Sanctioning and trustworthiness across ethnic groups: Experimental evidence from Afghanistan*

Vojtěch Bartoš †

Ian Levely‡

September 27, 2020

Abstract

Since social preferences towards individuals perceived as belonging to a different group are typically weaker, cooperation is more difficult in ethnically diverse settings. Using an economic experiment in Afghanistan, we show how the ability to impose financial penalties can help to overcome this. We use a trust game with two special features: investors communicate a desired back-transfer and, in some treatments, can choose whether to conditionally impose a small fine on trustees who do not comply with this request. We randomly paired subjects with either a co-ethnic or someone from a different ethnic group. We find that when investors do not have the ability to impose a fine, subjects are more trustworthy towards co-ethnics. When the fine is imposed by a co-ethnic, it has little effect. However, in cross-ethnic interactions, the fine increases trustworthiness, virtually eliminating in-group bias. Interestingly, this result is qualitatively similar when the fine is available to the investor but not used. These results suggest that institutions for enforcing cooperation are more effective when applied between, rather than within, ethnic groups, due to behavioral differences in how individuals respond to pecuniary sanctions.

Keywords: Sanctions, Cooperation, Crowding out, Moral incentives, Ethnicity, Afghanistan

JEL Classification: D01, D02, C93, J41

*We thank Nava Ashraf, Abigail Barr, Michal Bauer, Erwin Bulte, Subhasish M. Chowdhury, Davide Cantoni, Guillaume Frechette, Peter Katuščák, Klára Kalíšková, Friederike Lenel, Pieter Serneels, Maarten Voors and the seminar and conference participants at CERGE-EI, CESifo Behavioral Economics Conference, the CRC Rationality and Competition workshop, NYU, the Natural Experiments and Controlled Field Studies conference, Rutgers University, and University of Munich for their helpful comments, and Ahmad Qais Daneshjo and Hadia Essazada, for their excellent research assistance. The research was funded by GAUK (no. 46813), Czech Science Foundation (no. 13-20217S), the GDN (RRC13+11), and the German Science Foundation through CRC TRR 190. Disclaimer: Institutional Review Board approval has not been obtained because the institution, with which we were affiliated at the time of conducting the experiment (CERGE-EI and Charles University, Prague), did not have IRB. We declare that we have no relevant or material financial interests that relate to the research described in this paper. All opinions expressed are those of the author and have not been endorsed by CERGE-EI or the GDN.

†Corresponding author. Department of Economics, University of Munich, Geschwister-Scholl-Platz 1, D-80539 Munich, Germany (vojtech.bartos@econ.lmu.de).

‡Department of Political Economy, King's College London, Bush House NE, 30 Aldwych, London WC2B 4BG, UK.

1 Introduction

Cooperation is generally more difficult across ethnic lines. One reason is that individuals generally behave more pro-socially towards their in-group (Bernhard et al., 2006; Chen and Li, 2009), and trust and trustworthiness towards members of a different ethnic or social group is weaker (Glaeser et al., 2000). This has economic consequences, since both trust and trustworthiness are important elements of economic transactions (Arrow, 1972), and average levels of trust have been linked with economic development. Taken together, this helps to explain worse economic outcomes observed on average in ethnically diverse societies (Alesina and La Ferrara, 2005). Institutions that allow for punishment of shirkers can mitigate the negative effects of ethnic diversity, reducing reliance on voluntary cooperation in economic transactions (Habyarimana et al., 2007; Alexander and Christia, 2011; Masella et al., 2014). However, while the threat of punishment unambiguously improves cooperation among selfish individuals, it may reduce the intrinsic motivation of non-selfish agents to cooperate (Fehr and List, 2004; Falk and Kosfeld, 2006; Bowles and Polania-Reyes, 2012). Are such negative effects of punishment also group specific? Studying *how* group identity mediates behavioral reactions to punishment could lead to a better understanding of how institutions help facilitate cooperation in ethnically heterogeneous settings. This is especially relevant in developing and post-conflict settings where institutions are generally weak (Gennaioli and Rainer, 2007; Besley et al., 2010; Michalopoulos and Papaioannou, 2016; Ali et al., 2018).

In an experiment conducted in Mazar-i-Sharif, Afghanistan, we study how trustworthiness is affected by the availability and use of a financial sanction, and whether the effect varies with group identity. Our subjects are adult males who are members of either the Tajik or Hazara ethnic groups. In this setting, formal institutions are weak, making interpersonal trust essential in facilitating economic exchange. Ethnicity is also extremely salient, and communities and local institutions are *de facto* segregated.

We use a modified trust game, similar to Fehr and Rockenbach (2003). As in a standard trust game, an *investor* receives an endowment and can send any portion of it to a *trustee*. Whatever he sends is tripled by the experimenter. The trustee can then choose whether to send any portion of this back to the investor, and the amount they return serves as a measure of trustworthiness.[1] In contrast to the standard trust game, the investor in our experiment makes a request to the trustee regarding his

[1] In the absence of other-regarding preferences, a self-interested trustee will return nothing. Anticipating this, a self-interested investor will send nothing. However, the majority of investors across diverse settings and subjects do in fact send money, and the majority of trustees are to some extent trustworthy (Johnson and Mislin, 2011).

desired back-transfer. In one version of the game, (hereinafter the *"trust game"),* this is non-binding. In the *sanctioning game,* the investor has one more choice: he can impose a conditional fine on the trustee, which reduces the trustee's payoff if he returns less than the requested amount. The fine is sufficiently small that it is still in the trustee's best interest to return nothing, but by choosing to impose the conditional fine, an investor creates an additional motive for trustworthiness. However, if introducing financial incentives triggers different social preferences, in some circumstances, the fine could crowd-out intrinsic motivation. The game thus captures the interaction between financial and non-financial incentives. We compare the amount returned by trustees in the *trust game* and the *sanctioning game*. In the latter, we compare the amount returned by trustees when the fine is imposed and when the fine is available but not used.

The key innovation of this study is to test whether shared group identity moderates the response to financial sanctions. To accomplish this, we exogenously vary the ethnicity of players so that they interact with either someone from the same or a different ethnic group. Since our focus is the *reaction* to the fine, rather than the willingness to impose it, we carefully designed the experiment to isolate this. Trustees are presented with a series of choices. For some of these choices, the parameters— the amount sent, requested back-transfer, and whether the fine is used — are randomly varied, orthogonal to the ethnicity treatment.[2] Using this subset of choices allows us to causally study the effect of the decision to impose or not impose the fine on amount returned by trustees.

Our main finding is that imposing the conditional fine is more effective when the investor and trustee are from different ethnic groups. While the fines increase trustworthiness when imposed by someone from a different group, it has no effect when the investor is a co-ethnic. This is consistent with the fine producing two countervailing effects for in-group interactions: a financial effect, incentivizing higher returns, and a behavioral effect, in which the use of the fine crowds out intrinsic, pro-social motivation to return money to the investor. The result is a null effect of the sanction in co-ethnic interactions. When the investor and trustee come from different ethnic groups, there appears to be no such trade-off: the fine increases trustworthiness. Interestingly, we find qualitatively similar results regardless of whether the investor actually chose to apply the fine. When the investor came from a different ethnic group, we find higher trustworthiness in the *sanctioning game* when the fine was available but not used than we do in the *trust game*, when no fine was available.

[2]As explained in more detail in Section 4.3, trustees also received one decision from their partner, according to treatment, and were paid according to this decision. This was done to avoid deception; these decisions are not included in the main anslysis.

We observe an in-group bias in the *trust game*, with trustees returning more to members of their own ethnic group. However, the ability to impose a conditional fine closes this gap: in the *sanctioning game,* the amounts returned by trustees do not differ with the ethnic identity of the investor. These results have implications for understanding the relationship between group identity and institutions. Social preferences towards individuals without a shared group identity are often weaker, but certain institutions, especially those that facilitate punishment, can make cooperation easier.[3] Although introducing material incentives can crowd out intrinsic motivation, our results suggest that this is limited to interactions between members of the same group. This suggests that punishment is more effective as a means of enforcing cooperation when applied across ethnic lines.

While ethnically heterogeneous societies with low levels of trust benefit most from institutions that make enforcing contracts easier (Collier, 1999), the cooperation required to form institutions in the first place means that they might be the least likely to develop organically in such settings. We show how the behavioral reaction to sanctions partially mitigates this: at the margins, sanctioning institutions might have a higher return in diverse settings. Our results provide a mechanism that explains how strengthening formal institutions can improve inter-ethnic cooperation, and suggest that in diverse societies, the added value to enforcing contracts across ethnic groups is larger than doing so within ethnic groups.

2 Review of literature

This study links two topics in the experimental and behavioral economic literature: the role of group identity in shaping cooperative preferences and the effect of material incentives on pro-social behavior. In this section, we first outline existing literature on both topics separately, then discuss the few studies that directly examine punishment and inter-ethnic cooperation jointly. We build on this literature by showing how reactions to material sanctions differ with the group identity, and how this can improve inter-ethnic cooperation.

Numerous economic experiments have studied group bias in cooperation, using both organic group

[3]For example, Miguel (2004) argues that nation building in Tanzania has made inter-ethnic cooperation easier than in similar communities in Kenya. Hjort (2014) studies workers in a Kenyan factory, and finds that ethnically diverse teams are less productive, but a payment scheme that rewards teams rather than individuals, the difference disappears. In Sierra Leone, Glennerster et al. (2013) attribute a high capacity for collective action to the historical role of local authorities, whose jurisdiction crossed ethnic lines. By contrast, in Afghanistan most local institutions are specific to a particular ethnic group.

identity including ethnicity (Whitt and Wilson, 2007; Buchan et al., 2009; Meier et al., 2016), and group-identity induced in the laboratory (Tajfel et al., 1971; Charness et al., 2007; Chen and Li, 2009; Beekman et al., 2017). This literature demonstrates that parochial social preferences can impede cooperation with one's "out-group" (Bowles and Gintis, 2004). The trust game in particular has been used to study group identity: Falk and Zehnder (2013) find that residents of Zurich have higher trust and expect to receive higher back transfers from residents of their own neighborhoods. Fershtman and Gneezy (2001) demonstrate how ethnic stereotypes influence trust in Israel.[4]

A number of studies show that preferences for punishing norm violators are also sensitive to group identity: there is generally a greater willingness to punish members of different ethnic groups, and with greater severity (Shayo and Zussman, 2011; Rehavi and Starr, 2014). Data from third-party punishment experiments show that the identity of the victim also matters. Among traditional communities in Papua New Guinea, Bernhard et al. (2006) find that subjects are more willing to pay a cost to punish norm violators who harm an in-group member. Goette et al. (2006) and Goette et al. (2012) come to similar conclusions, studying group identity among Swiss soldiers, who were randomly assigned to platoons.

Additionally, we contribute to the literature on how material incentives crowd out intrinsic motivation. When non-material motivations for performing some costly action exist, explicit incentives can back-fire (Titmuss, 1971; Gneezy and Rustichini, 2000; Frey and Jegen, 2001; Falk and Kosfeld, 2006; Bowles, 2008; Ariely et al., 2009). For example, sanctions can signal mistrust or unfair intentions. This changes an individual's view of the other party, reducing altruistic or reciprocal motivation, and resulting in less cooperation (Falk and Kosfeld, 2006; Bowles, 2008). Our design is modeled after Fehr and Rockenbach (2003), who document the negative effects of a financial sanctions on trustworthiness. When investors in a trust game impose a fine, the amount returned by trustees decreases. These results replicate using samples of Costa Rican undergraduates and CEOs (Fehr and List, 2004). Conversely, these studies find that when an investor has the ability to sanction but intentionally refrains from doing so, this increases back-transfers. Forgoing the sanction sends an implicit signal of trust, which is rewarded by the trustee. Using a gift-exchange game, Falk and Kosfeld (2006) come to similar conclusions: principals are able to limit, but not eliminate, the ability of the agent to shirk. They find that doing so reduces the effort exerted by agents, on average.[5] Taken together, this literature suggests that imperfect contracts can be worse than no contract at all.

[4]Ethnic divisions do not necessarily limit cooperation, even in contacts where etnicity is salient, as shown in a recent experiment from Kenya (Berge et al., 2020). However, as the authors note, a lack of individual-level co-ethnic bias in preferences does not necessarily preclude institutionalized ethnic divisions.

[5]See also Riener and Widerhold (2016).

Three studies deal with the intersection of group identity, cooperation, and punishment. Habyarimana et al. (2007) conduct a public goods experiment in Uganda, finding that ethnically homogeneous pairs cooperate more efficiently than ethnically mixed pairs. They explain this by showing that heterogeneous groups lack informal mechanisms to discourage free-riding. When the experiment introduces a formal method for doing so, third-party punishment, the in-group advantage disappears. The second study, Alexander and Christia (2011), shows that the ability to punish in a public good game has a different effect on cooperation among ethnically heterogeneous and homogeneous groups of high-school students in Bosnia-Herzegovina. They find that, while ethnically mixed groups cooperate less, adding the ability to punish peers closes the gap.[6]

While both of these studies show that punishment (in the form of financial sanctions) can help mixed groups cooperate, it is unclear why. Possibly, punishment is relatively more useful for heterogeneous groups because there is an increase in (perceived) willingness to punish. Alternatively, an individual's reaction to receiving punishment could differ according to whether the person responsible is an in-group member.[7] A distinguishing feature of our study is that we concentrate on how group identity mediates the reaction to receiving a financial sanction, distinctly from the willingness to impose one. We show that the availability of a conditional fine in the *trust game* closes the gap between in-group and out-group trustworthiness, independent of the investor's behavior.

The third closely related study, Masella et al. (2014), is an experiment with students in Germany measuring aversion to control, using a design similar to Falk and Kosfeld (2006). They introduce minimal group identity in the laboratory and find that limiting the agent's ability to shirk reduces effort both within and between groups, but for different reasons. Control over the agent's choices is unexpected when exerted within groups, and crowds out intrinsic motivation. When used between groups, control creates hostility. In contrast, our study examines the effects of a type of punishment—conditional financial sanctions—on cooperative behavior.

[6]Interestingly, this is only true for subjects who attend integrated schools, and thus have experience interacting with students of a different ethnicity.

[7]Note that since punishment decisions are made in response to public good contributions, the two decisions are endogenous in these experiments. Moreover, since they do not use the strategy method, the observed frequency of punishment is not necessarily indicative of *willingness* to punish, since the perceived threat of punishment could lead to higher contributions.

3 Setting

There are several features of Afghanistan that make it an ideal setting for our study. It is an ethnically heterogenous country with salient ethnic divisions. This results in a separation of informal institutions across ethnic lines. While most of our subjects have little experience with formal institutions, recent state building efforts have aimed at formalization.

Ethnicity has been used as a political tool to divide Afghan society throughout recent history. Our subjects are Tajiks and Hazaras, who, along with Uzbeks, were singled out as groups distinct from the Pashtun majority by the royal government in the early twentieth century. Later, the Mujahideen fighting the Soviet occupation were organized along ethnic divisions. After the defeat of the Soviets, civil war ensued and the Mujahideen factions started fighting one another. The predominately Pashtun Taliban, a fundamentalist Sunni Muslim group, was responsible for the widespread persecution of ethnic minorities, especially the Hazara, including ethnic cleansing campaigns (Schetter, 2016).Ethnic affiliation continues to play a large role in Afghan politics at both national and local levels, and is salient in the ordinary lives of Afghans. After Pashtuns (43 percent), Tajiks (31 percent) and Hazaras (9 percent) constitute the second and third largest ethnic groups in Afghanistan (DHS, 2011). Hazaras are the largest predominantly Shia Muslim group. In the Balkh province, where the experiment was conducted, Tajiks are the largest group (44 percent). Among our sample, cohabitation is mostly peaceful—over 70 percent of the participants in our study reported having friends among people from a different ethnic group. Nonetheless, the communities we study are mostly self-segregated along ethnic lines. Beath et al. (2016) report that 75 percent of villages in their sample, which covers most of Afghanistan, are perfectly ethnically homogeneous.

This segregation means that the informal institutions affecting day-to-day interactions are also separated by ethnicity. For example, disputes within a single community are often settled by village or neighborhood's *kalantar* (community leader). Settling a dispute between members of different communities might involve mediation from *kalanatars* from both communities. Segregation is also present in trade and finance, both in the country and in international migration networks. Monsutti (2005, p. 238) writes that for the Hazara, "in the absence of genuine rule of law, [. . .] successful financial transactions depend upon trust and therefore great closeness among people involved in them," where closeness is mainly defined by ethnicity and religion. We find evidence of this in our survey data: while 18.6 percent of all subjects reported informal loans from co-ethnics, only 0.8 percent of respondents

owed money to members of a different ethnic group.

