Dhami, Sanjit; Wei, Mengxing

Working Paper

Norms, Emotions, and Culture in Human Cooperation and Punishment: Theory and Evidence

CESifo Working Paper, No. 10220

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Dhami, Sanjit; Wei, Mengxing (2023) : Norms, Emotions, and Culture in Human Cooperation and Punishment: Theory and Evidence, CESifo Working Paper, No. 10220, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/271864

CESifo WORKING PAPERS

10220
2023

January 2023

Norms, Emotions, and Culture in Human Cooperation and Punishment: Theory and Evidence

Sanjit Dhami, Mengxing Wei

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10220*

Norms, Emotions, and Culture in Human Cooperation and Punishment: Theory and Evidence

Abstract

We consider the psychological and social foundations of human contributions and punishments in a voluntary contributions mechanism with punishment (VCMP). We eliminate 'dynamic economic linkages' between the two stages of our 'modified' VCMP to rule out other potential explanations. We use a beliefs-based model, rooted in psychological game theory, to derive rigorous theoretical predictions that are then tested with pre-registered experiments in China and the UK. Social norms, culture, and endogenous emotions are the key determinants of contributions and punishments. The emotions of shame, frustration, and anger, play a key role in our theoretical and empirical analysis through 'dynamic psychological linkages'. We provide potential microfoundations for the inherent human tendency to follow social norms and punish norm violators, while respecting boundedly rational strategic decision making.

JEL-Codes: C910, C920, D010, D910.

Keywords: cooperation and punishment, emotions-shame, frustration, anger, social norms, culture, bounded rationality.

Sanjit Dhami
Division of Economics, School of Business
University of Leicester
London Road
United Kingdom – Leicester, LE2 1RQ
sd106@leicester.ac.uk

*Mengxing Wei**
School of Economics, The Laboratory for Economic Behaviors and Policy Simulation
Nankai University
94 Weijin Road
China – 300071 Tianjin
mengxing.wei@hotmail.com

*corresponding author

January 9, 2023

1 Introduction

Social dilemmas are situations where it is in the private interest to free-ride but it is in the social interest to cooperate. The leading example in economics is the *voluntary contribution mechanism with punishment* (VCMP). In a one-shot VCMP game, in Stage 1, players simultaneously choose contributions towards a pure public good. In Stage 2, having observed the contributions at the end of Stage 1, players simultaneously decide on the level of costly punishment to inflict on others. There is typically a high level of contributions in a repeated VCMP and punishment of low contributors in Western subject pools (Fehr and Gächter, 2000; Dhami, 2019, Vol. 2). However, our interest is in a one-shot VCMP.[1]

Several factors facilitate cooperation in a VCMP game.[2] Of these, *social norms*, backed by the potential to sanction and punish, provide a persuasive explanation of historic and contemporary human cooperation without recourse to an external agency (Ostrom, 1990; Ostrom et al., 1992; Bicchieri, 2006; Bowles and Gintis, 2011; Elster, 2011; Fehr and Schurtenberger, 2018a). Social norms may also serve to coordinate expectations towards a cooperative solution. *Shame*, experienced by norm violators is the typical emotional mechanism through which norm compliance is supported (Bicchieri, 2006; Bowles and Gintis, 2011; Elster, 2011). However, social norms may be difficult to enforce in the absence of punishments and sanctions (Fehr and Schurtenberger, 2018a,b).

There are several explanations of punishments in VCMP games. These include *negative reciprocity* (Fehr and Gächter, 2000); *inequity averse preferences* (Fehr and Schmidt, 1999); *frustration and anger* (Battigalli et al., 2019); *innate proclivity to punish norm offenders on account of indignation and outrage* (Bosman and van Winden, 2002; Bosman et al., 2005; Ben-Shakhar et al., 2007; Xiao and Houser, 2005; Hopfensitz and Reuben, 2009; Joffily et al., 2013); and a feeling that norm violators *deserve punishment* (Carlsmith et al., 2002). In neuroeconomics, punishment of norm violators also leads to the *revenge is sweet* finding (de Quervain et al., 2004).[3]

In this paper, we are interested in the psychological and social determinants of cooperation and punishments in a 'modified' one-shot VCMP game. Our 'modification' lies in minimizing/eliminating the *dynamic economic linkages* between the two stages of a VCMP by (i) having a separate endowment in each stage, and (ii) paying subjects only the payoffs from one of the two stages, chosen randomly. We propose that the determinants of cooperation and punishment lie in *social norms of contributions*, *cultural differences*, and *dynamic psychological linkages* between the two stages of a VCMP that rely on the emotions of frustration, anger, and

[1] In non-Western subject pools, antisocial punishments have been documented in repeated versions of the VCMP game (Herrmann et al., 2008). For a survey of the large literature on cooperation in repeated prisoner's dilemma games see Dal Bó et al. (2018). We consider cooperation and punishments in static games and abstract from reputational concerns.

[2] These factors include a combination of rewards and punishments, intergeneration advice, matching like-minded subjects who contribute high amounts, non-monetary punishments, and face to face communication; see Dhami (2019, Vol. 2) for a survey.

[3] This literature is mainly empirical, but there are two important exceptions; we consider the differences of this work from our paper in more detail later. Battigalli et al. (2019) do not deal with the class of games that we consider here. Fehr and Schmidt (1999) do not focus on a theory of social norms using models in psychological game theory.

shame. We develop an overarching theoretical framework that formally models the underlying norms and emotions, using the machinery of psychological game theory. We then successfully test our predictions using a stringent empirical test with data from Chinese and UK subjects.

1.1 Evidence on punishments and emotional states

Important insights already exist in psychology and neuroscience for why people might act differently when they are in anger, and the effect of anger on actions.[4] Self-reported anger, and anger measured by electrophysiological measures, such as the skin conductance rate (SCR), is associated with punishments in games where one party may perceive that it has been treated unfairly in the past.[5] Venting of anger in a two-stage VCMP game after players observe the contributions of others, but before they punish, reduces punishment and has a net positive effect on welfare (Dickinson and Masclet, 2015). The SCR only measures the level of emotional arousal, but not the kinds of emotions experienced (e.g., anger or frustration). Thus, short of neuroeconomic methods that correlate brain activation areas with a typology of emotions, particularly anger (Klimecki et al., 2019), self-reported measures of emotions are the only realistic possibility, and this is the method we follow.[6]

Exogenously induced emotional states, such as anger, induce subjects to contribute less and punish more (Bartke et al., 2019; Drouvelis and Grosskopf, 2016; Hopfensitz and Reuben, 2009). However, the affective states in our model are *endogenous*, and arise after observing the contributions of partners at the end of Stage 1. There is also other indirect evidence that affective states such as anger might be at play in determining punishments in VCMP.[7]

Emotions can provide a credible signal about punishments even in one shot games or those that are repeated only a few times (Hirshleifer, 1987; Frank, 1988). Fehr and Gächter (1999) give players hypothetical scenarios about the contribution of other players and their own contributions and asked players to rate the intensity of their emotions. They show that free-riding creates strong negative emotions that depend on the extent of free-riding, and this is anticipated by other players. They argue that "emotions are guarantors of credible threats."

1.2 Anger, frustration, punishment, and beliefs

Battigalli et al. (2019) identify the *expectations-reality gap* as an important link between emotions and punishment. In the class of leader-follower games, they argue that when others do not meet their expectations, players get frustrated. This triggers anger and blame which induces

[4]For surveys of the effect of emotions and moods on processing information, cognitive capacities, and on motivations, see Isen (1987), Drouvelis and Grosskopf (2016), Bartke et al. (2019), Dhami (2020, Vol. V7). For the neuroeconomic evidence, see Sanfey et al. (2003) and de Quervain et al. (2004).

[5]See, for instance, Bosman and van Winden (2002), Bosman et al. (2005), Ben-Shakhar et al. (2007), Puurtinen and Mappes (2009), and Hopfensitz and Reuben (2009).

[6]In the 'power to take' game, Ben-Shakhar et al. (2007) find that SCRs were highly correlated with self-reported measures of anger, but not with other self-reported emotions. More angry individuals destroyed more, and those who destroyed more, self-reported higher anger. Thus, anger was the primary emotion that supported a response to unfair behavior of the partner.

[7]Carpenter and Matthews (2012) consider both second and third party punishment. They conjecture that these punishments are driven, respectively, by anger and indignation. In a VCMP, Masclet and Villeval (2008) showed that low contributions invite punishments even when such punishments cannot reduce income inequality.

them to punish partners.[8] Empirical testing of this mechanism is at an early stage.[9]

In a VCMP, this insight implies that players who contribute below the expectations of group members in Stage 1, induce anger and frustration among group members who then wish to punish them in Stage 2. This gives rise to *dynamic psychological linkages* between the two stages, even if there are *no dynamic economic linkages* between the two stages, as is the case in our 'modified' VCMP. However, there are at least two important differences in our model from Battigalli et al. (2019). First, they only deal with leader-follower games of perfect information, while we have a two-stage game in which both players simultaneously choose their actions in each stage.[10] Second, we do not impose the conditions necessary for a sequential equilibrium (SE).[11] We believe that the most tenable assumption, based on the evidence from behavioral game theory, is that players best respond to their beliefs in each stage (Camerer, 2003; Dhami, 2020, Vol. 4; Battigalli and Dufwenberg, 2022). Our solution concept of a *psychological best response* is related to the notion of rationalizability (Khalmetski et al., 2015; Dhami et al., 2019; Dhami et al., 2022; Dhami et al., 2023, Battigalli and Dufwenberg, 2022). In this sense, we assume that our players are 'boundedly rational' and we believe that this provides a description of economic reality that is in much better conformity with the available evidence (Dhami and Sunstein, 2022).

Players might also be inequity averse (Fehr and Schmidt, 1999). High contributors in Stage 1 may wish to punish their partners in Stage 2 to reduce income inequality 'over the sum of payoffs for the two stages' (Fehr and Gächter, 2000).[12] There are no explicit emotions in inequity-averse preferences, so there are no explicit dynamic psychological linkages between the two Stages. It is not clear what the predictions of the inequity aversion model are in our 'modified' VCMP where there are no dynamic economic linkages. Hence, as far as we can see, our paper has no bearing on the claims/implications of the inequity aversion model. Furthermore, we derive beliefs-based predictions of our model that cannot arise under inequity-aversion; and these predictions are confirmed by the evidence.

[8] They write (p.16): "*Insights from psychology about the triggers and repercussions of anger are evocative... and the action tendency of anger is aggression and the urge to retaliate. Angry players may be willing to forgo material gains to punish others...*"

[9] Persson (2018) finds that the expectations-reality gap induces frustration and anger, but does not affect actions. Aina et al. (2020) report evidence supportive of the Battigalli et al. (2019) approach.

