx coefficients; se_dfdx in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Marginal probability that a player chose *Stag*. Probit regression analysis for *H* versus *L*.

first-period action, which obviously was not influenced by the subsequent history of the game. We refer to this first-period action as that player's "type".[10]

The results of probit regressions are displayed in Tables 3 and 4. For each variable, we report coefficients and clustered standard errors for average marginal effects.[11] For regressions in Table 3, the omitted case is that the player is low caste, and for Table 4, the omitted case is a low-caste player in same-status pair.

The regression results are consistent with earlier analysis based on Table 2. Initially, there is a modest and marginally significant difference in behavior between low- and high-caste individuals, in which high caste status is associated with a 16 percentage point lower likelihood of playing *Stag*. After the initial period, by far the largest difference in behavior occurs after the subject received the loser's payoff, which is column 4 of Tables 3 and 4.

[10]In our regression analysis, *type* is defined to be 1 if the player played *Stag* in period 1 *and* the period is greater than 1, and 0 otherwise.

[11]Players were organized into fixed pairings in groups of four, such that individuals A, B, C, and D were assigned to pairs (A, B) and (C, D) for periods 1–5 and to pairs (A, C) and (B, D) for periods 6–10. Standard errors are clustered by these four-tuples of players.

		Preceding outcome was:			
	Initial	(*Hare*,*Hare*)	(*Hare*,*Stag*)	(*Stag*,*Hare*)	(*Stag*,*Stag*)
HH	-0.133	-0.254	-0.159	-0.423**	0.0360
	(0.0869)	(0.201)	(0.126)	(0.119)	(0.0362)
H in *LH*	-0.0943	-0.121	-0.146	-0.184	0.0152
	(0.102)	(0.215)	(0.132)	(0.131)	(0.0576)
L in *LH*	0.0974	0.0114	-0.130	0.0741	0.0152
	(0.0970)	(0.173)	(0.133)	(0.122)	(0.0386)
Type	0.354***	-0.122	0.174	-0.0516	0.109**
	(0.0546)	(0.105)	(0.118)	(0.0645)	(0.0439)
Land	0.00584	-0.00514	-0.00156	0.00541	-0.00153
	(0.00400)	(0.00775)	(0.00682)	(0.00480)	(0.00242)
High school	0.0622	-0.124	-0.0840	0.183	-0.0777
	(0.0879)	(0.149)	(0.138)	(0.0981)	(0.0593)
Non-mud house	0.0519	0.0358	0.0459	0.0678	0.0234
	(0.0629)	(0.112)	(0.104)	(0.0842)	(0.0474)
N	242	154	181	181	452

dfdx coefficients; se_dfdx in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Marginal probability that a player chose *Stag*. Probit regression analysis for all pair statuses.

In this case, high caste status is associated with a decrease of 36 percentage points in the likelihood of playing *Stag* in the subsequent period. If one compares *LL* versus *HH*, then being in an *HH* pair is associated with an even larger average effect of 42 percentage points. Thus, controlling for covariates and the initial propensity to play *Stag* results in an even larger estimate of the effect of the player's caste status. Differences after other one-period histories are of a smaller magnitude and are statistically insignificant. In particular, for the regressions in Table 4 and for every history except (*Stag*,*Hare*), a t-test fails to reject the hypothesis that the coefficients on *HH*, *H* in *LH*, and *L* in *LH* are all zero ($p > 0.1$). Conversely, this null hypothesis is rejected for (*Stag*,*Hare*) with $p < 0.001$.

In all regressions, the coefficients on land, education, and housing are both individually and jointly insignificant. Thus, we find no evidence to support the hypothesis that it is covariates of caste, rather than caste culture itself, that explain the difference in behavior.

Finally, let us turn to the impact of social distance. The regressions in Table 4 have on the right-hand side the caste status of both players. If social distance were an impediment to efficient coordination, then we would expect that on average, being in a mixed-status pair would be associated with a lower likelihood of playing *Stag*. In fact, the regression results suggest almost the opposite. In all regressions except after (*Stag*,*Stag*), the coefficient on H in LH is larger than the coefficient on HH, and the coefficient on L in LH is positive after every regression except after (*Hare*,*Stag*) (recall that the omitted case is LL). This suggests that, if anything, mixed-status is associated with higher rates of playing *Stag* than is same-status.

Our empirical findings on individual behavior are summarized as follows:

Result 2 (Reaction to the loser's payoff)**.** *After receiving the loser's payoff, high-caste players are significantly less likely to play Stag than low-caste players. The caste gap in the probability of playing Stag in the initial period and after other one-period histories is of smaller magnitude and generally not statistically significant.*

Result 3 (No effect of social distance)**.** *Conditional on a player's own caste status, greater social distance from the other player never reduces the probability of playing Stag, and in some cases increases it.*

4 Understanding Differences in Caste Culture

Our hypothesis is that the differences in behavior after the loser's payoff and in the efficiency of coordination are linked through the different cultures of high and low castes. By and large, high castes possess a version of the culture of honor, while low castes do not. In this culture, to be in a subordinate position to others is an affront to one's honor that warrants a vigorous defense. Since the loser's payoff is a subordinate position in terms of stage payoffs, it would be natural for the high-caste subjects to interpret it as an insult to their honor, and to respond by punishing their partner with playing *Hare*.

In the introduction, we appealed to anecdotal evidence from the anthropology and sociology literatures that the culture of honor is a prominent part of high-caste culture but less prominent in low-caste culture. The case would be even stronger if there were direct quantitative evidence of a caste difference in how individuals respond to potential insults.

Specifically, are high-caste men more likely than low-caste men to respond with punishment when they suffer harm from someone whose intentions are ambiguous?

We implemented a follow-up survey to test for such differences in caste culture. The design follows a format developed by social psychologists for measuring moral attitudes (e.g., Haidt, Koller and Dias, 1993). We recruited 241 high- and low-caste men in the Spring of 2014, from the same region of Uttar Pradesh in which the Stag Hunt experiment was conducted. 121 of these individuals were high-caste and 120 were low-caste. [12] Individuals were presented with vignettes in which one individual, typically Dinesh, behaves in a way that causes harm to a second individual, Mahesh. In one of the vignettes, it is clear that Dinesh's intention is to harm Mahesh, whereas in the other scenarios, Dinesh's intentions are ambiguous. Mahesh subsequently responds in an aggressive and even violent manner. The respondents were asked whether or not they think that Mahesh's response was justified, and also what they would have done in Mahesh's place. We note that the names Dinesh and Mahesh are not associated with a particular caste.

Here we will summarize the vignettes, with a full description given in Appendix B. The control vignette V0 serves as a control: Dinesh robs Mahesh's house, after which Mahesh beats Dinesh, reports him to the police, and Dinesh receives a jail sentence. Here, there is no ambiguity in Dinesh's intentions, and Mahesh's (relatively) measured response was to first rough Mahesh up and then turn him over to the criminal justice system. We expect a high proportion of all individuals to agree that Mahesh's response was justified.

Vignettes V1–V3 introduce ambiguity in Dinesh's motives. His behavior results in harm to Mahesh, but Dinesh's intentions are unclear. In principle, the conflict could be interpreted as a misunderstanding. Moreover, Mahesh makes no attempt to avail himself of the legal system, but rather takes justice into his own hands by severely beating Dinesh. These vignettes are not directly related to issues of honor, but they do bear on how individuals construe others' intentions, as well as norms for what constitutes an appropriate punishment.