There is an ongoing attempt to strengthen formal institutions in the wake of the Afghan conflict. The United States alone spent over $100 billion on improving security and strengthening governance at both the national and local levels between 2001 and 2016 (SIGAR, 2017, p. 69). The World Bank and the Afghan national government are also implementing a nation-wide Citizens' Charter Afghanistan Project, a successor to the National Solidarity Program aimed at introducing formal local governance bodies, through which small infrastructure development grants are channeled. Between 2003 and 2013, over 32,000 Community Development Councils were introduced nationwide.

Despite this, our subjects have very little experience with formal governmental institutions: only 5.6 percent have ever signed a written contract, over half of the sample have never attended school, only 4.5 percent are employees of a registered private company or of a state institution, and only 1.5 percent of participants who are currently in debt owed money to a bank, micro-credit organization or other formal credit institution. When asked how they would respond to a hypothetical theft, only 10.7 percent say they would contact the police, compared to 40.0 percent who would contact their *kalantar*, and 37.2 percent who would contact a neighbor.

This situation in Afghanistan mirrors our experiment: we study how ethnic identity interacts with the introduction of a financial sanction, which serves as a proxy for a formal institution. The way that individuals respond to the presence of a financial sanction could depend on their culture and previous experience interacting with institutions. The existing studies that show how material incentives can crowd out pro-social motivations have been conducted with populations that have extensive experience with formal institutions. Understanding how individuals with little such experience is especially relevant to understanding how the interaction between group identity, prosociality and material incentives contribute to the development of formal institutions.

4 Design

4.1 Experimental games

To examine the effect of pecuniary sanctions on pro-social motivations, we use two experimental games following the design of Fehr and Rockenbach (2003). There are two anonymously matched players in

both games, an investor and a trustee, who both receive an initial endowment of ω= Afs 100, which was equivalent to around $2 US at the time of the experiment. An investor, i, then chooses whether to "trust" the trustee by transferring some portion of his endowment, $s_i \in [0, 10, 20, \ldots \omega]$. The amount sent is tripled by the experimenter, and the trustee receives $3s_i$. The trustee, t, then has the option of transferring some portion of what he receives, $r_t \in [0, 10, 20, \ldots 3s_i]$, back to the investor, thus sharing the benefits of the increased stake.

The payoffs for the investor and the trustee in the *trust game*, respectively, are:

$$\pi_i = \omega - s_i + r_t \tag{1}$$

$$\pi_t = \omega + 3s_i - r_t. \tag{2}$$

In contrast to a standard trust game, the investor also communicates a desired back transfer, $r_i^* \in [0, 10, 20, \ldots, 3s_i]$, to the trustee. In the *trust game*, this request is "cheap talk" and does not affect the payoffs of either party.

All subjects also played the *sanctioning game*, which adds one additional feature to the *trust game* with requested back transfers: the investor can choose whether to impose a fine, $f = 40$, dependent on whether the trustee's back transfer is less than the amount requested by the investor. Applying the fine is costless to the investor. We denote the decision to impose the fine as $p_i \in \{0, 1\}$, where $p_i = 1$ if the investor chooses to conditionally apply the fine and zero otherwise.

The payoff of the function for the trustee in the *sanctioning game* is given by:

$$\pi_t = \omega + 3s_i - r_t - fp_i(\mathbf{1}\{r_t < r_i^*\}) \tag{3}$$

and the payoff for the investor is identical as in the *trust game* (Equation 1).

From the trustee's perspective, there are three possible conditions across the two games: 1) the *trust game*, in which the option to impose the conditional fine is not available to the investor; and two conditions in the *sanctioning game*: 2) the *fine condition*, in which the fine was both available and applied; and 3) the *no-fine condition*, in which the fine was available, but not used.

In the *sanctioning game*, with the parameters we use,[8] the fine is too small to allow the investor

[8]These parameters are identical to Fehr and Rockenbach (2003) and Fehr and List (2004).

to capture the efficiency gains from a self-interested trustee in all but one, extreme case.[9] Assuming that the decision to impose the fine does not negatively affect trustworthiness, it provides a financial incentive, in addition to intrinsic motivation, for a trustee to meet the investor's request.

However, the fine could also affect trustworthiness by activating "state-dependent" preferences. First, the presence of the fine might change the nature of the relationship between the trustee and the investor, and thus activate a different set of norms or preferences than in the *trust game*. Second, since the investor chooses whether or not to apply the fine, this decision may signal something about his character or intentions, and this in turn may change the weight given to his payoff in the trustee's utility. By choosing to impose the fine, potentially the investor communicates a lack of trust, and this could impact back transfers. On the other hand, in the *no-fine condition*, the decision to refrain from imposing the fine might implicitly signal trust. This "good news" about the investor's beliefs and intentions could increase the amount returned by trustees (Bowles and Polania-Reyes, 2012). Our design allows us to compare trustworthiness in the *sanctioning* and *trust games*, and similarly compare results between the *fine* and *no-fine conditions* to study these effects.

If the fine crowds out trustworthiness, and the effect is large enough in magnitude, then it could induce a trustee to return less than he would in the *trust game* and *no-fine condition*. It is also possible that the fine "crowds-in" trustworthiness, by reinforcing norms or social preferences for behaving cooperatively or complying with requests.

Investors also played a *triple-dictator game*, which resembles the *trust game*, but in which the trustee—a passive receiver in this game—has no option to return any portion of the amount received. As in the *trust game*, investors are given endowments equal to trustees', $\omega = 100$, and the amount transferred was tripled by the experimenter. The game allows us to identify altruistic motivations and efficiency concerns independently of the beliefs and strategic concerns that affect the investor's behavior in the trust and *sanctioning games* (Fershtman and Gneezy, 2001; Cox, 2004; Bauer et al., 2018).

[9] If the investor sends $s_i = 10$, requests $r_i^* = 30$ and imposes the fine, then the trustee will maximize his profit by returning $r_t = r_i^* = 30$ to avoid paying the fine, $f = 40$. Thus, the maximum profit an investor can achieve when playing with a self-interested trustee is 10 Afs. Only one investor in our sample selected this strategy. If an investor sends $s_i = 20$ and requests $r_i^* = 40$, the trustee is indifferent between paying the fine and returning $r_t = r_i^* = 40$. Whenever $s_i > 20$, the trustee will always maximize his earnings by returning $r_i = 0$, regardless of amount requested and whether the fine is applied.

4.2 Treatments

In order to study how ethnicity affects trustworthiness and the response to sanctioning, we sampled only subjects who identify as either Tajik or Hazara, and held sessions with subjects exclusively from one group or the other. Treatment was assigned at the session level. Subjects were read a short profile describing their partner, which included the general selection criteria used for subject sampling, in addition to the fact that their partner lived in a community that was "mostly Tajik" or "mostly Hazara" according to treatment.[10] Thus we have four treatment arms in all, in a two-by-two design: the investor's ethnic identity was either Tajik or Hazara, and the trustee's ethnicity varied similarly. For most of the analysis, however, we condense this into an *in-group* treatment, in which both the investor and trustee share the same ethnic identity, and an *out-group* treatment in which their ethnic identities differed.

The profile read to subjects included additional information on the age range of subjects, that their partner was male, married and had at least one child, in order to avoid an experimental demand effect that might result from making the aim of the study too obvious. Since ethnicity is prominent in everyday life in Afghanistan, it is a reasonable assumption that including it in the description did not seem particularly out-of-place for subjects.

4.3 Procedures

In total we conducted 28 experimental sessions with 434 subjects in October and November 2013 in 7 predominantly Tajik and 6 predominantly Hazara peri-urban areas of Mazar-i-Sharif, which is located in the North of Afghanistan. The population is generally engaged in day labor or agriculture and communities are strongly ethnically homogeneous.

Subjects were randomly sampled according to their place of residence within the areas we selected. Individuals meeting our criteria (a married male between 18 to 60 years of age, with at least one child, and of a particular ethnic group) were invited to participate in the experiment. We used these criteria in order to focus on individuals with economic decision making power within their households. We studied males only, due to the cultural restrictions involved with working with female respondents in Afghanistan. The selection criteria were the same for both the investors and the trustees, and for both

[10] We did not deceive subjects: individuals were indeed matched with partners that fit the profile and they were paid according to that individual's decision.

ethnic groups. We were able to contact 76 percent of household heads sampled, and 85 percent of those interviewed matched our criteria and were invited to participate. There is no significant difference in willingness to participate across Tajik and Hazara communities (80 and 76 percent, respectively; two-sided t-test p=0.34). Supplementary Table A2 describes the sampling procedure in detail. The table also explains how the data used for the analysis were selected.

The experiment was conducted in groups of 15-20 subjects, who were informed that they would be matched with a person from a different community located in Mazar-i-Sharif, but that they would not know which specific community, nor would their partner be informed of their specific community. The profile describing the partner in the *trust game* was read several times throughout the experiment and 90 percent of the subjects correctly mentioned the ethnicity of their partners after the experiment when asked about their partner's characteristics. The treatment information was communicated during the group portion of the instructions, and thus our ethnic treatments are randomized at the session level. The other characteristics included in the profile remained constant for all treatments.

Roles in the game (i.e. investor and trustee) were assigned at the session level, and all subjects played both the *trust game* and the *sanctioning game*. We varied the order of the two games across sessions.[11] Following these two games, investors played the *triple dictator game* and trustees were informed of the possibility that they would receive money from the investors' dictator decisions.

Since the subject pool is largely illiterate, all instructions were given orally, using visual aids.[12] After a general introduction of the experiment and explanation of the task in a group setting, the subjects were seated in booths (See Supplementary Figure A1) where they made decisions in privacy—though not anonymous to research assistants. The choices were presented using simple visual aids in order to accommodate illiterate subjects (See Supplementary Figure A2).[13]

Trustees made several choices in each game. In addition to the actual investor choices in the *trust game* and the *sanctioning game*, we provided the trustees with four choices in the *sanctioning game* using randomly assigned parameters (s_i, r_i^*), including two choices each in the *fine* and *no-fine conditions*. Trustees were also presented with two choices using random parameters (s_i and r_i^*) in the *trust game*.[14] Subjects were told (truthfully) that each decision came from a participant from a

[11] In 75 percent of sessions trustees played the *sanctioning game* first, with the order reversed for the remaining sessions.

[12] Our script builds on the instructions originally used in Barr (2003) and Bauer et al. (2018). See Online Appendix C for the complete instructions (https://bit.ly/2FW72yG).

[13] Each decision was made by putting "banknotes" into envelopes representing money to be kept and sent/returned to one's partner. The endowments for each player were represented visually as well.

[14] The parameters within each game and condition were randomly selected from a pool of decisions made by investors in earlier sessions (or in pilot session). The distribution was the same for both treatments. We report the means of

previous session, but that only one of them came from the person we had described, and that they would be paid for this decision only. Since they did not know, ex-ante, which decision this was, they should have treated all decisions as if they had been intentionally chosen by their partner (i.e. reflecting the treatment).[15] [16] Supplementary Table A1 summarizes the structure of decisions that trustees made, the source of parameters for each decision, which decisions were paid and which are included in the analysis, and the values of the parameters by game and treatment.

There are three advantages to this method. First, the parameters communicated to subjects are orthogonal to the group treatment. This allows us to study trustees' responses to each parameter directly; if we were to examine only trustees' responses to investors' actual choices, the decision to impose the fine would plausibly be correlated with both the group treatment as well as the amount sent and the amount requested, and would thus bias our estimates. Second, exogenously varying the parameters of the game gives us the potential to explore a range of possible decision types, even if those decisions were not commonly chosen by investors. And third, collecting data from multiple decisions for each trustee considerably increases statistical power.[17]

At the end of each session we administered a short, one-on-one survey with all subjects, which included questions on demographic information, membership in various formal and informal organizations, experience with formal and informal credit markets, experience with writing or signing formal contracts, and hypothetical questions designed to elicit their degree of experience with attitudes towards formal institutions.

Each subject received a 100 Afs show-up fee. This is a substantial amount of money, compared to wages for a day of manual labor of around 150 Afs. Subjects were informed that the payoff they earned in the games would be distributed in two days to allow us to match their responses with their partner's.

---

parameters used by game, condition, and treatment in Supplementary Table A1.

[15] Note that since treatment was assigned at the session level, subjects had no knowledge of the other treatment, and therefore no specific reason to assume that some decisions came from investors from a different ethnicity than that described in the treatment.

[16] To give a concrete example, one subject was sequentially presented with this series of five choices: T1($s_i = 50$, $r_i^* = 100$), T2($s_i = 90$,$r_i^* = 180$), T3($s_i = 30$, $r_i^* = 60$), S1($s_i = 50$, $r_i^* = 100$, $p_i = 0$), S2($s_i = 60$, $r_i^* = 60$, $p_i = 0$), S3($s_i = 60$, $r_i^* = 60$, $p_i = 1$), S4($s_i = 30$, $r_i^* = 60$,$p_i = 1$), and S5($s_i = 80$,$r_i^* = 170$, $p_i = 0$). Decisions 3 and 8 came from a person matching the treatment profile, and the subject was paid for one of these. Since these parameters are endogenous, the corresponding trustee choices are excluded from the main analysis.

[17] We assigned the parameters according to two dimensions. (i) fair/unfair requests, depending on whether the proposed payoff for the trustee was at least as high as/lower than the sender's. (similar to Fehr and Rockenbach, 2003); and (ii) low/high-trust depending on whether $s_i < 50$ Afs or $s_i \geq 50$ Afs (i.e. half of the endowment). In the *trust game*, each trustee was presented with two randomly selected scenarios, each from a different category, but with different parameters. In the *sanctioning game*, each trustee was presented with four random decisions: two decisions for each of the two categories he received in the *trust game*, one with the fine imposed and one without. This procedure was used in order to limit within-subject variance.

5 Results

5.1 Group identity, fines, and trustworthiness

First, we analyze how both the availability and use of the fine affects the amount returned in the *trust game* across the *in-group* and *out-group* treatments. This allows us to test our main hypothesis: trustees react differently to the fine depending on whether the investor is a co-ethnic. Figure 1 summarizes for both the *trust* and *sanctioning games* (See also Supplementary Table A3). We limit our analysis to the decisions in which the parameters were randomly assigned to trustees by the experimenter, independent of the group treatment (decisions S1-S4; T1-T2), which allows us to interpret treatment effects causally. We begin by analyzing the amount returned by trustees in each game and treatment. The first two bars of Figure 1 demonstrate that trustees in the *trust game* return a significantly higher portion of what they receive in the *in-group* treatment than they do in the *out-group* treatment: 58.3 vs. 42.3 percent (p=0.00).[18] This indicates that ethnicity is indeed salient among the population that we study, and has an effect on trustworthiness.

In Table 2 we regress the percentage returned on treatment in order to confirm that this result is robust to controlling for the amount sent and requested back transfer. We include individual-level random effects with standard errors clustered at the session level. The model includes dummies for the *fine* and *no-fine conditions*, with the *trust game* as the excluded category. The coefficient for *in-group* in column 1 thus captures the treatment effect in the *trust game*.[19]

Next, we compare trustee behavior across treatments in the *sanctioning game*. Bars 3 and 4 of Figure 1 report the shares returned in the *fine condition*. Compared to the *trust game*, subjects in the *out-group* treatment send back more in the *fine condition*, returning 58.5 percent on average, an increase of 16.2 percentage points (p=0.00). In contrast, for the *in-group* treatment there is only a slight increase in the amount returned relative to the *trust game* (3.5 percentage points, p=0.00). Notably, under the *fine condition*, the gap between *in-group* and *out-group* treatments narrows to only

[18] All significance levels reported for comparison of means are from Somer's D tests clustered at the session level, and 95 percent confidence intervals reported for insignificant results are obtained using individual level random effects regressions with standard errors clustered at session level, unless otherwise noted.

[19] The results in column 1 of Table 2 are also robust to various alternative specifications. First, in Supplementary Table A4 we 1) control for enumerator fixed effects, 2) control for the order in which the games were played, 3) correct for the small number of clusters by using a linear model with multi-way bootstrapped clustered standard errors, 4) add individual fixed effects (without treatment), 5) exclude the *amount sent*, 6) exclude *share requested*, 7) include only the first choices for each game, and 8) include only decisions from the actual investors' decisions, rather than the exogenously assigned parameters. Second, the results are robust to using quantile regressions (Supplementary Table A5). Lastly, we restrict the analysis to only those sets of parameters that are represented for all games and sanctioning choices in the *sanctioning game* for a given ethnic treatment (Supplementary Table A6).

3.4 percentage points, which is no longer significant (p=0.46, CI -5.07 to 11.85). The regression results in column 1 of Table 2 show that the difference-in-differences between sanctioning and group treatment is statistically significant (*in-group x fine*). In columns 2-3 we divide the sample by treatment group and observe that while the fine increases back transfers relative to the *trust game* in the *out-group* treatment, there is no effect of the fine on back transfers in the *in-group* treatment.

