Sanctioning and Trustworthiness Across Ethnic Groups

Experimental Evidence From Afghanistan

Ian Levely (Wageningen University)
Vojtech Bartos (University of Munich)

Discussion Paper No. 107

July 24, 2018

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Dr. Klaus M. Schmidt, University of Munich, 80539 Munich, Germany
+49 (89) 2180 3405 | info@rationality-and-competition.de

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

Vojtěch Bartoš †

Ian Levely‡

July 23, 2018

Abstract

We show how sanctioning is more effective in increasing cooperation between groups than within groups. We study this using a trust game among ethnically diverse subjects in Afghanistan. In the experiment, we manipulate i) sanctioning and ii) ethnic identity. We find that sanctioning increases trustworthiness in cross-ethnic interactions, but not when applied by a co-ethnic. While we find higher in-group trustworthiness in the absence of sanctioning, the availability and use of the sanction closes this gap. This has important implications for understanding the effect of institutions in developing societies where ethnic identity is salient. Our results suggest that formal institutions for enforcing cooperation are more effective when applied between, rather than within, ethnic groups, due to behavioral differences in how individuals respond to 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. This research was supported by a grant from the CERGE-EI Foundation under a program of the Global Development Network, The Grant Agency of Charles University (46813), the Czech Science Foundation (13-20217S), and the Global Development Network (RRC13+11). All opinions expressed are those of the author and have not been endorsed by CERGE-EI or the GDN. Vojtěch Bartoš gratefully acknowledges support of the German Science Foundation through CRC TRR 190.

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

‡Development Economics Group, Wageningen University, De Leeuwenborch (building 201), Hollandseweg 1, 6706 KN Wageningen, The Netherlands. (ian.levely@wur.nl.).

1 Introduction

Ethnic diversity is often linked with worse economic outcomes (Alesina et al., 2003). One reason for this is that lower interpersonal trust between ethnic groups discourages trade and other cooperative interactions (Easterly and Levine, 1997; Alesina and La Ferrara, 2005). The capacity to sanction can mitigate the adverse effects of ethnicity, facilitating cooperation across ethnic divisions (Miguel and Gugerty, 2005; Habyarimana et al., 2007; Alexander and Christia, 2011; Glennerster et al., 2013). In the presence of strong institutions, societies can benefit from economies of scale to larger markets and public institutions, and complementarities in skills and endowments between ethnic groups (Easterly, 2001). Given this, facilitating the emergence of formal institutions is a top priority for many developing countries (Gennaioli and Rainer, 2007; Besley and Persson, 2010; Michalopoulos and Papaioannou, 2016; Ali et al., 2018).[1]

We use a lab-in-the-field experiment to investigate *how* sanctioning improves inter-ethnic cooperation. Well-functioning institutions for enforcing contracts and punishing shirkers may lead to more efficient outcomes by creating material incentives for cooperation. In other words, when trust and voluntary cooperation play a smaller role in economic transactions, so does in-group bias. Alternatively, sanctions may trigger different behavior depending on group identity. Previous literature has shown that willingness to punish norm violations is greater when the "victim" has a shared identity with the enforcer (Bernhard et al., 2006; Goette et al., 2006) and when the perpetrator is from a different ethnic group (Shayo and Zussman, 2011; Goette et al., 2012; Rehavi and Starr, 2014). Other studies show how the availability of sanctioning has a greater effect on cooperation in mixed groups (Habyarimana et al., 2007; Alexander and Christia, 2011), but do not concentrate on whether this is due to increased use of sanctions or due to a group-specifc reaction by individuals who receive or are threatened with sanctions. We contribute to this literature by showing that sanctions improve inter-group cooperation because individuals react to sanctions differently when they are applied by people from the same or a different ethnic group.

Our subjects are poor Afghans who have little experience with formal institutions, and who live in a multi-ethnic environment. We randomly assigned them to one of two treatments: they were

[1] State and nation building efforts are among top the priorities of international development agencies in post-conflict Afghanistan (Government of Afghanistan, 2008; Katzman and Thomas, 2017).

matched with someone from their own ethnic group (in-group treatment) or with someone from a different ethnic group (out-group treatment). They played a version of the trust game, a standard two-player experimental game, in which an "investor" can transfer money to a "trustee". The amount he sends is tripled, representing efficiency gains. The trustee can choose to return a portion of this to the investor, which represents trustworthiness. In contrast to the standard trust game, similar to Fehr and Rockenbach (2003), investors in our version have the ability to communicate their desired back-transfer to the trustee. In some decisions, the investor also has the choice to impose a conditional sanction. The trustee is aware of this choice. If the investor chooses to use the sanction, the trustee is informed that he will pay a small fine if he sends back less than what was requested.

We find that sanctioning increases trustworthiness in the out-group treatment, but has no effect on in-group trustworthiness. Without sanctioning, we find evidence of in-group bias among trustees: trustworthiness is lower in the out-group treatment. With sanctioning, however, the in-group bias disappears. Interestingly, the ability to sanction increases trustworthiness in the outgroup treatment even when the investor chooses not to apply it. We show that these results are consistent with a behavioral effect. The sanction crowds out trustworthiness in the in-group treatment, but not in the out-group treatment.

We contribute to a robust literature showing that i) not sharing a common group identity reduces cooperation and ii) certain institutions can mitigate these negative effects. For example, at the macro level, Miguel (2004) compares the effects of ethnic diversity in Kenya and Tanzania and argues that stronger nationalism in the former makes inter-ethnic cooperation easier. At the micro level, Hjort (2014) studies workers in a Kenyan flower factory, and finds that ethnically diverse teams are less productive–especially during a period of heightened inter-ethnic political tensions. When managers switched the payment scheme to reward teams rather than individuals, the difference disappeared.

In particular, institutions that allow for sanctioning have been shown to increase cooperation in ethnically diverse settings. Glennerster et al. (2013) study cooperation Sierra Leone. Despite being one of the most ethnically diverse and the least developed countries in the world, they find a surprisingly high capacity for collective action. They attribute this apparent contradiction to the historical role of local authorities, whose jurisdiction crossed ethnic lines. This is in contrast to many ethnically diverse societies, including Afghanistan, in which local institutions are specific to a particular ethnic group.

Numerous economic experiments have studied group bias, using organic group identity, including ethnicity (Whitt and Wilson, 2007; Buchan et al., 2009; Meier et al., 2014), as well as group-identity

induced in the lab (Tajfel et al., 1971; Charness et al., 2007; Chen and Li, 2009; Beekman et al., 2017). This literature explains how parochial social preferences can impede cooperation with out-group members (Bowles and Gintis, 2004).[2] 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 the same neighborhoods; Fershtman and Gneezy (2001) demonstrate how ethnic stereotypes influence trust in Israel.

Particularly relevant to this study are experiments showing that punishment can mitigate the effects of group identity. Habyarimana et al. (2007) conduct a lab-in-the-field experiment in Uganda, finding that ethnically homogeneous pairs cooperate more efficiently than ethnically mixed pairs. They explain this by showing that heterogenous groups are less able to sanction free riders. When the researchers make applying sanctions easy, by adding a mechanism for third-party punishment within the experiment, the ingroup advantage disappears. Alexander and Christia (2011) study how the ability to punish in a public good game affects cooperation differently among ethnically heterogeneous and homogeneous groups among high-school students in Bosnia-Herzogovina. They find that while ethnically mixed groups cooperate less, adding the ability to punish peers closes the gap. Interestingly, this is only true for subjects who attend integrated schools, and thus have experience interacting with students of a different ethnicity. Our study is similar, in that we examine how sanctioning can affect cooperation differently between and within ethnic groups. In their study, and other previous work, sanctioning institutions potentially increase cooperation between groups both because group identity affects willingness to impose sanctions and how individuals react to sanctions (i.e. behave differently when a sanction is available, or according to whether someone chooses to apply it). In contrast, we designed the experiment to isolate reactions to sanctions.[3]

[2] Cooperative behavior has been shown to be sensitive to group membership across many societies and from early ages Fehr et al. (2008); Bauer et al. (2014a,b).

[3] Cooperation in a public goods game with punishment, such as in Alexander and Christia (2011), is determined both by norms and preferences for cooperation, as well as the punishment one expects to incur for violating norms. Group identity treatments could mediate how the ability to punish peers affects contribution levels through i) different expectations of receiving punishment from the in-group and out-group, ii) by changing how individuals respond to the punishment that they (expect to) receive, or iii) by directly changing norms or preferences for cooperation. As explained in further detail in Sections 3.3 and 5, and in Online Appendix B, we designed the experiment in such a way that we can distinguish between the channels analogous to i) and ii) in the trust game.

Individuals are more willing to punish norm-violators from different ethnic groups, and with greater severity, than their co-ethnics. Shayo and Zussman (2011) and Rehavi and Starr (2014) find such an effect in in judicial records in Israel and the US, respectively. In deciding whether to punish, the identity of the "victim" also matters. In an experiment among traditional communities in Papua New Guinea, Bernhard et al. (2006) find that subjects are more willing to pay a cost to punish others when the norm violation harms an in-group member. Goette et al. (2006) come to similar conclusions, studying group identity among Swiss soldiers (by platoon). In a follow-up experiment with the same population, Goette et al. (2012) show how this depends on context. They punished soldiers from different platoons relatively more than members of their own, but only when platoons were asked to compete against one another beforehand. This was driven by an increase in anti-social punishment (i.e punishing cooperators). As a result, under competitive conditions, punishment led to *worse* outcomes between groups than within groups. This shows why it is important to understand precisely how sanctions affect cooperation across groups.

In the trust game (Berg et al., 1995), a self-interested investor will only send a positive amount if he expects the trustee to return a greater amount. Since a self-interested trustee will not return any money, the equilibrium is that the investor sends nothing. However, pro-social motivations such as altruism and reciprocity motivate trustees to return positive amounts, and consequently, investors to send a positive amount (Camerer, 2003). This game has been played in many settings with diverse subject pools, and the majority of investors do in fact send money, and the majority of trustees are to some extent trustworthy (Johnson and Mislin, 2011).

In our version, investors made three simultaneous choices: i) how much of their endowment to send to the trustee, ii) how much of the tripled amount to request back, and iii) whether to apply a conditional sanction. If they did so, the trustee would pay a small fine if they returned less than the amount requested by the investor. The amount of the fine was fixed. For comparison, we also had investors play a version of the game in which sanctioning was not possible (i.e. choices i. and ii., but not iii.).