The last scenario V4 directly relates to notions of honor. There are two versions of the vignette corresponding to the caste status of the individual being surveyed. For high-caste men, Mahesh Bania marries Dinesh Thakur's daughter. Even though Bania and Thakur are both high status, the latter is slightly higher rank. For low-caste respondents, Bania and Thakur were changed to Pasi and Chamar, respectively, which are both low-status castes with Chamar being slightly higher. Similar to the earlier vignettes, Dinesh responds by assaulting Mahesh in the village.

[12] Specifically, respondents were drawn from the same block and district near Lucknow. At least 11 of the 22 villages sampled for the punishment survey are also represented in the Stag Hunt sample.

	Response was justified				
	V0	V1	V2	V3	V4
L	0.90	0.03	0.20	0.10	0.22
H	0.93	0.38	0.43	0.33	0.70
$H-L$	0.03	0.35*	0.23*	0.23***	0.48***
N	61	59	60	59	203

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: Proportion of respondents who agree that the response was justified in the punishment survey.

Each individual was presented with one vignette from V0–V3 as well as V4. This was done to prevent cross-contamination in the response to the first four scenarios.

In Table 5, we report the proportions of individuals who thought that the response was justified.[13] We expect that both high- and low-caste men should respond affirmatively for V0, since there is no ambiguity in Dinesh's intentions and the punishment was obtained through the socially appropriate legal process. On the other hand, if the culture of honor makes individuals more likely to perceive an insult or more willing to punish severely, we would expect that high-caste men would be more likely to think the response was justified for V1–V4. Indeed, this is exactly what we find. For V0, the proportions that agree that the response was justified are not statistically different at all conventional levels of significance. For V1–V4, however, a one-sided t-test rejects the null hypothesis that more low-caste than high-caste men agree that the response was justified, at 5 percent for V1–V2 and at 0.1 percent for V3-V4. In the case of V4, over 70 percent of high-caste respondents agreed that the violent response was justified, compared to only 22 percent of low-caste respondents.

In addition to asking whether or not the response was justified, we asked the respondents what they would have done if they were the party who was wronged. For V4, a significant portion of the responses specifically referenced the preservation of honor as a motivation for responding in the same manner as the character in the vignette. Specifically, 32 of the 121 high-caste men and 11 of the 120 low-caste men thought the response in V4 was justified and mentioned honor as a justification.

The conclusions from the surveys on culture and locus of control are summarized in the following result:

[13]Of the 241 respondents, 38 were not able to answer for vignette V4, one was not able to answer for V1, and one was unable to answer for V3. Among the men who did respond for V4, 99 were low-caste and 104 were high-caste.

Result 4 (Attitudes towards punishment)**.** *High-caste men are more likely to agree that an offense warrants a severe punishment, even when the offender's intentions are ambiguous. In some situations, this is explicitly linked to the preservation of honor.*

5 A Model of Convention Formation with the Culture of Honor

In the preceding section, we argued that traits consistent with the culture of honor are more common among high-caste than among low-caste men. This may lead high-caste men to respond differently than their low-caste counterparts after receiving the loser's payoff. In particular, if high-caste men see the loser's payoff as an insult, then they might respond by playing *Stag* less, relative to low-caste players in the same situation. Importantly, this difference in behavior occurs only after miscoordination. If the subjects in the Stag Hunt experiment could perfectly play *Stag* from the get-go, then miscoordination might never occur, and we need not observe any difference in outcomes across groups. Thus, to explain how the culture of honor can lead to inefficient conventions, we need a theory that allows for miscoordination to occur while players converge to a convention. In this case, it is possible that the culture of honor may derail the process of learning an efficient convention and lead to an outcome with a high probability of (*Hare*, *Hare*).

This section develops such a theory. Our analysis is based on the well-known learning model of logistic fictitious play (Fudenberg and Levine, 1998). In this model, individuals use the history of their partner's actions to forecast future behavior. Thus, if my partner has played *Stag* frequently in the past, then I expect him to play *Stag* a lot in the future as well. This is of course a gross simplification of the players' thought process, but one that is likely to capture a first-order effect of the game's history on players' beliefs. As the game progresses and players collect more observations, they refine their forecasts of what their partners will do in the future, and in any given period, players' actions are noisy best responses given their current beliefs.

To this basic setup, we add one new feature: Players who subscribe to the culture of honor will interpret the loser's payoff as an insult rather than an accident, and insulted players prefer that their partners receive low payoffs in the following period. The addition of these simple state-dependent preferences turns out to have a large impact on which conventions can be learned. If the preference for punishment is sufficiently strong, it will be impossible to learn and sustain an efficient convention.

Here is the formal model. Individuals 1 and 2 repeatedly play the Stag Hunt for periods $t = 1, 2, \ldots$ In each period t, player $i \in \{1, 2\}$ is in one of two states—normal (N) or insulted (I)—denoted by $s_{i,t}$. Players are in state N unless they received the loser's payoff in the preceding period, in which case they are in state I. In state N, player i's utility is precisely his monetary payoff π_i as given in Figure 1. In state I, the player believes that his partner has "insulted" him, and utility changes to $\pi_i - \sigma_i \pi_j$, with $\sigma_i \geq 0$. If $\sigma_i = 0$, being insulted has no effect on preferences, but if $\sigma_i > 0$, an insulted player receives utility from his own monetary payoffs and disutility from his partner's monetary payoffs.[14] We think of the preference parameter σ_i as capturing the extent to which the player subscribes to the culture of honor, and the evidence cited in the Introduction suggests that high-caste subjects have, on average, higher σ_i than their low-caste counterparts. Importantly, σ_i has no effect on preferences after coordination occurs. It is only after *miscoordination* that players with a positive σ_i may prefer to punish their partner.

Player i's action in period t is denoted by $a_{i,t} \in \{Stag, Hare\}$. Player i maintains a belief that player j will play *Stag* with some probability $p_{i,t} \in (0, 1)$. This probability starts at the initial value $p_{i,1}$, and for $t > 1$, it is a weighted average of the empirical distribution of play together with the initial belief. Let $[a_{j,\tau} = Stag]$ be 1 if player j played *Stag* in period τ, and 0 otherwise. Then the forecast at period t is given by

$$p_{i,t} = \frac{1}{t}\left(p_{i,1} + \sum_{\tau=1}^{t-1} [a_{j,\tau} = Stag]\right),$$

The formula for $p_{i,t}$ intuitively says that if you have seen a player take an action frequently in the past, you expect him to take that action frequently in the future as well. This can be written recursively as

$$p_{i,t} - p_{i,t-1} = \frac{1}{t}\left([a_{j,t-1} = Stag] - p_{i,t-1}\right). \tag{1}$$

Over time, the initial component $p_{1,t}$ of the belief is downweighted, and the belief converges to the empirical distribution of the partner's actions.