Result 1: *The fine condition increases the average share returned, relative to the trust game, in the out-group treatment only. There is no change in the in-group treatment.*

We next turn to the *no-fine condition*, in which the investor had the option to apply the fine but refrained from doing so. Relative to the *trust game*, this condition could activate preferences related to the difference in institutional environment created by the availability of the fine. On the other hand, the investor potentially signals implicit trust by voluntarily refraining from imposing the fine (Bowles and Polania-Reyes, 2012). Bars 5-6 of Figure 1 present results from the *no-fine condition*. For trustees in the *out-group* treatment, the share returned in the *no-fine condition* falls between the levels for the *trust game* and *fine condition* at 49.3 percent, a decrease of 9.2 percentage points over the *fine condition* (p=0.01) and 6.9 percentage points higher than in the *trust game* (p=0.00). For the *in-group* treatment, however, the share returned in the *no-fine condition* was virtually the same as in both the *trust game* (p=0.75, CI -2.26 to 4.54) and in the *fine condition* (p=0.05). The regression results in Table 2 confirm this. In column 2 we observe that, in the *in-group* treatment, back transfers in the *no-fine condition* do not differ from those in the *trust game* (p=0.80), nor from those in the *fine condition* (p=0.76). For the *out-group*, on the other hand, *no-fine condition* increases the share returned by 5.5 percentage points relative to the *trust game* (p=0.02).

Result 2: *The no-fine condition increases the average share returned, relative to the trust game, in the out-group treatment only. There is no change in the in-group treatment.*

Since there is no financial effect to consider, the change in trustworthiness that results from the *no-fine condition* is necessarily a behavioral one. The literature provides two potential explanations for Result 2. First, the ability to impose a fine might change the relationship between investor and trustee, and activate a different set of preferences. Secondly, by voluntarily abstaining from imposing the fine, the investor might signal something about his character or intentions. Higher back transfers

in this case might be a response to this "good news."[20]

In the regressions we control for the *amount sent* and the *share requested* by the investor. This helps to account for any differences between treatments and conditions in the parameters that we assigned (See Supplementary Table A1). However, the *share requested* has different implications across games and conditions. The *share requested* changes incentives in the *fine condition* only. While it is payoff neutral in both the *trust game* and *no-fine condition*, the request may nonetheless communicate qualitatively different information about the investor's intentions. To address this, we interact *share requested* with *fine* and *no-fine* in columns 4-5 of Table 2, with the sample split by group treatments.[21]

In general, the results show that trustees respond positively to requests: the percentage returned is consistently increasing in the *share requested* in both treatments and across all conditions. In the *trust game*, requesting an additional 1 percent of the tripled amount leads to a 0.4 percent increase in back-transfers in the *in-group* treatment, and a slightly lower effect in the *out-group* treatment (0.3 percent).[22] In the *fine condition,* requests are more effective in increasing back-transfers. However, the effect is not group specific. This is in contrast to the effect of requests on the *share returned* in the *no-fine condition*. For the *in-group* treatment, requests were followed more closely in the *no-fine condition* than in the *trust game*. In the *out-group* treatment, the opposite is true (although the *no-fine x share requested* coefficient is not statistically significant, p=0.19). This suggests that requests in the *no-fine condition* may send a qualitatively different signal about the investor's intentions in the *in-group* and *out-group* treatments.

After accounting for the different effects of *share requested* by game and condition, we find evidence that, on average, the *fine condition* crowds out trustworthiness in the *in-group* treatment and has only a small and statistically insignificant effect for the *out-group* treatment. The *no-fine condition* is similar to the *fine condition* for the *ingroup* subjects, but increases back transfers relative to the *trust game* and *fine condition* for the *out-group* treatment. However, it is likely that the response to *share requested* is non-linear and depends on whether the requested amount is considered fair or unfair. We explore this in the next subsection.

[20] See Houser et al. (2008), who explore intentions in the *sanctioning game* in detail.

[21] There are potentially too few sessions in order for parametric clustering of standard errors to produce consistent results in columns 4-5 of Table 2. Results are similar using the same model without clustering and using a linear regression model (i.e. without random effects) with multi-way bootstrapped standard errors. Available upon request.

[22] The difference is not statistically significant (p=0.23).

5.2 Fairness of requests

Both theory (Rabin, 1993; Fehr and Schmidt, 1999) and experimental evidence (e.g. Herrmann et al., 2008; Henrich et al., 2010) suggest that individuals tend to reward fair behavior and to punish behavior they consider unfair. In line with this, Fehr and Rockenbach (2003) find that responses to the *fine condition* differ between "fair" and "unfair" requests. We use their definition of a fair request: the amount sent and requested by the investor is such that the payoffs of the investor and the trustee are either equal or are in favor of the trustee, which implies that $r_i^*/3s_i \leq 0.67$.[23]

In columns 1 to 3 in Panel A of Table 3 we present results for fair requests only (as before, considering only decisions made using randomly assigned parameters). In column 2, which includes only fair requests for *in-group* subjects, the effect of the *fine condition*, reduces the share returned, relative to the *trust game*, by 3.8 percentage points (p=0.08).

We also find stronger results for the *no-fine condition* for *in-group* subjects when we consider only fair requests: there is a corresponding decrease of 4.3 percentage points in the share returned relative to the *trust game* (p=0.08). Interestingly, there is no significant difference between the *fine* and *no-fine conditions* (p=0.67, CI: -1.9 to 2.9). This suggests that, in the *in-group treatment,* it is the *ability* of investors to impose a fine that reduces back-transfers, rather than any signal sent by the decision to impose the fine. Plausibly the possibility of the fine in the *sanctioning game* provides a situational cue that activates a different set of (more selfish) preferences than the *trust game.*[24]

For the *out-group* and fair requests, the effect is reversed, and the *fine condition* has a similar effect on the sub-sample of fair decisions as it does overall, increasing the share returned by 9.7 percentage points (p=0.03). Likewise, and contrary to the results for the *in-group* treatment, the *no-fine* condition is associated with a 6.3 percentage-point increase in the share returned (p=0.02), relative to the *trust game* (Table 3, column 3 of Panel A). As with the *in-group* treatment, there is no statistically significant difference between the *fine* and *no-fine conditions* (p=0.47, CI: -5.68 to 12.41).

Next, we turn to unfair allocations, in which the amount requested leaves the investor with a higher payoff than the trustee. There is no effect in the *in-group* treatment. In column 5 of Panel A we find that the effect of *sanctioning* relative to the *trust game*, while positive (4.0 percentage points), is not statistically significant for the *in-group* treatment (p=0.20, CI -2.1 to 10.2). Nor is there any

[23]Note that we assumed in our design, ex ante, that trustee's decisions might qualitatively differ with respect to this dimension, and we provided subjects with an equal number of each type of decisions, giving us a roughly balanced number of observations in each category. In total we have 547 observations of fair requests and 452 observations of unfair requests.

[24]See Bowles and Polania-Reyes (2012), who refer to this phenomenon as "moral disengagement."

statistically significant difference between the *fine condition* and the *no-fine conditions* (p=0.98, CI -6.3 to 6.5). In column 6, we find that results for the "unfair" allocations in the *out-group* are qualitatively similar to the fair decisions.

These results indicate that reactions to the *fine* and *no-fine conditions* differ according to the fairness of requests. In the *in-group* treatment, when requests are fair, we find clear evidence that both the *fine* and *no-fine conditions* crowd out trustworthiness. We see that the vast majority of decisions made by investors in both treatments, for both the *trust game* and the *sanctioning game*, are classified as fair.[25] It is therefore possible that trustees had less clear interpretations as to the investor's intentions when faced with unfair decisions, and indeed this is reflected by the larger standard errors for most coefficients when trustees faced unfair requests. For the *out-group* treatment, our main findings are consistent regardless of whether the request was fair or unfair.

5.3 The behavioral effect of fines

We find that financial sanctions increase trustworthiness in the *out-group*, but not *in-group* treatment. The fact that we do not see a response to the fine for subjects in the *in-group* treatment on average does not necessarily indicate that it has no underlying effects, however, as the financial effect of the fine may cancel out the behavioral effect. In fact, there are several scenarios consistent with our results: First, the *fine condition* could influence behavior by crowding out intrinsic motivation for trustworthiness, but less so in the *out-group* treatment. Second, the fine could complement or "crowd-in" non-financial incentives in the *out-group* treatment, but have no effect on the *in-group* treatment. And third, the fine could have an opposite behavioral effect in each group treatment, reinforcing pro-social norms for the *out-group,* but crowding out moral incentives for the *in-group.* Understanding this underlying mechanism would make it possible to draw broader conclusions from the results of the experiment.

If a trustee's utility is maximized by choosing to return more than the requested amount, $r_t > r_i^*$, in the *trust game* or the *no-fine condition,* then the introduction of the fine provides no material incentive for the subject to change his behavior.[26]

On the other hand, if the fine crowds out non-financial incentives for cooperation, then the trustee will maximize his utility by returning a smaller amount when the fine is imposed, since he then puts

[25] Fair requests were made by subjects in the *in-group* and *out-group*, respectively, by 83.6 and 92.3 percent of subjects in *trust game*, and 78.8 and, 91.7 percent of subjects in the *sanctioning game.*

[26] For a more formal discussion, refer to the theoretical framework we present in Online Appendix B.

less weight on the investor's payoff. If the magnitude of this behavioral response were large enough, holding all other parameters constant, the trustee would no longer return more than the requested amount, and therefore on average we would expect to see a drop in the frequency of trustees returning $r_t > r_i^*$ in the *fine condition*, relative to the *trust game*. Alternatively, if the fine reinforces (crowds in) existing norms of reciprocity or altruism, then the frequency of decisions in which $r_t > r_i^*$ could be larger in the *fine condition* than in the *trust game*, following similar logic.

The frequencies of each type of decision in the *trust game*, and *fine* and *no-fine conditions*, by group treatment, are shown in Supplementary Figure A3. In the *trust game*, 32.5 and 15.4 percent of subjects returned more than the requested amount in the *in-group* and *out-group* treatments, respectively. In the *in-group* treatment, the frequency of decisions in which the amount returned was more than the requested amount drops by 9.4 percentage points in the *fine condition* relative to the *trust game*. This provides support for our finding that imposing the fine does in fact crowd out trustworthiness in the *in-group* treatment. The difference is statistically significant (p=0.00). In the *out-group* treatment, however, there is actually an increase of 6.8 percentage points in the frequency of decisions in which the amount returned exceeds the amount requested in the *fine condition* relative to the *trust game* (p=0.07).

There is virtually no change in the proportion of subjects returning more than requested between the *trust game* and *no-fine condition* in either the *in-group* and *out-group* treatments (p=0.64, CI -0.12 to 0.07, and p=0.48, CI -0.04 to 0.09, respectively).[27]

Although the analysis of trustees who returned more than the amount requested offers the cleanest evidence of crowding out, the individuals in this group may not be representative of the population as a whole, and financial sanctions are typically used in cases when individuals would have otherwise not cooperated. In Online Appendix B we formally show that, absent any state-dependent preferences and assuming that (state-independent) altruism is stronger towards in-group members, the frequency of individuals in the *in-group* treatment returning less than the requested amount ($r_t < r_i^*$) in the *fine condition* should drop relatively more than in the *out-group* treatment. The basic intuition is that the higher utility a trustee receives from his partner's payoff, the more he is willing to increase his back transfers to the requested amount, in order to avoid paying the fine. With sufficiently low

[27]Results of a random effects probit model presented in columns 1 to 3 of Panel B in Table 3 directionally match the findings presented in the main text (controlling for observables). While the increase in the frequency of returning more than requested in the *fine condition*, relative to the *trust game*, is statistically significant for the *out-group*, the decrease in similar decisions for the *in-group* is not statistically significant. Also, after controlling for observables, we find an increase in the frequency of returning more than requested in the *no-fine condition*, relative to the *trust game*, for the *out-group* treatment only. The results are strongest in the case of fair requests (See Supplementary Table A9).

altruism, the trustee would prefer to pay the fine, and decrease his back transfer to compensate for the loss incurred, reducing the investor's payoff as a result. Since we assume that altruism is higher in the *in-group* treatment, we predict that the fine should be more effective in lowering the frequency of decisions in which the trustee returns less than requested in the *in-group* treatment than in the *out-group* treatment.

This is not what we find: we observe a drop in the frequency of decisions in which trustees return less than the requested amount from the *trust game* to the *fine condition* in the *out-group* treatment by nearly half, from 56.8 percent to 28.7 percent of total decisions, (p=0.00). The decrease in the *in-group* treatment is comparatively smaller: from 42.3 percent in the *trust game* to 32.5 percent in the *fine condition* (p=0.05, see Supplementary Figure A3). The results are confirmed using a probit estimation that includes observables. See columns 4 to 6 in Panel B of Table 3.

To summarize, the results for the *out-group* suggest that financial sanctions actually reinforce behavioral motivations to return a higher amount, but crowd out trustworthiness for those in the *in-group* treatment.

5.4 Investor results

Since the components of the investor's decision in the *sanctioning game*—the choice of whether to impose the fine, the amount sent and requested back transfer—are endogenous, it is difficult to draw independent conclusions from any one measure in isolation. Given this, we briefly consider social efficiency and frequency of imposing the fine across treatments.

First, we do not find differences in the amount sent in the *trust game* by treatment. Since the amount sent in each of the three games was tripled, this can be interpreted as a measure of social efficiency. In the *trust game*, investors sent 57.21 Afs and 56.19 Afs on average in the *in-group* and *out-group* treatments, respectively (p=0.85, CI -11.24 to 9.20).[28] Does the option to impose a fine affect the amount sent? In the *sanctioning game*, results are nearly identical to the *trust game*: investors sent 55.96 Afs in the *in-group* treatment and 56.67 Afs in the *out-group* treatment, on average (p=0.80, CI: -7.71 to 9.12). Nor do we find a treatment effect when we compare the difference in what each investor sent in the *sanctioning game* minus the amount sent in the *trust game* (p=0.68, CI: -0.22 to

[28]For investor results, all significance levels reported for comparison of means are from Somer's D tests clustered at the session level, and 95 percent confidence intervals reported for insignificant results are obtained using individual OLS regressions with standard errors clustered at session level.

0.14). [29]

Investors in both treatments do, however, send less in the *dictator game* than in the *trust* and *sanctioning games*. This shows that subjects react strategically to the experimental environment, though not to the presence of the sanctioning mechanism.[30] Moreover, while there is no difference in the overall size of the surplus, investors differentiate between *in-group* and *out-group* trustees in their requested back transfers. In the *trust game*, investors in the *in-group* treatment request 52.0 percent of the tripled amount sent ($r_i^*/3s_i$) compared to 42.5 in the *out-group* treatment (p=0.01). Similarly, in the *sanctioning game* investors requested 54.4 percent and 49.0 percent of the amount sent from *in-group* and *out-group* trustees, respectively (p=0.13, CI: -0.13 to 0.02). The difference in shares requested also results in differences in requested, expected, and realized profits (see Supplementary Table A7 and Supplementary Figure A5). While the share of profits for investors relative to trustees is higher in the *in-group* treatment in the *trust game*, the treatment gap is reduced in the *sanctioning game*.

Second, we turn to the frequency of punishment across treatments. Given that the trustee results demonstrate that there is a higher marginal return to imposing the fine in the *out-group* treatment, one might expect this to be reflected in the investor results. However, we find no treatment difference in the use of the fine by investors: 36.5 percent of investors in the *in-group* treatment and 38.1 percent of investors in the *out-group* treatment chose to impose the fine (p=0.89, CI: -0.21 to 0.24). Overall, the majority of those who imposed the fine in both treatments expected trustees to comply with their request. In the *out-group* treatment, 39.5 percent of those who imposed the fine expected trustees to be fined, compared to 28.1 percent of those who used the fine in the *in-group*, the difference is not statistically significant (p=0.14, CI: -0.27 to 0.04), although restricting the sample substantially reduces power.

Altogether, the investor results must be interpreted with caution. Investor choices reflect both profit-maximizing strategy and preferences. Even if subjects conjectured that imposing the fine would be more effective in increasing back transfers in the *out-group* treatment, they may care more about punishing perceived norm violators from their own group. Thus, though subjects in the *in-group* and

[29] We control for individual characteristics in Supplementary Table A8. Taken together, this indicates that social efficiency is not affected by group treatment, in either game.