Trustees made several decisions. In each decision they received all of the information corresponding to the investor's choices, then chose how much to return. As with investors, they played versions of the game both in which sanctioning was and was not possible. All information was public: trustees knew how much had been sent by the investor, the amount requested back, if sanctioning was available, and if so, whether the investor had chosen to apply the sanction.

Our central hypothesis is that ethnic identity influences how individuals respond to sanctions. We test this causally, independent of investors' choices, across treatments. The amount sent, requested and whether the sanction was applied by the investor are endogenous, and likely vary systematically by treatment. Therefore, we presented each trustee with multiple decisions. One of these decisions used parameters from the trustee's partner, and they were paid for that decision. For the other decisions, we varied the parameters orthogonally to treatment. Trustees did not know ex-ante which decision represented their actual partner's choice, and were instructed to treat each decision as if it were. This allows us to study reaction to sanctioning by ethnic group, causally, and without using deception.

The conventional prediction is that the sanction will lead to increased back transfers by disciplining trustees who fail to comply with the investor's request. In the game, the fine is small enough that trustees are almost always better off financially if they do not return anything to the investor. But, if an individual has an intrinsic motivation for trustworthiness, such as reciprocity, the threat of paying a fine should only lead to increased compliance. However, if other-regarding preferences are "state-dependent" (Bowles and Polania-Reyes, 2012), the presence of the sanction not only changes the financial incentives, but also affects an individual's intrinsic motivation. In many cases, material incentives have been shown to crowd out preferences for cooperation (Titmuss, 1971; Gneezy and Rustichini, 2000; Frey and Jegen, 2001; Bowles, 2008). This can make sanctioning less effective, or in extreme cases, counter-productive.

A number of laboratory experiments have examined this phenomenon in detail. Our design is similar to Fehr and Rockenbach (2003) and Fehr and List (2004). Both studies find that when investors apply the sanction, the amount returned by trustees decreases. Ostensibly, by using the sanction, the investor signals a lack of trust. This decreases the trustee's motivation for trustworthiness. 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.[4]

Falk and Kosfeld (2006) use a different experimental design to study a similar effect: "aversion to control." They demonstrate that when a principal restricts an agent's choice set, this crowds out her motivation to exert effort, which is costly to the agent but beneficial for the principal. Particularly

[4] Fehr and Rockenbach (2003) also find that the effect of sanctioning depends on whether the request is fair. Sanctions crowd out trustworthiness when the requested back-transfer is high, and thus leaves the trustee relatively worse off. This is not the case when the requests are fair.

relevant to this study, Masella et al. (2014) run a similar experiment with students in Germany. They introduce minimal group identity in the lab, and find that that control reduces effort both within and between groups, but for different reasons. Control is unexpected when applied within groups, and has a crowding out effect. When used between groups, control creates hostility. Similarly, Riener and Widerhold (2016) show that control crowds out effort to a larger degree when pairs have first completed a team-building exercise together.

A key difference between these studies and ours is the subject pool. We conducted the experiment with residents of peri-urban areas of Mazar-e-Sharif, Afghanistan. This is a country currently undergoing a coordinated transition to stronger and more prevalent formal institutions. Ethnicity is extremely salient in Afghan society, and plays a large role in everyday life. We study two ethnic groups in particular: the Hazara and Tajiks. The two groups live in mostly (unofficially) segregated communities. This also means that the majority of informal and semi-formal institutions are also de facto segregated by ethnicity. Community leaders serve small, ethnically homogeneous neighborhoods and are responsible for resolving many local disputes. Moreover, most Tajiks are Sunni Muslims and the majority of Hazaras are Shia, which means that they attend different mosques and celebrate different religious holidays.

This study contributes to our limited understanding of how individuals in diverse societies react to the introduction of new institutions. This both provides a deeper understanding of how instituions might have develped in ethnically diverse societies, and has important implications for policy makers. While ethnically heterogeneous societies with low levels of trust benefit most from institutions that make enforcing contracts easier (Collier, 1999), the cooperation requred to form instituions in the first place means that they might be the least likely to develop them organically. We show how the behavioral reaction to sanctions partially mitigates this: at the margins, sanctioning institutions might have a higher return in diverse settings. Many developing and post-conflict countries that are characterized by both salient ethnic divisions as well as weak, inefficient and incomplete contract enforcement mechanisms. Our results provide a mechansim that explains how strengthening formal insitutions can improve inter-ethnic cooperation, and suggest that in such societies, the added value to enforcing contracts across ethnic groups is larger than doing so within ethnic groups.

2 Setting

Ethnicity has been used as a political tool to divide Afghan society throughout recent history. Tajiks, Hazaras, and Uzbeks in particular were singled out as groups distinct from the Pashtun majority by the royal government in the early twentieth century, and 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. Later, the predominately Pashtun Taliban, a fundamentalist Sunni Muslim group, was responsible for widespread persecution, including ethnic cleansing campaigns, against minorities, especially the Hazara (Schetter, 2016).

Ethnic affiliation continues to play a large role in Afghan politics both at the national and local levels, and is salient in the ordinary lives of Afghans. Even though cohabitation is mostly peaceful—over 70% of the participants in our study reported having friends among people from a different ethnic group—the communities we study are mostly self-segregated along ethnic lines. Disputes within communities are often settled by *kalantars* (community leaders), and in disputes with a member of another community, kalanatars from both communities are typically involved in mediation. Thus, the informal institutions affecting day-to-day interactions are also separated by ethnicity. Beath et al. (2016) report that 75% of villages in their sample, which covers most of Afghanistan, are perfectly ethnically homogeneous. 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 Hazra, "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.

Afghanistan is an interesting setting to study the interaction of ethnicity and institutions because it is an ethnically heterogeneous post-conflict country with a low level of development, currently undergoing a rapid transition to formal institutions. 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 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 a very low levels of experience with formal governmental institutions: less than 6 percent have ever signed a written contract, over half of the sample never attended school, only 5 percent of the participants were employees of a registered private company or of a state institution, and only 1.5 percent of participants who were currently in debt owed money to a bank, microcredit organization or other formal credit institution. When asked how they would respond to a hypothetical theft, only around 11 percent said they would contact the police, compared to 40 percent who responded that they would contact their kalantar, and 36 percent who would contact a neighbor.

3 Design

3.1 Experimental games

To examine the effect of pecuniary sanctions on prosocial 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, $s_i \in [0, 10, 20, \ldots \omega]$ of his endowment. 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 baseline condition, 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 sanction, $f = 40$, dependent on whether the trustee's back transfer is less than the amount requested by the investor. Applying the

sanction is costless to the investor. We denote the decision to impose the sanction as $p_i \in [0, 1]$, where $p_i = 1$ if the investor chooses to conditionally apply the sanction 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).

In the sanctioning game, with the parameters we use—which are identical to Fehr and Rockenbach (2003) and Fehr and List (2004)—the sanction is too small to allow the investor to capture the efficiency gains from a self-interested trustee in all but one, extreme case.[5] However, assuming that the decision to impose the sanction does not negatively affect trustworthiness, there is no reason for an investor to refrain from using it, as doing so provides a financial incentive, in addition to intrinsic motivation, for trustees to transfer an amount at least as high as the investor's request.

However, the sanction could also affect trustworthiness by activating "state-dependent" preferences. First, the presence of the sanction 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. Secondly, since the investor chooses whether or not to apply the sanction, it 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 apply the sanction, potentially the investor communicates a lack of trust, and this could impact back transfers. On the other hand, in the no-sanctioning condition, when the sanction was available but the investor chose not to impose it, this might constitute an implicit signal of trust. This "good news" about the investor's beliefs and intentions could increase the amount returned by trustees (Bowles and Polania-Reyes, 2012). We can compare trustworthiness in the sanctioning and trust games, and similarly compare results between the sanctioning and no-sanctioning conditions to study these effects.

If the sanction crowds out trustworthiness, and the effect is large enough in magnitude, then the

[5] If the investor sends $s_i = 10$, requests $r_i^* = 30$ and imposes the sanction, 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 actually 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 $r_t = r_i^* = 40$. If the trustee complies, the investor makes a profit of 20 Afs. Whenever $s_i > 20$, the trustee will always maximize his earnings by returning $r_i = 0$, regardless of amount requested and whether the sanction is applied.

sanction could cause a trustee to return less than he would in either the trust game or in the no-sanctioning condition. It is also possible that the sanction "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 were 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., 2017).

3.2 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. After Pashtuns (43%), Tajiks (31%) and Hazaras (9%) constitute the second and third largest ethnic groups in Afghanistan (DHS, 2011). 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.[6] 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. Ethnicity is prominent in everyday life in Afghanistan, and it is a reasonable assumption that including it in the description did not seem particularly out-of-place for subjects.

[6] 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.

3.3 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-e-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 this 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 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 response rates across Tajik and Hazara communities (80 and 76 percent, respectively; p=0.34). Supplementary Table A1 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-e-Sharif, but that they would not know which community, specifically, 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.[7] 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.

[7] In 75% of sessions trustees played the sanctioning game first, with the order reversed for the remaining sessions.

Since the subject pool is largely illiterate, all instructions were given orally, using visual aids.[8] After a general introduction of the experiment and explanation of the task in a group setting, the subjects were seated in private booths (See Figure A1) where they made their decisions in privacy—though not anonymous to research assistants.

We had trustees make several choices in each game. We provided them with a set of four, randomly assigned sets of parameters (s_i, r_i^*, and p_i) for the sanctioning game (including exactly two with the sanction applied), and two sets of parameters per subject for the trust game. [9] Subjects were told (truthfully) that each decision came from a participant from a 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).[10] [11] The choices were presented using simple visual aids (See Figure A2). Individual decisions were made by putting "banknotes" into envelopes representing own and matched partner's budget, a task easy to comprehend even to illiterate subjects. Table 1 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.

There are three main 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

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

[9] The parameters within each category were randomly selected from a pool of decisions made by investors in earlier sessions or in practice rounds in each respective game. This was the same for both treatments.

[10] 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 the that described in the treatment.

[11] To give a numeric 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($s_i = 80$,$r_i^* = 170$, $p_i = 0$). The T stands for a trust game and S stands for a sanctioning game, while the numbers represent the order in which the choices appeared. Between the Trust and Sanctioning game, there was a short break during which insturctions for the second game were presented.

directly; if we were to examine only trustees' responses to investors' actual choices, sanctioning 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. Secondly, 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.[12]

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.