Player i plays *Stag* with probability $b_i(p|s)$ when his state is s and his forecast is p. This probability is a "smoothed" best response defined by

$$b_i(p|s) = \frac{\exp\left(u_i(Stag|p, s)\right)}{\exp\left(u_i(Stag|p, s)\right) + \exp\left(u_i(Hare|p, s)\right)}, \tag{2}$$

[14] This corresponds to spiteful preferences in the sense of Fehr and Schmidt (1999) and Charness and Rabin (2002).

where $u_i(a|p,s) = p\,u_i(a, Stag|s) + (1-p)u_i(a, Hare|s)$ is a player's expected utility from the action a. The function b_i captures in reduced form the idea that while there is noise in behavior, individuals are more likely to take the action that gives greater expected utility. Since being in the insulted state lowers the the expected utility from playing *Stag*, it follows immediately that $b_i(p|N) > b_i(p|I)$, i.e., *Stag* is played with lower probability when insulted than when not insulted.

We are interested in how the culture of honor impacts players' ability to learn and sustain an efficient convention. In the model, this question can be expressed as: how do higher values of σ affect the long-run probability of the outcome (*Stag*,*Stag*)? Appendix A contains a formal analysis, and here we will summarize the main results and intuitions. The core finding is that as σ_i increases for either player, the long-run probability of (*Stag*,*Stag*) must decrease. For sufficiently large σ, the only long-run equilibrium will involve a high probability of (*Hare*,*Hare*). In practice, this means that if the culture of honor is sufficiently important, then efficient behavior cannot be sustained.

On to the analysis. Suppose players are in a long-run steady state in which beliefs are not changing. Then it would have to be the case that beliefs about others' actions are equal to the empirical distribution that is generated by best responses. Thus, it is helpful to think about the induced distribution of actions for a hypothetical pair of steady-state beliefs (p_i, p_j). Since players with $\sigma_i > 0$ play *Stag* with lower probability when insulted, the average probability of playing *Stag* depends on the relative frequency of states I and N. Indeed, the long-run distribution of $s_{i,t}$ depends on the beliefs and preferences of both players. Let $\Pi(s_i, s_j|p_i, p_j)$ denote the long-run probability of states (s_i, s_j) when players beliefs are fixed at (p_i, p_j). This distribution also depends on the preference parameters (σ_i, σ_j), but we suppress this dependence to keep our notation tight. It is a fact that this distribution exists and is unique for all beliefs and preference parameters.

This distribution of states can be used to calculate the average probabilities of playing *Stag*, given beliefs and types. Let $b_i(p_i, p_j)$ denote the probability with which a player i chooses *Stag* when beliefs are fixed at (p_i, p_j), averaged across states:[15]

$$b_i(p_i, p_j) = \Pi(I, N|p_i, p_j)b_i(p_i|I) + \big(1 - \Pi(I, N|p_i, p_j)\big)b_i(p^i|N). \tag{3}$$

[15]Since a player enters state I only after receiving the loser's payoff, and only one player can receive the loser's payoff at a time, it is impossible for *both* players to be in state I.

In a steady state in which beliefs are not changing, it must be that beliefs are equal to the long-run empirical average. In other words,

$$p_i = b_j(p_j, p_i) \ \forall i = 1, 2, j \neq i. \tag{4}$$

A belief that satisfies this condition is called *consistent*, and a *learning equilibrium* is a pair (p_1, p_2) such that both beliefs are consistent. We argue in Appendix A that for each p_j, there will exist a unique p_i that is consistent and satisfies (4). We define this consistent belief to be $p_i^*(p_j)$. The learning equilibria lie at the intersection of p_i^* and $p_j^{*,-1}$.

In addition, the average best response $b_i(p_i, p_j)$ and consistent belief functions p_i^* are decreasing in both σ_i and in σ_j. The reason is that as the culture of honor becomes stronger, the probability of playing *Stag* when insulted decreases. Starting from a regime in which *Stag* is played with high probability, this tends to increase the likelihood of insults, which further depresses the probability of playing *Stag*.

We have been thinking about the behavior induced by a fixed set of beliefs, but in fact beliefs are evolving over time according to (1). The literature on stochastic approximation has shown that in the long-run, beliefs trend towards an average probability of *Stag* that is precisely (3). Moreover, since the weight of $\frac{1}{t}$ is decreasing, the difference equation (1) "slows down", and asymptotically beliefs evolve according to the differential equation

$$\frac{dp_i}{dt} = b_j(p_j, p_i) - p_i. \tag{5}$$

Thus, we can understand long-run behavior by looking at paths of this system of differential equations. Note that in a learning equilibrium, $\frac{dp_i}{dt} = 0$ for all i, so that beliefs are not changing. In principle, there can be trajectories that do not converge to a learning equilibrium, but for our game, beliefs will in fact converge to a learning equilibrium. Generically, there are either one or three learning equilibria, the highest and lowest of which are stable.

In Figure 5, we give examples of the consistent belief functions for $\sigma_1, \sigma_2 \in \{0, 0.3, 0.6\}$. The small grey arrows indicate the direction in which the system (5) is trending. For example, the arrows are pointing into the lowest and highest crossings of the consistent belief functions, indicating that they are the stable learning equilibria. For the cases where there are two stable equilibria, we have drawn in the boundary of the "basins of attraction" of the two equilibria.[16]

[16] A more general version of this model would parametrize the level of noise by multiplying the utilities inside exponentials in (2) by a level of precision $\lambda > 0$. Throughout our analysis, we have fixed $\lambda = 1$. For a given choice of (σ_1, σ_2), as λ becomes large, the best response functions converge to those of the static model in which the game is played once, so that both efficient and inefficient equilibria exist.

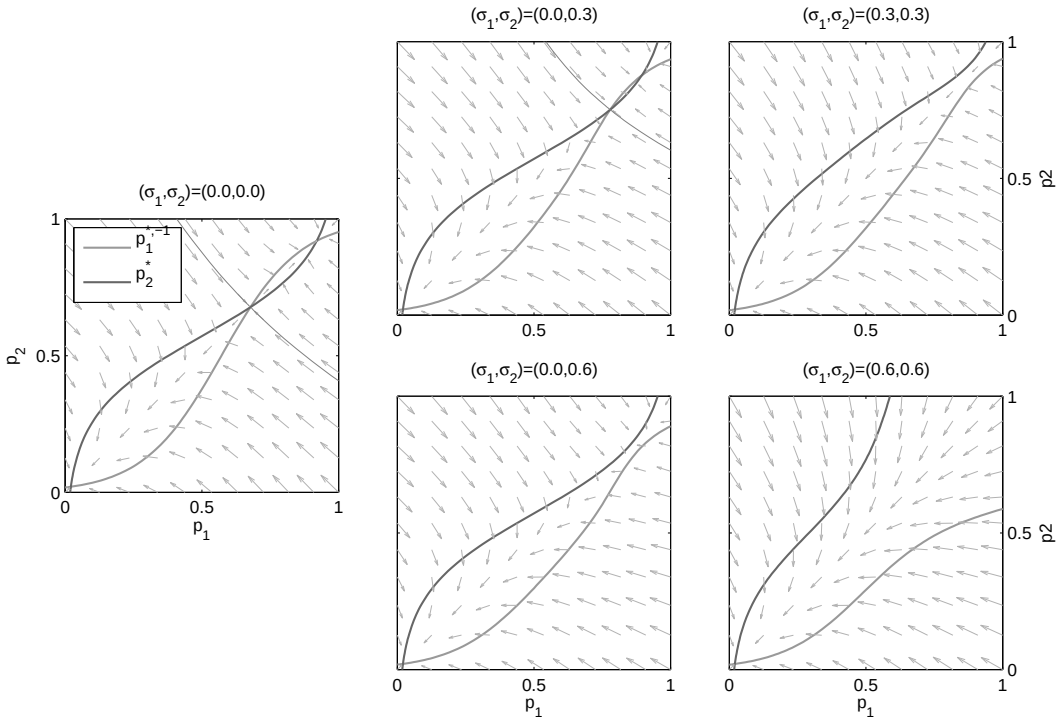

Figure 5: Comparison of learning equilibria for $\sigma_i \in \{0, 0.3, 0.6\}$.