[10] Battigalli et al. (2019, p.31), write: "*An important task for future work is to explore whether and how frustration and anger may matter in other games, e.g., where many players move simultaneously in the first stage, with multiple stages, or with incomplete information.*" They also suggest (2019, p.31) that: "*In addition, one may want to explore other solution concepts than SE [sequential equilibrium].*"

[11] Battigalli and Dufwenberg (2022) write: "*Economists frequently take for granted that ad hoc notions of equilibrium (whereby players are assumed to have correct beliefs) meaningfully describe strategic interaction. Often such assumption are not well justified... Only in rare cases is it justified to assume that a sequential equilibrium will be played, most notably when this solution concept yields the same prediction as rationalizability, or self-confirming equilibrium.*" This assertion is well supported in behavioral game theory (Camerer, 2003; Bellemare et al., 2011; Mauersberger and Nagel, 2018; Eyster, 2019; Dhami, 2020, Vol. 4; Dhami, 2020, Vol. 5; Battigalli and Dufwenberg, 2022).

[12] Frustration based on inequity aversion may also be critical in the determination of punishment and cooperation in close primate relatives (Vale and Brosnan, 2017).

1.3 Social norms

Evolutionary and cultural selection has self-selected individuals who engage in costly second party and third party punishment of norm violators and have an intrinsic desire to follow norms (Ostrom, 2000; Bowles and Gintis, 2011; Gintis, 2017). This also arguably underlies the presence of formal and institutional punishment. It has been suggested that the violation of social norms elicits anger and that anger is important to uphold human cooperation because it produces a credible threat that violators will be punished (Trivers, 1971; Fessler and Haley, 2003). Our work is also related to providing microfoundations to this idea.

Successful norms require a combination of three elements that create a shared understanding of social behavior.[13] (1) *Empirical expectations*: Beliefs about the likely actions of other group members, based on direct observations of the "actual behavior" of others. (2) *Normative expectations*: Beliefs about the "normatively desired behavior" expected by the social group. (3) *Sanctions/Punishments* of norm violators.[14] In the absence of punishments, normative expectations may be much less potent (Fehr and Schurtenberger, 2018b). Hence, we employ all three components of social norms and ensure that empirical and normative expectations are aligned. We also formally model the insight of Bicchieri (2006) and Elster (2011) that subjects experience the emotion of shame from violating a social norm. By varying a public signal of the normative expectations of the social group, we provide a direct test of our theory.

1.4 Cross-cultural differences

Differences in cultural values are likely to shape individual prosocial and cooperative behavior (Hofstede, 1980; Henrich et al., 2010; Luria et al., 2015; Bartling et al., 2015; Martí-Vilar et al., 2019).[15] For this reason, in our experiments, we pit the predictions of our theoretical model against data from Chinese and UK subjects. Systematic differences between the two societies are still in the process of being understood and our paper also contributes to this literature.

1.5 Our paper and main results

We play a VCMP game with Chinese and UK subjects in separate experiments conducted in the two countries. We begin by deriving the theoretical predictions from a model that combines the literature on social norms with the literature on emotions. We use the machinery of psychological game theory that uses belief hierarchies to rigorously model underlying social norms and emotions. Our theoretical model then closely guides our empirical design.

[13] For a brief guide to the literature, including the recent literature on measurement of social norms which is not the subject of our study (and for that reason we do not cite it), see Fehr and Schurtenberger (2018a), and Dhami (2019, Vol. 2, Section 5.7). See also Bicchieri (2006) for an extensive exposition of these ideas.

[14] The alignment of empirical and normative expectations is essential in the formation of successful social norms. When empirical and normative expectations are in conflict, empirical expectations are relatively more important (Bicchieri, 2006; Bicchieri and Xiao, 2009). d'Adda et al. (2020) and Bicchieri (2017) consider uncertainty or partial knowledge about the relevant norm. We also allow for this possibility.

[15] Henrich et al. (2010) make a powerful case for cross cultural comparisons of results from experiments. After arguing that most established results in the social and behavioral sciences arise from an unusual subject pool of WEIRD (Western, Educated, Industrial, Rich, Democratic) subjects, they write (p.1): "Overall, these empirical patterns suggests that we need to be less cavalier in addressing questions of human nature on the basis of data drawn from this particularly thin, and rather unusual, slice of humanity."

Prior to their contribution decisions in Stage 1, subjects receive signals of empirical expectations and normative expectations from their social group. In Stage 1, subjects may suffer from the emotion of shame if they contribute below a contributions norm, as inferred from a public signal of normative expectations from the social group. In Stage 2, subjects may suffer from frustration and anger if they find that the Stage 1 contributions of the partner fell below expectations (expectations-reality gap). We have 5 different treatments that (i) switch-on/switch-off social norms, and vary the signals of normative expectations (low and high signals); and (ii) switch-on/switch-off punishments (this takes the form of removing Stage 2 from a VCMP). The contrasts between the treatments then allow us to test our theoretical predictions.

The results are as follows.

1. *Results on Stage 1 contributions*: We predict, and confirm, that the availability of the punishment option, and a higher signal of normative expectations enhance contributions in 7/8 cases, as predicted by our model. There are culture-specific differences between Chinese and British subjects in the reaction to norms and to the availability of the punishment option. For instance, on average, in the absence of a punishment mechanism, Chinese subjects are more sensitive to changes in the signal of normative expectations from the social/peer group, relative to the British subjects. Social norms influence a mediator variable, expected contributions of the partner, which in turn significantly influences the contribution choices of players. We also observe significant cultural-gender differences. The underlying beliefs mirror observed actions and there are important cultural differences between the Chinese and the British subjects.

2. *Results on Stage 2 punishment decisions*: Players punish their partners more, the more the partner's contributions fall short of their expectations, i.e., the more frustrated they are. This arises in the absence of any dynamic economic linkages between the two stages of the VCMP, and confirms the frustration-aversion channel identified in Battigalli et al. (2019). In the presence of social norms, the average punishment chosen by the Chinese subjects is higher than that of the British subjects. But in the absence of social norms, there are no significant differences in punishment between the two. Punishment is also found to be increasing in the expected punishment choice of the partner. This suggests a form of *contemporaneous anticipated revenge*. Subjects expected to be punished less, the higher are their contributions relative to the partner. But they expected to be punished more, the greater was the shortfall in their contributions relative to the social norm. There are important cultural-gender differences in the beliefs about expected punishment, and in the chosen level of punishment.

3. *Results on emotions and the desire to punish*: When partners contribute below the expectations of players, the latter feel not just anger and frustration, but also indignation and dissatisfaction at the partner. More frustrated and angry subjects punished more, confirming our theoretical predictions. We find evidence for the *venting hypothesis* (Dickinson and Masclet, 2015). Players experience a decline in the strength of frustration, anger, and indignation after punishing the partner at the end of Stage 2, relative to the strength of

emotions they felt at the end of Stage 1.

Frustration, anger, indignation, and dissatisfaction are positively correlated with the choice of punishment. However, two positive emotions, elation and satisfaction, are negatively correlated with the choice of punishment. The reduction in 'satisfaction' arising from 'costly' punishment of the partner is statistically significant for Chinese subjects, but not the UK subjects. However, the 'aversion from expected punishment' received from the partner is statistically significant in explaining the reduction in 'elation' for UK subjects, but not the Chinese subjects.

The plan of the paper is as follows. Section 2 describes the model, which includes a description of the preferences, beliefs, belief-updating, and the sequence of moves. Section 3 describes Stage 1 and Stage 2 preferences. Section 4 explains our solution concept (psychological best response to beliefs), and derives the optimal contribution and punishment decisions. Section 5 describes our experimental design. Section 6 gives the results on the choice of punishments and the beliefs about punishments; while Section 7 gives the analogous results on contributions and beliefs about contributions. Section 8 gives the results on emotions and punishments. Finally Section 9 concludes. All proofs are in the Appendix. The supplementary section provides theoretical extensions, such as to sequential conditional reciprocity, further statistical analysis, and the experimental instructions.

2 The Model

Consider a two-stage public goods game with punishment (or a VCMP), and two players $N = \{1, 2\}$. In some treatments, the second stage is missing, so there is no punishment option. We use the index $i = 1, 2$ for a player, and the index j for the partner, $j \neq i$.

2.1 Stage 1 utility

Each player has an identical initial endowment of $y > 0$, and they simultaneously choose contributions $g_i \in [0, y], i = 1, 2$, towards a pure public good, G. The remaining endowment is used for private consumption, $c_{i1} = y - g_i$. The production technology of the public good is linear, $G = g_i + g_j$. The utility function of player $i = 1, 2$ in Stage 1, is denoted by u_{i1}.[16] It is additively separable in private consumption, c_{i1}, and public goods consumption, G,

$$u_{i1}(c_{i1}, G) = v_i(c_{i1}) + rG;\ 0 < r < 1,\ i = 1, 2, \tag{2.1}$$

where r is the return on a unit of the public good that accrues to each player; $v_i : [0, y] \longrightarrow \Re$ is strictly increasing and strictly concave, so $v_i' > 0$, $v_i'' < 0$; and $v_i(0) = 0$. Substituting $c_{i1} = y - g_i$ and $G = g_i + g_j$ in (2.1), the Stage 1 utility of player is

$$u_{i1}(g_i, g_j) = v_i(y - g_i) + r(g_i + g_j),\ i = 1, 2. \tag{2.2}$$

At the end of Stage 1, player $i = 1, 2$ receives the private and public consumption bundle (c_{i1}, G) and the contributions of both players are publicly announced.

[16] We distinguish between *utility* or *economic utility* (lowercase 'u') and *psychological utility* (uppercase 'U') in this paper. We formalize the psychological utilities of players in Section 3.

2.2 Stage 2 utility

At the beginning of Stage 2, player $i = 1, 2$ receives a fresh endowment, $y > 0$, independent of the Stage 1 outcomes; this is a substantively important difference in our model from previous work in this area. Both players simultaneously decide on the level of costly punishment, $p_i \in [0, \overline{p}]$, $i = 1, 2$ to inflict on each other, where $\overline{p} > 0$ is the maximum punishment. Inflicting a unit of punishment on the partner costs a player, $0 < \kappa < 1$; $1/\kappa$ is known as the efficiency of punishment. The remaining endowment, net of the cost of punishment, κp_i, and the punishment inflicted by the partner, p_j, constitutes second stage consumption, c_{i2}, of player i. Hence,

$$c_{i2} + \kappa p_i = y - p_j. \tag{2.3}$$

The Stage 2 utility of player $i = 1, 2$ is $u_{i2}(c_{i2}, p_i, p_j) = v_i(c_{i2})$, where v_i is defined in (2.1). Using (2.3), we get

$$u_{i2}(p_i, p_j) = v_i(y - \kappa p_i - p_j). \tag{2.4}$$

2.3 Dynamic economic and psychological linkages

In the typical experiments on public goods games with punishment (Fehr and Schmidt, 1999; Fehr and Gächter, 2000), an initial endowment y is given only once to players, at the beginning of Stage 1. It is used for contributions in Stage 1 and for Stage 2 punishments. For instance, Fehr and Schmidt (1999, p. 439) use the following representative version of payoffs in this literature for player i that adds the payoffs from the two stages of the VCMP.

$$u_i = y - g_i + rG - \kappa p_i - p_j.$$

This allows for dynamic economic linkages between the two stages. Using this structure, the inequity aversion model of Fehr and Schmidt (1999) provides a powerful motive for players to punish each other and reduce the (sum of) payoff differences from both stages. The explanation of punishment in public goods games is challenging for several other models of other-regarding preferences.