4 Results

First, we analyze how both the availability and use of sanctions affect the amount returned in the trust game across the in-group and out-group treatments. This allows us to test our main hypothesis: group identity plays a role in how individuals react to sanctioning. In Subsection 4.2 we discuss investors' behavior.

[12] We assigned the parameters according to two dimensions: the level of investor's trust (s_i) and the requested back transfer (r_i^*). We classified requests as "fair" or "unfair," depending on whether the payoff for the trustee was at least as high as or less than the sender's, respectively. This is based on the categories defined in Fehr and Rockenbach (2003). We classified allocations as low/high-trust, if the amount sent was less than/more than 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 sanction applied and one in which the investor chose not to apply of the sanction. This procedure was used in order to limit within-subject variance.

4.1 Trustee experimental results

Table 3 and Figure 1 summarize the results for both the trust and sanctioning games. 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 return a significantly higher portion of what they receive in the in-group treatment than they do in the out-group treatment. When trustees were paired with an investor from the same ethnic group, the share returned is on average 58.35 percent of the total amount received from investors, compared to only 42.34 percent in the out-group treatment (p=0.00).[13] This indicates that ethnicity is indeed salient among the population that we study, and has an effect on trustworthiness.

In Table 4 we regress treatment on the percentage returned 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 sanctioning and no-sanctioning 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.[14]

Next, we compare trustee behavior across treatments in the sanctioning game. Bars 3 and 4 of

[13] All significance levels reported for comparison of means are from Wilcoxon rank-sum tests, unless otherwise noted.

[14] The results of the regression model estimated in Table 4 are also robust to various alternative specifications. First, Supplementary Table A2 shows that the results are robust to 1) controlling for enumerator fixed effects, 2) controlling for the effects of the order in which the games were played, 3) correcting for the small number of clusters, 4) individual fixed effects (without treatment) 5) excluding the amount sent and 6) share requested variables, 7) including only the first choices for each game where the effects of intentions are assumed to be strongest, and 8) also including the reactions to the actual investors' decisions, rather than to the exogenously assigned parameters used in all other specifications. Second, the results are robust to using quantile regressions (25th, 50th, and 75th quantiles; see Supplementary Table A3). Lastly, as noted earlier, while sets of parameters were exogenously assigned across treatments, they were drawn from the distributions of the actual investors' choices for each game separately. 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 sanctioning condition. We then further 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. See Supplementary Table A4.

Figure 1 report the shares returned in the sanctioning condition—i.e when the sanction was available and used—by subjects in the in-group and out-group treatments, respectively. Compared to the trust game, subjects in the out-group treatment send back more in the sanctioning condition, returning 58.52 percent on average, an increase of 16.18 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 (3.52 percentage points), and the difference is only marginally significant (p=0.09). Notably, under the sanctioning condition, the gap between in-group and out-group treatments narrows to only 3.35 percentage points, which is no longer significant (p=0.23). The regression results in column 1 of Table 4 show that the difference-in-differences between sanctioning and group treatment is statistically significant (in-group*sanctioning). In columns 2-3 we divide the sample by treatment group and observe, that while sanctioning increases back transfers relative to the trust game in the out-group treatment, there is no effect of sanctioning on back transfers in the in-group treatment.

Result 1: *Sanctioning 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-sanctioning condition, in which the investor had the option to apply the sanction 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 sanctioning. On the other hand, the investor potentially signals implicit trust by voluntarily refraining from using the sanction (Bowles and Polania-Reyes, 2012). Bars 5-6 of Figure 1 present results from the no sanctioning condition. For trustees in the out-group treatment, the share returned in the no sanctioning condition falls between the levels for the trust and sanctioning conditions at 49.28 percent, a decrease of 9.24 percentage points over the sanctioning condition (p=0.00) and 6.94 percentage points higher than in the trust game (p=0.00). For the in-group treatment, however, the share returned in the no sanctioning condition was virtually the same as in both the trust game (p=0.81) and in the sanctioning condition, (p=0.22). The regression results in Table 4 confirm this. In column 2 we observe that, in the in-group treatment, back transfers in the no sanctioning condition do not differ from those in the trust game (p=0.80), nor from those in the sanctioning condition (p=0.76). For the out-group, on the other hand, "no sanctioning" increases the share returned by 5.45 percentage points relative to the trust game (p=0.02).

Result 2: *The no sanctioning 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 sanctioning condition is necessarily a behavioral one. The literature provide two potential explanations for Result 2. First, the possibility of sanctioning might change the relationship between investor and trustee, and activate a different set of preferences. Secondly, by voluntarily abstaining from using the sanction, the investor might signal something about his character or intentions. Higher back transfers in this case might be a response to this "good news."[15] Our results differ from those of previous studies using similar designs (Fehr and Rockenbach, 2003; Fehr and List, 2004) in that the sanctioning condition does not lower the average amount returned. While this does not necessarily rule out the possibility that either the sanctioning or no sanctioning conditions crowd out intrinsic motivation for cooperation, it does suggest that any such effect is substantially smaller than in previous studies.

Our main result is that applying the sanction has a higher marginal effect in cross-ethnic situations than in ethnically homogenous ones. This is true when comparing the sanctioning condition with both the trust game and no sanctioning conditions.

4.2 Investor experimental results

4.2.1 Efficiency

We now turn to investors' decisions in the trust, sanctioning and triple-dictator games, which are presented in Table 5. To begin with, we examine social efficiency in the three games. Since the amount sent in each of the three games was tripled, comparing the social efficiency achieved by subjects is straightforward: the more money sent by investors, the greater the surplus.

First, we compare the amount sent in each game across treatment. 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=0.75). In the sanctioning game, results are nearly identical: investors sent 55.96 Afs in the in-group treatment and 56.67 Afs in the out-group treatment, on average (p=0.70). Neither is there a statistically significant difference in dictator game allocations across treatments (p=0.55).

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

Next, we compare the difference in the amount sent between games. Comparing the amounts sent in the Trust and Sanctioning games gives an indication of how the ability to sanction affects social efficiency. The difference is small and not statistically significant for either treatment (p=0.60 and p=0.82 for the in-group and out-group treatments, respectively). Nor is the difference in differences in the amount sent between the trust and sanctioning games and treatment statistically significant (p=0.66).

Together, this indicates that social efficiency is not affected by group treatment, in either game, nor is it affected by the ability to sanction, in either treatment. We confirm this through regression analysis, controlling for individual characteristics in panel A of Supplementary Table A5. Investors in both treatments did, however, send more in both the trust and sanctioning games than in the dictator game, which shows that subjects do react strategically to the experimental environment, just not to the presence of the sanctioning mechanism.[16]

4.2.2 Requested, expected and realized profits

The lack of a statistically significant difference in amounts sent is puzzling on the surface, as investors might plausibly anticipate higher levels of trustworthiness in the in-group treatment and would thus achieve Pareto dominant outcomes by sending more. However, while there is no difference in the overall size of the surplus, investors do use a different strategy when dealing with in-group members by requesting a higher share of the surplus. In the trust game, investors in the in-group treatment request 51.97 percent of the tripled amount sent ($r_i^*/3s_i$) compared to 42.46 in the out-group treatment (p=0.00).

Since the investor's choice has multiple elements (which are endogenous), it is difficult to consider this result on its own. Given this, we combine the elements of an investors decision and consider the overall payoff that an investor's allocation and requested back transfer implies for himself ($\omega_i - s_i +$

[16] We do not find any difference in the amount sent between treatments in the dictator game (p=0.55), 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., 2014b). Potentially, these effects counteract one another.

r_i^*). We can meaningfully compare this measure of "requested profit" across games and treatments. Combining responses from incentivized questions that elicited investors' beliefs about trustees' expected back transfers with data from trustees' actual decisions, we construct similar measures of expected and realized profits. We compare these measures across games and treatments in Supplementary Figure A3 and Table 5.

In the Trust game, investors assigned to the in-group treatment requested a higher profit, on average, than in the out-group (p=0.00). In both treatments, the average request still leaves a higher profit for the trustee. The higher requested back transfers by investors in the in-group treatment leads to a lower percentage of fair requests in the in-group treatment (84 vs. 93 percent, p=0.06). Investors in the in-group treatment also expected to earn more than those in the out-group (10.81 Afs, p=0.09). Given the higher back transfers among in-group trustees that we find when analyzing randomly assigned parameters, it is unsurprising that investors in the in-group treatment earned 25.39 Afs more than out-group investors in the trust game (p=0.00). Note that since the overall surplus is virtually identical across treatments, the higher investor profits in the in-group treatment come at the expense of trustees.[17] (See Panel B of Supplementary Table A5 for regression analysis).

The treatment differences in requested, expected and realized profits in the sanctioning game look much like those in the trust game. Again, investors request a larger share when playing with an in-group partner (p=0.09). Subjects in the in-group treatment expected to earn 9.30 Afs more than subjects in the ougroup treatment in the sanctioning game (p=0.13), and did in fact earn 11.15 Afs more, though the treatment difference in realized profits is much smaller than in the trust game and not statistically significant (p=0.16).

We observe that requested profits are higher overall in the sanctioning game than in the trust game, and this is driven primarily by subjects in the out-group treatment, who requested 6.51 percentage points more of the tripled amount sent in the sanctioning game than in the trust game (p=0.09). The difference is smaller for the in-group treatment, at 2.69 percentage points, and is not statistically

[17] Investor profits are calculated using the average amount returned in all decisions in that game and group treatment (i.e. Tajik-Tajik, Tajik-Hazara, etc...) in which trustees were presented with the parameters matching the investor's choice (i.e. amount sent, requested back, and whether the sanction was applied). This includes the decision of the individual with whom the investor was matched for payment, and also decisions made by other trustees who were presented with those parameters in any of the choices they made, including when parameters were randomly assigned. Trustee profits in Supplementary Figure A3 are calculated analogously.

different from zero (p=0.30), though the difference in differences between game and treatment is statistically insignificant (p=0.13).

4.2.3 Sanctioning

While these results are broadly consistent with the trustee results presented in the previous section, we have not yet considered the essential feature of the sanctioning game. Given that the trustee results demonstrate that there is a higher marginal return to applying the sanction 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 sanction by investors: 36.54 percent of investors in the in-group treatment and 38.10 percent of investors in the out-group treatment chose to apply the sanction (p=0.53). Overall, the majority of those who applied the sanction in both treatments expected trustees to comply with their request. In the out-group treatment, 39.47 percent of those who applied the sanction expected trustees to be fined, compared to 28.13 percent of those who used the sanction in the in-group, the difference is not statistically significant (p=0.33). In both treatments, investors who applied the sanction were more successful, earning higher profits (for the combined sample, p=0.12).