In the left-most panel, neither player subscribes to the culture of honor. In this case, there are two stable learning equilibria, one of which involves a high probability of playing *Stag* of approximately 0.96. The other panels show what happens as we increase σ. In the center panels, only player 2 subscribes to the culture of honor. This means that player 1's consistent belief is lower, since player 2 plays *Stag* less often when he is insulted. For $\sigma_2 = 0.3$, there still exists a stable equilibrium in which *Stag* is played with high probability of about 0.90, as depicted in center-top. But as σ_2 increases further, the consistent belief functions decrease so that at $\sigma_2 = 0.6$ (center-bottom), there is only a single learning equilibrium, in which *Stag* is played with probability 0.02. This outcome is even worse than the *worst* stable equilibrium when $\sigma_1 = \sigma_2 = 0$. The right panels show that when both players subscribe to the culture of honor, the situation deteriorates further: for neither $\sigma = 0.3$ or $\sigma = 0.6$ do efficient equilibria exist.

More generally, we can define an efficient learning equilibrium to be one in which the probability of playing *Stag* is at least the probability of playing *Stag* in the worst stable equilibrium with $\sigma_1 = \sigma_2 = 0$. This is a fairly modest criterion for efficient behavior, since the lowest stable probability of *Stag* when neither player has the culture of honor is less than 0.021. In Figure 6, we have divided the set of possible σ's in $[0, 1] \times [0, 1]$ into regions in which

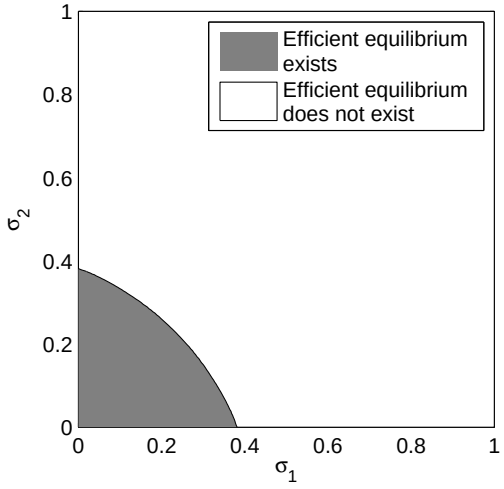

Figure 6: Values of (σ_1, σ_2) for which an efficient equilibrium exists.

an efficient equilibrium does and does not exist. This Figure shows that efficient equilibria only exist when the culture of honor for both players is relatively weak. We summarize this result as:

Result 5 (Efficiency of learning equilibria)**.** *If the culture of honor is relatively weak, there exist learning equilibria in which the outcome is close to efficiency. However, if the culture of honor is sufficiently strong for either player, the only learning equilibria involve a high degree of inefficient coordination.*

Thus, the model shows that Results 1 and 2 of the experiment can both be explained by state-dependent preferences of the high-caste players. Such preferences would cause lower probabilities of *Stag* after receiving the loser's payoff, which in turn affects the long-run probabilities of the outcome (*Stag*,*Stag*). If low-caste players have normal preferences, there is no reason why they cannot have high beliefs and hence high probabilities of (*Stag*,*Stag*) in the long-run. On the other hand, if high-caste players have high σ_i on average, and therefore a stronger preference for punishing their partner after the loser's payoff, then it is less likely that they will be able to sustain a high probability of playing *Stag*. The fact that some pairs are able to coordinate on an efficient convention while others are not is consistent with variation in σ_i among high-caste players.

6 Discussion

6.1 A Difference in Trust?

We have argued that the outcome of the Stag Hunt experiment can be explained by state-dependent preferences of high-caste men that are shaped by the culture of honor. It is only after a particular *strategic* outcome—the loser's payoff—that high-caste men become concerned about the distribution of payoffs. An alternative explanation of the lower rate of efficient coordination among high-caste players is a caste difference in how much players "trust" their partner to take the efficient action. Indeed, some of the results in Table 2 are consistent with the hypothesis that high-caste subjects are less trusting than their low-caste counterparts: In period 1, before there is any history of play to influence decisions, 57 percent of low-caste players chose *Stag*, as compared to only 37 percent of high-caste players.

A canonical game for assessing trust is the investment game of Berg, Dickhaut and McCabe (1995) (see Algan and Cahuc (2013) for a review of the trust literature). Two players are anonymously paired, and each player is given an initial endowment of money. One of the players (the principal) chooses a portion of his endowment to send to the the other player (the agent), and the principal is told that the experimenter will multiply the investment by three. For example, if the principal invests 1 rupee, then the agent receives 3 rupees. Finally, the agent chooses a non-negative amount of money to return to the principal. A self-interested agent will always seize the investment, so the unique Nash equilibrium calls for the principal to invest nothing. We interpret a positive investment as an indication that the principal trusts the agent to pay a return.

We implemented a binary choice version of this game in 2007 in the district of Unnao, Uttar Pradesh. Unnao is approximately 40 miles from the district in which we conducted the Stag Hunt experiment. The players in a given pairing were drawn from different villages within the district. Each player was given an endowment of 50 rupees, which was equivalent to the daily wage for an unskilled worker. The principal had to choose between investing all or none of the 50 rupees. If the principal chose to invest, the agent received 150 rupees so that he had a total of 200, and then the agent chose whether or not to return the 100.

There were four pairings that differed in the caste statuses of the two players. In *LL* pairs, both players were low-caste, and *HH* pairs, both players were high-caste. For *LH* pairs, the principal was low-caste and the agent was high-caste, and finally in *HL* pairs, the principal was high-caste and the agent was low-caste. Statuses were indicated to the players by the use of names that were known to be exclusive to particular caste, as we verified in a pre-experiment test in the district. The numbers of pairs by caste status were: 26 (*LL*), 34 (*LH*), 30 (*HH*), and 30 (*HL*).

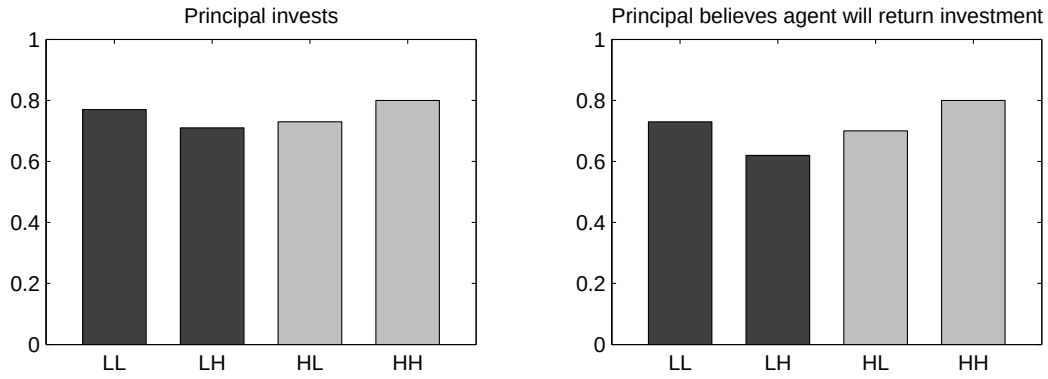

Figure 7: The left panel depicts by caste status the proportion of principals who chose to invest. The right panel gives the proportion of principals who believed that the agent would return the investment.