[30] We do not find any difference in the amount sent between treatments in the *dictator game* (p=0.70, CI: -9.51 to 13.66), which might seem surprising given the treatment differences for other experimental outcomes. However, note that receivers in the *dictator game* were endowed. This means that while sending a higher amount benefits the receiver—and therefore we might expect to see higher allocations in the *in-group* treatment—higher allocations also increase inequality, which might lead to comparatively *lower* allocations in the *in-group* treatment (Bernhard et al., 2006; Bauer et al., 2014). Potentially, these effects counteract one another.

out-group impose the fine with similar frequencies, they may do so for different reasons.[31]

6 Discussion

The different effect that financial sanctions have in the *in-group* and *out-group* treatments could result from several underlying mechanisms. First, it is possible that there is simply more trustworthiness in the *in-group* to crowd out. If the behavioral effect of imposing the fine exhibits diminishing marginal returns, and baseline trustworthiness is lower in the *out-group* treatment, the effect of the fine would be comparatively smaller. However, this seems unlikely given the magnitude of the effect. In the *fine condition*, we see that the gap between *in-group* and *out-group* subjects virtually disappears, and we find evidence that sanctions actually increase *out-group* trustworthiness. If the treatment difference in response to the fine were driven by initial levels of trustworthiness alone, this would not be the case.

Second, differences in responses to the *fine* and *no-fine conditions* may be driven by different perceptions of fairness between the *in-group* and *out-group* treatments. If this were the case, we would expect a different reaction to the fine condition, even if the underlying preferences related to sanctioning were identical. If true, we should see similar effects of the *fine* and *no-fine conditions* on the *in-group* and *out-group* treatments after adjusting the definition of fairness. We test several cut-offs for fair requests, defined by the requested back transfer share, and do not observe any clear pattern in response to different thresholds in either treatment (See Supplementary Figure A4). This allows us to rule this out as the sole explanation.

Finally, we are left with the interpretation that group identity affects behavioral responses to financial sanctions in a more fundamental way. Our results suggest that state-dependent, other-regarding preferences differ according to group identity. This complements previous findings that group identity, and ethnic identity in particular, plays an important role in defining other-regarding preferences (Fershtman and Gneezy, 2001; Bernhard et al., 2006).

Thus far, we have considered only the reduced form of our treatments (i.e. *in-group* and *out-group* rather than Tajik-Hazara, Tajik-Tajik etc. . .). To help rule out the possibility that we capture a specific dynamic between Hazaras towards Tajik or visa versa, we split the sample by trustee's ethnicity

[31] This would be consistent with Bernhard et al. (2006) and of Goette et al. (2006). Alternatively, in line with Fearon and Laitin (1996), some investors in the *out-group* treatment might refrain from imposing the fine out of fear of damaging relations with the other ethnic group.

and estimate the same model as in Table 2. The results are very similar in both cases and consistent with the pooled results (see Supplementary Tables A10 and A11).[32] This is in contrast to Fershtman and Gneezy (2001) who use a trust game to measure ethnic discrimination. In their case, the results suggest the presence of stereotypes about a particular ethnic group, by both in-group and out-group members, whereas in our case, the results are evidence that preferences related to sanctioning are parochial (Bowles and Gintis, 2004; Bernhard et al., 2006).[33]

We acknowledge that the relationship between formal sanctioning mechanisms and ethnic identity may differ in other settings, and more research should be done in this area, to understand these cultural differences. Further, since we only study the behavior of males, due to cultural restrictions on interactions with women, we cannot comment on possible gender differences.

The generalizability of studies on behavioral responses to institutions is at the heart of our contribution. In addition to our findings that ethnic identity mediates the effect of material incentives on moral incentives, subjects in our experiment exhibit behavior that differs from previous studies using similar games (Fehr and Rockenbach, 2003; Fehr and List, 2004): the *fine condition* does not lower the average amount returned, on average. While this does not necessarily rule out the possibility that either the *fine* crowds out intrinsic motivation to cooperate, as we discuss in section 5.3, it does suggest that any such effect is substantially smaller than in previous studies.[34] Despite this, investors in both treatments choose to impose the fine much less frequently (37.2 percent) compared to Fehr and Rockenbach (2003) and Fehr and List (2004), who find that the fine was imposed in by 60-80 percent of investors.[35] Fehr and Rockenbach (2003) find that for unfair requests, imposing the fine led to much lower back-transfers (relative to the *no-fine condition*). We do not find evidence of a relationship between the fairness of requests and the decision to impose the fine. Our sample has very little experience interacting with formal institutions and we conjecture that this may underlie these differences in behavior. Experience and culture may help to shape preferences in relation to institutional settings, and thus the results from one particular context might not be universally applicable.

[32] Likewise, the difference-in-difference for investor profits between treatment and game holds independently for the sub-samples of Tajik (p=0.07) and Hazara investors (p=0.09).

[33] Bartoš (2016) runs dictator and third-party punishment games in Northern Afghanistan, with a sample that includes both Hazara and Tajik subjects, and finds no difference in average behavior towards in-group counterparts between the two ethnic groups. This provides further evidence that stereotypes about members of a particular ethnicity are unlikely to account for our results.

[34] In Supplementary Figure A6 we present these results side-by-side with our own.

[35] The fine was imposed by 67 percent of German students (Fehr and Rockenbach, 2003), and by 79 percent of students and 61 percent of CEOs in Costa Rica (Fehr and List, 2004).

7 Conclusion

When formal institutions are weak, societies rely on informal cooperation to a greater extent. Group affiliation thus often plays a significant role in shaping economic interactions. While previous studies have focused on the co-evolution of institutions, culture, and preferences (Boyd et al., 2003; Boyd and Richerson, 2009; Henrich et al., 2010), or have examined the long-run impacts of institutional setting on preferences and norms (Lowes et al., 2017), there is less evidence on the role of ethnicity. In this study, we ask how behavior related to a particular institution—an imperfect sanctioning regime—differs according to whether individuals share an ethnic identity. By employing an experiment with two ethnic groups in Northern Afghanistan—the Tajik and Hazara—we causally study how the effectiveness of imposing a financial sanction differs with group identity. We find that ethnic identity affects individual responses to a financial sanction: while the fine increase trustworthiness when applied across ethnic groups, there is no effect, on average, when it is applied by a co-ethnic. When no fine is available, trustworthiness is significantly higher in the *in-group* treatment, but imposing a conditional fine closes this gap.

We presented trustees with multiple experimentally manipulated scenarios. This allows us to causally test how trustworthiness is affected by the fine, independent of differences in investor behavior in each treatment. We are able to infer that the use of financial sanctions crowds out trustworthiness in the *in-group* treatment, but crowds in trustworthiness in the *out-group* treatment. The null effect of imposing the fine on trustworthiness, relative to the *trust game*, that we observe in the *in-group* treatment is consistent with the crowding out of intrinsic motivation to cooperate. Imposing the fine potentially increases cooperation in both treatments by changing incentives, but is counteracted by a behavioral effect in the *in-group* treatment only. While this suggests that the fine crowds-out trustworthiness less than in previous studies, the result is qualitatively similar to Fehr and Rockenbach (2003) and Fehr and List (2004).

Interestingly, when the fine is available but not imposed the effect is similar: trustworthiness in the *out-group* treatment increases, relative to the trust game, but there is a null effect for the *in-group* treatment. This suggests that it might be the ability to impose the fine that crowds out trustworthiness in the in-group treatment, rather than the investor's decision to use it. Here again, there may be to opposing forces at play. On the one hand, giving an *in-group* investor the ability to impose the fine could change the way that trustees perceive the relationship and activate a different set of (more selfish)

preferences. On the other hand, an investor's choice to voluntarily refrain from using the fine might send a positive signal about the investor's character.

Although investors in the *out-group* treatment could potentially achieve higher payoffs by sending more in the *sanctioning game* than in the *trust game*, this is not what we observe. Instead, investors who are paired with trustees from a different ethnic group capture a larger share of the surplus by requesting higher amounts in the *sanctioning game* than in the *trust game*. While we do not find a difference in the frequency of imposing the fine—as one might expect, given that it is more effective in the *out-group* treatment—this decision is influenced by both expected returns and preferences, and these motives may cancel each other out.

Our results complement previous findings from experiments (Habyarimana et al., 2007) and observational studies (Miguel and Gugerty, 2005), which show that ethnic diversity makes the provision of public goods more difficult, and are broadly consistent with previous work that demonstrates how willingness to punish differs across ethnic lines (Alexander and Christia, 2011; Bernhard et al., 2006). We find that while trustworthiness is lower when individuals come from different ethnic groups, introducing conditional financial sanctions eliminates the within-ethnic group advantage. Our study is unique in that we are able to show that this is not due simply to a difference in the willingness to punish.

This has important implications for understanding and predicting how institutional change will affect ethnically heterogeneous societies and helps to explain some previous observations about ethnicity, cooperation, and institutional setting. In line with the empirical findings of Easterly (2001), we show that formal institutions may moderate the adverse effects of ethnic heterogeneity (Fafchamps, 2000; Biggs et al., 2002). In terms of policy, our results provide tentative evidence that effort may be best spent strengthening formal mechanisms for instituting sanctions between, rather than within communities.

While there is an established theoretical explanation for why altruism towards members of one's own social group is stronger (Bowles and Gintis, 2004), it is less clear why the reaction to sanctioning is group specific. We conjecture that attitudes towards group identity and sanctioning might develop in communities where ethnicity is salient, as a strategy for avoiding inter-group conflict. Fearon and Laitin (1996) outline a theory that predicts higher costs of conflict between members of opposing groups, as inter-personal disputes can spiral into conflicts involving the entire groups. Given this, peace is maintained by avoiding such conflicts and by ignoring transgressions from members of the

other ethnic group, leaving an individual's co-ethnics to "police their own." In contexts similar to ours, such attitudes may have developed either through the conscious promotion of norms or through an evolutionary process of cultural transmission (Boyd and Richerson, 2009).

Ethnic diversity can be a mixed blessing: on one hand, it may lower trust and the ability to effectively cooperate, and increase chances of conflict, but on the other hand, complementarities between ethnic groups can lead to increased specialization. Well-functioning institutions can limit the negative impact of the former, allowing societies to take advantage of the positive aspects of diversity (Collier, 1999; Easterly, 2001; Alesina and La Ferrara, 2005). We offer an additional argument for how such institutions can have a greater benefit when applied across ethnic groups: while financial sanctions crowd out moral incentives to cooperate within ethnic groups, they actually reinforce trustworthiness between individuals from different ethnic groups.

References

Alesina, Alberto and Eliana La Ferrara, "Ethnic Diversity and Economic Performance," *Journal of Economic Literature*, 2005, *43* (3), 762–800.

Alexander, Marcus and Fotini Christia, "Context Modularity of Human Altruism," *Science*, 2011, *334* (6061), 1392–1394.

Ali, Merima, Odd Helge Fjeldstad, Boqian Jiang, and Abdulaziz B. Shifa, "Colonial Legacy, State-building and the Salience of Ethnicity in Sub-saharan Africa," *Economic Journal*, 2018, *forthcoming.*

Ariely, Dan, Anat Bracha, and Stephan Meier, "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *American Economic Review*, March 2009, *99* (1), 544–55.

Arrow, Kenneth J, "Gifts and Exchanges," *Philosophy & Public Affairs*, 1972, *1* (4), 343–362.

Barr, Abigail, "Trust and Expected Trustworthiness: Experimental Evidence from Zimbabwean Villages," *Economic Journal*, 2003, *113* (489), 614–630.

Bartoš, Vojtěch, "Seasonal Scarcity and Sharing Norms," 2016.

Bauer, Michal, Julie Chytilová, and Barbara Pertold-Gebicka, "Parental Background and Other-regarding Preferences in Children," *Experimental Economics*, 2014, *17* (1), 24–46.

_ , Nathan Fiala, and Ian Levely, "Trusting Former Rebels: An Experimental Approach to Understanding Reintegration After Civil War," *Economic Journal*, 2018, *128* (613), 1786–1819.

Beath, Andrew, Fotini Christia, Georgy Egorov, and Ruben Enikolopov, "Electoral Rules and Political Selection: Theory and Evidence from a Field Experiment in Afghanistan," *Review of Economic Studies*, 2016, *83* (3), 932–968.

Beekman, Gonne, Stephen L. Cheung, and Ian Levely, "The Effect of Conflict History on Cooperation within and between Groups: Evidence from a Laboratory Experiment," *Journal of Economic Psychology*, 2017, *63*, 168–183.

Berge, Lars Ivar Oppedal, Kjetil Bjorvatn, Simon Galle, Edward Miguel, Daniel N. Posner, Bertil Tungodden, and Kelly Zhang, "Ethnically Biased? Experimental Evidence from Kenya," *Journal of the European Economic Association*, 2020, *18* (1), 134–164.

Bernhard, Helen, Urs Fischbacher, and Ernst Fehr, "Parochial Altruism in Humans," *Nature*, 2006, *442* (7105), 912–915.

Besley, Timothy, Torsten Persson, and Daniel M Sturm, "Political Competition, Policy and Growth: Theory and Evidence from the United States," *Review of Economic Studies*, 2010, *77* (3), 1329–1352.

Biggs, Tyler, Mayank Raturi, and Pradeep Srivastava, "Ethnic Networks and Access to Credit: Evidence from the Manufacturing Sector in Kenya," *Journal of Economic Behavior & Organization*, 2002, *49* (4), 473–486.

Bowles, Samuel, "Policies Designed for Self-interested Citizens May Undermine "The Moral Sentiments": Evidence from Economic Experiments," *Science*, 2008, *320* (5883), 1605–1609.

_ **and Herbert Gintis**, "Persistent Parochialism: Trust and Exclusion in Ethnic Networks," *Journal of Economic Behavior & Organization*, 2004, *55*, 1–23.

_ **and Sandra Polania-Reyes**, "Economic Incentives and Social Preferences: Substitutes or Complements?," *Journal of Economic Literature*, 2012, *50* (2), 368–425.

Boyd, Robert and Peter J Richerson, "Culture and the Evolution of Human Cooperation.," *Philosophical Transactions of the Royal Society B: Biological Sciences*, nov 2009, *364* (1533), 3281–8.

_ **, Herbert Gintis, Samuel Bowles, and Peter J Richerson**, "The Evolution of Altruistic Punishment," *Proceedings of the National Academy of Sciences of the United States of America*, 2003, *100* (6), 3531–3535.

Buchan, Nancy R., Gianluca Grimalda, Rick Wilson, Marilynn Brewer, Enrique Fatas, and Margaret Foddy, "Globalization and Human Cooperation," *Proceedings of the National Academy of Sciences*, 2009, *106* (11), 4138–4142.

Charness, Gary, Luca Rigotti, and Aldo Rustichini, "Individual Behavior and Group Membership," *American Economic Review*, 2007, *97* (4), 1340–1352.

Chen, Yan and Sherry Xin Li, "Group Identity and Social Preferences," *American Economic Review*, 2009, *99* (1), 431–57.

Collier, Paul, "The Political Economy of Ethnicity," in Boris Pleskovic and Joseph E. Stiglitz, eds., *Annual World Bank Conference on Development Economics 1998*, Washington, D.C.: World Bank Publications, 1999, pp. 387–399.

Cox, James C., "How to Identify Trust and Reciprocity," *Games and Economic Behavior*, 2004, *46* (2), 260–281.

DHS, "Afghanistan Mortality Survey 2010," in "in," Calverton, Maryland, USA: Afghan Public Health Institute at the Ministry of Public Health - APHI/MoPH, Central Statistics Organization - CSO/Afghanistan, ICF Macro, Indian Institute of Health Management Research - IIHMR, & World Health Organization Regional Office for the Eastern Mediterranean - WHO/EMRO, 2011.

Easterly, William, "Can Institutions Resolve Ethnic Conflict?," *Economic Development and Cultural Change*, 2001, *49* (4), 687–706.

Fafchamps, Marcel, "Ethnicity and Credit in African Manufacturing," *Journal of Development Economics*, 2000, *61* (1), 205–235.

Falk, Armin and Christian Zehnder, "A City-wide Experiment on Trust Discrimination," *Journal of Public Economics*, 2013, *100*, 15–27.

_ **and Michael Kosfeld**, "The Hidden Costs of Control," *American Economic Review*, 2006, *96* (5), 1611–1630.

Fearon, James D and David D Laitin, "Explaining Interethnic Cooperation," *American Political Science Review*, 1996, *90* (4), 715–735.

Fehr, Ernst and Bettina Rockenbach, "Detrimental Effects of Sanctions on Human Altruism," *Nature*, 2003, *422* (6928), 137–140.

_ **and John A List**, "The Hidden Costs and Returns of Incentives - Trust and Trustworthiness among CEOs," *Journal of the European Economic Association*, 2004, *2* (5), 743–771.

_ **and Klaus M Schmidt**, "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 1999, *114* (3), 817–868.