Result 3: *Sanctions are applied with a similar frequency across ethnic treatments.*

Since the components of the investor's decision in the sanctioning game—the choice of whether to apply the sanction, the amount sent and requested back transfer—are endogenous, it is difficult to draw independent conclusions from any one in isolation. Moreover, investor behavior is not only strategic, but also reflects preferences. Even if subjects conjectured that sanctions 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. This would be consistent with the results of Bernhard et al. (2006) and of Goette et al. (2006). On the other hand, in line with Fearon and Laitin (1996), some investors in the out-group treatment might refrain from sanctioning due to the fear of damaging relations with the other ethnic group. Thus, though subjects in the in-group and out-group apply the sanction with similar frequencies, they may do so for different reasons.

5 Discussion and further results

Our results suggest that sanctions are more effective in increasing cooperation between ethnic groups

than within ethnic groups. While sanctioning increases back-transfers in the out-group treatment, back-transfers in the in-group treatment do not statistically differ between the the trust game and sanctioning condition. This is true both when the sanction is applied by the investor and when the sanction is available but not applied. In this section we further analyze trustees' decisions, and attempt to identify the behavioral mechanism underlying these results. First, we consider whether the fairness of the requested back-transfer influences how trustees react to sanctioning, and whether this differs by treatment. Second, in Section 5.2 we explore the behavioral motivations that potentially drive our results. Lastly, in Section 5.3 we discuss the mechanisms behind the treatment differences in behavioral reactions to sanctions.

5.1 Sanctioning and fairness

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 that is considered unfair. Fehr and Rockenbach (2003) find that responses to the sanctioning condition differ between "fair" and "unfair" requests. Here, we also 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 $r_i^*/3s_i \leq 0.67$.[18]

In columns 1 to 3 in Panel A of Table 6 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 sanctioning condition, reduces the share returned, relative to the trust game, by 3.76 percentage points (p=0.08).

We also find stronger results for the no sanctioning condition for in-group subjects when we consider only fair requests: there is a corresponding decrease of 4.28 percentage points in the share returned relative to the trust game (p=0.08). Interestingly, there is no signifcant difference between the sanctioning and no sanctioning conditions (p=0.67). This implies that it is not the intentions that an investor communicates by his decision to apply the sanction, but rather the sanctioning game itself

[18] 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.

which crowds out trustworthiness in the in-group treatment.

For the out-group, the effect is reversed, and the sanctioning condition has a similar effect on the sub-sample of fair decisions as it does overall, increasing the share returned by 9.67 percentage points (p=0.03). Likewise, and contrary to the results for the in-group treatment, the no sanctioning condition is associated with a 6.30 percentage-point increase in the share returned (p=0.02), relative to the trust game, for fair decisions in the out-group treatment (Table 6, column 3 of Panel A). As with the in-group treatment, there is no statistically significant difference between the sanctioning and no sanctioning conditions for out-group subjects when requests are fair (p=0.47).

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 (4.02 percentage points), is not statistically significant for the in-group treatment (p=0.20). Nor is there any statistically significant difference between the sanctioning condition and the no sanctioning conditions (p=0.98). 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 the reaction to the sanctioning and no sanctioning conditions does differ according to the fairness of requests. In the in-group treatment, when requests are fair, we find clear evidence that both the sanctioning and no sanctioning conditions crowd out trustworthiness. As discussed in Section 4.2, the vast majority of decisions made by investors in both treatments, for both the trust game and the sanctioning game, are classified as fair. 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.

5.2 The behavioral effect of sanctions

We find that sanctioning increases trustworthiness in the out-group, but not in-group treatment. The fact that we do not see a response to the sanction for subjects in the in-group treatment does not necessarily indicate that it has no underlying effects, however, as the financial effect of the sanction may cancel out the behavioral effect. In fact, there are several scenarios consistent with our results: First, the sanctioning condition could influence behavior by crowding out intrinsic motivation for

trustworthiness, but less so in the out-group treatment. Second, the sanction 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 sanction 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 sanctioning only affects back-transfers by imposing financial incentives, we would expect an increase in a trustee's utility is maximized by returning $r_t > r_i^*$ in the trust game, then—holding other parameters constant—the introduction of the sanction provides no financial incentive to change one's behavior, since neither the trustee's nor investor's payoff would be affected given the status quo.[19]

On the other hand, if sanctions crowd out non-financial incentives for cooperation, then the trustee will maximize his utility by returning a smaller amount when the sanction is present, since he then puts 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 sanctioning condition, relative to the trust game. Alternatively, if the sanction 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 sanctioning condition than in the trust game, following similar logic.

The frequencies of each type of decision in the trust game, and sanctioning and no sanctioning conditions, by group treatment, are shown in Supplementary Figure A4. In the trust game, 32.52 and 15.38 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 9.39 percentage points in the sanctioning condition relative to the trust game. This provides support for our finding that sanctions do in fact crowd out trustworthiness in the in-group treatment. The difference is (marginally) statistically significant (p=0.06). In the out-group treatment, however, there is actually an increase of 6.77 percentage points in the frequency of decisions in which the amount returned exceeds the amount requested in the sanctioning condition relative to the trust game, though the difference is statistically insignificant (p=0.11).

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

There is virtually no change in the fraction of subjects returning more than requested between the trust game and no sanctioning condition in either the in-group and out-group treatments (p=0.67 and p=0.51, respectively).[20]

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 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 sanctioning 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 altruism, the trustee would prefer to pay the fine, and decrease his back transfer to compensate for the loss incurred by the sanction, and reducing the investor's payoff as a result. Thus we predict that the sanction 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, in fact, what we find: we observe that there is a drop in the frequency of decisions in which trustees return less than the requested amount from the trust game to the sanctioning condition in the out-group treatment by nearly half, from 56.80 percent to 28.74 percent of total decisions, (p=0.00). The decrease in the in-group treatment is comparatively smaller: from 42.33 percent in the trust game to 32.50 percent in the sanctioning condition (p=0.07, see Supplementary Figure A4). The results are confirmed using a probit estimation that includes observables. See columns 4 to 6 in Panel B of Table 6.

To summarize, the results for the out-group suggest that sanctions actually reinforce behavioral

[20] Results of a random effects probit model presented in columns 1 to 3 of Panel B in Table 6 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 sanctioning 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 sanctioning 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 A6).

motivations to return a higher amount, but crowd out trustworthiness for those in the in-group treatment.

5.3 Potential Mechanisms

Differences in state-dependent preferences towards in-group and ougroup members that we observe among trustees could result from several underlying mechanisms. Firstly, it is possible that we observe more crowding out of behavioral trustworthiness in the in-group treatment simply because there is more trustworthiness to crowd out. Potentially there is a diminishing marginal return on the behavioral effect of sanctions, and this is smaller in the out-group treatment, since trustworthiness is lower in general. However, this seems unlikely given the magnitude of the effect. In the sanctioning condition, we see that the gap between in-group and out-group subjects virtually disappears, and in fact we find evidence that sanctions actually increase trustworthiness. If the treatment difference in response to sanctions were driven by initial levels of trustworthiness alone—rather than a difference in state-dependent preferences—this would not be the case.

Another possibility is that there are different perceptions of fairness between the in-group and out-group treatments, and this leads to the differences in responses to the sanctioning and no sanctioning conditions. Potentially, state-dependent preferences are independent of group identity, but do depend on whether the request is considered fair. If a given request is considered fair by the trustee if the investor is from a different ethnic group, but unfair if the investor is a co-ethnic, then we would expect a different reaction to the sanctioning conditions, even if the underlying preferences related to sanctioning were identical. To this end, we test several definitions for fair requests and do not observe any clear pattern in response to different thresholds in either treatment. Coefficients from regressions on sub-samples defined by the requested back transfer are presented in Supplementary Figure A5. If differences in perceptions of fairness alone were driving the treatment difference, we would expect to see a similar effect of the sanctioning and no sanctioning conditions on the in-group and out-group treatments after adjusting the definition of fairness. The fact that we see no such pattern indicates that this is not the case.

Finally, we are left with the interpretation that group identity affects behavioral responses to 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. . .). This assumes that it is only relevant whether subjects come from the same or different ethnic groups, rather than the specific group identities. 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 and estimate the same model as in Table 4. The results are very similar in both cases and consistent with the pooled results (see Supplementary Tables A7 and A8).[21] 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).[22]

We acknowledge that is difficult to generalize from our sample to other inter-ethnic, or inter-cultural settings. Although we do find similar behavior in both Hazara and Tajik subjects in our sample, these two ethnic groups share many cultural similarities and live in the same setting. It is possible that the relationship between formal sanctioning mechanisms and ethnic identity would differ in other settings, and more research should be done in this area, to understand these cultural differences. In addition, due to the cultural issues associated with interactions with women, our sample is entirely male. It is possible that gender differences exist in the effects we describe. 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: we find that sanctions crowd out trustworthiness to a lesser degree than in previous experiments (Fehr and Rockenbach, 2003; Fehr and List, 2004).[23] Our sample has very little experience interacting with formal institutions—only about 6 percent have ever signed a contract—and we conjecture that this may underlie this difference

[21] 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).

[22] 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.

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

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.

6 Conclusion

When formal institutions are weak, societies must rely on informal cooperation to a greater extent, and this means that group affiliation 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 artefactual field experiment with two ethnic groups in Northern Afghanistan—the Tajik and Hazara—we are able make causal inferences about how the effectiveness of sanctioning differs with group identity. We find that ethnic identity does indeed affect the way that individuals respond to sanctions: while sanctions increase trustworthiness when applied across ethnic groups, there is no effect, on average, when sanctions are applied by a co-ethnic. When sanctioning is not possible, trustworthiness is significantly higher in the in-group treatment, but sanctioning closes this gap.