The left panel of Figure 7 presents the proportion of principals who invested. There are no significant differences in this proportion across different pairs. In fact, trust as measured by the proportion of principals who invest is higher for *HH* (80 percent) than for the other types of pairs.

We also asked the principals directly: Do you believe that the agent will return the 100 rupees if you invest? The right panel of Figure 7 reports the proportion that said yes. Among principals who invested, there are again no significant differences in beliefs by the caste status of the principal or the agent. Finally, we asked the agents: Do you believe the principal will invest the 50 rupees? In a dprobit regression (not reported here), all coefficients are close to zero and insignificant; the results are robust to controls for the agents' wealth.

In sum, the investment game experiment provides no evidence to support the hypothesis H are less trusting than L, and trust differences cannot account for the caste differences in behavior in the Stag Hunt experiment.

6.2 A difference in self-efficacy?

We have argued that the different responses to the punishment survey by high- and low-caste men are indicative of different mental frameworks for identifying offenses and/or different rules of thumb for punishment. This is because high-caste men are much more likely than low-caste men to say that a severe punishment for harmful behavior is justified, even when the intentions of the offender were not clearly malicious. An alternative explanation would be that high- and low-caste cultures have the same rules for understanding others' intentions and for how to punish, but that low-caste individuals simply feel that they are unable to respond to transgressions like those depicted in the vignettes. The belief in the ability to

Does this statement describe you?	Yes		No		N
	L	H	L	H	
I have no trouble making and keeping friends.	0.97	0.97	0.00	0.03	62
If I need help in carrying off a plan of mine, it's usually difficult to get others to help.	0.06	0.13	0.84	0.77	62
I often find it hard to get my point of view across to others.	0.03	0.17	0.78	0.77	62
In attempting to smooth over a disagreement, I usually make it worse.	0.09	0.07	0.84	0.83	62

Table 6: Responses to the survey on self-efficacy.

have an effect on one's environment has been variously referred to as locus of control and self-efficacy in the psychology literature. If low-caste individuals systematically had lower sense of self-efficacy, then that would explain the results.

We conducted an additional survey involving 62 different individuals, 30 high-caste and 32 low-caste, to test for caste differences in self-efficacy. We employed a survey design that is borrowed from the psychology literature (cf. Paulhus, 1983): The men were presented with a series of four statements that take positions on different aspects of one's locus of control, e.g., "I have no trouble making and keeping friends." The respondents were then asked whether or not they felt that the statements described them.

The statements and the responses are given in Table 6. We report the proportions of men who responded yes or no to whether or not the statements described them, with the remaining respondents giving equivocal answers or being unable to respond. Not only did the response rates not vary significantly across castes, but the point estimates indicate that if anything, low-caste men have greater locus of control than their high-caste counterparts. For example, 84 percent of low-caste men disagreed with S2, that they have a hard time finding friends to help them, compared to only 77 percent of high-caste men. Thus, self-efficacy cannot explain the caste difference in responses in the punishment survey.

7 Conclusion

The broad question that this paper addresses is how conventions emerge. Much of what we value in society depends on coordination, and development itself has been viewed as a process of coordination. To shed light on the influence of culture on this process, we ran an experi-

ment in India with low-caste and high-caste men. The low-caste pairs generally coordinated on a convention under which their endowment grew by 70 percent (6 to 10 rupees), whereas the high-caste pairs showed evidence of adopting a convention under which their endowment grew by only 17 percent (6 to 7 rupees). The evidence suggests that this caste gap is due to a novel obstacle to efficient coordination: Because they come from a culture of honor, high-caste individuals may interpret the loser's payoff and, more generally, miscoordination, as an insult, to which they respond by not trying as hard for efficient coordination in the future. In simplest terms, a mindset associated with the culture of honor is "cross me, and I'll punish you." This mindset means that an accidental miscoordination can lead to more miscoordination and misunderstanding and an unraveling of cooperative behavior. The lack of efficient coordination among the high castes, which are the dominant social group in the region, is consistent with the observed failures of coordination and "institutional inertia" in the state of Uttar Pradesh.

Events and actions do not always speak for themselves but instead depend on framing for their meaning. In such cases, social behavior depends on subjective interpretations. Experiments that expose individuals to the same situation under alternative frames have demonstrated that frame shifts can induce large shifts in economic behavior. These findings make it plausible that a difference in mental frameworks, associated with a difference in culture, is also capable of producing a large difference in economic behavior.

We considered and found little support for two alternative explanations for the observed caste difference in coordination. First, there is a statistically insignificant difference in trust in the initial period of the game. To assess trust directly, we report results of an investment game experiment in a nearby district in Uttar Pradesh. We find no difference in trust between high- and low-caste men in the investment game. Second, all of our results are robust to controls for education and wealth, which suggests that non-cultural covariates of caste cannot explain the caste gap in coordination.

Our evidence that mental frameworks may powerfully impede coordination has policy implications. It is not enough to create opportunities for development. We also have to remove the psychological barriers to taking advantage of those opportunities, and culture can at times be one such barrier. There is clear evidence from interventions that mental frameworks are malleable.[17] For instance, Beaman et al. (2009) find that exposure to

[17]Algan et al. (2012) evaluate a randomized control trial in which disruptive young boys in the treated group were regularly placed in small groups with pro-social children. The intervention helped the subjects learn trust, empathy, and self-control. At age 26, the treated subjects, compared to the control group, were much more likely to be employed and much less likely to have a criminal record. Experimental studies also show that people who were taught to reframe, in a self-distancing perspective, a negative experience that had made them feel rejected or enraged, were less likely to experience high distress and less likely to reciprocate hostile behavior (Ayduk and Kross, 2008, 2010; Kross and Ayduk, 2011).

women village leaders through a policy of reservations for women in village elections in India improved men's ability to judge fairly the quality of female leaders. Blattman, Hartman and Blair (2012) find that an advocacy and training program in mediation gave individuals new conceptual tools with which to understand conflicts—for instance, seeing the interaction as positive-sum instead of zero-sum—which increased the resolution of land disputes. The frameworks within which people view the world are an under-studied aspect of economic behavior, and they are not set in stone.

References

Alesina, Alberto, Paola Giuliano, and Nathan Nunn, "On the Origins of Gender Roles: Women and the Plough," *Quarterly Journal of Economics*, 2013, *128* (2).

Algan, Y and P Cahuc, "Trust and Human Development: Overview and Policy," in Philippe Aghion and Steven Durlauf, eds., *Handbook of Economic Growth*, Vol. 1, Elsevier, 2013.