However, in our 'modified' VCMP we have eliminated the dynamic economic linkages between the two stages. One of the stages is randomly chosen at the end to pay off the players. Subjects either receive the Stage 1 payoffs, $y - g_i + rG$, or the Stage 2 payoffs, $y - \kappa p_i - p_j$, but not from both stages. The endowments are also separate in both stages. To apply inequity-averse preferences to our VCMP, one requires further auxiliary assumptions that somehow combine the Stage 1 and Stage 2 payoffs. It is not clear what these testable auxiliary assumptions might be. Thus, our results have no direct bearing on inequity-averse preferences, but allow us to explore complementary emotions-based explanations of punishments in a modified VCMP where the two stages are decoupled.

A central comparative static prediction of our model, which we test successfully in our data, is the effect of the signal of normative expectations of the social/peer group, s, that works through the underlying belief hierarchies of players. However, this signal is not predicted to have any effect in the inequity aversion model, where belief hierarchies do not play any role.

2.4 Sequence of moves

Pairs of subjects were randomly matched together to play the VCMP. The following sequence of moves are used in our theoretical model, and were closely implemented in our preregistered experiments.

Stage 1 (Public goods contributions)

1. Each player is given an endowment $y > 0$. Players observe a public signal $s \in S$ of the *normative expectations* of their peer/social group about the levels of contributions others 'ought' to make (see Section 2.5, below); $S \in [0, y]$ is the set of all possible public signals.
2. Players make an incentive compatible guess of the expected contributions of the partner.[17]
3. Players simultaneously choose their actual contributions, following which the contributions of the players, and the respective material payoffs are publicly announced.
4. The following emotions of players are elicited on a 7 point Likert scale: Frustration, Anger, Indignation, Shame, Elation, Satisfaction, Dissatisfaction.

Stage 2 (Punishment decisions[18])

1. Each player is given a fresh endowment $y > 0$ and asked to guess the punishments they expect the partner to choose, in an incentive compatible manner.[19]
2. Players simultaneously choose costly punishments to levy on each other.
3. The same emotions as those at the end of Stage 1 are again elicited once a player chooses the punishment level, but before knowing the punishment chosen by the partner.

One of the two stages is chosen at random and only the material payoffs in that stage are paid to the subjects. Subjects do not receive the material payoffs for the other stage.

2.5 Beliefs of the players

Our beliefs-based model requires the formal modeling of *belief hierarchies*. (1) We need *positive belief hierarchies* to formally model emotions such as frustration and anger (see Sections 2.5.1 and 3.1). (2) *Normative belief hierarchies* are required to formally model social norms and the emotion of shame (see Sections 2.5.2 and 3.2).

2.5.1 Positive beliefs

The initial beliefs of players about the partner's expected contributions to the public good in Stage 1, before any player chooses their contributions, are known as *first order positive beliefs*. Denote the probability distribution of these beliefs for player i by $f_i^1 : [0, y] \to [0, 1]$; where

[17] Correct guesses are rewarded with extra tokens; see Section 5 and the experimental instructions.

[18] We also have treatments in which there is only Stage 1, but no Stage 2, i.e., no punishment option. Proposition 2(c) gives the relevant result on Stage 1 contributions when there is no Stage 2.

[19] In the experiments, this is implemented as follows. Among all correct guesses of the partner's punishment decisions, one is chosen randomly and given an additional prize of 5 tokens. If nobody guessed correctly, then the closest guess is given 2 extra tokens. If there are several such guesses one is chosen randomly.

superscript '1' denotes the order of the beliefs, and subscript i denotes the index for the player. The corresponding cumulative distribution is given by $F_i^1 : [0, y] \to [0, 1]$. Thus, $f_i^1(\widetilde{g}_j)$ is the unconditional probability assigned by player i that the contribution of player j equals $\widetilde{g}_j$, before the contribution decision is made.

2.5.2 Normative beliefs

Two sorts of beliefs play a key role in a rigorous analysis of norms (Bicchieri, 2006; Elster, 2011; Fehr and Schurtenberger, 2018a). The *empirical expectations* of player A are A's beliefs about the likely "actions" of others (say, player B) before they observe the actions. These beliefs are typically based on the "actually observed" actions of others in the past.[20] The *first order normative expectations* (or normative beliefs) of a social/peer group about any group member, say, A, are their beliefs about the actions that A 'ought' to take.[21] Successful social norms require that the empirical expectations and the second order normative beliefs of players should not be in conflict. Thus, what players observe others actually doing is also what they believe is the normative injunction of the social group; we ensure this is the case in our experiments.[22] Social norms also often require that the actions of players are observed by their peers, so that norm violators face, and expect, social sanctions from the peer group (Bicchieri, 2006; Fehr and Schurtenberger, 2018a); this is also a feature of our VCMP. We now formalize this concepts.

The relevant social/peer group has beliefs, or injunctions, about how much group member i "ought" to contribute to the public good. These are the *first order normative beliefs* of the peer group, and they may be possibly, but not necessarily, degenerate.[23] Since these beliefs are not directly observed by the players (or observed with noise/error), players need to form "beliefs about the first order normative beliefs of the social/peer group." Such beliefs of the players are known as *second order normative beliefs*. The probability density of unconditional second order normative beliefs held by player i is given by $h_i^2 : [0, y] \to [0, 1]$, and the corresponding cumulative distribution is given by $H_i^2 : [0, y] \to [0, 1]$. Thus, prior to the contribution decisions, $h_i^2(\widetilde{g})$ is the unconditional probability assigned by player i that the peer group expects player i "ought" to contribute $\widetilde{g}$.

2.5.3 Belief updating

Both players observe an identical public signal $s \in S$ of the unknown underlying first order normative belief distribution of the peer group. The signal, s, could be the mean, median, or

[20] In our experiments, we directly inform players of the actions of their peers in similar public goods games in the past. Hence, we do not need to formalize empirical expectations any further.

[21] Empirical and normative expectations are also sometimes referred to as *descriptive norms* and *injunctive norms*. The beliefs of player A about the first order normative beliefs of the social group are known as the *second order normative beliefs* of player A. For the formal terminology and a brief introduction to the literature on norms, see Dhami (2019, Vol. 2, Section 5.7).

[22] These expectations would be in conflict, for instance, if the empirical expectation is the action a_1 but players expect that the normative injunction is the action $a_2 \neq a_1$. In corrupt societies, one observes corruption (empirical expectations), but the expectation about the normative injunction is that people 'ought' not to engage in corruption (Bicchieri, 2006). When these expectations are in conflict, empirical expectations have been shown to play a relatively more powerful role (Bicchieri and Xiao, 2009).

[23] An example of degenerate beliefs is that the social group might expect that group members should, with probability 1, contribute 12 out of their 20 tokens of endowment towards the public good.

mode of the underlying first order normative belief distribution. Once s is announced to all players (and this is mutual knowledge), then player $i = 1, 2$ updates the unconditional belief distributions as follows. This occurs prior to the Stage 1 contribution decision.

1. The first order unconditional positive belief density $f_i^1(.)$ is updated to the conditional first order belief density $f_i^1(. \mid s) : [0, y] \to [0, 1]$; the cumulative distribution is $F_i^1(. \mid s) : [0, y] \to [0, 1]$. Thus, $f_i^1(\widetilde{g}_j \mid s)$ represents the first order positive beliefs of player i about the contribution expected from player j in Stage 1, conditional on having observed the public signal s, and prior to the contributions decision being made.

2. The second order unconditional normative belief density, $h_i^2(.)$, is updated to the second order conditional normative belief density $h_i^2(. \mid s) : [0, y] \to [0, 1]$; the cumulative conditional distribution is $H_i^2(. \mid s) : [0, y] \to [0, 1]$. Thus, $h_i^2(\widetilde{g} \mid s)$ is the probability assigned by player i, conditional on observing the public signal s, that the social/peer group expects player i 'ought' to contribute an amount $\widetilde{g}$ in Stage 1, prior to the contribution decision.

Our first assumption on beliefs is a purely technical assumption.

Assumption 1. *All unconditional cumulative distributions, $F_i^1(g_j)$, $H_i^2(g_j)$, are continuous functions of g_j, hence, integrable. All cumulative conditional distributions, $F_i^1(g_j \mid s)$, $H_i^2(g_j \mid s)$, are continuously differentiable with respect to s.*

2.5.4 Assumptions on positive and normative beliefs

We make only one assumption each on positive and normative beliefs. Both assumptions require first order stochastic dominance in beliefs with respect to the signal, s, which is a minimal distributional assumption. In each case, we need not specify whether s is the mean, median, or the mode of the underlying distribution.

Define the conditional expectation of player i about the contributions of player j in Stage 1, prior to the contributions decision, by

$$E_i(g_j \mid s) = \int_0^y g_j dF_i^1(g_j \mid s). \tag{2.5}$$

Assumption 2. *We assume that $\frac{\partial E_i(g_j|s)}{\partial s} \geq 0$ for all $s \in S$.*

From Assumption 2, when the signal of normative expectation of the social/peer group, s, is higher, then players expect their partners, on average, to contribute (weakly) more.[24] Assumption 2 is implied by the usual first order stochastic dominance assumption on the first order beliefs of players: $\frac{\partial F_i^1(g_j|s)}{\partial s} \leq 0$ for all $g_j \in (0, y)$, $s \in S$.[25]

Assumption 3. *We assume that $\frac{\partial H_i^2(g_j|s)}{\partial s} \leq 0$ for all $g_j \in (0, y)$, $s \in S$.*

[24]The converse assumption that if the relevant social group expects higher contributions, then the partners would, on average, contribute strictly less is not borne out by the evidence on social norms, unless the social norms are too demanding (Bicchieri, 2006; Dhami, 2019, Vol. 2). In our experiments, the signals of social norms, s are generated from previous pilots and do not impose particularly demanding normative expectations.

[25]In other words, a higher public signal, s, makes it more likely, in the minds of a player, that the partner will make (weakly) higher contributions. The proof of this assertion is in the supplementary section.

From Assumption 3, a higher signal of normative expectations $s \in S$ induces first order stochastic dominance in H_i^2. Thus, on observing a higher signal s, player i believes that the relevant social/peer group expects player i 'ought' to make (weakly) higher contributions. The contrary assumption ($\frac{\partial H_i^2(g_j|s)}{\partial s} > 0$) is implausible and not supported by our data.[26]

3 Psychological utility under social norms and frustration-aversion

In this section we formulate the psychological utilities of the players and show, formally, how emotions such as shame, frustration, and anger play a key role in our analysis.