We presented trustees with multiple scenarios, manipulated in such a way that we can causally test how trustworthiness is affected by sanctioning independent of differences in investor behavior in each treatment. This allows us to infer that sanctioning crowds out trustworthiness in the in-group treatment, but crowds in trustworthiness in the out-group treatment. The null effect of sanctioning on trustworthiness that we observe in the in-group treatment is consistent with the crowding out of intrinsic motivation for cooperation. The financial effect of the sanction potentially increases cooperation in both the in-group and out-group treatments, but is counteracted by a behavioral effect in the in-group treatment only. This could explain the lack of difference between the trust game and sanctioning condition for the in-group treatment. While this suggests that sanctioning treatment crowds-out trustworthiness less than in previous studies, the result is qualitatively similar to Fehr and Rockenbach (2003) and Fehr and List (2004).

Interestingly, 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 attempt (and succeed) to 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 applying the sanction—as one might expect, given that it is more effective in the out-group treatment—this decision is influenced by both expected returns as well as 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 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) shows that while access to informal credit in Zimbabwe is strictly limited to co-ethnic business partners, formal credit institutions do not discriminate by ethnicity, and Biggs et al. (2002) examine access to credit in Kenya to find similar results. As a matter of policy, the results of our study provide tentative evidence that perhaps effort is best spent strengthening formal mechanisms for instituting sanctions between communities, 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 there exists a group specific difference in the reaction to sanctioning. We conjecture that attitudes towards group identity and sanctioning might develop in communities where ethnicity is salient, as a best response to avoiding inter-group conflict. Fearon and Laitin (1996) outline a theory that predicts higher costs of conflict between members of opposing groups, as these inter-personal disputes can spiral into more costly conflicts involving the entire groups. Given this, peace is maintained by avoiding such conflicts when possible, and by ignoring transgressions from members of the other ethnic group, leaving an individual's respective co-ethnics to "police their own." In contexts such as the one we study, such attitudes may have developed either through the conscious promotion of certain 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 for cooperation 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.

_ , Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg, "Fractionalization," *Journal of Economic Growth*, 2003, *8*, 155–194.

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.*

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, Alessandra Cassar, Julie Chytilová, and Joseph Henrich, "War's Enduring Effects on the Development of Egalitarian Motivations and In-Group Biases," *Psychological Science*, 2014, *25* (1), 47–57.

_ , 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*, 2017, *forthcoming.*

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.

Berg, Joyce, John Dickhaut, and Kevin McCabe, "Trust, Reciprocity, and Social History," *Games and Economic Behavior*, 1995, *10* (1), 122–142.

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

Besley, Timothy and Torsten Persson, "State Capacity, Conflict, and Development," *Econometrica*, 2010, *78* (1), 1–34.

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.

Camerer, Colin F., *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton, NJ: Princeton University Press, 2003.

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.

_ **and Ross Levine**, "Africa's Growth Tragedy: Policies and Ethnic Divisions," *Quarterly Journal of Economics*, 1997, *112* (4), 1203–1250.

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.

_ **, Helen Bernhard, and Bettina Rockenbach**, "Egalitarianism in Young Children.," *Nature*, 2008, *454* (7208), 1079–1083.

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.

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, "A Fine Is a Price," *Journal of Legal Studies*, 2000, *29* (1), 1–17.

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.

Government of Afghanistan, "Afghanistan - National Development Strategy 2008-2013 (1387-1391)," Technical Report, Islamic Republic of Afghanistan, Kabul, Afghanistan 2008.

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.

Katzman, Kenneth and Clayton Thomas, "Afghanistan: Post-Taliban Governance, Security, and U.S. Policy," Technical Report, Congressional Research Service, Washington, DC 2017.

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, and Antonio Vaccaro, "Trust and In-Group Favoritism in a Culture of Crime," 2014.

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

Miguel, Edward 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.

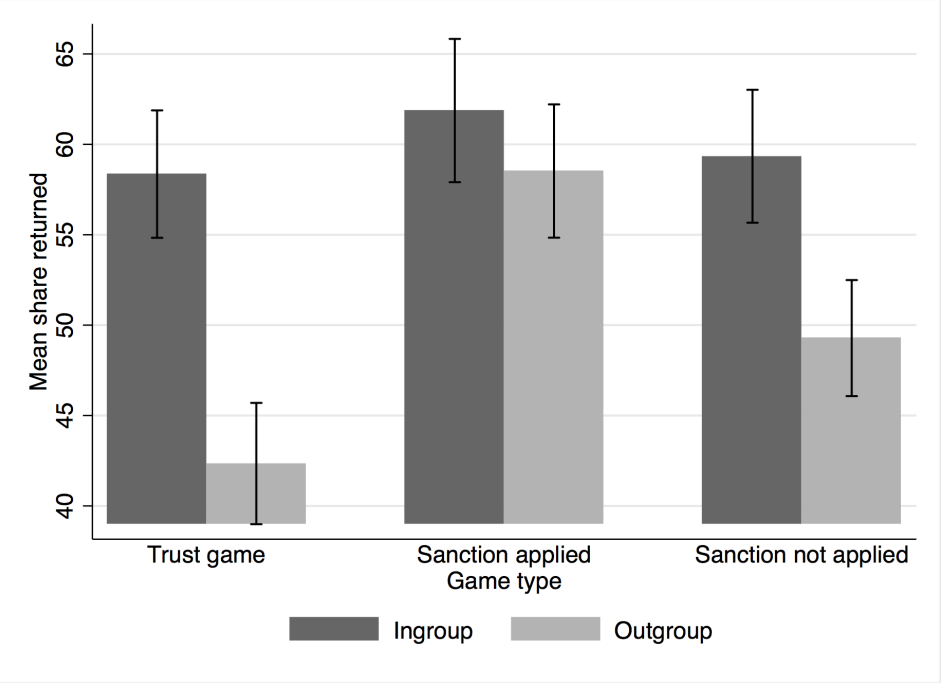
Tajfel, Henri, M G Billig, R P Bundy, and Claude Flament, "Social Categorization and Intergroup Behaviour," *European Journal of Social Psychology*, 1971, *1* (2), 149–178.

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. Error bars represent 95 percent confidence intervals.

Table 1: Summary of trustees decisions

Decision number[a]	Source of parameters[b]	Independent of treatment & included in analysis	Payoff relevant
Panel A: Trust Game			
T1	Randomly assigned	X	
T2	Randomly assigned	X	
T3	Matched partner		X
Panel B: Sanctioning Game			
S1	Randomly assigned	X	
S2	Randomly assigned	X	
S3	Randomly assigned	X	
S4	Randomly assigned	X	
S5	Matched partner		X

Note: Trustees were presented with a series of decisions in each game. They were told that only one would be payoff relevant, but that they would not know which when deciding (i.e. they were not informed that their partner's was the last one). Therefore they should have treated all decisions as potentially reflecting an ingroup or outgroup investor's intentions, according to treatment. 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 sanctioning by treatment, independent of the other parameters.

[a] The order of the games was randomized.

[b] The randomly assigned parameters are: the amount sent (s_i), the requested backtransfer (r_i^*), and in the sanctioning game, whether or not the sanction was applied (p_i).

Table 2: Individual characteristics by role and treatment

Sample	*Investors*			*Trustees*		
	Ingroup (1)	*Outgroup* (2)	Difference (1)-(2) (T-test p-value) (3)	*Ingroup* (4)	*Outgroup* (5)	Difference (4)-(5) (T-test 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 ingroup and the outgroup 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 outgroup 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 3: Trustee behavior in games: aggregate and by treatment, strategy method allocations

Sample	*Total* (1)	*Ingroup* (2)	*Outgroup* (3)	Difference (2)-(3) (Wilcoxon p-value) (4)
Trust game				
Share returned	50.20	58.35	42.34	16.01
	(23.79)	(22.78)	(22.09)	(0.00)
Observations	332	163	169	
Sanctioning game				
Sanctioning condition				
Share returned	60.16	61.87	58.52	3.35
	(24.77)	(25.38)	(24.14)	(0.23)
Observations	327	160	167	
No sanctioning condition				
Share returned	54.25	59.34	49.28	10.06
	(23.28)	(24.14)	(21.33)	(0.00)
Observations	340	168	172	

Note: Means reported in Columns 1-3. Standard deviations in parentheses. Randomly assigned parameters only. Column 4 reports the difference in means between the ingroup and the outgroup treatment. P-values of a Wilcoxon rank-sum test are reported in parentheses in Column 4. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table 4: Effect of sanctions on share returned in trust and sanctioning games across treatments

Sample	*All*	*Ingroup*	*Outgroup*
Dependent variable	Share returned		
	(1)	(2)	(3)
Ingroup	15.38***		
	(4.09)		
Sanctioning condition	13.85***	-0.01	13.92***
	(5.02)	(1.68)	(5.33)
Ingroup x Sanctioning condition	-13.34**		
	(5.23)		
No sanctioning condition	5.36***	-0.44	5.45**
	(1.99)	(1.71)	(2.27)
Ingroup x No sanctioning condition	-5.65**		
	(2.47)		
Sent	-0.16***	-0.13***	-0.19***
	(0.05)	(0.04)	(0.07)
Share requested	0.39***	0.50***	0.29***
	(0.06)	(0.09)	(0.05)
Control variables	Yes	Yes	Yes
Constant	22.66***	29.18***	32.47***
	(6.73)	(7.54)	(8.83)
Observations	999	491	508
Number of IDs	167	82	85
	F-test		
	H_0 : Sanctioning equals no sanctioning		
Ingroup p-value	0.55	0.76	
Outgroup p-value	0.05		0.05

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 sanctioning and no sanctioning condition coefficients. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table 5: Investor summary statistics: trust, sanctioning and dictator games