Fershtman, Chaim and Uri Gneezy, "Discrimination in a Segmented Society: An Experimental Approach," *Quarterly Journal of Economics*, 2001, *116* (1), 351–377.

Frey, Bruno S. and Reto Jegen, "Motivation Crowding Theory," *Journal of Economic Surveys*, 2001, *15* (5), 589–611.

Gennaioli, Nicola and Ilia Rainer, "The Modern Impact of Precolonial Centralization in Africa," *Journal of Economic Growth*, 2007, *12* (3), 185–234.

Glaeser, Edward L., David I. Laibson, Jose A. Scheinkman, and Christine L. Soutter, "Measuring Trust*," *The Quarterly Journal of Economics*, 08 2000, *115* (3), 811–846.

Glennerster, Rachel, Edward Miguel, and Alexander D. Rothenberg, "Collective Action in Diverse Sierra Leone Communities," *Economic Journal*, 2013, *123* (568), 285–316.

Gneezy, Uri and Aldo Rustichini, "Pay Enough or Don't Pay at All," *Quarterly Journal of Economics*, 2000, *115* (3), 791–810.

Goette, Lorenz, David Huffman, and Stephan Meier, "The Impact of Social Ties on Group Interactions: Evidence from Minimal Groups and Randomly Assigned Real Groups," *American Economic Journal: Microeconomics*, 2012, *4* (1), 101–115.

_ , Dustin Huffman, and Stephan Meier, "The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups," *American Economic Review*, 2006, *96* (2), 212–216.

Habyarimana, James, Macartan Humphreys, Daniel N Posner, and Jeremy M Weinstein, "Why Does Ethnic Diversity Undermine Public Goods Provision?," *American Political Science Review*, 2007, *101* (04), 709–725.

Henrich, Joseph, Jean Ensminger, Richard McElreath, Abigail Barr, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwins Gwako, Natalie Henrich, Carolyn Lesorogol, Frank Marlowe, David Tracer, and John Ziker, "Markets, Religion, Community Size, and the Evolution of Fairness and Punishment.," *Science*, 2010, *327* (5972), 1480–4.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter, "Antisocial Punishment Across Societies.," *Science*, 2008, *319* (5868), 1362–1367.

Hjort, Jonas, "Ethnic Divisions and Production in Firms," *Quarterly Journal of Economics*, 2014, *129* (4), 1899–1946.

Houser, Daniel, Erte Xiao, Kevin McCabe, and Vernon Smith, "When Punishment Fails: Research on Sanctions, Intentions and Non-cooperation," *Games and Economic Behavior*, 2008, *62* (2), 509–532.

Johnson, Noel D and Alexandra A Mislin, "Trust Games: A Meta-analysis," *Journal of Economic Psychology*, 2011, *32* (5), 865–889.

Lowes, Sara, Nathan Nunn, James A Robinson, and Jonathan Weigel, "The Evolution of Culture and Institutions: Evidence from the Kuba Kingdon," *Econometrica*, 2017, *85* (4), 1065–1091.

Masella, Paolo, Stephan Meier, and Philipp Zahn, "Incentives and Group Identity," *Games and Economic Behavior*, 2014, *86*, 12–25.

Meier, Stephan, Lamar Pierce, Antonio Vaccaro, and Barbara La Carae, "Trust and In-Group Favoritism in a Culture of Crime," *Journal of Economic Behavior & Organization*, 2016, *132* (A), 78–92.

Michalopoulos, Stelios and Elias Papaioannou, "The Long-run Effects of the Scramble for Africa," *American Economic Review*, 2016, *106* (7), 1802–1848.

Miguel, Edward, "Tribe or Nation? Nation Building and Public Goods in Kenya versus Tanzania," *World Politics*, 2004, *56* (3), 327–362.

_ **and Mary Kay Gugerty**, "Ethnic Diversity, Social Sanctions, and Public Goods in Kenya," *Journal of Public Economics*, 2005, *89* (11-12), 2325–2368.

Monsutti, Alessandro, *War and Migration: Social Networks and Economic Strategies of the Hazaras of Afghanistan*, New York, NY: Routledge, 2005.

Rabin, Matthew, "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, 1993, *83* (5), 1281–1302.

Rehavi, Marit M. and Sonja B. Starr, "Racial Disparity in Federal Criminal Sentences," *Journal of Political Economy*, 2014, *122* (6), 1320–1354.

Riener, Gerhard and Simon Widerhold, "Team Building and Hidden Costs of Control," *Journal of Economic Behavior & Organization*, 2016, *123*, 1–18.

Schetter, Conrad, "Playing the Ethnic Card: On the Ethnicization of Afghan Politics," *Studies in Ethnicity and Nationalism*, 2016, *16* (3), 460–477.

Shayo, Moses and Asaf Zussman, "Judicial Ingroup Bias in the Shadow of Terrorism," *Quarterly Journal of Economics*, 2011, *126* (3), 1447–1484.

SIGAR, "Quarterly Report to the United States Congress," Technical Report, Special Inspector General for Afghanistan Reconstruction, Arlington, VA, USA 2017.

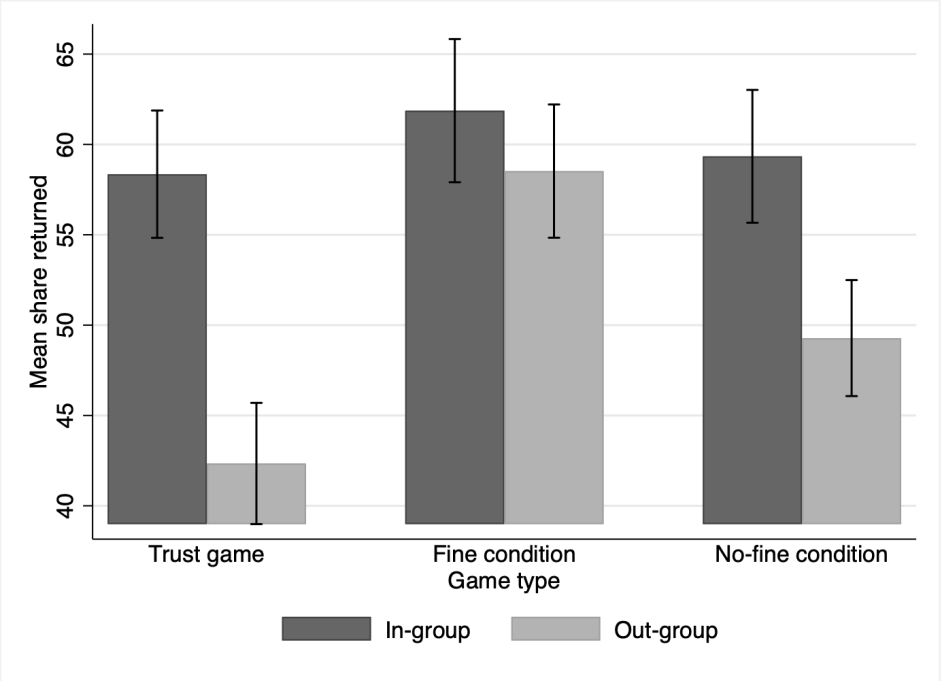
Tajfel, Henri, M G Billig, R P Bundy, and Claude Flament, "Social Categorization and Intergroup Behaviour," *European Journal of Social Psychology*, 1971, *1* (2), 149–178.

Titmuss, Richard M., *The Gift Relationship: From Blood Donations to Social Policy*, New York, NY: Pantheon Books, 1971.

Whitt, Sam and Rick K. Wilson, "The Dictator Game, Fairness and Ethnicity in Postwar Bosnia," *American Journal of Political Science*, 2007, *51* (3), 655–668.

Tables and Figures

Figure 1: Trustees' average share returned by game and treatment



Notes: Mean back transfers in the trust and sanctioning games by treatment for randomly assigned parameters only. The share returned is the percentage of the tripled amount received that the trustee transfers back to the investor. The Fine and no-fine conditions indicate whether the fine was imposed by the investor in the sanctioning game. Error bars represent 95 percent confidence intervals of the sample means, assuming a student distribution and treating each decision as an independent observation.

Table 1: Individual characteristics by role and treatment

Sample	*Investors*			*Trustees*		
	In-group (1)	*Out-group* (2)	Difference (1)-(2) (Wilcoxon p-value) (3)	*In-group* (4)	*Out-group* (5)	Difference (4)-(5) (Wilcoxon p-value) (6)
Share Hazara	0.59	0.62	-0.03	0.55	0.55	-0.00
	(0.49)	(0.49)	(0.65)	(0.50)	(0.50)	(0.96)
Age	40.96	40.98	-0.01	38.95	37.25	1.70
	(13.82)	(13.21)	(0.91)	(13.07)	(13.06)	(0.35)
Household members	7.69	7.42	0.28	8.07	8.61	-0.54
	(3.23)	(2.95)	(0.98)	(3.15)	(5.67)	(0.88)
Can read letter (d)	0.32	0.28	0.04	0.37	0.51	-0.14
	(0.47)	(0.45)	(0.55)	(0.48)	(0.50)	(0.07)
Years living in Mazar	12.91	11.63	1.28	18.80	16.81	1.99
	(12.90)	(13.18)	(0.43)	(15.41)	(16.07)	(0.30)
Income (Afs)	1691.83	1724.40	-32.58	1811.34	25011.76	-23200.42
	(4005.29)	(3206.39)	(0.65)	(3697.60)	(216789.10)	(0.11)
Written contract in the past (d)	0.08	0.05	0.03	0.06	0.04	0.03
	(0.27)	(0.22)	(0.43)	(0.24)	(0.19)	(0.44)
Others can be trusted (d)	0.55	0.75	-0.20	0.52	0.62	-0.10
	(0.50)	(0.44)	(0.00)	(0.50)	(0.49)	(0.20)
Others are fair (d)	0.39	0.38	0.01	0.27	0.34	-0.07
	(0.49)	(0.49)	(0.85)	(0.45)	(0.48)	(0.31)
Others are selfish (d)	0.78	0.61	0.17	0.71	0.68	0.02
	(0.42)	(0.49)	(0.01)	(0.46)	(0.47)	(0.73)
Observations	104	84		82	85	

Note: Means reported in Columns 1, 2, 4, and 5. Standard deviations in parentheses. Columns 3 and 6 report the difference in means between the in-group and the out-group treatment. P-values of a Wilcoxon rank-sum test are reported in parentheses in columns 3 and 6. (d) denotes a dummy variable. The high income for trustees' in the out-group treatment is driven by one individual who reported income of 2,000,000 Afs. Excluding this individual reduces the average income for this group to 1,500 Afs (SD=3,057.19). *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table 2: Effect of fine on share returned in trust and sanctioning games across treatments

Sample	*All*	*In-group*	*Out-group*	*In-group*	*Out-group*
Dependent variable	Share returned				
	(1)	(2)	(3)	(4)	(5)
In-group	15.38***				
	(4.09)				
Fine condition	13.85***	-0.01	13.92***	-12.46***	2.49
	(5.02)	(1.68)	(5.33)	(4.75)	(6.27)
In-group x Fine condition	-13.34**				
	(5.23)				
No-fine condition	5.36***	-0.44	5.45**	-13.43**	11.76**
	(1.99)	(1.71)	(2.27)	(5.58)	(4.69)
In-group x No-fine condition	-5.65**				
	(2.47)				
Share requested	0.39***	0.50***	0.29***	0.39***	0.27***
	(0.06)	(0.09)	(0.05)	(0.09)	(0.03)
Fine x Share requested				0.20***	0.18
				(0.08)	(0.12)
No-fine x Share requested				0.21***	-0.10
				(0.07)	(0.08)
Sent	-0.16***	-0.13***	-0.19***	-0.12***	-0.19**
	(0.05)	(0.04)	(0.07)	(0.04)	(0.07)
Constant	22.66***	29.18***	32.47***	34.95***	33.44***
	(6.73)	(7.54)	(8.83)	(8.07)	(10.36)
Observations	999	491	508	491	508
Number of subjects	167	82	85	82	85
	F-test H_0 : Fine equals no-fine				
In-group p-value	0.55	0.76		0.83	
Out-group p-value	0.05		0.05		0.06

Note: Individual level random effects regression coefficients. Standard errors in parentheses (clustering at session level). Randomly assigned parameters only. The Fine and no-fine conditions indicate whether the fine was imposed by the investor in the sanctioning game. The excluded category is the trust game. In each regression we control for trustee's ethnicity, age, number of household members, a dummy for literacy, years spent living continuously in Mazar-e-Sharif, log of income (Afs), a dummy for whether the individual had ever signed a contract and an index of perceptions of trust and fairness towards others (3 questions). The F-test compares the fine and no-fine condition coefficients. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table 3: Fines, fairness, and behavioral effects

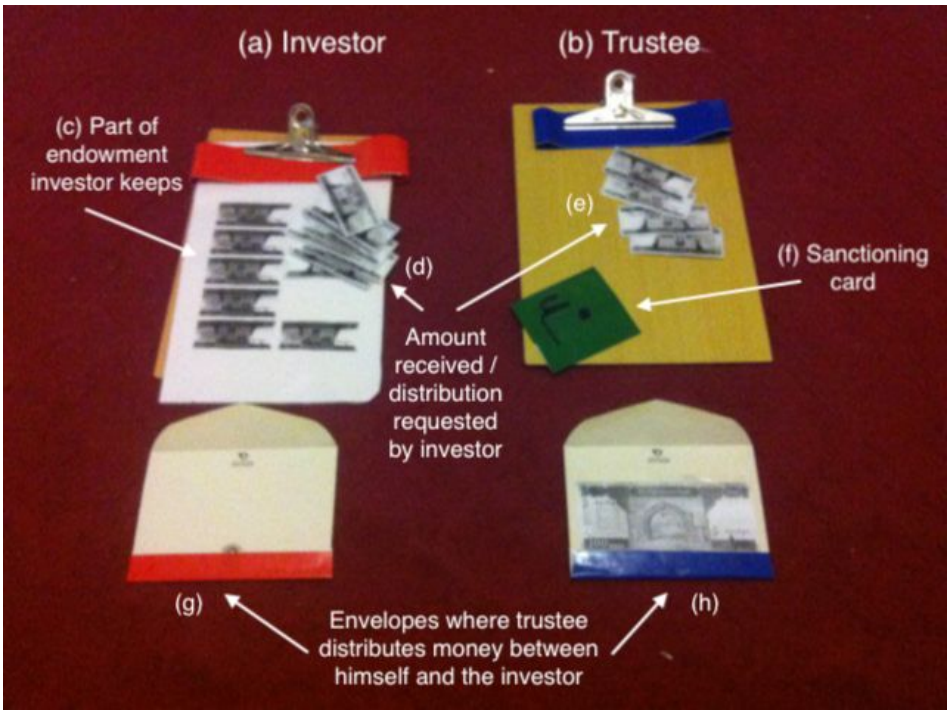
Panel A: Fairness						
Sample	*Fair request*			*Unfair requests*		
	All	*In-group*	*Out-group*	*All*	*In-group*	*Out-group*
Dependent variable	Share returned					
	(1)	(2)	(3)	(4)	(5)	(6)
In-group	12.95***			18.29***		
	(3.55)			(7.00)		
Fine condition	9.45**	-3.76*	9.67**	19.29***	4.02	17.95**
	(4.11)	(2.16)	(4.55)	(7.37)	(3.13)	(7.82)
In-group x Fine condition	-12.87***			-15.46*		
	(4.58)			(8.19)		
No-fine condition	5.99**	-4.28*	6.30**	5.31	3.95*	4.02
	(2.55)	(2.47)	(2.65)	(4.27)	(2.32)	(4.60)
In-group x No-fine condition	-10.32***			-2.02		
	(3.41)			(4.63)		
Constant	26.43***	29.38***	34.02***	32.75**	42.76**	45.00***
	(5.95)	(10.00)	(8.21)	(13.72)	(21.05)	(13.35)
Observations	547	265	282	452	226	226
Number of id	157	77	80	127	62	65
Panel B: Motivations						
Sample	*All*	*In-group*	*Out-group*	*All*	*In-group*	*Out-group*
Dependent variable	Returned more than requested (d)			Returned less than requested		
	(1)	(2)	(3)	(4)	(5)	(6)
In-group	0.19***			-0.17***		
	(0.05)			(0.06)		
Fine condition	0.12***	-0.04	0.10***	-0.32***	-0.14**	-0.29***
	(0.04)	(0.04)	(0.04)	(0.05)	(0.06)	(0.04)
In-group x Fine condition	-0.14**			0.17**		
	(0.06)			(0.08)		
No-fine condition	0.07*	-0.00	0.06*	0.01	-0.10*	0.01
	(0.04)	(0.05)	(0.03)	(0.05)	(0.05)	(0.05)
In-group x No-fine condition	-0.07			-0.11		
	(0.06)			(0.07)		
Observations	999	491	508	999	491	508