Algan, Yann, Elizabeth Beasley, Richard E Tremblay, and Franck Vittaro, "The Long-Term Impact of Social Skills Training at School Entry: A randomized controlled trial.," *Sciences Po Working Paper*, 2012.

Andersen, Steffen, Erwin Bulte, Uri Gneezy, and John A List, "Do women supply more public goods than men? Preliminary experimental evidence from matrilineal and patriarchal societies," *The American Economic Review*, 2008, pp. 376–381.

Andreoni, James, "Warm-glow versus cold-prickle: The effects of positive and negative framing on cooperation in experiments," *The Quarterly Journal of Economics*, 1995, *110* (1), 1–21.

Ariely, Dan, *Predictably irrational: The hidden forces that shape our decisions*, New York: Harper Collins, 2009.

Ayduk, Özlem and Ethan Kross, "Enhancing the pace of recovery: Self-distanced analysis of negative experiences reduces blood pressure reactivity," *Psychological Science*, 2008, *19* (3), 229–231.

_ and _, "From a distance: Implications of spontaneous self-distancing for adaptive self-reflection," *Journal of personality and social psychology*, 2010, *98* (5), 809.

Battalio, Raymond, Larry Samuelson, and John Van Huyck, "Optimization incentives and coordination failure in laboratory stag hunt games," *Econometrica*, 2001, *69* (3), 749–764.

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova, "Powerful women: Does exposure reduce bias?," *The Quarterly Journal of Economics*, 2009, *124* (4), 1497–1540.

Benaïm, Michel and Morris W Hirsch, "Mixed equilibria and dynamical systems arising from fictitious play in perturbed games," *Games and Economic Behavior*, 1999, *29* (1), 36–72.

Berg, Joyce, John Dickhaut, and Kevin McCabe, "Trust, reciprocity, and social history," *Games and economic behavior*, 1995, *10* (1), 122–142.

Bernhard, Helen, Ernst Fehr, and Urs Fischbacher, "Group affiliation and altruistic norm enforcement," *The American Economic Review*, 2006, *96* (2), 217–221.

Blattman, Christopher, Alexandra Hartman, and Robert Blair, "Building institutions at the micro-level: Results from a field experiment in property dispute and conflict resolution," *Available at SSRN 2158966*, 2012.

Charness, Gary and Matthew Rabin, "Understanding social preferences with simple tests," *The Quarterly Journal of Economics*, 2002, *117* (3), 817–869.

_ **and Uri Gneezy**, "What's in a name? Anonymity and social distance in dictator and ultimatum games," *Journal of Economic Behavior & Organization*, 2008, *68* (1), 29–35.

Chen, Roy and Yan Chen, "The potential of social identity for equilibrium selection," *The American Economic Review*, 2011, *101* (6), 2562–2589.

Chen, Yan and Sherry Xin Li, "Group identity and social preferences," *The American Economic Review*, 2009, *99* (1), 431–457.

Cohen, Dov, Joseph Vandello, Sylvia Puente, and Adrian Rantilla, "'When You Call Me That, Smile!' How Norms for Politeness, Interaction Styles, and Aggression Work Together in Southern Culture," *Social Psychology Quarterly*, 1999, pp. 257–275.

Derné, Steve, "Beyond institutional and impulsive conceptions of self: Family structure and the socially anchored real self," *Ethos*, 1992, *20* (3), 259–288.

DiMaggio, Paul, "Culture and cognition," *Annual review of sociology*, 1997, pp. 263–287.

Drèze, Jean and Haris Gazdar, "Uttar Pradesh: The burden of inertia," in Jean Drèze and Amartya Sen, eds., *Indian development: Selected regional perspectives*, Oxford: Oxford University Press, 1997, pp. 33–128.

_ **and Naresh Sharma**, "Palanpur: population, society, economy," in Peter Lanjouw and Nicholas Stern, eds., *Economic Development in Palanpur over Five Decades*, Oxford: Oxford University Press, 1998, pp. 3–113.

Fehr, Ernst and Klaus M Schmidt, "A theory of fairness, competition, and cooperation," *The Quarterly Journal of Economics*, 1999, *114* (3), 817–868.

_ **, Karla Hoff, and Mayuresh Kshetramade**, "Spite and development," *The American Economic Review*, 2008, *98* (2), 494–499.

Fernández, Raquel, "Cultural change as learning: The evolution of female labor force participation over a century," *The American Economic Review*, 2013, *103* (1), 472–500.

Fisman, Raymond and Edward Miguel, "Corruption, norms, and legal enforcement: Evidence from diplomatic parking tickets," *Journal of Political Economy*, 2007, *115* (6), 1020–1048.

Fudenberg, Drew and David K Levine, *The theory of learning in games*, Vol. 2, Cambridge: MIT press, 1998.

Goffman, Erving, *Frame analysis: An essay on the organization of experience*, Boston: Harvard University Press, 1974.

Greif, Avner, "Cultural beliefs and the organization of society: A historical and theoretical reflection on collectivist and individualist societies," *Journal of Political Economy*, 1994, pp. 912–950.

_ **and Guido Tabellini**, "Cultural and institutional bifurcation: China and Europe compared," *The American Economic Review*, 2010, *100* (2), 135–140.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales, "Long term persistence," Technical Report, National Bureau of Economic Research 2008.

Haidt, Jonathan, Silvia Helena Koller, and Maria G Dias, "Affect, culture, and morality, or is it wrong to eat your dog?," *Journal of personality and social psychology*, 1993, *65* (4), 613.

Harrison, Glenn W and John A List, "Field experiments," *Journal of Economic Literature*, 2004, pp. 1009–1055.

Heine, Steven J, *Cultural psychology*, WW Norton New York, 2012.

Henrich, Joseph and Jean Ensminger, "Theoretical foundations: The coevolution of social norms, intrinsic motivation, markets and the institutions of complex societies," *mimeo*, 2011.

Hitchcock, John T, "The Idea of the Martial Rājpūt," *The Journal of American Folklore*, 1958, *71* (281), 216–223.

Hoff, Karla, "Beyond Rosenstein-Rodan: The modern theory of coordination problems in development," in "Proceedings of the Annual World Bank Conference on Development Economics 2000" The World Bank Washington, D.C. 2001, pp. 145–188.

_ **and Joseph E Stiglitz**, "After the big bang? Obstacles to the emergence of the rule of law in post-communist societies," *American Economic Review*, 2002, *94* (3), 753–763.

_ **, Mayuresh Kshetramade, and Ernst Fehr**, "Caste and Punishment: the Legacy of Caste Culture in Norm Enforcement," *The Economic Journal*, 2011, *121* (556), F449–F475.

Huyck, John B Van, Joseph P Cook, and Raymond C Battalio, "Adaptive behavior and coordination failure," *Journal of Economic Behavior & Organization*, 1997, *32* (4), 483–503.

_ , Raymond C Battalio, and Richard O Beil, "Tacit coordination games, strategic uncertainty, and coordination failure," *The American Economic Review*, 1990, *80* (1), 234–248.

_ , _ , and _ , "Strategic uncertainty, equilibrium selection, and coordination failure in average opinion games," *The Quarterly Journal of Economics*, 1991, *106* (3), 885–910.

Jackson, Matthew O and Yiqing Xing, "Culture-dependent strategies in coordination games," *Proceedings of the National Academy of Sciences*, 2014, *111* (Supplement 3), 10889–10896.