Sample	Total (1)	Ingroup (2)	Outgroup (3)	Difference (2)-(3) (Wilcoxon p-value) (4)
Tust game				
Amount sent	56.76	57.21	56.19	1.02
	(22.46)	(23.67)	(20.99)	(0.75)
Share requested back	0.48	0.52	0.42	0.10***
	(0.22)	(0.22)	(0.20)	(0.00)
Requested profit	122.71	130.67	112.86	17.81***
	(42.09)	(42.86)	(39.17)	(0.00)
Expected profit	117.29	122.12	111.31	10.81*
	(38.90)	(39.13)	(37.98)	(0.09)
Realized profit	119.98	131.69	106.30	25.39***
	(34.19)	(34.05)	(29.06)	(0.00)
Sanctioning game				
Sanction applied	0.37	0.37	0.38	-0.02
	(0.48)	(0.48)	(0.49)	(0.83)
Amount sent	56.28	55.96	56.67	-0.71
	(20.91)	(21.97)	(19.66)	(0.70)
Share requested back	0.52	0.54	0.49	0.05*
	(0.21)	(0.22)	(0.20)	(0.09)
Requested profit	130.05	134.81	124.17	10.64
	(43.71)	(48.45)	(36.61)	(0.41)
Expected profit	122.62	126.80	117.50	9.30
	(34.95)	(37.16)	(31.50)	(0.13)
Realized profit[a]	128.73	133.73	122.58	11.15
	(36.03)	(41.67)	(26.59)	(0.16)
Dictator game				
Amount sent	45.16	44.23	46.31	-2.08
	(25.57)	(26.50)	(24.48)	(0.55)
Differences				
Trust (Trust Sent-Dictator Sent)	11.60	12.98	9.88	3.10
	(25.56)	(28.18)	(21.93)	(0.87)
Sanction efficiency gain	-0.48	-1.25	0.48	-1.73
(Sanctioning sent- Trust sent)	(21.82)	(23.92)	(19.01)	(0.66)
Sanction-Trust requested profits	7.34	4.13	11.31	-7.18
	(51.70)	(52.61)	(50.58)	(0.13)
Sanction-Trust expected profits	5.45	4.85	6.19	-1.34
	(43.12)	(44.61)	(41.48)	(0.62)
Sanction-Trust realized profits[a]	8.79	3.15	15.76	-12.61**
	(46.88)	(53.46)	(36.40)	(0.02)
Observations	188	104	84	

Note: Means reported in Columns 1 and 2. Standard deviations in parentheses. Column 3 reports the difference in means between the ingroup and the outgroup treatment. P-values of a Wilcoxon rank-sum test are reported in parentheses in Column 3. [a] For realized profits, the sample size is 86 for the ingroup treatment and 70 for the outgroup treatment. This is due to two cancelled sessions that resulted in observations unmatched with receivers.

Table 6: Sanctions, fairness, and behavioral effects

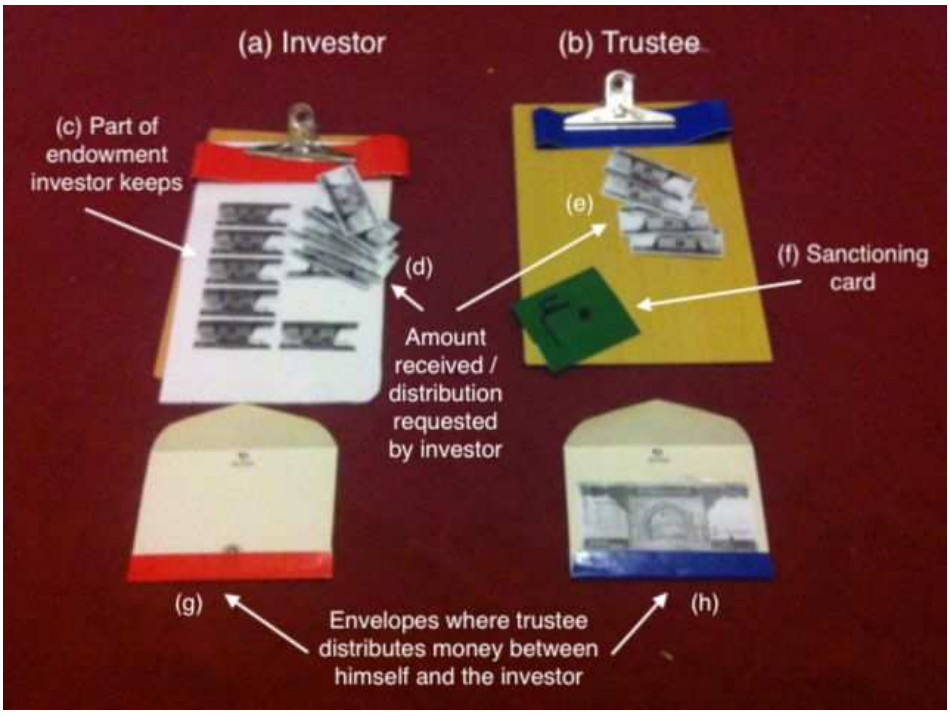
Panel A: Fairness						
Sample	*Fair request*			*Unfair requests*		
Dependent variable	*All*	*Ingroup*	*Outgroup*	*All*	*Ingroup*	*Outgroup*
	Share returned					
	(1)	(2)	(3)	(4)	(5)	(6)
Ingroup	12.95***			18.29***		
	(3.55)			(7.00)		
Sanctioning condition	9.45**	-3.76*	9.67**	19.29***	4.02	17.95**
	(4.11)	(2.16)	(4.55)	(7.37)	(3.13)	(7.82)
Ingroup * Sanctioning condition	-12.87***			-15.46*		
	(4.58)			(8.19)		
No sanctioning 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)
Ingroup * No sanctioning 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*	*Ingroup*	*Outgroup*	*All*	*Ingroup*	*Outgroup*
Dependent variable	Returned more than requested (d)			Returned less than requested		
	(1)	(2)	(3)	(4)	(5)	(6)
Ingroup	0.19***			-0.17***		
	(0.05)			(0.06)		
Sanctioning 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)
Ingroup * Sanctioning condition	-0.14**			0.17**		
	(0.06)			(0.08)		
No sanctioning 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)
Ingroup * No sanctioning 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. 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)

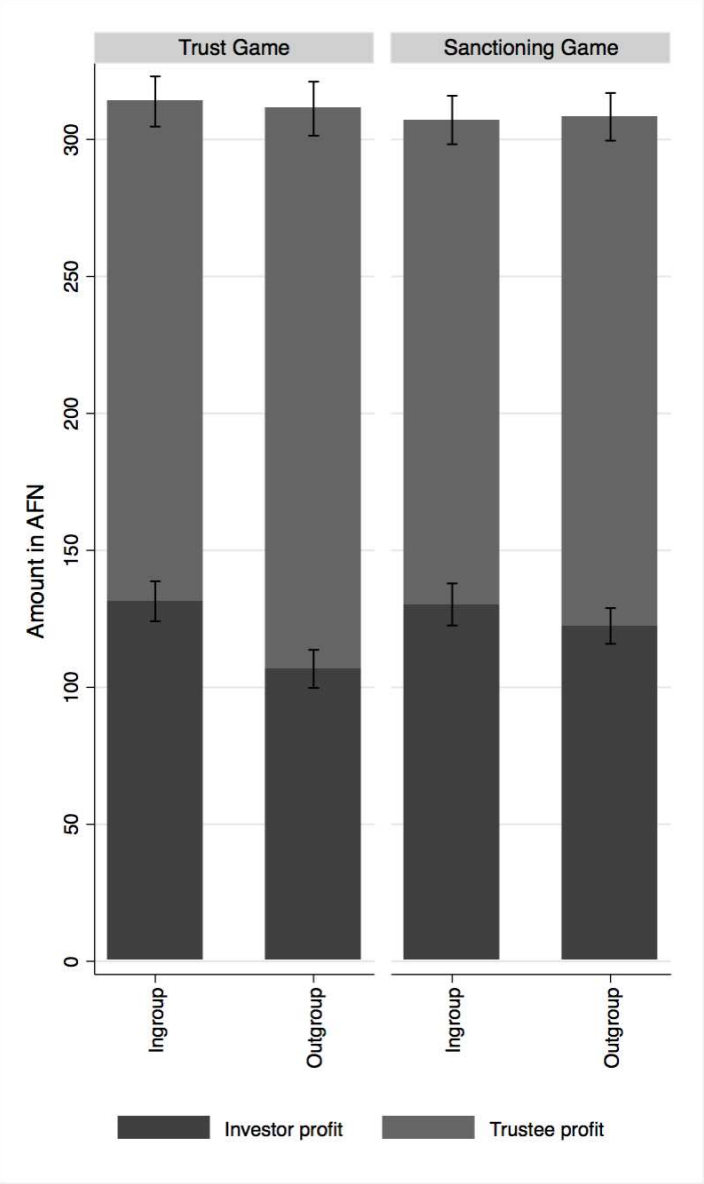
Figure A1: Participants in an individual session in a pop-up field laboratory

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



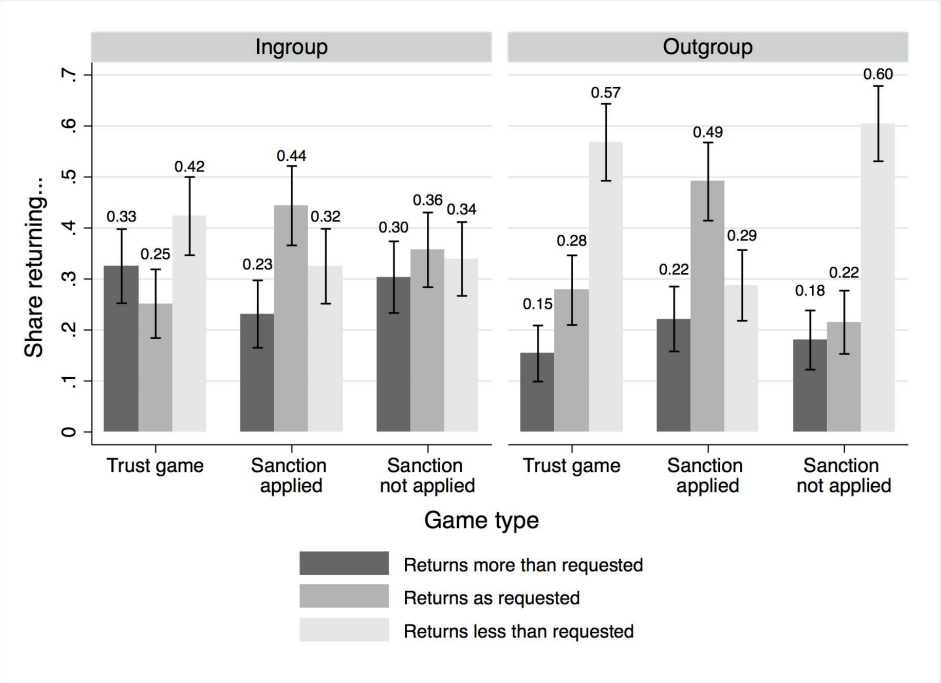
Note: The figure shows the decision-making environment for a trustee in the sanctioning game, which was designed to be easy for illiterate subjects to understand. The investor's choices were communicated using visual aids. In this example, the trustee faced a choice in which the investor had decided 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 1. Trust game choices were presented in the same way, but without the sanctioning card (f).