Note: Panel A presents individual level random effects regression coefficients. Panel B presents marginal effects reported for individual level random effects probit regressions. Standard errors in parentheses (clustering at session level). Randomly assigned parameters only. The Fine and no-fine conditions indicate whether the fine was imposed by the investor in the sanctioning game. The excluded category is the trust game. In each regression we control for amount sent, share requested, trustee's ethnicity, age, number of household members, a dummy for literacy, years spent living continuously in Mazar-e-Sharif, log of income (Afs), a dummy for whether the individual had ever signed a contract and an index of perceptions of trust and fairness towards others (3 questions). *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

A Supplementary Tables and Figures (for online publication only)

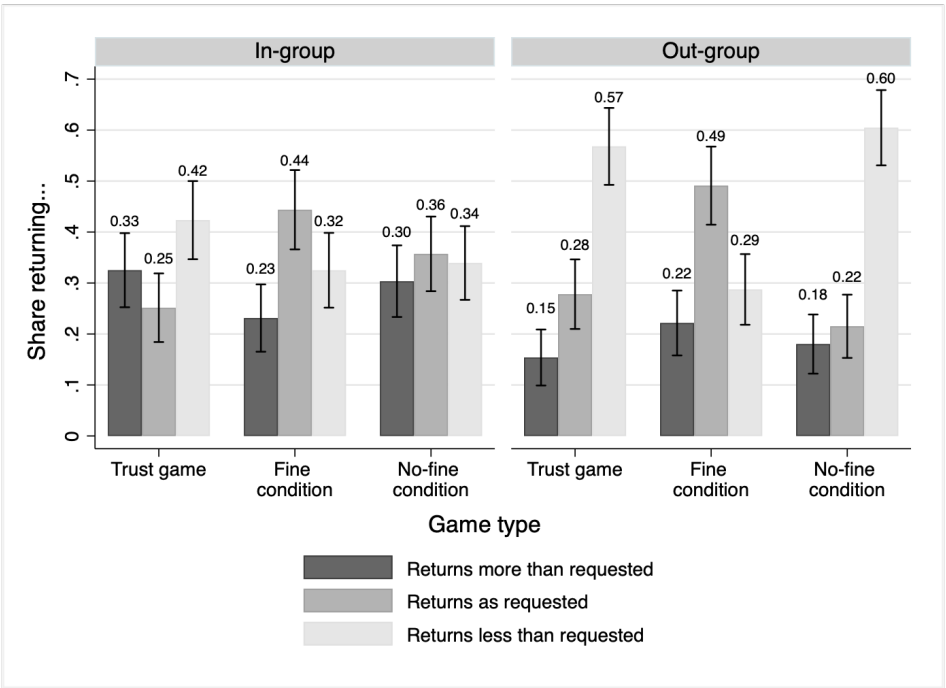
Figure A1: Participants in an individual session in a pop-up field laboratory

Figure A2: Trustee's decision-making environment with visual aids



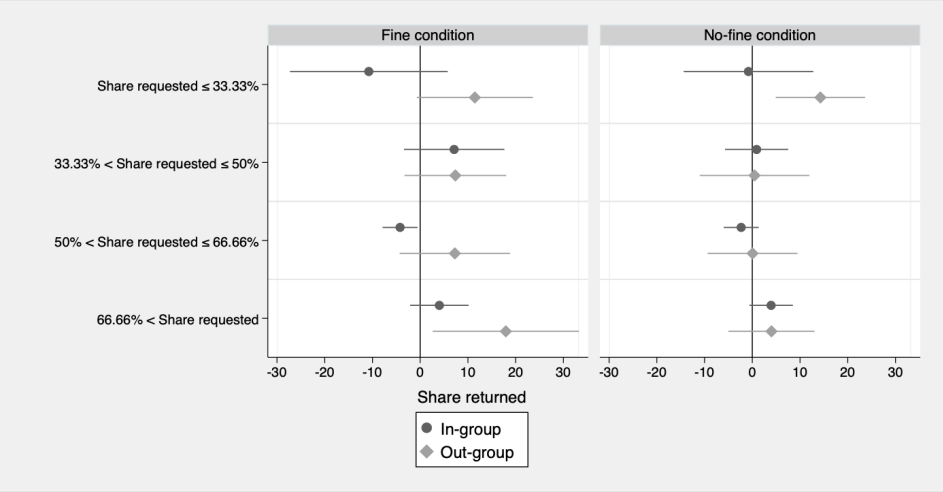
Note: The figure shows the decision-making environment for a trustee in the sanctioning game, which was designed to be easy for illiterate subjects to understand. The investor's choices were communicated using visual aids. In this example, the trustee faced a choice in which the investor had decided to i) send $s_i = 40$ Afs, ii) request to $r_i^* = 80$ Afs back, and iii) to apply the sanction ($p_i = 1$). The red clipboard (a) represents the investor, and the blue clipboard (b) represents the trustee. The card with 6 x 10-Afs banknotes (c) represents the part of the original endowment that the investor kept for himself ($\omega - s_i = 60$ Afs). The amount sent was tripled, and the trustee received 120 Afs. The 12 x 10-Afs banknotes loose banknotes on both clipboards represent this amount, and their placement on the clipboards represents the distribution requested by the investor: 4 x 10-Afs banknotes for the trustee (d) and 8 x 10-Afs for the investor (e). The green card (f) shows that the investor decided to apply the sanction. If the card were turned to the other side (yellow), this would indicate that the sanction was available but not applied. The trustee were asked to allocate the loose banknotes to the empty envelopes below the clipboards as he pleased. Money placed in the red envelope (g) would be returned to the investor, and money put into the blue envelope (h) would go to the trustee. The 100 Afs banknote attached to the blue envelope stands for trustee's endowment, which he always receives. After the trustee made his choice, an experimenter collected the envelopes and privately recorded the data. The experimenter then presented the subject with a new choice, using a different set of parameters following the structure in Table A1. Trust game choices were presented in the same way, but without the sanctioning card (f).

Figure A3: Trustees' decisions to meet investors' requests by game and treatment.



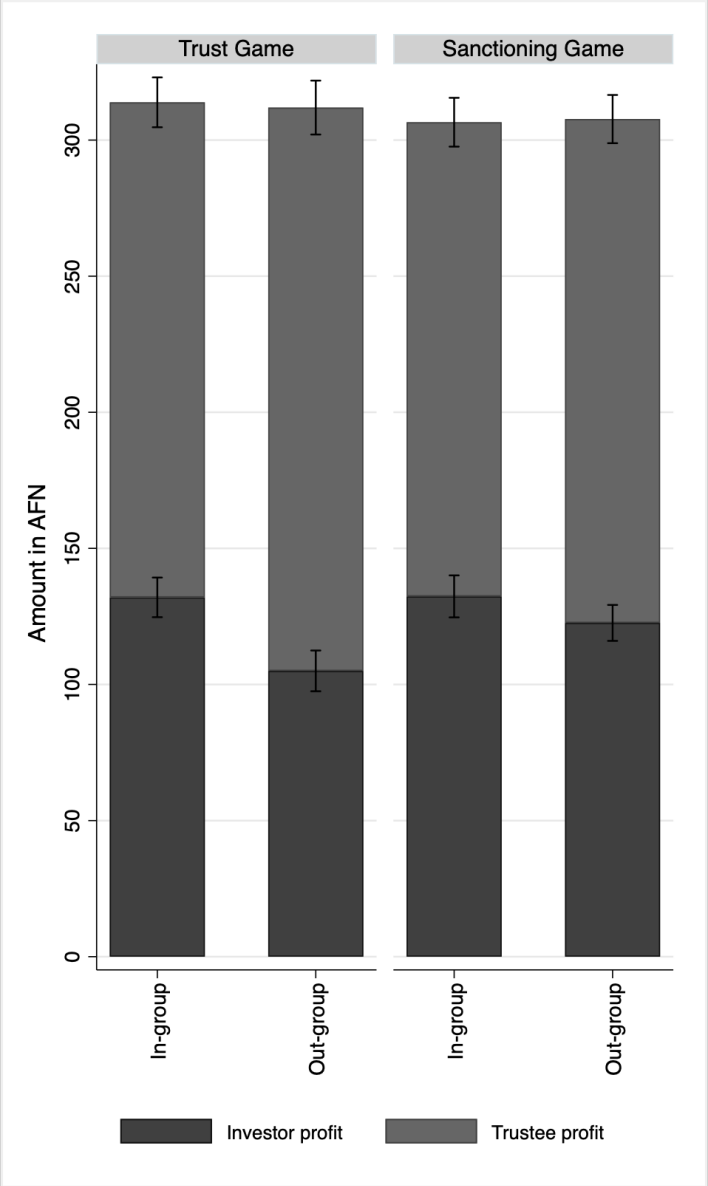
Notes: Mean back transfers in the trust and sanctioning games by treatment for randomly assigned parameters only. Returns more/as/less than requested are dummy variables representing choices in which the amount returned by the trustee exceeds/equals/is lower than the amount requested back by the investor. Error bars represent 95 percent confidence intervals.

Figure A4: Coefficient plot: share returned regressed on fine and no-fine conditions, by different ranges of share requested



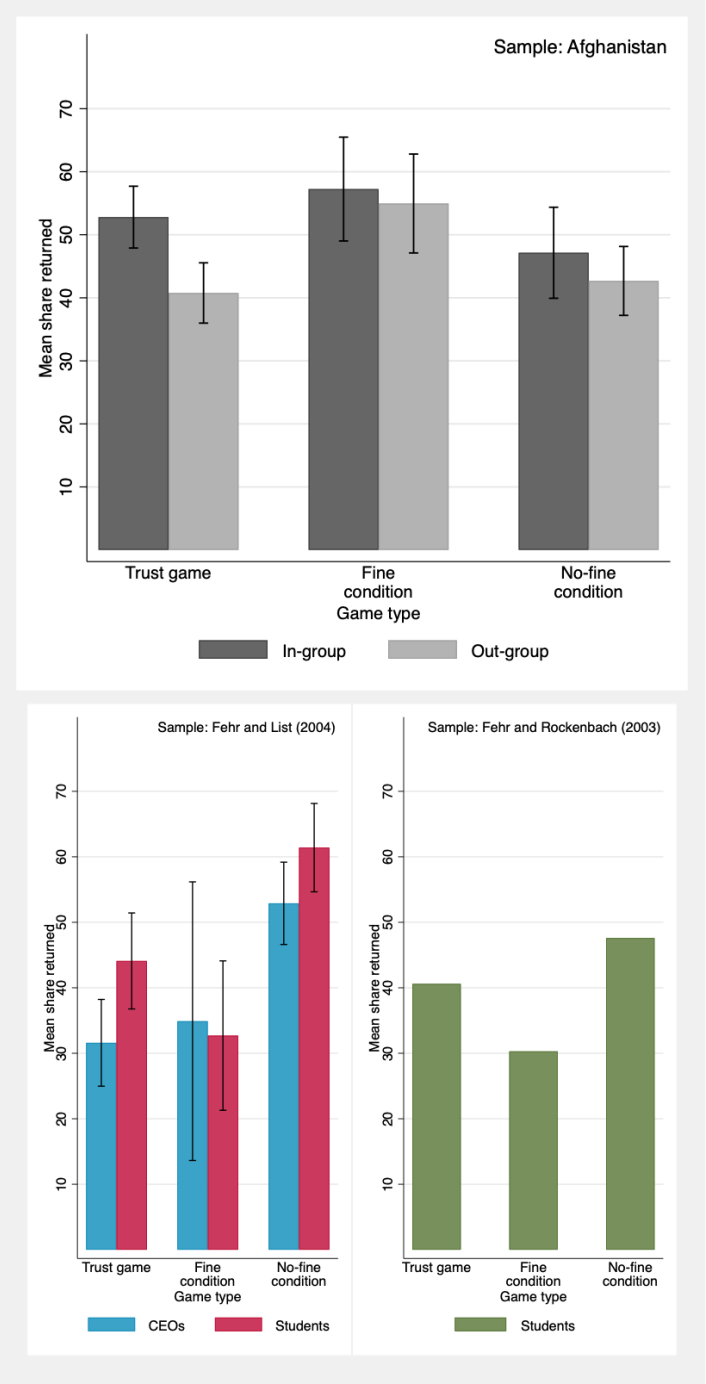
Notes: Regression results from model estimated in column 1 of Table 2. The dependent variable is the share returned in the fine and no-fine conditions. Each mark represents the coefficient for the given treatment, for a range of requests (implying various thresholds for what might be considered fair requests). Error bars represent 95 percent confidence intervals.

Figure A5: Investor and trustee profits by game and treatment



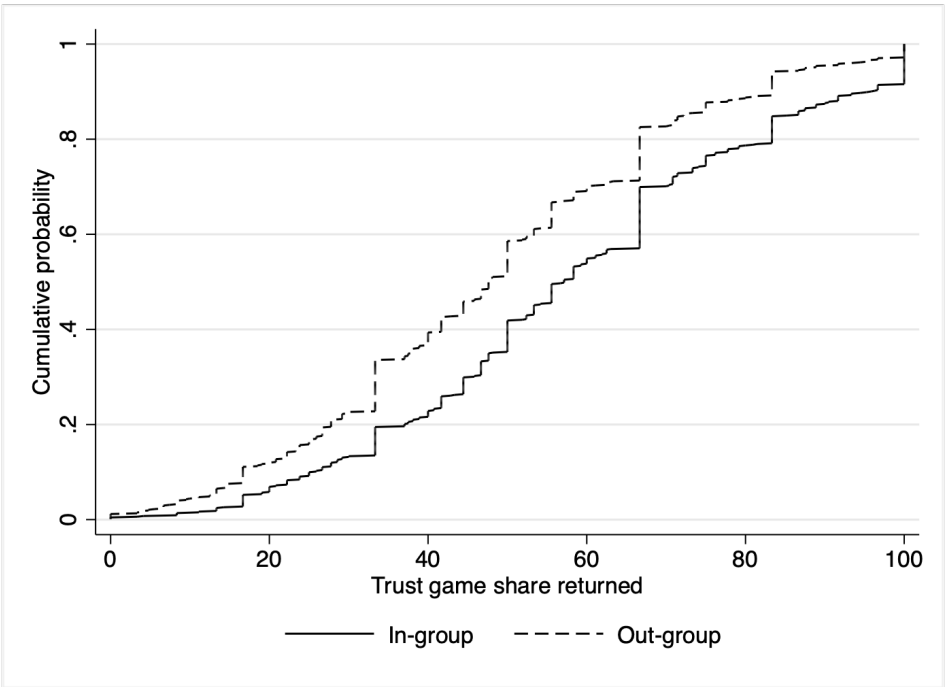
Notes: Mean cumulative profits in Afs. The investor profits are calculated as a combination of an investor's initial endowment minus the amount sent plus the average amount returned by the trustee, for all decisions in that game and group treatment in which trustees were presented with the parameters matching the investor's choice (i.e. amount sent, requested back, and whether the fine was imposed). Trustee profits are generated analogously. Error bars represent 95 percent confidence intervals.

Figure A6: Comparison of results with previous studies using similar designs



Notes: Mean back-transfers in the trust and sanctioning games generated using actual investor choices only (i.e. excluding choices made for randomly assigned parameters). Share returned is the percentage of the tripled amount received that the trustee sends back to the investor. Error bars represent 95 percent confidence intervals. Results in Fehr and Rockenbach (2003) do not allow us to calculate the confidence intervals.

Figure A7: Cumulative distribution functions of share returned by group treatment



Notes: Share returned in the trust game by group treatment for randomly assigned parameters only. The share returned is the percentage of the tripled amount received that the trustee transfers back to the investor.

Table A1: Summary of trustees decisions, and randomly assigned parameters by game, condition, and treatment

Panel A: Trustee decisions			
Decision number	Source of parameters	Independent of treatment & included in analysis	Payoff relevant
Trust game			
T1-T2	Randomly assigned	X	
T3	Matched partner		X
Sanctioning game			
S1-S4	Randomly assigned	X	
S5	Matched partner		X

Panel B: Randomly assigned parameters by treatment			
	In-group (1)	Out-group (2)	Difference (1)-(2) (Willcoxon p-value) (3)
Trust game			
Amount sent (s_i)	61.10	61.54	-0.43
	(26.62)	(26.75)	(0.85)
Share requested (r_i^*)	61.45	60.10	1.35
	(26.49)	(30.12)	(0.69)
Observations	163	169	
Fine condition			
Amount sent (s_i)	55.94	55.15	-0.79
	(25.58)	(25.27)	(0.79)
Share requested (r_i^*)	67.51	63.48	4.03
	(22.90)	(24.30)	(0.23)
Observations	160	167	
No-fine condition			
Amount sent (s_i)	58.04	57.09	0.94
	(25.58)	(25.68)	(0.77)
Share requested (r_i^*)	63.37	62.84	0.53
	(22.90)	(24.26)	(0.81)
Observations	168	172	

Note: The order of the games was randomized. Panel A reports the set of decisions in each game presented to trustees. In decisions S1-S4, fine was randomly assigned to exactly two out of the four decisions for each trustee. Decisions T3 and S5, in which the parameters were not independent of treatment are excluded from the main analysis. This allows us to measure causal effects of fines by treatment, independent of the other parameters. Panel B reports means of the parameters. Standard deviations in parentheses. Column 3 reports the difference in means between the in-group and the out-group treatment. P-values of a Wilcoxon rank-sum test are reported in parentheses in column 3.