3.1 Stage 1 preferences

The *psychological utility* of player i in Stage 1 is given by

$$U_{i1}(g_i, g_j) = u_{i1}(g_i, g_j) - \mu_i \phi_i^S(g_i, s); i = 1, 2; \ \mu_i \geq 0. \tag{3.1}$$

In (3.1), the first term on the RHS is the 'economic utility' given in (2.2). The second term, that we explain in more detail below, captures the utility loss to player i from violation of a social norm of contributions *if and only if* the following conditions for norm compliance are met. (i) The normative and empirical expectations are not in conflict, and (ii) player i's shortfall in contributions relative to the normative expectations of the social group becomes common knowledge, and the social group has the ability and means to punish player i.[27]

The conditions for norm-compliance are satisfied in our experiments. In the absence of these conditions, there is no difference between economic utility and psychological utility in (3.1) (Bicchieri, 2006; Fehr and Schurtenberger, 2018a,b). When these conditions are satisfied, the second term on the RHS of (3.1) reflects *shame-aversion* from non-compliance with the normative expectations of the social group.[28] The parameter μ_i captures the relative weight that player i assigns to shame-aversion. In the special case $\mu_i = 0$, player i is not shame-averse, even when the conditions for norm compliance hold.

We now explain the shame-aversion motive. The function $\phi_i^S(g_i, s)$ is defined as follows.

$$\phi_i^S(g_i, s) = \int_{g'=g_i}^{y} (g' - g_i)\, dH_i^2(g' \mid s),\ s \in [0, y]\,, \ \ i = 1, 2. \tag{3.2}$$

In (3.2), $g_i \in [0, y]$ is the actual contribution chosen by player $i = 1, 2$. Conditional on the signal, s, player i believes that the social/peer group expects player i to choose a contribution $g' \geq g_i$ with probability $h_i^2(g' \mid s)$. In the interval $g' \in (g_i, y]$, $g_i < g'$, thus, the RHS of (3.2) measures the expected loss from *shame-aversion*, due to falling below the normative expectations of the

[26] If we assume 'strict' (and not 'weak') first order dominance in Assumptions 2 and 3, then all the comparative static results in our paper with respect to s take the form of strict inequalities.

[27] Following Fessler (2004), it is sufficient to have the following three rounds of knowledge about the norm violation for the emotion of shame. (1) One knows that one has violated the norm. (2) Others in the social group know that one has violated the norm. (3) One knows that others know that one has violated the norm.

[28] We could also have added, to the shame-aversion motive, the *approval-seeking motive* that gives utility to players from exceeding the normative expectations of the social/peer group (Dhami et al. 2022). However, this does not add any new insights to our comparative static results.

social/peer group, as perceived by player i. Substituting (2.2) and (3.2) in (3.1) we get the Stage 1 psychological utility of player i as

$$U_{i1}(g_i, g_j \mid H_i^2) = v_i(y-g_i)+r(g_i+g_j)-\mu_i\left[\int_{g'=g_i}^{y}\left(g'-g_i\right)dH_i^2(g' \mid s)\right]; i=1,2; \mu_i \geq 0. \quad (3.3)$$

Given the linear public goods production technology, the first order condition of player i in Stage 1, found by differentiating (3.3) with respect to g_i, is independent of g_j. Our solution concept requires players to play a best response to their beliefs at each stage (Section 4.2 below). In order to facilitate that exposition, denote by $E_i(g_j \mid s)$, the contributions that player i expects player j to make, after observing the public signal s but before making the contributions decision. It is convenient to rewrite Stage 1 utility (defined in (3.3)) as

$$U_{i1}(g_i \mid E_i(g_j \mid s), H_i^2) = v_i(y-g_i)+r(g_i+E_i(g_j \mid s))-\mu_i\left[\int_{g'=g_i}^{y}\left(g'-g_i\right)dH_i^2(g' \mid s)\right]; i=1,2; \mu_i \geq 0 \quad (3.4)$$

3.2 Stage 2 preferences

At the beginning of Stage 2, the history of the game is summarized in (i) the Stage 1 signal of normative expectations of the social/peer group, s, (ii) the initial conditional expectations of contributions each player has from the partner, $E_i(g_j \mid s)$, prior to the contributions decision, and (iii) the publicly revealed actual contributions of the two players, g_i, g_j. The main psychological component of Stage 2 preferences is frustration-aversion.[29]

Following Battigalli et al. (2019), define the frustration function, ϕ_i^F, of player i, at the beginning of Stage 2, by

$$\phi_i^F\left(E_i(g_j \mid s), g_j\right) = \left[E_i(g_j \mid s) - g_j\right]^+, \quad (3.5)$$

where $z = x^+$ means that $z = x$ if $x > 0$ and $z = 0$ if $x \leq 0$. Thus, player i gets frustrated at the end of Stage 1 (or equivalently, the beginning of Stage 2) if the partner's contribution, g_j, is below player $i's$ initial conditional expectations, $E_i(g_j \mid s)$, i.e., $E_i(g_j \mid s) > g_j$. If on the other hand $E_i(g_j \mid s) \leq g_j$, then player i faces no frustration.

Remark 1. *The empirical evidence strongly supports the* false consensus effect.[30] *Namely, in forming expectations about the actions and beliefs of other players, subjects in experiments assign to the other players their own actions and beliefs. Under the false consensus effect, which we find in our data, we expect player i to expect the other player to make a contribution that is close to player $i's$ own contribution. Hence, we expect $E_i(g_j \mid s)$ to be close to g_i, and in the extreme case of false consensus we expect $E_i(g_j \mid s) = g_i$.*[31]

[29] In the supplementary section, we show how preferences can be extended to incorporate *conditional sequential reciprocity*, using the framework of Dufwenberg and Kirchsteiger (2004).

[30] See Dhami (2020 Vol. 4) for a discussion of the false consensus effect and the references. Ellingsen et al. (2010) showed the presence of the false consensus effect in models of psychological game theory.

[31] It might be tempting to infer in this extreme case ($E_i(g_j \mid s) = g_i$) that since player i become frustrated when $g_i - g_j < 0$, this is a form of inequity averse preferences. However, as noted earlier, such an inference is not supported because there are no dynamic economic linkages between the two stages of our VCMP, and models of inequity aversion do not allow for beliefs to enter into the utility functions.

Battigalli et al. (2019) distinguish between three different notions of blame. Here we use their first notion, which they call *simple blame*, in which the level of blame B_i assigned by player i towards player j equals the frustration experienced by player i, hence,

$$B_i\left(E_i(g_j \mid s), g_j\right) = \phi_i^F\left(E_i(g_j \mid s), g_j\right). \tag{3.6}$$

The Stage 2 psychological utility, U_{i2}, of player $i = 1, 2$ is given by

$$U_{i2}(p_i, p_j) = u_{i2}(p_i, p_j) - \lambda_i B_i\left(E_i(g_j \mid s), g_j\right)(y - \kappa p_j - p_i);\ \lambda_i \geq 0. \tag{3.7}$$

In (3.7), the first term on the RHS gives the second stage economic utility, defined in (2.4). The second term captures frustration-aversion with relative weight $\lambda_i \geq 0$, the *frustration-aversion parameter*. This term has an effect on the utility of player i, only if player i blames player j, $B_i > 0$, and $\lambda_i > 0$ (see (3.5), (3.6)). If $B_i > 0$, then player i responds with anger. The action tendency of anger is revenge and retaliation. Hence, player i responds by negatively internalizing the Stage 2 material payoff of player j, $(y - \kappa p_j - p_i)$.[32]

Substituting (3.5), (3.6), in (3.7), the Stage 2 optimization problem of player i is[33]

$$p_i^* \in \underset{\{p_i \in [0,\bar{p}]\}}{\arg\max}\, U_{i2}(p_i, p_j) = v_i\left(y - \kappa p_i - p_j\right) - \lambda_i\left[E_i(g_j \mid s) - g_j\right]^+\left(y - \kappa p_j - p_i\right); \lambda_i \geq 0. \tag{3.8}$$

Remark 2. *The presence of the term $[E_i(g_j \mid s) - g_j]$ in (3.8) creates dynamic psychological linkages between the two stages of our VCMP, despite the absence of dynamic economic linkages between the two stages.*

4 Optimal choice of contributions and punishments

Our model, based on belief-hierarchies that directly enter into the utility function, is in the class of models of psychological game theory (Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009; Battigalli and Dufwenberg, 2022). In such models, typically, players (1) play their best response to their beliefs, and (2) there is mutual consistency of beliefs and actions. Best response to one's beliefs is not controversial. However, the bulk of the evidence shows that 'consistency between beliefs and equilibrium actions' required in variations of sequential Nash equilibrium does not hold in the early rounds of most games; nor is there any guarantee that it holds in games that are repeated and learning is allowed (Camerer, 2003; Dhami, 2020, Vol. 4; Dhami, 2020, Vol. 5). For this reason, as in models of non-equilibrium beliefs (e.g., level-k models,

[32] Player i cannot be expected to know the utility function of player j, $v_j\left(y - \kappa p_j - p_i\right)$, defined in (2.4). But the material payoff of player j, $y - \kappa p_j - p_i$, can be readily calculated.

[33] It is often argued that biological and cultural evolution has self-selected individuals with an innate propensity to follow social norms (Ostrom, 2000; Bowles and Gintis, 2011; Gintis, 2017). Incorporating this channel would require an extra term in (4.2) of the form, say, $\gamma\left(s - g_j\right)^+\left(y - \kappa p_j - p_i\right)$, which creates an innate propensity to punish norm offenders, i.e., partners that contribute below the public signal of normative expectations ($g_j < s$). It can be shown that this leads to the comparative static result that optimal punishment p_i^* is directly increasing in $s - g_j$, when $g_j < s$. However, this effect is already partly accounted for by the second term on the RHS of (3.8). From Assumption 2, $\frac{\partial E_i(g_j|s)}{\partial s} \geq 0$ for all $s \in S$. Hence, a higher signal of normative expectations, s, increases the expected contributions, $E_i\left(g_j \mid s\right)$, from the partner. Thus, there is likely to be high correlation between $[E_i(g_j \mid s) - g_j]^+$and $(s - g_j)^+$, which we also find in our data. Indeed, this suggests that the inherent tendency to punish norm violators may arise from frustration-aversion, thereby furnishing it with potential microfoundations.

cognitive hierarchy models, evidential equilibrium, and models of cursed equilibrium) we do not require the mutual consistency of beliefs and actions.[34]

In the Stage 2 objective function in (3.8), at the time of choosing p_i, player i does not yet know the punishment chosen by the partner, p_j. However, conditional on the observed contributions at the end of Stage 1, player i has the following "point" expectations

$$E_i p_j \equiv E_i(p_j \mid g_i, g_j) \tag{4.1}$$

about the punishment chosen by player j. We elicit these expectations in our experiment, and test our predictions with respect to them.