Figure A3: Investor and trustee profits by game and treatment



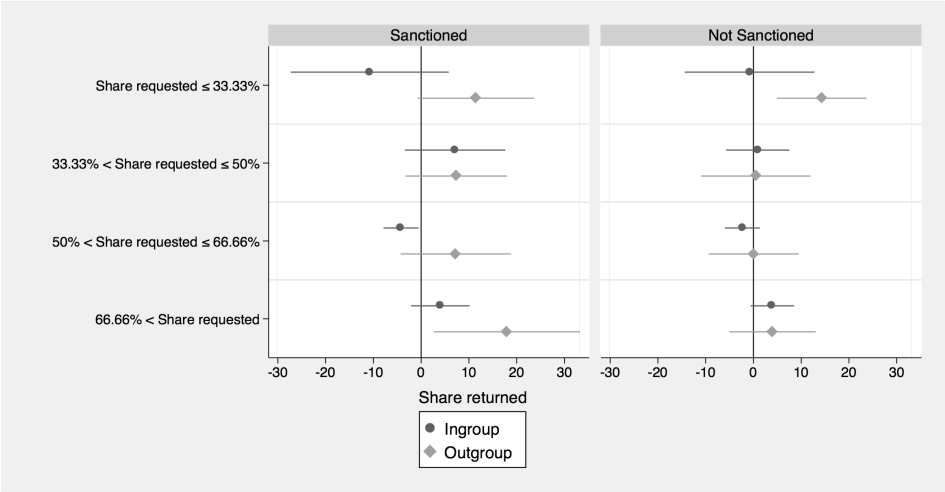
Notes: Mean cumulative profits in Afs. The investor profits are calculated as a combination of an investor's initial endowment minus the amount sent plus the average amount returned by the trustee, for all decisions in that game and group treatment in which trustees were presented with the parameters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

Figure A4: Trustees' decisions to meet investors' requests by game and treatment.



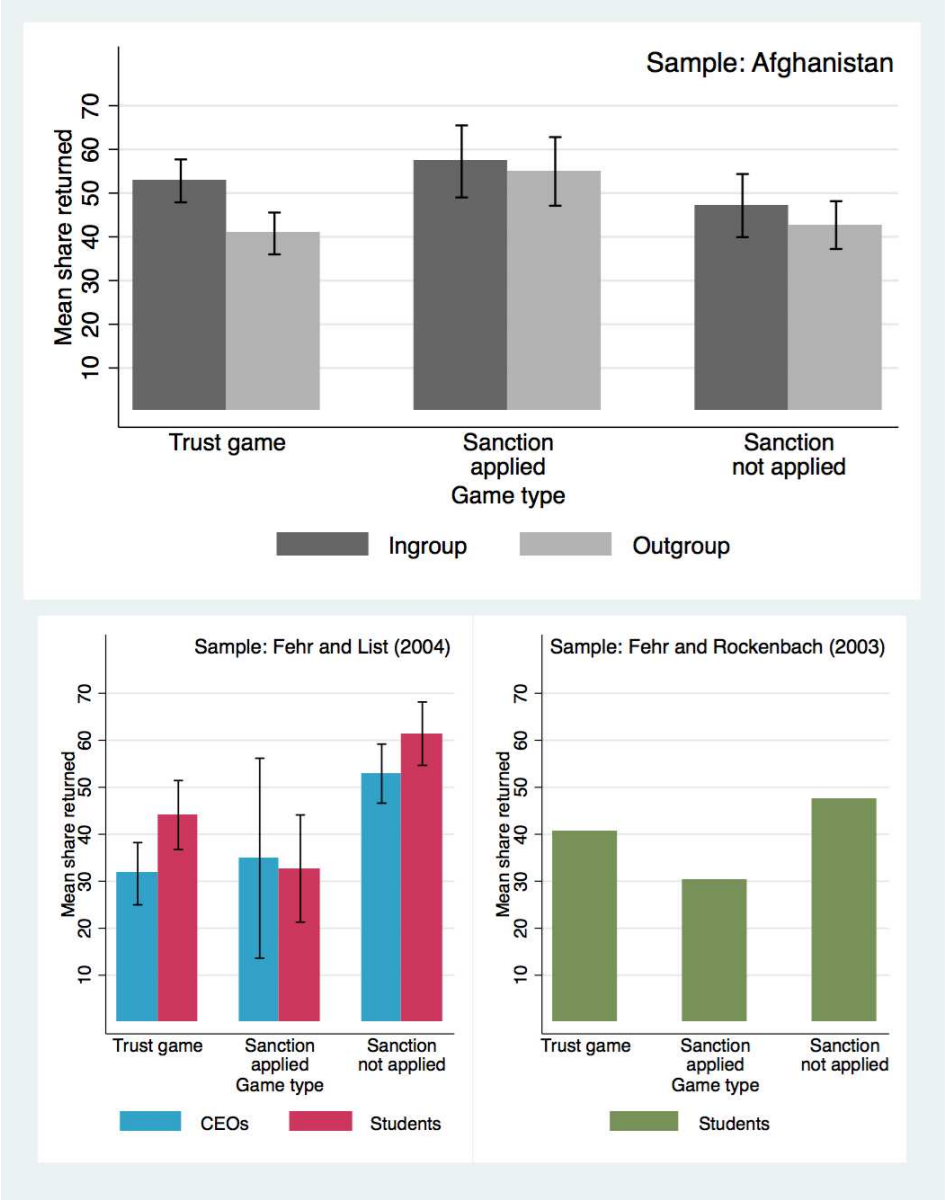
Notes: Mean back transfers in the trust and sanctioning games by treatment for randomly assigned parameters only. Returns more/as/less than requested are dummy variables representing choices in which the amount returned by the trustee exceeds/equals/is lower than the amount requested back by the investor.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

Figure A5: Coefficient plot: share returned regressed on sanctioning and no sanctioning conditions, by different ranges of share requested



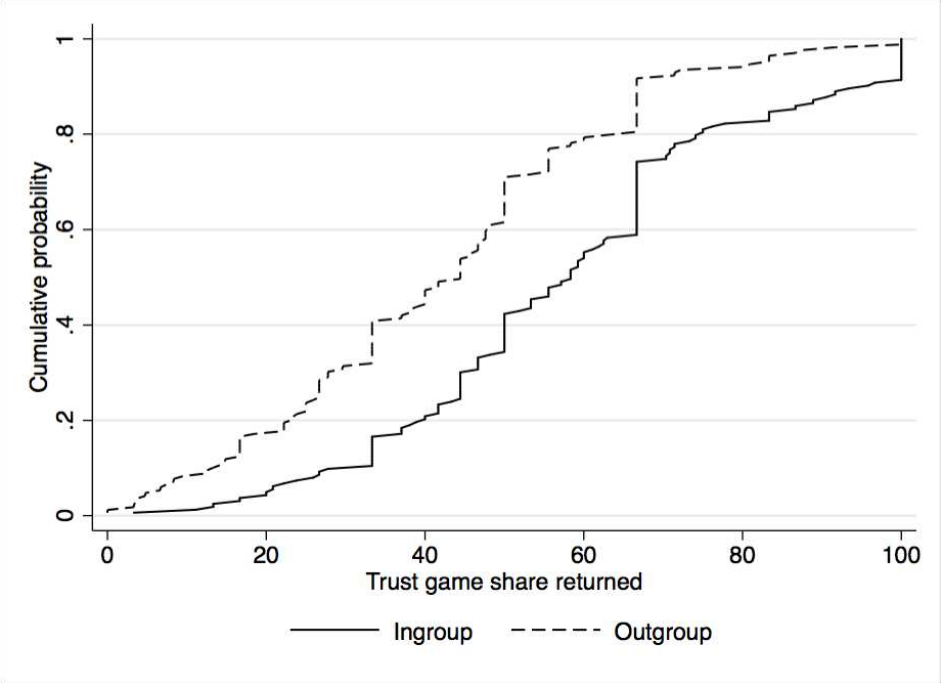
Notes: Regression results from model estimated in column 1 of Table 4. The dependent variable is the share returned in the sanctioning or no sanctioning conditions. Each mark represents the coefficient for the given treatment, for a range of requests (implying various thresholds for what might be considered fair requests). Error bars represent 95 percent confidence intervals.

Figure 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: 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	*Ingroup:* 54 Tajik (4 sessions), 70 Hazara (4 sessions) *Outgroup:* 36 Tajik (2 sessions), 53 Hazara (3 sessions)	*Ingroup:* 59 Tajik (4 sessions), 49 Hazara (3 sessions) *Outgroup:* 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	*Ingroup:* 43 Tajik (4 sessions), 61 Hazara (4 sessions) *Outgroup:* 32 Tajik (2 sessions), 52 Hazara (3 sessions)	*Ingroup:* 37 Tajik (4 sessions), 45 Hazara (3 sessions) *Outgroup:* 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 A2: Effect of sanctions 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)
Ingroup	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)
Sanctioning 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)
Ingroup * Sanctioning 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 sanctioning 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)
Ingroup * No sanctioning 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 : Sanctioning equals no sanctioning								
Ingroup p-value	0.55	0.58	0.55	0.95	0.54	0.49	0.11	0.95	0.17
Outgroup 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 sanctioning and no sanctioning condition coefficients. *** Significant at the 1 percent level, ** Significant at the 5 percent level, * Significant at the 10 percent level.