Table A2: Sampling strategy and selection of data for analysis

COMMUNITY SCREENING	Telephone interviews with community leaders determining ethnic homogeneity and pre-approval of experiments in personal meetings with community leaders.	
RECRUITING PARTICIPANTS		
Step	**Both investor and trustee (roles yet to be determined)**	
Communities selected	7 Tajik communities; 6 Hazara communities (high degree of ethnic homogeneity as reported by community leaders in a community screening interview)	
Potential participants screened	Random walk method within communities; interviews with household heads.	
Participants selected	Individual screening survey: all Tajik or Hazara (depending on community from which selected) married males aged between 18-60 years with at least one child invited for experiments.	
PARTICIPANTS REGISTERING TO THE SESSION		
Step	**Investor**	**Trustee**
Participants registered in sessions	*In-group:* 54 Tajik (4 sessions), 70 Hazara (4 sessions) *Out-group:* 36 Tajik (2 sessions), 53 Hazara (3 sessions)	*In-group:* 59 Tajik (4 sessions), 49 Hazara (3 sessions) *Out-group:* 57 Tajiks (4 sessions), 56 Hazaras (4 sessions)
	Total: 213 IDs/observations	*Total:* 221 IDs (corresponds to 1768 observations)
OBSERVATIONS USED IN THE MAIN ANALYSIS		
Step	**Investor**	**Trustee**
Observations used in the analysis	*In-group:* 43 Tajik (4 sessions), 61 Hazara (4 sessions) *Out-group:* 32 Tajik (2 sessions), 52 Hazara (3 sessions)	*In-group:* 37 Tajik (4 sessions), 45 Hazara (3 sessions) *Out-group:* 38 Tajiks (4 sessions), 47 Hazaras (4 sessions)
	Total: 188 IDs/observations used in main analysis	*Total:* 167 IDs (1328 observations, out of which 999 randomly assigned parameters observations used in main analysis)
REASONS FOR DROPPING OBSERVATIONS		
	Investor	**Trustee**
	25 observations dropped because of inconsistency in ethnicity reported in screening and post-experimental surveys	32 IDs (256 observations, 192 randomly assigned parameters observations) dropped because of inconsistency in ethnicity reported in screening and post-experimental surveys
		12 IDs (96 observations, 72 randomly assigned parameters observations) dropped because of incorrect experiment procedure or completely missing games data
		10 IDs (80 observations, 60 randomly assigned parameters observations) dropped because of missing survey data used for controls in main regressions
		For 7 IDs we drop 8 observations due to partially missing games data (out of which we drop 3 randomly assigned parameters observations)

Table A3: Trustee behavior in games: aggregate and by treatment, strategy method allocations

Sample	*Total* (1)	*In-group* (2)	*Out-group* (3)	Difference (2)-(3) Sommer's D (p-value) [Clustered SE] (4)
Trust game				
Share returned	50.20	58.35	42.34	16.01***
	(23.79)	(22.78)	(22.09)	0.40 (0.00)
				[4.54]
Observations	332	163	169	
Sanctioning game				
Fine condition				
Share returned	60.16	61.87	58.52	3.35
	(24.77)	(25.38)	(24.14)	0.08 (0.46)
				[4.32]
Observations	327	160	167	
No-fine condition				
Share returned	54.25	59.34	49.28	10.06***
	(23.28)	(24.14)	(21.33)	0.25 (0.00)
				[2.87]
Observations	340	168	172	

Note: The Fine and no-fine conditions indicate whether the fine was imposed by the investor in the sanctioning game. Means reported in Columns 1-3. Standard deviations in parentheses. Randomly assigned parameters only. Column 4 reports 1) the difference in means between the in-group and the out-group treatment, 2) Somer's D with variance clustered at the session level and p-values in parentheses, and 3) standard errors obtained using individual level random effects regressions with standard errors clustered at session level. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table A4: Effect of fine on share returned in trust and sanctioning games across treatments: robustness checks

Sample	*Full sample*							*Early choices*	*Actual investor choices*
Dependent variable	Share returned								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
In-group	15.38***	15.13***	15.61***	15.41***		15.37***	15.55***	14.87***	10.01**
	(4.09)	(4.05)	(3.66)	(5.97)		(4.19)	(4.43)	(4.06)	(4.81)
Fine cond.	13.85***	13.83***	13.86***	13.96*	13.78**	14.73***	16.07***	13.14**	12.31***
	(5.02)	(5.02)	(5.03)	(7.42)	(4.99)	(5.15)	(5.96)	(5.41)	(4.14)
In-group x Fine cond.	-13.34**	-13.35**	-13.35**	-13.44*	-13.30**	-13.56**	-12.49**	-17.49***	-11.01**
	(5.23)	(5.22)	(5.24)	(7.30)	(5.19)	(5.40)	(6.03)	(6.33)	(4.51)
No-fine cond.	5.36***	5.38***	5.37***	5.29**	5.38**	6.06***	7.14**	5.70*	0.91
	(1.99)	(1.99)	(1.99)	(2.37)	(1.94)	(2.08)	(3.11)	(2.93)	(3.45)
In-group x No-fine cond.	-5.65**	-5.63**	-5.65**	-5.47*	-5.73**	-5.84**	-6.07*	-10.43*	-5.47
	(2.47)	(2.47)	(2.46)	(3.10)	(2.42)	(2.57)	(3.54)	(5.76)	(4.81)
Sent	-0.16***	-0.16***	-0.15***	-0.15***	-0.16***			-0.28***	-0.04
	(0.05)	(0.05)	(0.05)	(0.06)	(0.05)			(0.06)	(0.03)
Share requested	0.39***	0.39***	0.39***	0.37***	0.40***	0.40***		0.36***	0.48***
	(0.06)	(0.06)	(0.06)	(0.00)	(0.05)	(0.06)		(0.06)	(0.05)
Enumerator "H"		-7.41***							
		(2.46)							
Sanctioning game first			5.79*						
			(2.98)						
Constant	22.66***	25.96***	16.95**	23.19***	36.00***	13.35**	42.85***	37.60***	10.15*
	(6.73)	(7.25)	(7.71)	(8.73)	(4.84)	(6.48)	(5.68)	(8.04)	(5.60)
Observations	999	999	999	999	999	999	999	332	329
Number of id	167	167	167		167	167	167	167	166
R-squared				0.27	0.29				
	F-test: H_0 : Fine equals no-fine								
In-group p-value	0.55	0.58	0.55	0.95	0.54	0.49	0.11	0.95	0.17
Out-group p-value	0.05	0.05	0.05	0.27	0.07	0.04	0.03	0.14	0.00

Note: Individual level random effects regression coefficients except for columns 4 and 5 in which we use a linear regressions with multi-way bootstrapped clustered standard errors and an individual level fixed effect, respectively. Standard errors in parentheses (clustering at session level, with the exception of column 4). Randomly assigned parameters only with the exception of column 9 in which actual investor choices are used. Column 8 restricts the sample to the first choices in either the trust or the sanctioning games where we hypothesise that the investor intentions are be strongest. In each regression, with the exception of column 5, we control for trustee's ethnicity, age, number of household members, a dummy for literacy, years spent living continuously in Mazar-e-Sharif, log of income (Afs), a dummy for whether the individual had ever signed a contract and an index of perceptions of trust and fairness towards others (3 questions). Enumerator "H" is a dummy for a Hazara enumerator, our second enumerator was a Tajik. The F-test compares the fine and no-fine condition coefficients. *** Significant at the 1 percent level, ** Significant at the 5 percent level, * Significant at the 10 percent level.

Table A5: Effect of fine on share returned in trust and sanctioning games across treatments: quantile regressions

Sample	*Full sample*			*Fair requests*		*Unfair requests*	
	All	*In-group*	*Out-group*	*In-group*	*Out-group*	*In-group*	*Out-group*
Dependent variable	Share returned						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: 25th quantile							
In-group	14.89***						
	(3.14)						
Fine condition	15.94***	2.57	14.63***	-4.50	10.34***	10.89*	20.12***
	(3.77)	(3.66)	(3.91)	(3.26)	(3.94)	(5.64)	(7.53)
In-group x Fine condition	-15.93***						
	(5.46)						
No-fine condition	4.87**	0.64	6.57***	-4.60	4.26	10.89**	1.53
	(2.20)	(2.99)	(2.39)	(2.80)	(2.70)	(4.92)	(4.80)
In-group x No-fine condition	-5.12						
	(3.71)						
Constant	3.99	3.95	20.59***	-6.59	16.23**	43.95	35.82*
	(5.24)	(10.16)	(7.13)	(10.98)	(7.66)	(26.87)	(19.94)
Observations	999	491	508	265	282	226	226
Panel B: 50th quantile							
In-group	16.21***						
	(2.42)						
Fine condition	15.84***	2.21	15.29***	-2.57	9.75***	12.62***	27.69***
	(2.13)	(1.86)	(2.68)	(1.80)	(2.43)	(3.48)	(3.23)
In-group * Fine condition	-14.38***						
	(3.16)						
No-fine condition	1.13	0.56	3.00	-4.83**	1.97	8.43***	0.79
	(2.32)	(2.03)	(2.67)	(1.98)	(2.62)	(2.87)	(4.41)
In-group * No-fine condition	-3.01						
	(3.46)						
Constant	20.20***	18.09***	27.55***	31.82***	15.79**	18.42	51.05***
	(4.18)	(5.85)	(7.11)	(8.09)	(7.84)	(19.90)	(16.96)
Observations	999	491	508	265	282	226	226
Panel C: 75th quantile							
In-group	11.92***						
	(2.99)						
Fine condition	9.76***	-3.08*	12.06***	-5.52*	7.75**	0.90	18.99***
	(2.25)	(1.85)	(2.38)	(3.02)	(3.47)	(2.36)	(5.22)
In-group * Fine condition	-9.78**						
	(3.95)						
No-fine condition	3.12	0.00	5.23	-5.35*	1.33	2.71	9.00
	(2.89)	(2.24)	(3.50)	(2.98)	(3.51)	(2.69)	(6.47)
In-group * No-fine condition	-2.93						
	(4.15)						
Constant	23.94***	34.85***	32.87***	50.80***	54.44***	9.76	18.71
	(5.29)	(6.89)	(7.33)	(10.17)	(9.67)	(7.02)	(20.26)
Observations	999	491	508	265	282	226	226

Note: Quantile regression coefficients. Robust standard errors in parentheses. Randomly assigned parameters only. Panels A, B, and C present results of quantile regressions on the 25th, 50th, and 75th quantiles, respectively. In each regression we control for investor's amount sent and share returned, and trustee's ethnicity, age, number of household members, a dummy for literacy, years spent living continuously in Mazar-e-Sharif, log of income (Afs), a dummy for whether the individual had ever signed a contract and an index of perceptions of trust and fairness towards others (3 questions). *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table A6: Effect of fine on shares returned in trust and sanctioning games across treatments: average amount returned by condition, parameters, and treatment

Sample	*Full sample*			*Fair requests*		*Unfair requests*	
	All	*In-group*	*Out-group*	*In-group*	*Out-group*	*In-group*	*Out-group*
Dependent variable	Share returned						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A							
In-group	14.65***						
	(3.27)						
Fine condition	13.81***	1.90	13.64***	-1.09	12.47***	1.19	17.12***
	(3.21)	(3.73)	(3.21)	(4.13)	(3.74)	(4.83)	(5.03)
In-group x Fine condition	-11.98**						
	(4.94)						
No-fine condition	6.91**	1.02	6.57**	1.05	8.65***	-0.10	3.06
	(2.78)	(3.47)	(2.71)	(4.06)	(3.01)	(5.17)	(5.25)
In-group x No-fine condition	-5.50						
	(4.43)						
Constant	35.47***	53.58***	37.03***	33.76**	28.38***	68.53***	36.26**
	(6.00)	(13.74)	(6.55)	(16.63)	(8.11)	(25.09)	(13.78)
Observations	431	203	228	137	155	66	73
R-squared	0.11	0.05	0.13	0.12	0.14	0.19	0.30
Panel B							
In-group	12.02**						
	(5.56)						
Fine condition	9.77*	-1.80	8.62	-6.94	8.58	7.67	16.39**
	(5.77)	(5.97)	(5.62)	(7.52)	(7.36)	(7.19)	(6.18)
In-group x Fine condition	-12.26						
	(8.28)						
No-fine condition	5.68	2.30	4.84	1.20	7.79	7.17	2.63
	(4.63)	(5.26)	(4.56)	(6.57)	(5.51)	(7.86)	(5.08)
In-group x No-fine condition	-2.82						
	(7.10)						
Constant	43.81***	51.02**	61.14***	38.81	52.39***	41.92	-11.50
	(14.51)	(24.16)	(14.04)	(35.15)	(15.30)	(53.55)	(28.92)
Observations	129	63	66	39	42	24	24
R-squared	0.16	0.19	0.21	0.33	0.21	0.38	0.64

Note: OLS coefficients. Robust standard errors in parentheses. Randomly assigned parameters only. To show that differences in parameters assigned across games do not affect the results, we take the average share returned by trustees for every given combination of amount sent and requested back transfer, by ethnic treatment, game, and fine condition. In panel A the share returned represents an average share returned by the trustees responding to a given combination of amount sent and amount requested back by ethnic treatment, game, and fine choice in the sanctioning game. In panel B, we further restrict the sample to those observations used in Panel A for which we observe the combination of amount sent and amount requested back for both games and both fine choices in the sanctioning game. In each regression we control for trustee's ethnicity, age, number of household members, a dummy for literacy, years spent living continuously in Mazar-e-Sharif, log of income (Afs), a dummy for whether the individual had ever signed a contract and an index of perceptions of trust and fairness towards others (3 questions). *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table A7: Investor summary statistics: trust, sanctioning, and dictator games

Sample	Total (1)	In-group (2)	Out-group (3)	Difference (2)-(3) Sommer's D (p-value) [Clustered SE] (4)
Tust game				
Amount sent	56.76	57.21	56.19	1.02
	(22.46)	(23.67)	(20.99)	0.03 (0.85)
				[4.69]
Share requested back	0.48	0.52	0.42	0.10**
	(0.22)	(0.22)	(0.20)	0.25 (0.01)
				[0.03]
Requested profit	122.71	130.67	112.86	17.81**
	(42.09)	(42.86)	(39.17)	0.25 (0.02)
				[5.58]
Expected profit	117.29	122.12	111.31	10.81
	(38.90)	(39.13)	(37.98)	0.15 (0.25)
				[6.92]
Realized profit	119.98	131.69	106.30	25.39***
	(34.19)	(34.05)	(29.06)	0.44 (0.00)
				[6.54]
Sanctioning game				
Fine imposed	0.37	0.37	0.38	-0.02
	(0.48)	(0.48)	(0.49)	-0.02 (0.89)
				[0.10]
Amount sent	56.28	55.96	56.67	-0.71
	(20.91)	(21.97)	(19.66)	-0.03 (0.80)
				[3.86]
Share requested back	0.52	0.54	0.49	0.05
	(0.21)	(0.22)	(0.20)	0.14 (0.13)
				[0.03]
Requested profit	130.05	134.81	124.17	10.64
	(43.71)	(48.45)	(36.61)	0.07 (0.48)
				[6.78]
Expected profit	122.62	126.80	117.50	9.30
	(34.95)	(37.16)	(31.50)	0.13 (0.17)
				[4.94]
Realized profit[a]	128.73	133.73	122.58	11.15**
	(36.03)	(41.67)	(26.59)	0.16 (0.04)
				[4.19]
Dictator game				
Amount sent	45.16	44.23	46.31	-2.08
	(25.57)	(26.50)	(24.48)	-0.05 (0.70)
				[5.32]
Observations	188	104	84	

Note: Means reported in Columns 1 and 2. Standard deviations in parentheses. Column 3 reports 1) the difference in means between the in-group and the out-group treatment, 2) Somer's D with variance clustered at the session level and p-values in parentheses, and 3) standard errors obtained using OLS regressions with standard errors clustered at session level. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

[a] For realized profits, the sample size is 86 for the in-group treatment and 70 for the out-group treatment. This is due to two cancelled sessions that resulted in observations unmatched with receivers.