Kahneman, Daniel, *Thinking, fast and slow*, New York: Farrar, Straus and Giroux, 2011.

_ and Amos Tversky, "Prospect theory: An analysis of decision under risk," *Econometrica*, 1979, pp. 263–291.

Khare, R. S., *The untouchable as himself: Ideology, identity, and pragmatism among the Lucknow Chamars*, Cambridge: Cambridge University Press, 1984.

Kross, Ethan and Ozlem Ayduk, "Making meaning out of negative experiences by self-distancing," *Current Directions in Psychological Science*, 2011, *20* (3), 187–191.

Kushner, Harold J and George Yin, *Stochastic approximation and recursive algorithms and applications*, Vol. 35, New York: Springer-Verlag, 2003.

Lazear, Edward P, "Culture and Language," *Journal of Political Economy*, 1999, *107* (S6).

Mandelbaum, David G, *Women's seclusion and men's honor: Sex roles in North India, Bangladesh, and Pakistan*, Tucson: University of Arizona Press, 1993.

Matsuyama, Kiminori, "Economic development as coordination problems," in Masahiko Aoki, Hyung-Ki Kim, and Masahiro Okuno-Fujiwara, eds., *The role of government in East Asian economic development: comparative institutional analysis*, New York: Oxford University Press, 1996.

Merton, Robert K, *Social theory and social structure*, Free Press, New York, 1968.

Nisbett, Richard E and Dov Cohen, *Culture of honor: The psychology of violence in the South.*, Colorado: Westview Press, 1996.

North, Douglass C, *Institutions, institutional change and economic performance*, Cambridge: Cambridge University Press, 1990.

_ , *Understanding the process of economic change*, Academic Foundation, 2006.

Paulhus, Delroy, "Sphere-specific measures of perceived control.," *Journal of Personality and Social Psychology*, 1983, *44* (6), 1253.

Rabin, Matthew, "Incorporating fairness into game theory and economics," *The American Economic Review*, 1993, pp. 1281–1302.

Ray, Debraj, "What's new in development economics?," *The American Economist*, 2000, *44* (2), 3–16.

Sapienza, Paola, Luigi Zingales, and Luigi Guiso, "Does culture affect economic outcomes?," Technical Report, National Bureau of Economic Research 2006.

Schelling, Thomas C, *Micromotives and Macrobehavior*, New York: WW Norton, 1978.

Skyrms, Brian, *The stag hunt and the evolution of social structure*, Cambridge: Cambridge University Press, 2003.

Swidler, Ann, "Culture in action: Symbols and strategies," *American Sociological Review*, 1986, *51* (2), 273–286.

Tabellini, Guido, "The scope of cooperation: Values and incentives," *The Quarterly Journal of Economics*, 2008, *123* (3), 905–950.

Zerubavel, Eviatar, *Social mindscapes: An invitation to cognitive sociology*, Cambridge: Harvard University Press, 1999.

A Formal Analysis of Learning Model

This Appendix contains the formal analysis of the model of Section 5.

First we make some preliminary observations about b. Using the utilities in Figure 1, we can calculate $u_i(\mathit{Stag}|p, I) = 10(1-\sigma_i)p + (3-7\sigma_i)(1-p)$ and $u_i(\mathit{Hare}|p, I) = (7-3\sigma_i)p + 7(1-\sigma_i)(1-p)$. Hence, $u_i(\mathit{Stag}|p, I) - u_i(\mathit{Hare}|p, I) = 7(1-\sigma_i)p - 4$ and so $b_i(p|N) = (1+\exp(4-7p))^{-1}$ and $b_i(p|I) = (1+\exp(4-7(1-\sigma_i)p))^{-1}$. We note for future reference that $b_i(p|I)$ is decreasing in σ_i, so that $\Delta_i(p) = b_i(p|N) - b_i(p|I)$ is increasing in σ_i.

We begin by solving explicitly for $\Pi(s_i, s_j|p_i, p_j)$ and $b(p_i, p_j)$ for a fixed σ_1, σ_2. In the following expressions, we will economize on notation by writing b_i for $b_i(p_i|N)$ and Δ_i for $\Delta_i(p_i)$. Π and b_i are given by

$$\Pi(I, N|p_i, p_j) = \frac{b_i\left(1 - b_j + \Delta_j(1-b_i)\right)}{(1+\Delta_i)(1+\Delta_j) - b_i\Delta_j(1+\Delta_i) - b_j\Delta_i(1+\Delta_j)} \tag{6a}$$

$$b_i(p_i, p_j) = b_i\frac{1 + \Delta_j\left(1 - b_i - b_j\Delta_i\right)}{(1+\Delta_i)(1+\Delta_j) - b_i\Delta_j(1+\Delta_i) - b_j\Delta_i(1+\Delta_j)} \tag{6b}$$

Π can be derived from the system of equations

$$\begin{aligned}\Pi(I, N|p_i, p_j) = {} & \Pi(I, N|p_i, p_j)b_i(p_i|I)(1 - b_j(p_j)) \\ & + \Pi(N, I|p_i, p_j)b_i(p_i|N)(1 - b_j(p_j|I)) \\ & + (1 - \Pi(I, N|p_i, p_j) - \Pi(N, I|p_i, p_j))b_i(p_i|N)(1 - b_j(p_j|N))\end{aligned}$$

and the analogous equation flipping the identities i and j. The details are left to the reader. From these formulae, the expression for $b_i(p_i, p_j)$ can be derived from

$$b_i(p_i, p_j) = b_i(p_i|N) - \Delta_i(p_i)\Pi(I, N|p_i, p_j)$$

Using this expression, we can derive comparative statics for b_i in σ_i and σ_j. Note that $b_i(p_i, p_j)$ can be rewritten

$$b_i(p_i, p_j) = b_i\frac{\underbrace{1 + \Delta_j\left(1 - b_i - b_j\Delta_i\right)}_{X}}{\underbrace{1 + \Delta_j\left(1 - b_i - b_j\Delta_i\right)}_{X} + \underbrace{\Delta_i\left(1 - b_j + \Delta_j(1-b_i)\right)}_{Y}}.$$

An increase in σ_i or σ_j only affects $b_i(p_i, p_j)$ through increases in Δ_i and Δ_j, respectively. Thus, we can look for comparative statics of $\frac{X}{X+Y}$ as either Δ_i or Δ_j increase. Note that $\frac{X}{X+Y}$ decreases only if $\frac{Y}{X}$ increases. It is easy to see that X is decreasing in Δ_i and Y is

increasing in Δ_i (since $1-b_i$, $1-b_j$, and Δ_j are all positive). Hence, an increase in σ_i must lead to a decrease in $b_i(p_i, p_j)$. For Δ_j, the argument is only slightly more involved. To show that $b_i(p_i, p_j)$ is decreasing in σ_j, it is sufficient to show that $\frac{Y}{X}$ is increasing in Δ_j, which is the case if $X\frac{dY}{d\Delta_j} - Y\frac{dX}{d\Delta_j} > 0$. But this expression reduces to $Y\Delta_i b_j + \Delta_i(1-b_i)b_j(1-\Delta_i\Delta_j)$, which must be positive since $\Delta_i\Delta_j < 1$. Thus, $b_i(p_i, p_j)$ is decreasing in both σ_i and σ_j.