Table A3: Effect of sanctions on share returned in trust and sanctioning games across treatments: quantile regressions

Sample	*Full sample*			*Fair requests*		*Unfair requests*	
	All	*Ingroup*	*Outgroup*	*Ingroup*	*Outgroup*	*Ingroup*	*Outgroup*
Dependent variable	Share returned						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: 25th quantile							
Ingroup	15.13***						
	(3.24)						
Sanctioning condition	16.49***	-0.04	13.48***	-2.48	9.24**	6.94	16.79**
	(3.43)	(3.53)	(4.00)	(2.23)	(3.81)	(6.75)	(7.22)
Ingroup * Sanctioning condition	-16.87***						
	(5.49)						
No sanctioning condition	3.97*	-0.26	4.98*	-3.47	2.38	7.83	0.42
	(2.26)	(2.94)	(2.57)	(2.29)	(2.90)	(4.75)	(4.51)
Ingroup * No sanctioning condition	-3.59						
	(3.80)						
Constant	10.85**	8.61	29.28***	4.60	25.42***	49.09*	45.09**
	(4.92)	(7.48)	(7.50)	(6.10)	(7.48)	(25.86)	(21.38)
Observations	999	491	508	265	282	226	226
Panel B: 50th quantile							
Ingroup	15.11***						
	(2.48)						
Sanctioning condition	16.23***	2.40	15.83***	-2.61	7.94***	12.98***	28.09***
	(2.00)	(1.80)	(2.81)	(1.75)	(2.21)	(4.65)	(3.34)
Ingroup * Sanctioning condition	-14.70***						
	(3.05)						
No sanctioning condition	1.01	0.89	1.69	-5.04***	2.07	9.06*	3.11
	(2.49)	(1.79)	(2.69)	(1.73)	(2.28)	(4.92)	(3.76)
Ingroup * No sanctioning condition	-1.64						
	(3.64)						
Constant	22.98***	18.16***	35.29***	32.30***	27.00***	18.14	49.62***
	(3.61)	(4.12)	(6.53)	(6.07)	(7.09)	(23.37)	(15.73)
Observations	999	491	508	265	282	226	226
Panel C: 75th quantile							
Ingroup	13.06***						
	(2.87)						
Sanctioning condition	10.27***	-3.17*	12.91***	-5.25*	7.53**	0.81	26.00***
	(2.23)	(1.92)	(2.78)	(2.97)	(3.23)	(2.95)	(6.65)
Ingroup * Sanctioning condition	-11.58***						
	(3.74)						
No sanctioning condition	2.77	0.00	6.44*	-4.99*	2.31	2.62	18.28**
	(3.23)	(2.29)	(3.58)	(2.89)	(3.60)	(3.19)	(8.11)
Ingroup * No sanctioning condition	-2.99						
	(4.50)						
Constant	27.16***	33.78***	44.97***	51.06***	59.61***	10.26	19.68
	(5.18)	(6.05)	(7.94)	(8.32)	(8.63)	(9.13)	(20.72)
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 A4: Effect of sanctions 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	*Ingroup*	*Outgroup*	*Ingroup*	*Outgroup*	*Ingroup*	*Outgroup*
Dependent variable	Share returned						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A							
Ingroup	14.76***						
	(3.26)						
Sanctioning condition	13.89***	1.81	13.86***	-1.06	12.57***	1.26	18.70***
	(3.20)	(3.73)	(3.18)	(4.08)	(3.70)	(4.74)	(4.92)
Ingroup x Sanctioning condition	-12.02**						
	(4.93)						
No sanctioning condition	6.97**	0.92	6.76**	1.08	8.84***	-0.03	3.85
	(2.79)	(3.48)	(2.75)	(4.02)	(3.08)	(5.14)	(5.27)
Ingroup x No sanctioning condition	-5.55						
	(4.43)						
Constant	37.07***	49.80***	41.70***	34.51***	33.79***	78.91***	45.49***
	(5.62)	(12.16)	(6.32)	(12.89)	(7.54)	(21.71)	(13.28)
Observations	431	203	228	137	155	66	73
R-squared	0.11	0.05	0.12	0.12	0.12	0.18	0.25
Panel B							
Ingroup	11.68**						
	(5.54)						
Sanctioning condition	9.75*	-0.56	8.65	-5.38	8.59	6.67	17.36**
	(5.74)	(6.08)	(5.53)	(7.51)	(7.29)	(7.01)	(6.31)
Ingroup x Sanctioning condition	-11.91						
	(8.15)						
No sanctioning condition	5.63	3.08	4.83	2.00	7.79	6.11	3.33
	(4.65)	(5.52)	(4.50)	(6.77)	(5.48)	(7.29)	(6.31)
Ingroup x No sanctioning condition	-2.67						
	(7.12)						
Constant	40.39***	36.10	63.26***	18.75	50.15***	54.41	14.37
	(14.34)	(24.41)	(13.78)	(26.06)	(13.89)	(49.03)	(29.02)
Observations	129	63	66	39	42	24	24
R-squared	0.15	0.16	0.21	0.31	0.21	0.28	0.58

Note: OLS coefficients. Robust standard errors in parentheses. Randomly assigned parameters only. 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 sanctioning 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 sanctioning 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 A5: 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	1.72	16.19**	9.58	-11.59**
	(3.70)	(5.46)	(6.34)	(4.94)
Observations	185	185	185	141
R-squared	0.08	0.04	0.04	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 A6: Effect of punishment on amounts returned relative to request (by fairness)

Panel A: Returned more than requested						
Sample	*Fair requests*			*Unfair requests*		
	All	*Ingroup*	*Outgroup*	*All*	*Ingroup*	*Outgroup*
Dependent variable	Returned more than requested (d)					
	(1)	(2)	(3)	(4)	(5)	(6)
Ingroup	0.30***			0.08		
	(0.07)			(0.06)		
Sanctioning condition	0.16**	-0.11*	0.15**	0.06	0.02	0.07*
	(0.07)	(0.07)	(0.06)	(0.05)	(0.06)	(0.04)
Ingroup * Sanctioning condition	-0.25***			-0.04		
	(0.09)			(0.07)		
No sanctioning condition	0.12*	-0.15**	0.11**	0.03	0.11**	0.05
	(0.06)	(0.07)	(0.05)	(0.07)	(0.05)	(0.05)
Ingroup * No sanctioning 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	*Ingroup*	*Outgroup*	*All*	*Ingroup*	*Outgroup*
Dependent variable	Returned less than requested (d)					
	(1)	(2)	(3)	(4)	(5)	(6)
Ingroup	-0.15**			-0.22*		
	(0.06)			(0.11)		
Sanctioning condition	-0.27***	-0.06	-0.27***	-0.47***	-0.30***	-0.38***
	(0.06)	(0.06)	(0.06)	(0.09)	(0.11)	(0.07)
Ingroup * Sanctioning condition	0.21**			0.19		
	(0.09)			(0.14)		
No sanctioning condition	0.05	-0.01	0.04	-0.10	-0.27***	-0.09
	(0.06)	(0.06)	(0.06)	(0.09)	(0.10)	(0.08)
Ingroup * No sanctioning condition	-0.05			-0.13		
	(0.08)			(0.13)		
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 A7: Trustees' behavior in games: by ethnicity and treatment

	Panel A: Tajik participants		
Sample	*Ingroup* (1)	*Outgroup* (2)	Difference (1)-(2) (Wilcoxon p-value) (3)
Trust game			
Share returned	58.66	42.58	16.08
	(23.95)	(20.51)	(0.00)
Observations	74	76	
Sanctioning game			
Sanctioning condition			
Share returned	61.93	56.44	5.50
	(24.60)	(24.43)	(0.18)
Observations	72	74	
Sanction not applied			
Share returned	59.42	45.88	13.54
	(22.94)	(21.46)	(0.00)
Observations	76	78	
	Panel B: Hazara participants		
Sample	*Ingroup* (1)	*Outgroup* (2)	Difference (1)-(2) (Wilcoxon p-value) (3)
Trust game			
Share returned	58.10	42.15	15.95
	(21.90)	(23.41)	(0.00)
Observations	89	93	
Sanctioning game			
Sanctioning condition			
Share returned	61.82	60.19	1.63
	(26.14)	(23.90)	(0.66)
Observations	88	93	
Sanction not applied			
Share returned	59.28	52.11	7.17
	(25.21)	(20.91)	(0.06)
Observations	92	94	

Note: Means reported in Columns 1 and 2. Standard deviations in parentheses. Randomly assigned parameters only. Column 3 reports the difference in means between the ingroup and the outgroup treatment. P-values of a Wilcoxon rank-sum test are reported in parentheses in Column 3. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table A8: Effect of sanctions on share returned by ethnicity across treatments

Sample	*Tajik*			*Hazara*		
	All	*Ingroup*	*Outgroup*	*All*	*Ingroup*	*Outgroup*
Dependent variable	Share returned					
	(1)	(2)	(3)	(4)	(5)	(6)
Ingroup	15.45***			14.24*		
	(3.49)			(7.63)		
Sanctioning 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)
Ingroup x Sanctioning condition	-13.69**			-12.71		
	(6.79)			(8.09)		
No sanctioning 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)
Ingroup x No sanctioning 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 : Sanctioning equals no sanctioning					
Ingroup p-value	0.73	0.76		0.10	0.63	
Outgroup 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 sanctioning and no sanctioning condition coefficients. *** 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 sanction on trustees' decisions. We adapt the model of state-dependent preferences described in Bowles and Polania-Reyes (2012) to the trust game with sanctions, with the aim of providing a framework for the interpretation of our results. Our goal in this study is to determine whether sanctions crowd out (or crowd in) trustworthy behavior, and if so, whether the effect differs between the in-group and out-group treatments. The sanction, 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 sanction), and by activating preferences related to the sanction. 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 sanction was available or used.[24] When $\beta_t^g > 0$, this term captures t's state-independent altruism towards i. We assume that β_t^g varies between individuals,

[24] 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, sanctioning is orthogonal to the other parameters in our experiment and this allows us to consider the effect of sanctioning independently.

and that β_t^{In} and β_t^{Out} are identically distributed around different means.[25] 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 the application of the sanction—again, we allow this parameter to vary with group treatment, g. The parameter λ_t^g encompasses several motivations as discussed in 3.1. 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 decisons.

***Propostion 1**: If $\lambda_t^g = 0$, a trustee who retruns $r_t > r_i^*$ 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 sanction affect t's choice? Since the sanction 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 sanctioning condition and trust game (and no sanctioning 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 sanctioning condition and trust game (no sanctioning 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 sanction crowds in or crowds out trustworthiness, respectively).

[25] 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 sanction will lead to a larger decrease the frequency of subjects who return less than the requested amount ($r_t < r_i^*$) in the Ingroup treatment than in the Outgroup 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 sanction, the trustee's utility would be maximized by returning less than the requested amount). How does introducing the sanction 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 sanction, 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 sanctioning condition): $\pi_t(r_t = r_t \tilde{-} f | p_i = 1, \tilde{r_t} < r_i^*) = \pi_t(\tilde{r_t}|p_i = 0, \tilde{r_t} < r_i^*)$. 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.[26]

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 sanction, relative to the trust game (or no-sanctioning 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 sanction 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 A4 and Table 6 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 the sanction, systematically, by treatment.

[26] 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 sanctioning depend on β_t^g in a similar fashion. It is also possible that the best response might be unchanged under sanctioning.