Table A8: Investor behavior, beliefs and profit by treatment

		Profit		
	Amount sent	Requested	Expected	Realized
	(1)	(2)	(3)	(4)
Panel A				
	Trust game			
Ingroup	0.45	17.47**	10.86	24.40***
	(4.39)	(5.95)	(6.95)	(5.49)
Observations	185	185	185	164
R-squared	0.02	0.09	0.08	0.21
Panel B				
	Sanctioning game			
Ingroup	-0.35	12.88*	11.22*	9.79**
	(3.13)	(6.68)	(5.68)	(3.89)
Observations	185	185	184	154
R-squared	0.06	0.03	0.06	0.06
Panel C				
	Difference: Trust-Dictator			
Ingroup	1.72	16.19**	9.58	23.93***
	(3.70)	(5.46)	(6.34)	(6.79)
Observations	185	185	185	164
R-squared	0.08	0.04	0.04	0.14
Panel D				
	Difference: Sanctioning-Trust			
Ingroup	-0.80	-4.59	0.48	-11.59**
	(3.15)	(8.22)	(6.23)	(4.94)
Observations	185	185	185	141
R-squared	0.05	0.04	0.03	0.06

Note: Note: OLS coefficients. Robust standard errors, clustered at the session level in parentheses. In each regression we control for investor's ethnicity, age, number of household members, a dummy for literacy, years spent living continuously in Mazar-e-Sharif, log of income (Afs), a dummy for whether the individual had ever signed a contract and an index of perceptions of trust and fairness towards others (3 questions). *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table A9: Effect of fine on amounts returned relative to request (by fairness)

Panel A: Returned more than requested						
Sample	*Fair requests*			*Unfair requests*		
	All	*In-group*	*Outgroup*	*All*	*In-group*	*Outgroup*
Dependent variable	Returned more than requested (d)					
	(1)	(2)	(3)	(4)	(5)	(6)
In-group	0.30***			0.08		
	(0.10)			(0.07)		
Fine condition	0.16***	-0.11*	0.15***	0.06	0.02	0.07**
	(0.04)	(0.06)	(0.04)	(0.04)	(0.05)	(0.03)
In-group * Fine condition	-0.25***			-0.04		
	(0.07)			(0.06)		
No-fine condition	0.12*	-0.15**	0.11**	0.03	0.11*	0.05
	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)	(0.05)
In-group * No-fine condition	-0.27***			0.06		
	(0.09)			(0.08)		
Observations	547	265	282	452	226	226
Panel B: Returned less than requested						
Sample	*Fair requests*			*Unfair requests*		
	All	*In-group*	*Outgroup*	*All*	*In-group*	*Outgroup*
Dependent variable	Returned less than requested (d)					
	(1)	(2)	(3)	(4)	(5)	(6)
In-group	-0.15*			-0.22		
	(0.08)			(0.18)		
Fine condition	-0.27***	-0.06	-0.27***	-0.47***	-0.30**	-0.38***
	(0.08)	(0.06)	(0.07)	(0.13)	(0.14)	(0.10)
In-group * Fine condition	0.21**			0.19		
	(0.10)			(0.17)		
No-fine condition	0.05	-0.01	0.04	-0.10	-0.27**	-0.09
	(0.06)	(0.06)	(0.06)	(0.08)	(0.11)	(0.06)
In-group * No-fine condition	-0.05			-0.13		
	(0.09)			(0.11)		
Observations	547	265	282	452	226	226

Note: Marginal effects reported for individual level random effects probit regressions. Standard errors in parentheses (clustering at session level). Randomly assigned parameters only. In each regression we control for amount sent, share requested, trustee's ethnicity, age, number of household members, a dummy for literacy, years spent living continuously in Mazar-e-Sharif, log of income (Afs), a dummy for whether the individual had ever signed a contract and an index of perceptions of trust and fairness towards others (3 questions). *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table A10: Trustees' behavior in games: by ethnicity and treatment

	Panel A: Tajik participants		
Sample	*In-group* (1)	*Out-group* (2)	Difference (1)-(2) Somer's D (p-value) [Clustered SE] (3)
Trust game			
Share returned	58.66	42.58	16.08***
	(23.95)	(20.51)	-0.39 (0.00)
			[3.89]
Observations	74	76	
Sanctioning game			
Fine condition			
Share returned	61.93	56.44	5.50
	(24.60)	(24.43)	-0.13 (0.33)
			[5.30]
Observations	72	74	
No-fine condition			
Share returned	59.42	45.88	13.54***
	(22.94)	(21.46)	-0.36 (0.00)
			[3.77]
Observations	76	78	
	Panel B: Hazara participants		
Sample	*In-group* (1)	*Out-group* (2)	Difference (1)-(2) Somer's D (p-value) [Clustered SE] (3)
Trust game			
Share returned	58.10	42.15	15.95
	(21.90)	(23.41)	-0.41 (0.10)
			[8.02]
Observations	89	93	
Sanctioning game			
Fine condition			
Share returned	61.82	60.19	1.63
	(26.14)	(23.90)	-0.04 (0.83)
			[6.63]
Observations	88	93	
No-fine condition			
Share returned	59.28	52.11	7.17
	(25.21)	(20.91)	-0.16 (0.18)
			[3.67]
Observations	92	94	

Note: Means reported in Columns 1-3. Standard deviations in parentheses. Randomly assigned parameters only. Column 4 reports 1) the difference in means between the in-group and the out-group treatment, 2) Somer's D with variance clustered at the session level and p-values in parentheses, and 3) standard errors obtained using individual level random effects regressions with standard errors clustered at session level. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table A11: Effect of fine on share returned by ethnicity across treatments

Sample	*Tajik*			*Hazara*		
	All	*In-group*	*Out-group*	*All*	*In-group*	*Out-group*
Dependent variable	Share returned					
	(1)	(2)	(3)	(4)	(5)	(6)
In-group	15.45***			14.24*		
	(3.49)			(7.63)		
Fine condition	13.66**	-0.30	13.51*	13.81*	0.43	14.28
	(6.53)	(1.84)	(6.99)	(7.83)	(3.01)	(8.80)
In-group x Fine condition	-13.69**			-12.71		
	(6.79)			(8.09)		
No-fine condition	3.38	-1.25	3.06	6.98**	0.18	7.43**
	(2.13)	(1.57)	(2.45)	(2.75)	(3.02)	(3.25)
In-group x No-fine condition	-4.42*			-6.69*		
	(2.53)			(3.45)		
Sent	-0.11***	-0.06***	-0.16***	-0.20**	-0.20***	-0.22
	(0.03)	(0.02)	(0.02)	(0.08)	(0.05)	(0.15)
Share requested	0.40***	0.50***	0.29***	0.38***	0.51***	0.28***
	(0.05)	(0.07)	(0.03)	(0.10)	(0.16)	(0.09)
Constant	22.27***	29.32***	24.06***	30.52***	20.14***	42.23***
	(7.07)	(8.34)	(5.48)	(10.88)	(4.05)	(11.26)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes
Observations	450	222	228	549	269	280
Number of IDs	75	37	38	92	45	47
	F-test H_0 : Fine equals no-fine					
In-group p-value	0.73	0.76		0.10	0.63	
Out-group p-value	0.18		0.19	0.19		0.23

Note: Individual level random effects regression coefficients. Standard errors in parentheses (clustering at session level). Randomly assigned parameters only. In each regression we control for trustee's ethnicity, age, number of household members, a dummy for literacy, years spent living continuously in Mazar-e-Sharif, log of income (Afs), a dummy for whether the individual had ever signed a contract and an index of perceptions of trust and fairness towards others (3 questions). The F-test compares the fine and no-fine condition coefficients. Pooling models (1) and (4) into a single regression yields no significant ethnic treatment differences between the coefficients of in-group, fine and no-fine conditions (results available upon request). *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

B Theoretical background (for online publication only)

In this section we present a theoretical framework for interpreting the effects of the financial sanction on trustees' decisions. We adapt the model of state-dependent preferences described in Bowles and Polania-Reyes (2012) to the *sanctioning game*, with the aim of providing a framework for the interpretation of our results. Our goal in this study is to determine whether financial sanctions crowd out (or crowd in) trustworthy behavior, and if so, whether the effect differs between the *in-group* and *out-group* treatments. The fine, however, can be expected to influence the amount returned by trustees in two distinct ways: by changing the trustee's payoff function (i.e. the financial effect of the fine), and by activating preferences related to the fine. The key intuition of the model is that, while it is difficult to disentangle these two effects, we can nonetheless make inferences about the treatment effect on state-dependent preferences by considering the frequency of certain choices by condition and treatment. We can therefore demonstrate that our results are driven by a more nuanced effect than simply greater altruism towards *in-group* members.

We assume that the trustee's utility in the *trust* and *sanctioning games* is influenced by a combination of material and other-regarding preferences:

$$U_t^g(\pi_t,\ \alpha_t^g \pi_i) \tag{4}$$

where π_t and π_i represent trustee t's and investor i's payoff function from Equation 1 and 2, respectively and utility is increasing concavely in both arguments. The term α_t^g represents the trustee's other regarding preferences towards an investor in group g, such that

$$\alpha_t^g = \beta_t^g + p_i \lambda_t^g, \tag{5}$$

where β_t^g captures t's social preferences, conditional on whether t and i have a shared group affiliation, $g \in \{In, Out\}$, but unconditional on whether the fine was available or imposed.[36] When $\beta_t^g > 0$, this term captures t's state-independent altruism towards i. We assume that β_t^g varies between individuals,

[36] This model assumes that the amount sent by the investor and the requested back transfer are held constant. It is likely that these parameters meaningfully interact with sanctioning as well. However, we omit them here for simplicity. Moreover, by randomly assigning parameters, imposing a conditional fine is orthogonal to the other parameters in our experiment and this allows us to consider the effect of imposing a conditional fine independently.

and that β_t^{In} and β_t^{Out} are identically distributed around different means.[37] Based on both the literature showing in-group bias, as well as our own data, in which we find that back transfers are on average higher in the *in-group* treatment, we assume that $\overline{\beta^{In}} > \overline{\beta^{Out}}$. The parameter λ_t^g represents a set of the trustee's state-dependent preferences that change with imposing the conditional fine—again, we allow this parameter to vary with group treatment, g. The parameter λ_t^g encompasses several motivations as discussed in 4.1. Our ultimate goal in this analysis is to make inferences about λ_t^{In} relative to λ_t^{Out}.

We can do so by comparing the relative frequencies of certain types of decisions.

***Propostion 1**: If $\lambda_t^g = 0$, a trustee who returns $r_t > r_i^*$ in the trust game will also return $r_t > r_i^*$ in the sanctioning game.*

Proof:

A trustee maximizes 4 by choosing a value of r_t such that $\frac{\partial U_t}{\partial \pi_t} = \frac{\partial U_t}{\partial \pi_i}$. Let $\tilde{r_t}$ be the back transfer that maximizes utility when $p_i = 0$. Note that $\tilde{r_t}$ is increasing in β_t^g (i.e more altruistic trustees have higher preferred back transfers). A trustee returns $r_t > r_i^*$ iff $\tilde{r_t} > r_t^*$. How does introducing the conditional fine affect t's choice? Since the fine only enters the payoff function if $r_t < r_i^*$, by definition of $\tilde{r_t}$, the best response when $p_i = 1$ is: $r_t = \tilde{r_t} > r_t^*$. ■

Thus, holding other parameters constant, we expect the frequency of decisions for which $r_t > r_i^*$ to be equal across the *fine condition* and *trust game* (and *no-fine condition*) when $\lambda_t^g = 0$. According to 4, any change in the difference in the frequency of decisions for which $r_t > r_i^*$ between the *fine condition* and *trust game* (*no-fine condition*) suggests that $\lambda_t^g \neq 0$. If the frequency of such decisions increases when $p_i = 1$, this indicates that $\lambda_t^g > 0$, and similarly, if the frequency decreases, we assume $\lambda_t^g < 0$, (i.e. that the conditional fine crowds in or crowds out trustworthiness, respectively).

[37] This assumption seems reasonable. The standard deviations do not differ between treatments (p=0.69), nor do distributions (Kolmogorov–Smirnov, p=0.14). Supplementary Figure A7 plots the cumulative distribution of trustworthiness in the *trust game* by treatment, and demonstrates that trustworthiness is consistently higher in the *in-group* treatment.

Proposition 2: *If $\lambda_t^g = 0$, introducing the conditional fine will lead to a larger decrease the frequency of subjects who return less than the requested amount ($r_t < r_i^*$) in the in-group treatment than in the out-group treatment:*

$$\left[\boldsymbol{P}(r_t < r_i^* | p_i = 0) - \boldsymbol{P}(r_t < r_i^* | p_i = 1)\right]^{In} < \left[\boldsymbol{P}(r_t < r_i^* | p_i = 0) - \boldsymbol{P}(r_t < r_i^* | p_i = 1)\right]^{Out}.$$

Proof:

Take the case when $\tilde{r_t} < r_i^*$ (i.e. when, absent the fine, the trustee's utility would be maximized by returning less than the requested amount). How does introducing the conditional fine affect back transfers in this scenario?

Consider two possible responses: first, the trustee can avoid the fine by increasing his back transfer such that $r_t = r_i^*$. This decreases the trustee's payoff relative to $\pi_t(\tilde{r_t} | p_i = 0)$, but increases the investor's payoff by an equal amount. We can think of this as the "price" of complying with the investor's request: $r^* - \tilde{r_t}$. How does paying this "price" affect utility? Since $\tilde{r_t}$ is by definition the utility-maximizing level of back transfer in the absence of the fine, complying with the investor's request will necessarily result in a utility loss, relative to the case when $p_i = 0$. The size of the utility loss is mediated by β_t^g, as this determines the weight that the trustee puts on the investor's increased payoff; this utility loss is less severe as β_t^g increases.

Second, consider a salient alternative response: the trustee returns $r_t < r_i^*$, pays the fine, f, but reduces the back transfer by the amount of the fine, such that $r_t = \tilde{r_t} - f$. For this response, the payoff to the trustee remains unchanged, relative to the *trust game* (or *no-fine condition*): $\pi_t(r_t = \tilde{r_t} - f | p_i = 1, \tilde{r_t} < r_i^*) = \pi_t(\tilde{r_t} | p_i = 0, \tilde{r_t} < r_i^*)$. Again there is a utility loss for the trustee, mediated by β_t^g, but in this case the utility loss comes from the reduction in t's payoff, and thus the utility loss *increasing* in β_t^g.

Holding $\tilde{r_t}$ constant, trustees will prefer the first option iff β_t^g is sufficiently large. Additionally, since $\tilde{r_t}$ is increasing in β_t^g, the "price" of complying with the investor's requested back transfer, $r^* - \tilde{r_t}$, is also decreasing in β_t^g, with a similar implication that the trustee will prefer to increase his back

transfer to r_i^* whenever β_t^g is sufficiently large.[38]

If we consider the distribution of β_t across a population, as the mean increases, the difference between $p(r_t < r^*|p_i = 0) - p(r_t < r^*|p_i = 1)$ will also increase; in a population that has a higher average β_t there is a higher expected frequency of subjects who increase their back transfers to $r_t = r^*$ in the presence of the conditional fine, relative to the *trust game* (or *no-fine condition*). ■

If we had precise estimates of $\tilde{r_t}$ and β_t^g, we could derive a trustee's best response to p_i, r_i^* and attribute the residual to λ_i^g. Unfortunately, our design does not allow for this. However, if we assume $\beta_t^{In} > \beta_t^{Out}$, we predict that comparatively more subjects in the *in-group* treatment will increase back transfers in response to the conditional fine than subjects in the *out-group*: i.e. $\big[\boldsymbol{P}(r_t < r^*|p_i = 0) - \boldsymbol{P}(r_t < r^*|p_i = 1)\big]^{In} < \big[\boldsymbol{P}(r_t < r^*|p_i = 0) - \boldsymbol{P}(r_t < r^*|p_i = 1)\big]^{Out}$. In fact, in Supplementary Figure A3 and Table 3 we find the opposite. This suggests that $\lambda^{In} < \lambda^{Out}$. In other words, we find evidence that the treatment difference that we observe among trustees is not due to altruism alone, but rather that subjects react differently to financial sanctions, systematically, by treatment.

[38] In some cases, the best response when $p_i = 1$ is $\tilde{r_t} > r_t > (\tilde{r_t} - f)$, in which back-transfers under financial sanctions depend on β_t^g in a similar fashion. It is also possible that the best response might be unchanged under financial sanctions.