An important property for characterizing the model is that the functions $p_i - b_j(p_j, p_i)$ and $b_i(p_i, p_j)$ should both be increasing in p_i. This property does not hold for all possible specifications of the players' utilities, but it is true for the particular payoffs used in our experiment. An analytical proof of this fact would be both lengthy and unedifying, so we will have simply verified numerically that it holds in all of the values of σ_i utilized our examples, and we suspect it holds more generally. For the rest of the analysis, this property is treated as an assumption.

Kushner and Yin (2003) Chapter 8 Theorem 4.3 shows that the forecasts converge to a trajectory of the dynamical system (5). A consequence of the fact that $p_i - b_j(p_j, p_i)$ is increasing is that $\frac{\partial}{\partial p_i}\left(\bar{b}(p^i, p^{-i}) - p^i\right) < 0$. Hence, the divergence of the vector field for $(\dot{p^1}, \dot{p^2})$ is negative, and the vector field is area decreasing. Therefore, no non-empty open sets are invariant under the mean dynamical system, and beliefs must converge to a stable steady state of (5) (cf. Benaïm and Hirsch, 1999; Fudenberg and Levine, 1998).

In addition, the fact that $p_i - b_j(p_j, p_i)$ is strictly increasing means that there exists a unique solution to the functional equation $p_i^*(p_j) = b_j(p_j, p_i^*(p_j))$. For $b_j(p_j, p_i) \in (0,1)$ and hence $0 - b_j(p_j, 0) < 0$ and $1 - b_j(p_j, 1) > 0$, and the expression is strictly increasing and continuous, so by the intermediate value theorem, a solution to $p_i - b_j(p_j, p_i) = 0$ must uniquely exist. For $p_i < p_i^*(p_j)$, $p_i < b_j(p_j, p_i)$, and for $p_i > p_i^*(p_j)$, $p_i > b_j(p_j, p_i)$.

Moreover, because $b_j(p_j, p_i)$ is strictly increasing in p_j, it must be that p_i^* is strictly increasing in p_j as well and has a strictly increasing inverse $p_i^{*,-1}$. For if $p_j' > p_j$, then $0 = p_i^*(p_j) - b_j(p_j, p_i^*(p_j)) > p_i^*(p_j) - b_j(p_j', p_i^*(p_j))$, and since $p_i - b_j(p_j', p_i)$ is strictly increasing, we must have that $p_i^*(p_j') > p_i^*(p_j)$. By a similar argument, since $b_i(p_i, p_j)$ is decreasing in σ_i and σ_j, p_i^* must be decreasing in σ_i and σ_j as well.

Finally, we need the following technical result to characterize the equilibria:

Lemma 1. *Let $h^1, h^2, l^1, l^2 : [0,1] \to (0,1)$ be continuous and strictly increasing functions, with $h^1 \geq l^1$ and $h^2 > l^2$. Let $\bar{x}_h$ denote the largest solution to $h^1(x) = h^{2,-1}(x)$ and let $\underline{x}_l$ be the smallest solution to $l^1(x) = l^{2,-1}(x)$. Then for any solutions x_h and x_l to $h^1(x) = h^{2,-1}(x)$ and $l^1(x) = l^{2,-1}(x)$, respectively, we must have $x_l < \bar{x}_h$ and $\underline{x}_l < x_h$.*

Proof. Let $\bar{x} = (h^2)^{-1}(1)$. Observe that since $h^2 > l^2$, $(l^2)^{-1} > (h^2)^{-1}$, as $h^2(x) = l^2(x') = y$ implies that $x < x'$.

For $x > \bar{x}_h$, it must be that $h^1(x) < (h^2)^{-1}(x)$. For $h^1(\bar{x}) < (h^2)^{-1}(\bar{x}) = 1$, so if $h^1(x) > (h^2)^{-1}(x)$, then by continuity there is an $x' \in (x, \bar{x})$ such that $h^1(x') = (h^2)^{-1}(x')$, and $x' > \bar{x}_h$, which contradicts that $\bar{x}_h$ is the largest solution. But if $\bar{x}_h \leq x_l$, then $(h^2)^{-1}(x_l) < (l^2)^{-1}(x_l) = l^1(x_l) < h^1(x_l)$.

The proof that $\underline{x}_l < x_h$ is essentially the same. □

This lemma implies that the set of learning equilibria are strictly decreasing in the weak set order: the functions p_1^* and p_2^* are strictly decreasing in (σ_1, σ_2), so that both $\bar{p}(\sigma_1, \sigma_2)$ and $\underline{p}(\sigma_1, \sigma_2)$ are decreasing as well.

We can verify numerically that $p_1^*(\cdot|1,1)$ and $p_2^{*,-1}(\cdot|1,1)$ have a single point over intersection. By continuity, there will exist an open set around $(1,1)$, denoted $\underline{\Sigma} \in [0,1]^2$, such that if $(\sigma_1, \sigma_2) \in \underline{\Sigma}$, there is a unique point of intersection $\underline{p}(\sigma_1, \sigma_2)$ that is lower than the lowest point of intersection $\underline{p}(0,0)$.

B Vignettes for Section 4

All vignettes except Vignette 4 use caste neutral names. In the control vignette 0, the offense is unambiguous. The law is violated. The law prescribes the appropriate response. The vignettes 1-3 have some ambiguity, and thus there is greater scope for cultural attitudes towards punishment to operate. It is a difference in caste culture that we seek to evaluate. Vignette 4 is about inter-caste marriage, a central issue in the culture of honor in India. No mention of caste is made except in Vignette 4.

V0: (Control) A thief named Dinesh robs Mahesh's house at night. Mahesh finds out, beats up the thief, and reports him to the police. The thief is locked up in the police lock-up for a few days.

V1: Dinesh digs a canal through the field of Mahesh to have water reach his field. Dinesh does not obtain permission from Mahesh before doing this. He digs over night and Mahesh finds out the following morning, only after the canal has been dug. After this incident, Mahesh met Dinesh in public, had an argument, and badly beat up Dinesh.

V2: Dinesh lets his cattle graze on Mahesh's fields at night. Mahesh's crop is ready for harvest. Dinesh does not inform Mahesh, nor does he obtain permission from Mahesh before letting his cattle graze in Mahesh's fields. Mahesh finds out only after the crop has been eaten by Dinesh's animals. After this incident, Mahesh met Dinesh in public, had an argument, and badly beat up Dinesh.

V3: Mahesh was going back home at night for a feast in the village. On the way, he met Dinesh sitting with his friends. Dinesh called Mahesh names and beat him up. After this incident, Mahesh met Dinesh in public, had an argument, and badly beat up Dinesh.

V4a: Dinesh Thakur's daughter marries Mahesh Bania. After this incident, Dinesh met Mahesh in public, had an argument, and badly beat up Mahesh.

V4b: Dinesh Pasi's daughter marries Mahesh Chamar. After this incident, Dinesh met Mahesh in public, had an argument, and badly beat up Mahesh.

Note: V4a is for high-caste subjects and V4b is for low-caste subjects. Thakur and Bania are high castes, with Bania being a lower rank. Similarly, Pasi and Chamar are low castes, with Chamar being a lower rank.

Questions for the respondents:

(1) Was Mahesh's action justified?

(2) If you were in Mahesh's place, what would you have done?