Remark 3. *We are not interested in how the subjective expectations of players are formed. This is a separate research agenda. Subjective expectations might be formed by cognitive shortcuts, subject-specific intuition or experience; subject-specific moods and optimism. We are only interested in the behavior of players, conditional on these subjective expectations, and we can, and do, test the implications.*

Given that we only assume that players play a best response to their expectations and beliefs in each Stage (see Section 4.2 below), player i follows the cognitive shortcut of substituting in (3.8) the unobserved punishments of the partner by their expectations in (4.1). Thus, we can rewrite (3.8) as

$$p_i^* \in \underset{\{p_i \in [0,\overline{p}]\}}{\arg\max}\, U_{i2}(p_i \mid E_i p_j) = v_i\left(y - \kappa p_i - E_i p_j\right) - \lambda_i \left[E_i(g_j \mid s) - g_j\right]^+ \left(y - \kappa E_i p_j - p_i\right); \lambda_i \geq 0. \tag{4.2}$$

4.1 Stage 1 continuation payoffs of boundedly rational players

In Stage 1, when choosing optimal contributions, player i needs to know the Stage 2 continuation payoff V_{i2}. However, in Stage 2, the utility function in (4.2) requires both players to have observed the Stage 1 contributions of each other, and based on these observations, player i forms the expectation $E_i p_j$. But at the beginning of Stage 1, when V_{i2} is formed, the contributions are unknown. Hence, player i needs to form inferences about V_{i2} based on the initial Stage 1 information set. Mutual consistency of beliefs and actions, as in a sequential equilibrium (SE), makes this problem relatively simple, if not trivial, by using equilibrium actions and forcing beliefs and equilibrium actions to match. In particular, the rationality and cognitive requirements in a SE in psychological game theory are even more stringent due to beliefs that enter directly into the utility function. As noted above in the introductory discussion to Section

[34] For useful surveys of the evidence, see Mauersberger and Nagel (2018), and Dhami (2020, Vol. 4). In particular, Bellemare et al. (2011) show that there is a lack of consistency between actions, first-order beliefs, and second-order beliefs in their data. See also Section 9 in Battigalli and Dufwenberg (2022) for a critical discussion of the solution concepts in psychological games and a recognition of the importance of non-equilibrium beliefs. For applications of models of psychological game theory that do not require consistency between beliefs and actions, see Khalmetski et al. (2015), Dhami et al. (2019), Dhami et al. (2022), Dhami et al. (2023).

4, this is rejected by the evidence.[35] Yet, boundedly rational players in experiments will attempt to calculate the subjective effects of their Stage 1 actions on their Stage 2 continuation utility. We impose the following plausible restrictions directly on V_{i2} in the following assumption.

Assumption 4. *We make the following assumptions on the continuation payoff of player i, V_{i2}, conditional on the information set of player i at the time of choosing the Stage 1 contribution.*

$$(i)\ \frac{\partial^2 V_{i2}}{\partial g_i^2} \leq 0;\ (ii)\ \frac{\partial V_{i2}}{\partial g_i} \geq 0.\ (iii)\ \frac{\partial}{\partial s}\left(\frac{\partial V_{i2}}{\partial g_i}\right) \geq 0.$$

Assumption 4 allows for subjective differences in the continuation payoffs of the players. But it imposes three common restrictions for all players. Assumption 4(i) is a purely technical assumption that ensures a unique solution to Stage 1 contributions. It can, however, be relaxed, in which case locally optimal results can be stated. Assumption 4(ii) requires that player i believes that an increase in Stage 1 contributions (weakly) increases the Stage 2 continuation utility. This is reasonable; players who contribute more in Stage 1 believe that they are less likely to be punished in Stage 2 by their partners, hence increasing Stage 2 utility.[36] This transmission mechanism is confirmed by our empirical results in Section 8. Assumption 4(iii) requires that when the signal of normative expectations, s, is high, the marginal effect of contributions on continuation utility, V_{i2}, is (weakly) high. This too is reasonable for the following reason. When the social group has higher normative expectations (higher s), a lower contribution is more likely to invite greater frustration from the partner, who is likely to respond by punishing more, reducing V_{i2}. Thus, the marginal effects of a unit increase in g_i on V_{i2} are likely to be greater relative to the case of a lower signal.

4.2 Psychological best responses

Players play a psychological best response if they maximize their psychological utility in each stage, conditional on their subjective beliefs, and their current information set.

Definition 1. *A psychological best response for player i $(i = 1, 2)$ is a pair of Stage 1 contributions and Stage 2 punishment levels (g_i^*, p_i^*) $g_i^* \in [0, y]$, $p_i^* \in [0, \overline{p}]$, with the following properties: (i) In Stage 2, p_i^* maximizes $U_{i2}(p_i \mid E_i p_j)$ in (4.2), given the Stage 1 observed contributions g_i, g_j, and the expectations of player i about the Stage 2 punishment chosen by player j, $E_i p_j$. (ii) In Stage 1, g_i^* maximizes $U_{i1}(g_i \mid E_i(g_j \mid s), H_i^2) + V_{i2}(.)$, where $V_{i2}(.)$ is the second stage continuation utility of player i and $U_{i1}(g_i \mid E_i(g_j \mid s), H_i^2)$ is defined in (3.4), conditional on the second order beliefs of player i about the normative expectations of the social group, H_i^2, and the expectations about the Stage 1 contribution of player j, $E_i(g_j \mid s)$.*

[35] The supplementary section discusses this issue further and sketches the incredible cognitive requirements for a SE in our model that, in addition to forcing consistency between equilibrium actions, the beliefs, and the normative expectations, requires players to form third order belief distributions over the unknown second order belief distributions of the opponents. One will also need to specify the signs of the first and second order partial derivatives of these third order belief distributions to get any sensible comparative static results. Our reading of the extensive evidence on this issue does not lead us to believe that players in one shot experiments behave in a manner that fulfills the requirements of a SE (Camerer, 2003; Bellemare et al., 2011; Mauersberger and Nagel, 2018; Eyster, 2019; Dhami, 2020, Vol. 4, 5; Battigalli and Dufwenberg, 2022).

[36] The converse assumption, $\frac{\partial V_{i2}}{\partial g_i} < 0$, potentially leads to the result that the presence of a punishment mechanism reduces Stage 1 contributions in VCMPs, which is known to be empirically false (Dhami, 2019, Vol. 2).

4.3 Stage 2 optimal choices

We first solve for the Stage 2 *psychological best response* for player i. Differentiating the Stage 2 objective function of player i in (4.2), we get

$$\frac{dU_{i2}(p_i \mid E_i p_j)}{dp_i} = -\kappa v_i'\left(y - \kappa p_i - E_i p_j\right) + \lambda_i \left[E_i(g_j \mid s) - g_j\right]^+ ; \; i = 1, 2. \tag{4.3}$$

The first term on the RHS in (4.3) gives the marginal cost to player i, of a unit of punishment. If $E_i(g_j \mid s) > g_j$, and $\lambda_i > 0$, then player i is frustrated, and blames player j. In this case, a marginal increase in p_i increases the marginal utility of player i, otherwise it has no effect.

Proposition 1. *(Comparative statics of punishment) There exists a unique solution to the problem in (4.2) given by p_i^*. Suppose that $\lambda_i > 0$, so that players are frustration-averse.*
(a) (i) If Assumption 2 holds, then p_i^* is increasing in the signal of normative expectations, s, received in Stage 1. p_i^* is strictly increasing in (ii) The extent of frustration, as captured by $[E_i(g_j \mid s) - g_j]^+$, and (iii) the frustration-aversion parameter, λ_i.
(b) p_i^* is strictly decreasing in the following. (i) The Stage 1 contributions of the partner, g_j. (ii) The expected punishment from the partner, $E_i p_j$.

Example 1. *Consider the special case where $v_i(x) = lnx$. Define*

$$\widetilde{p} = \frac{y - E_i p_j}{\kappa} - \frac{1}{\lambda_i \left[E_i(g_j \mid s) - g_j\right]^+}, \tag{4.4}$$

such that $\frac{dU_{i2}(\widetilde{p} \mid E_i p_j)}{dp_i} = 0$. Thus, the optimal punishment chosen by player i is

$$p_i^*(s, E_i p_j, g_j) = \begin{cases} 0 & if \quad \widetilde{p} < 0 \\ \widetilde{p} & if \quad 0 \leq \widetilde{p} \leq \overline{p} \\ \overline{p} & if \quad \widetilde{p} > \overline{p} \end{cases}. \tag{4.5}$$

We can directly verify all the results in Proposition 1 from (4.4), (4.5).

Discussion of Proposition 1: The existence of the optimal solution in Proposition 1 ensures that the condition in Definition 1(i) holds; furthermore the solution is unique. Assumption 2 implies that an increase in s increases the expectation that player i has about the contribution of player j, hence, potentially increasing frustration, and optimal punishment (Proposition 1(ai)). Parts (aii) and (aiii) follow directly from (4.3) by increasing the marginal benefit of punishment when one is frustrated. Punishments are decreasing in the contributions of the partner (Proposition 1(bi)) because an increase in the partner's contributions reduces frustration. The prediction, $\frac{dp_i^*}{dE_i p_j} < 0$, in Proposition 1(bii) is rejected by our evidence. It relies on the following mechanism that is common to most standard models in VCMP: As $E_i p_j$ increases, player $i's$ consumption falls and the marginal utility of consumption, $v_i'\left(y - \kappa p_i - E_i p_j\right)$, increases. Hence, the opportunity cost of punishment increases, reducing the optimal level of punishment. Subjects do not appear to reason in this manner in our data. They appear to demonstrate *revenge for a concurrently anticipated event* and punish the partner more when they expect higher punishment, i.e., $\frac{dp_i^*}{dE_i p_j} > 0$. A model of concurrent Stage 2 reciprocity, as in Rabin (1993), can provide a potential explanation; for specific evidence of this mechanism in public goods games, and the relevant theory, see Dhami et al. (2019).

4.4 Stage 1 optimal choice of contributions

The first stage optimization problem of player i is given by

$$g_i^* \in \underset{\{g_i \in [0,y]\}}{\arg\max}\, U_{i1}(g_i \mid E_i(g_j \mid s), H_i^2) = v_i\left(y - g_i\right) + r\left(g_i + E_i(g_j \mid s)\right) - \mu_i \left[\int_{g'=g_i}^{y} \left(g' - g_i\right) dH_i^2(g' \mid s)\right] + \delta V_{i2};\ \mu_i \geq 0, \delta \in [0,1]\,. \quad (4.6)$$

On the RHS of (4.6), the first three terms capture the Stage 1 psychological utility defined in (3.4). V_{i2} is the Stage 2 continuation payoff of player i defined in Section 4.1. The parameter δ in (4.6) is a discount factor; fully myopic players set $\delta = 0$.

Lemma 1.

$$\int_{g'=g_i}^{y} \left(g' - g_i\right) dH_i^2(g' \mid s) = y - g_i - \int_{g'=g_i}^{y} H_i^2(g' \mid s) dg'.$$

Differentiating (4.6) with respect to g_i and using Lemma 1

$$\frac{\partial U_{i1}}{\partial g_i} = -v_i'\left(y - g_i\right) + r + \mu_i\left(1 - H_i^2(g_i \mid s)\right) + \delta \frac{\partial V_{i2}}{\partial g_i}. \quad (4.7)$$

The RHS of (4.7) captures the marginal effects as Stage 1 contributions increase by a unit. The first term is the fall in marginal utility of Stage 1 consumption; the second term is the marginal return on the public good; the third term is the marginal reduction in shame; the fourth term is the marginal effect on the Stage 2 continuation payoff.

Proposition 2. *Suppose that Assumption 4 holds.*
(ai) There exists a unique solution, g_i^*, to the first stage optimization problem in (4.6).
(aii) Suppose that Assumption 3 also holds. Then g_i^* is increasing in the signal of normative expectations, s.
(b) Optimal contributions, g_i^*, are higher when a punishment mechanism is present, relative to when the mechanism is absent.
(c) Suppose that players are myopic, so that $\delta = 0$. Then, g_i^* is increasing in the signal of normative expectations, s.

Discussion of Proposition 2: Assumption 4(i) is sufficient for the existence of a unique solution to the Stage 1 contributions (Proposition 2(ai)); existence also ensures that the condition in Definition 1(ii) holds. Proposition 2(aii) allows us to compare the contribution decisions in different treatments when the signal of normative expectations, s, is low and when it is high. A higher signal s induces subjects to contribute more for two reasons. (1) From Assumption 3, it is more likely that the social group normatively expects higher contributions; hence, in order to avoid shame aversion, subjects have an incentive to contribute more. (2) An increase in s increases the positive marginal effect of contributions on the Stage 2 continuation payoffs (Assumption 4(iii)). Proposition 2(b) enables us to study treatment contrasts when the punishment mechanism is present and when it is absent. Proposition 2(c) gives the benchmark case of myopic subjects who do not take account of Stage 2. This also addresses our treatments where there is only a contributions decision (Stage 1) but no punishment decision (Stage 2).

5 The experimental design

We conducted our virtual lab experiments with 278 students in Nankai University in China (June, 2021), and with 256 students in University of Nottingham in UK (October, 2021).[37] In the Nottingham experiment, 56 subjects were not UK nationals. Since we are also interested in the cultural differences between UK and China subjects, hence we used data for the remaining 200 subjects who were UK nationals. Each of our 5 treatments had 2–3 sessions, and there were 16–32 subjects in each session. No subject attended more than one session.

	C + P	C only
EE/NE	$T2(9), T3(12)$	$T4(9), T5(12)$
No EE/NE	$T1$	$--$

Table 1: The design of treatments.

Table 1 shows our between-subjects design of 5 treatments. Along the rows, in one dimension *empirical expectations* (EE) and *normative expectations* (NE) were present (labeled EE/NE); and in the other dimension they were absent (labeled No EE/NE). Along the columns, in one dimension, we had a VCMP with both Stage 1 and Stage 2, i.e., contributions and punishments (labeled C + P); and along the second dimension, a VCM with only Stage 1, the contributions stage, but no punishment stage (labeled C only). There are no treatments at the intersection of 'C only' and 'No EE/NE'; this is the standard one-shot public goods game without punishment where the results are well known. However, we needed treatment T1 in order to (i) contrast the effects of the presence/absence of social norms in the presence of the punishment option (C + P, but varying the rows), and (ii) to perform a pilot experiment to generate the data on empirical and normative expectations for use in the other treatments.

We had two public signals of normative expectations; a low signal $s = 9$ (treatments T2(9) and T4(9)) and a high signal $s = 12$ (treatments T3(12), T5(12)). In order to prevent issues of subject deception, we used an actual incentivized pilot experiment, using Treatment T1, which gave rise to the actual signals $s = 9$ and $s = 12$ that we used in our main experiments.[38]

The experiments were closely guided by our theoretical model, and followed the sequence of moves described in Section 2.4. An endowment of 20 tokens in each stage was given to each of the players. Subjects were required to pass a test of understanding of the experimental instructions before being allowed to make their choices. The full experimental instructions are given in the supplementary section, however, we comment on some specific features below.

The maximum punishment allowed in Stage 2 was $\overline{p} = 5$ tokens. For every token used for punishment, the target of the punishment lost 3 tokens, so the unit cost of punishment is

[37] This study is pre-registered; see https://doi.org/10.1257/rct.7110. Subjects in both countries were from various disciplines. Some students quit the experiment before filling in the post-experimental survey (1 in Nankai and 4 in Nottingham). Their data was removed in the analysis.

[38] We are grateful to Gary Charness and Chris Starmer for alerting us to the possibility of subject deception with hypothetical public signals. Since treatment T1 had no empirical and normative expectations, we used it to collect the data on the signals $s = 9$ and $s = 12$ during our pilot session, in the post-experimental survey questions. Furthermore, since the UK and Chinese subjects face identical values of s and the China experiments were done earlier than the UK experiments, the data for signal s were collected in China.

$\kappa = 1/3$ tokens. The empirical expectations and normative expectations were conveyed through the following instructions.[39]

You are provided with the following data from a previous similar experiment (you may take this as the behavior/opinion of your social or peer group): (1) Most individuals contributed more than 7 tokens. (2) Most individuals said that others who play this experiment 'ought' to contribute at least s tokens, or that it would be "socially desirable" to contribute at least s tokens. [The signal s was either 9 or 12 depending on the treatment.]

Subjects are asked to state on a 7 point Likert scale, the intensity of the emotions they experienced at the end of each Stage. Each of these emotions was briefly, and precisely, defined in the experimental instructions to ensure greater objectivity.

The experiment took around 30 minutes. The Chinese subjects earned 43 Yuan on average, and the UK subjects earned 11 pounds on average; in June 2021, 1 Chinese Yuan was approximately 0.11 pounds. All subjects were paid in private after the experiment. The earnings in each case are around 3 times the local minimum half-hourly wage so that the real earnings were similar across the two subject pools. One of the two stages, Stage 1 or Stage 2, was randomly chosen and the decisions made by the subjects in that stage are paid out. The other stage was not paid. Hence, we decouple the two stages of the VCMP for the reasons discussed earlier (see Section 2.3), while in VCMP experiments, the sum of earnings over both stages is paid out.

We use the following list of independent variables in our empirical results.

Business: Dummy variable that equals 1 for business/economics subjects, and 0 otherwise.

China: Dummy variable that equals 1 for Chinese subjects and 0 for UK subjects.

$E_i g_j$: Short form for $E_i(g_j \mid s)$, player $i's$ expected contribution from player j in Stage 1.

Experience: Dummy variable that equals 1 if the subject has attended similar experiments before, and 0 otherwise.[40]

Norm: Dummy variable for the signal s of normative expectations of the social/peer group; it equals 1 for $s = 12$, and 0 for $s = 9$.

Male: Dummy variable that equals 1 for male subjects and 0 otherwise.

Punishment : Treatment dummy that equals 1 in the presence of punishment (treatments T1, T2, T3) and 0 in the absence of the punishment (treatments T4, T5).

6 Analysis of Stage 2 punishments

6.1 Descriptive statistics

Figure 1 shows basic data on the distribution of punishments, across all treatments with the punishment option (T1, T2, T3), separated by Chinese and UK subjects. On the vertical axis, in each of the three diagrams in Figure 1, we have the level of punishment, $p_i \in [0, 5]$.

[39] Recall that we used the 'actual' expectations of the relevant social group from a previous pilot in order to avoid the charge of subject deception that may have arisen with hypothetical expectations. This leads to two constraints. First the 'actual' signals of normative expectations, $s = 9$ and $s = 12$, were not as widely spaced as hypothetical signals could have been. Second, we need careful wording to ensure that the empirical and normative expectations are not in obvious conflict.

[40] 29% (= 81/277) Chinese subjects and 78% (= 156/200) British subjects reported having participated in experiments before.

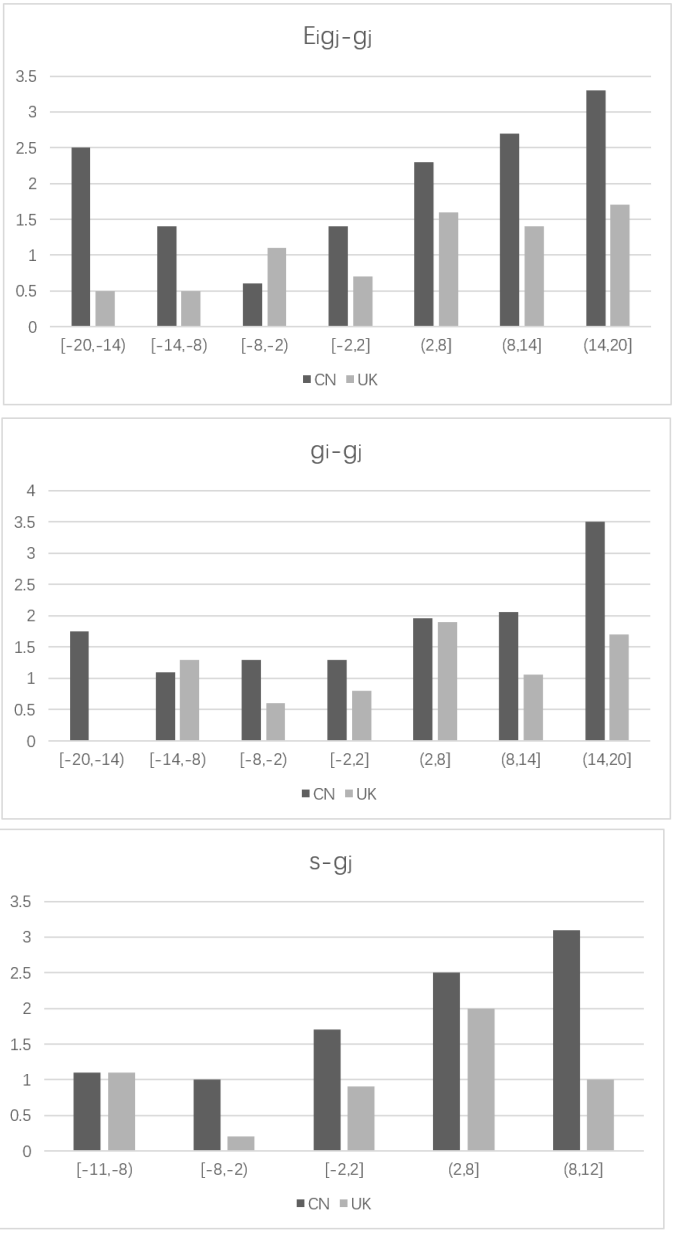

Figure 1: Punishment choices corresponding to different levels of the deviations of the contributions of player j, g_j, from (i) the expected contributions of player i, $E_i g_j$ (upper panel), (ii) from the actual contributions of player i, g_i (middle panel), and (iii) from the signal of normative expectations, s (bottom panel).

Along the horizontal axis, we have $E_i g_j - g_j$ (upper panel); $g_i - g_j$ (middle panel); and $s - g_j$ (bottom panel). We have followed the presentation scheme in Fehr and Gächter (2000).[41] In general, as the partner's contribution increasingly falls short of, respectively, $E_i g_j$, g_i, and s, the punishment meted out to the partner increases.

The distributions are quite comparable. We have already noted that the false consensus effect is likely to lead to $E_i g_j$ and g_i to be highly correlated (see Remark 1); recall that $E_i g_j$ is the short form for $E_i(g_j \mid s)$. The Spearman correlation between $E_i g_j$ and g_i is significantly positive for both British and Chinese subjects in each of the 5 treatments (all p-values = 0.000). The values of the Spearman correlation coefficient between $E_i g_j$ and g_i, denoted by r, and the corresponding treatments (T1–T5) in each culture as follows. For Chinese subjects, $r = 0.73$ in T1; $r = 0.66$ in T2; $r = 0.76$ in T3; $r = 0.67$ in T4; $r = 0.74$ in T5. For British subjects, these values are: $r = 0.61$ in T1; $r = 0.65$ in T2; $r = 0.82$ in T3; $r = 0.65$ in T4; $r = 0.57$ in T5.

In the sequence of moves, the public signal, s, is announced before players form their expectations $E_i g_j$, hence, $E_i g_j$ and s are hypothesized to be highly correlated as well.[42] The correlation between $E_i g_j$ and s is significantly positive for both Chinese and British subjects; the Spearman coefficients are 0.30 ($p = 0.000$) and 0.18 ($p = 0.013$), respectively.

The distributions of punishments between Chinese and British subjects are *not* significantly different in any treatment (see the supplementary section). Figure 2 shows a comparison of the 'average punishments' chosen by Chinese and UK subjects in the three treatments that allowed punishments (T1, T2, and T3). In the presence of social norms (EE/NE), i.e., in the treatments T2, and T3, the average punishment chosen by the Chinese subjects is higher than that of the British subjects[43]. However, in the absence of social norms (No EE/NE), i.e., treatment T1, there are no statistically significant cultural difference in average punishments.

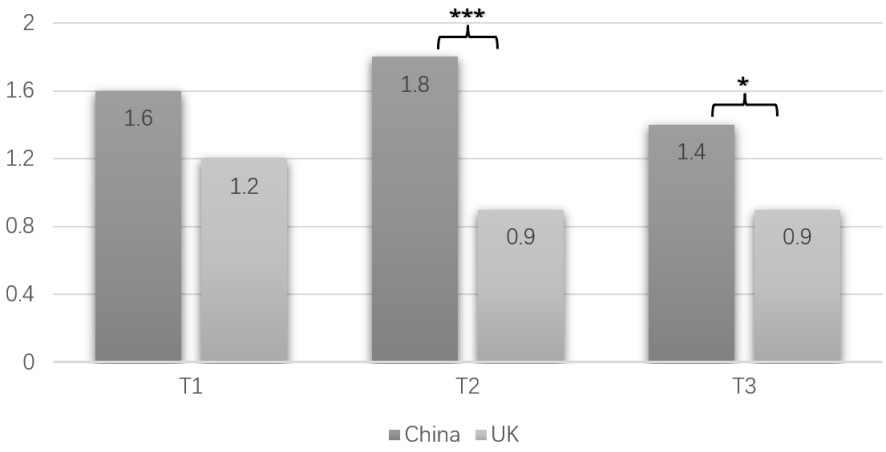

Figure 2: Average punishments chosen by Chinese and UK subjects. Superscripts * and *** denote the statistical significance of the t test at 10% and 1% levels, respectively.

[41] For instance, consider the uppermost panel. $(14, 20]$ represents all the cases where the difference $E_i g_j - g_j$ is in the interval $(14, 20]$ tokens. At the other extreme, $[-20, -14)$ represents all the cases where the difference $E_i g_j - g_j$ is in the interval $[-20, -14)$ tokens.

[42] A potential explanation is in terms of the anchoring heuristic (Dhami and Sunstein, 2022).

[43] The Mann-Whitney test of the difference in punishments chosen in treatments T2 and T3 gives a p-value equal to 0.055 (Chinese subjects) and 0.146 (UK subjects).

Experimental design features limit strict comparability of results across studies. In our experiment, when the partner contributes relatively less ($g_j - g_i < 0$), the average punishment chosen by the British and Chinese subjects, respectively, was 1.5 and 1.9 tokens out of a maximum of 5 tokens. When the partner contributes relatively more ($g_j - g_i \geq 0$), the corresponding figures were 0.68 and 1.3 tokens, respectively. By contrast, in the two cases $g_j - g_i < 0$ and $g_j - g_i \geq 0$, Cubitt et al. (2011) report the average punishment to be 0.85 and 0.14 tokens; and Weber et al. (2021) report the corresponding figures to be 0.6 and 0.2 tokens. These studies also set the punishment choices between 0 and 5 tokens, but the efficiency of punishment ($\kappa = 0.5$) is different from ours. There are also differences in the number of group members; 3 in Cubitt et al. (2010) and 4 in Weber et al. (2021). Our results on punishment are similar to the results in Walker and Halloran (2004), Fehr and Gächter (2000), and Gächter and Hermann (2009) in the sense that there is a decreasing trend of punishment when the deviation ($g_j - g_i$) decreases (see Figure 1). However, these papers do not report the absolute values of the average punishment.

6.2 Determinants of punishment

Table 2 reports the results of robust OLS regressions to analyze the determinants of punishment in treatments T2 and T3.[44] As noted earlier, due to the false consensus effect, the variables $E_i g_j - g_j$ and $g_i - g_j$ are highly correlated (Spearman correlation p-value < 0.01), hence, each model in Table 2 incorporates one of the two variables at a time. Both variables significantly increase punishment, consistent with our explanation of punishments/sanctions in terms of frustration-aversion (see Proposition 1(aii)).[45] Proposition 1(aii) applies in the absence of dynamic economic linkages between the two stages of a VCMP; a setting in which other theories may fail to make a similar prediction without further auxiliary assumptions. The correlation between $E_i g_j$ and g_i due to the false consensus effect explains why $g_i - g_j$ is also a significant determinant of punishments.

The expected choice of punishment by the partner, $E_i p_j$, significantly and positively increased punishment. In other words, the higher is the expected punishment from the partner, the higher does a player punish the partner. This contradicts our prediction, $\frac{dp_i^*}{dE_i p_j} < 0$, in Proposition 1(bii), which is based on a standard neoclassical diminishing marginal utility argument. As noted earlier, the most likely explanation is a form of *contemporaneous anticipated revenge* (Rabin, 1993), which has been modeled and confirmed by the evidence in public goods games (Dhami et al., 2019).

Chinese subjects punish relatively more. The significance and the negative sign of the interaction term 'norm$\times E_i p_j$' has the interpretation that when the norm of contributions is higher, contemporaneous anticipated revenge plays a weaker role. This is reasonable: When the social group expects higher contributions, then anticipated punishment by the partner is more justified, if one contributed a lower amount. We tried other interactions, such as norm$\times China$

[44] Robust OLS employs robust standard errors. The Tobit models (censored on both sides and clustered on individual subject) produced similar results for all the regressions reported in our paper. Hence, throughout, we report the robust OLS results.

[45] When we run a constrained robust OLS regression such that $E_i g_j > g_j$ (i.e., players are frustrated), the coefficient of $E_i g_j - g_j$ is significant for Chinese subjects (OLS coefficient is 0.16, p-value= 0.014), however it is not significant for British subjects.

Table 2: Determinants of punishment. Superscripts **,*** denote the statistical significance at 5%, and 1%, respectively.

	Model 1	Model 2
$E_i g_j - g_j$	0.08***	
	[0.021]	
$g_i - g_j$		0.06***
		[0.019]
norm	0.37	0.35
	[0.276]	[0.281]
$E_i p_j$	0.70***	0.71***
	[0.089]	[0.093]
China	0.61**	0.69***
	[0.258]	[0.255]
age	-0.00	-0.01
	[0.039]	[0.039]
male	0.05	0.07
	[0.228]	[0.232]
experience	0.03	0.00
	[0.273]	[0.278]
business	-0.07	-0.10
	[0.208]	[0.209]
norm$\times E_i p_j$	-0.33**	-0.28**
	[0.140]	[0.141]
constant	-0.13	-0.14
	[0.848]	[0.844]
F-stat	11.47***	10.18***
Adjusted R^2	0.34	0.32
No. Obs.	184	184

and norm×*male* etc., but these are not significant.

6.3 Beliefs on punishment

The distribution of beliefs of Chinese and UK subjects on the punishment expected from the partner, $E_i p_j$, are not significantly different. Neither are the average beliefs about $E_i p_j$ different between the two cultures. However, subjects expected to be punished more, the greater was the shortfall in their contributions relative to the social norm, $(s - g_i)$. This is an important finding because it supports our transmission channel for the implementation of norms. For the statistical evidence on all three of these assertions, see the supplementary section.

7 Analysis of the Stage 1 contributions

7.1 Descriptive statistics on contributions

The distribution of contributions for Chinese and UK subjects are not significantly different for each of the treatments and for the data pooled across all treatments; see the supplementary section. In order to explore the effects of the presence/absence of (i) the punishment option, and (ii) social norms, we conduct two kinds of analyses. The distributional comparison is reported in the supplementary section. We report below, a comparison of the average differences between treatments that is based on our predictions in Proposition 2. To understand the comparisons below, the reader might wish to consult Table 1.

1. In order to test the effects of punishment on contributions, we hold fixed the presence of norms (first row of Table 1 titled EE/NE) and compare the distributions of contributions in the treatments-with-punishment (T2, T3) and the treatments-without-punishment (T4, T5). We compare T2 vs T4 (fixing the public signal $s = 9$) and T3 vs T5 (fixing the public signal $s = 12$).

2. To test the effects of the size of social norms ($s = 9$ versus $s = 12$), we control for the absence/presence of punishment. Thus, we consider the contrasts T2 vs T3 (holding fixed the presence of punishment, i.e., the dimension $C + P$) and T4 vs T5 (holding fixed the absence of punishment, i.e., the dimension C).[46]

Figure 3 shows the average contributions of Chinese and British subjects in all 5 treatments in a *between-cultures* comparison. When a punishment mechanism exists (treatments T1, T2, and T3), the average contributions between Chinese and UK subjects are not significantly different. However, when the punishment mechanism is absent (treatments T4 and T5), there are significant differences. In the treatment with the low signal of normative expectation, $s = 9$ (T4), Chinese subjects contributed less than the British subjects. But in the treatment with the high signal of normative expectations, $s = 12$ (T5), the Chinese subjects contributed higher than the British subjects.[47] In other words, in the absence of a punishment mechanism, the

[46] If we compare the pooled data for the treatments-with-punishment (T2 + T3) and the treatments-without-punishment (T4 + T5), then the contribution distributions are not significantly different (KS test p-value=0.178).

[47] The Mann-Whitney test between the contributions of Chinese and British subjects in the treatments T4 and T5 gives p-values equal to 0.107 and 0.077, respectively.

Chinese subjects are more sensitive to changes in the signal of normative expectations from the social/peer group, relative to the British subjects. This is also confirmed in the results in Figure 4.

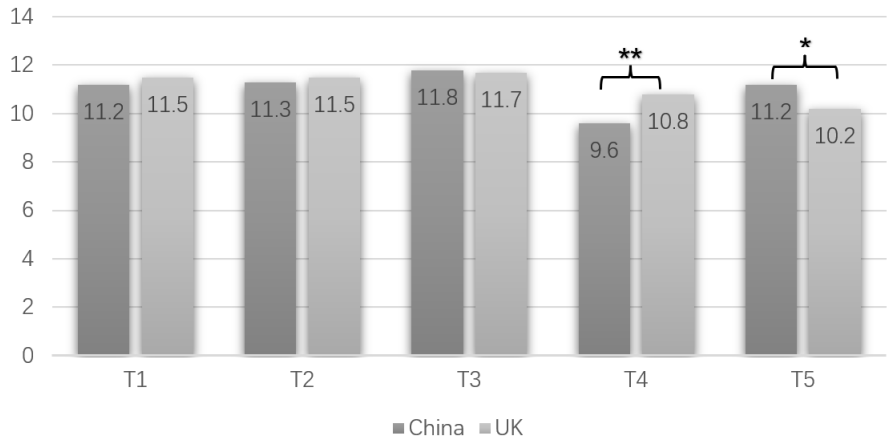

Figure 3: Average contribution differences between Chinese and UK subjects for each treatment. Superscripts * and ** denote the statistical significance of t test at 10% and 5%, respectively.

Figure 4 shows the average contributions in each treatment separately for the Chinese and the UK subjects in a *within-culture* comparison. We report 4 contrasts for both cultures: T3 vs T5 and T2 vs T4 (for the effects of presence/absence of punishment, keeping the size of normative signal fixed); and T2 vs T3 and T4 vs T5 (for the effects of changes in the normative signals, keeping fixed the presence/absence of punishment). Proposition 2(b) predicts the effect of punishments on contributions. It requires the contribution levels in these contrasts to be: T3 $>$ T5 and T2 $>$ T4. Proposition 2(aii) tests for the effects of social norms on contributions. It requires that the contribution levels in the contrast should be: T2 $<$ T3 and T4 $<$ T5. We find that these predictions hold for 7/8 cases. The only exception is for UK subjects in the contrast T4 $<$ T5.

Since our normative signals, $s = 9$ and $s = 12$, are close to each other (these are based on actual pilots and are not hypothetical) we do not get statistical significance in every case. The statistically significant cases are indicated in Figure 4.

1. For the Chinese subjects, the statistically significant cases are: T2 $>$ T4 (the punishment mechanism improves contributions when the signal of normative expectations is low, $s = 9$) and T5 $>$ T4 (the high signal of normative expectations, $s = 12$, increases contribution in the absence of a punishment mechanism, relative to a low signal $s = 9$).[48]

2. For the UK subjects, the statistically significant case is T3 $>$ T5 (the punishment mechanism improves contributions when the signal of normative expectations is high, $s = 12$).[49]

[48]The Mann-Whitney test of the differences in contributions of Chinese subjects in treatments (i) T2 vs T4, (ii) T4 vs T5, respectively, gives a p-value of 0.065 and 0.007.

[49]The Mann-Whitney test of the differences in contributions of UK subjects in treatments T3 vs T5 gives a p-value of 0.045.

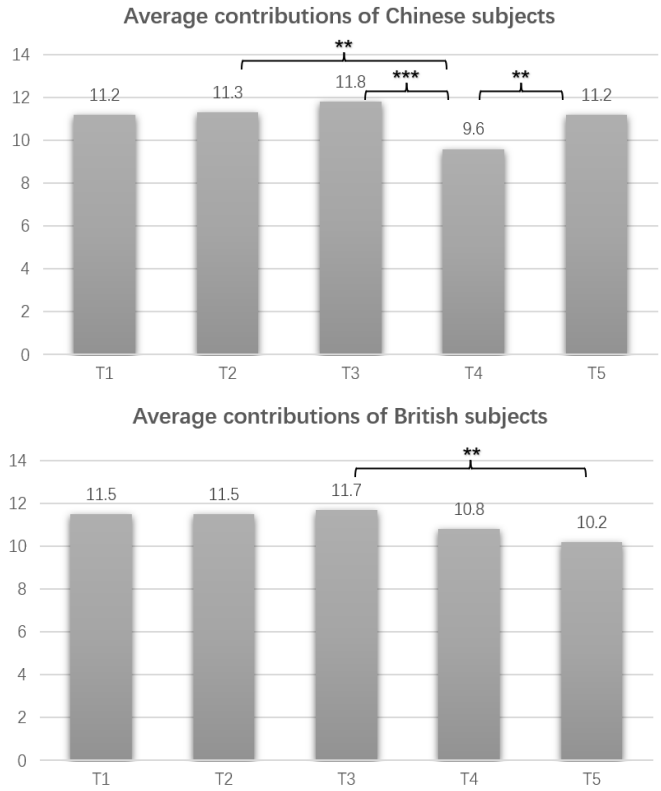

Figure 4: Comparison of the contributions between Chinese and British subjects. Superscripts ** and *** denote the statistical significance of the t test at 5% and 1%, respectively.

7.2 Determinants of Contributions

Table 3 reports the results of robust OLS regression models to explore the determinants of contributions. Model 1 uses the data from all 5 treatments T1,...,T5. Model 2, because it introduces a dummy variable for high and low normative signals, omits treatment T1, because social norms are absent in treatment T1. In the two models in Table 3, we did not simultaneously use the independent variables $E_i g_j$ and $norm$[50], because they are highly correlated (Spearman correlation coefficient= 0.26, p-value= 0.000). We explore the reason for the high correlation in terms of the mediating effect of s on $E_i g_j$ below.

The regression results from Model 1 show that expectations of higher contributions from the partner, $E_i g_j$, significantly improve contributions.[51] The existence of the punishment mechanism (through the dummy variable 'punishment'), holding other things fixed, improves the level of contributions. Thus, even though players know that there are no dynamic economic linkages between the two stages, they appear to forecast, correctly, dynamic psychological linkages. Namely, that frustration-aversion is likely to trigger Stage 2 contributions from the partners, if they contributed low amounts. From Model 2, there is no statistically significant "direct" effect of the dummy variable for the signal of normative expectations, s, high or low. This is likely

[50] Recall from Section 5 that the variable 'norm' takes a value 1 when $s = 12$ and a value 0 when $s = 9$.

[51] This mechanism is consistent with *contemporaneous anticipated conditional reciprocity* (Rabin, 1993). Dhami et al. (2019) demonstrated theoretically, and empirically verified, this effect in public goods games. Their demonstration can also be used to explain our empirical results.

Table 3: Determinants of Stage 1 contributions. Superscripts *, **, and *** denote the statistical significance at 10%, 5% and 1%, respectively.

	Model 1	Model 2
Treatment	T1,T2,T3,T4,T5	T2,T3,T4,T5
$E_i g_j$	0.98***	
	[0.048]	
norm		0.56
		[0.493]
punishment	0.54*	1.07**
	[0.319]	[0.494]
China	0.24	2.06***
	[0.446]	[0.679]
age	0.07	0.03
	[0.095]	[0.119]
male	0.96*	3.02***
	[0.515]	[0.769]
experience	0.30	0.16
	[0.320]	[0.533]
business	-0.50	-0.91
	[0.364]	[0.579]
China × male	-1.67**	-4.05***
	[0.673]	[1.053]
constant	-1.81*	8.36***
	[0.673]	[2.606]
F-stat	91.19***	3.27***
Adjusted R^2	0.55	0.04
No. Obs.	477	371

because the low and the high signals of normative expectations, $s = 9, s = 12$, are quite close to each other, having been generated from previous pilots.

Since $E_i g_j$ and the dummy variable '*norm*' are highly correlated, we conjecture that there might be a *mediating effect* of $E_i g_j$. In other words, the independent variable, *norm*, which precedes $E_i g_j$ in the sequence of moves (see Section 2.4), may influence the mediator variable, $E_i g_j$, which in turn influences the dependent variable, the choice of contributions. In order to test this conjecture, we first ran the Sobel test, and we find that the mediating effect of $E_i g_j$ is significantly positive ($z = 2.6$, p-value$= 0.009$). Then we used the non-parametric Preacher and Hayes bootstrapping method; the indirect effect[52] is 0.97. The 95% confidence interval is $[0.21, 1.70]$, which does not contain zero, hence, the indirect effect is statistically significant. Thus, a high signal of normative expectations increases the subject's expected contributions from the partner, which in turn increases the subject's own contribution choices.

There are cultural gender differences in both models. In Model 1, the differences are as follows. Chinese males contributed less than Chinese females ($0.96 - 1.67 = -0.71$), while British males contributed significantly more than British females (0.96).[53] Chinese males contributed significantly less than British males ($0.24 - 1.67 = -1.43$). Chinese females contributed more than British females, but the difference (0.24) is not statistically significant. Similar differences are found for Model 2, except that the differences between British males and British females become even greater (3.02); Chinese females now contribute even more relative to British females (2.06); Chinese males contribute even lower relative to British males ($2.06 - 4.05 = -1.99$); and Chinese males contributed even lower relative to Chinese females ($3.02 - 4.05 = -1.03$). These results speak to the literature on behavioral differences between WEIRD and non-WEIRD societies.

7.3 Beliefs on contributions

There are no statistically significant differences in the belief distributions between Chinese and UK subjects (see supplementary section).

Figure 5 shows an *across-cultures* comparison of average beliefs of the contributions of Chinese and British subjects in the 5 treatments. The cultural differences in average beliefs are not statistically significant, except in the following two cases where the Chinese subjects expected relatively higher contributions. (i) Treatment T2 (presence of punishment and low signal of norms $s = 9$). (ii) Treatment T5 (absence of punishment mechanism and a high signal of norms $s = 12$). Thus, in the absence of a punishment mechanism, the Chinese subjects expected relatively higher contributions only when the social norm is stronger ($s = 12$). However, in the presence of the punishment mechanism, the Chinese subjects expected relatively higher contributions even under weaker norms ($s = 9$).

In Figure 6, we compare the *within-culture* differences in the average expectation of the partner's contributions across the different treatments. The statistically significant differences

[52] The indirect effect measures the extent to which the dependent variable, the contribution choice, changes when the independent variable, *norm*, is held fixed and the mediator variable, $E_i g_j$, changes by the amount it would have changed had the independent variable, *norm*, increased by one unit.

[53] The three-way interaction, China × male × norm, had an insignificant effect on contributions.

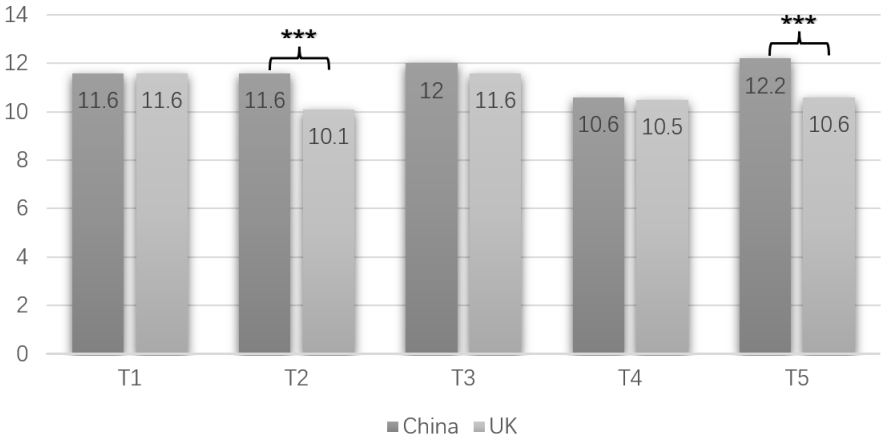

Figure 5: Average belief of contribution.

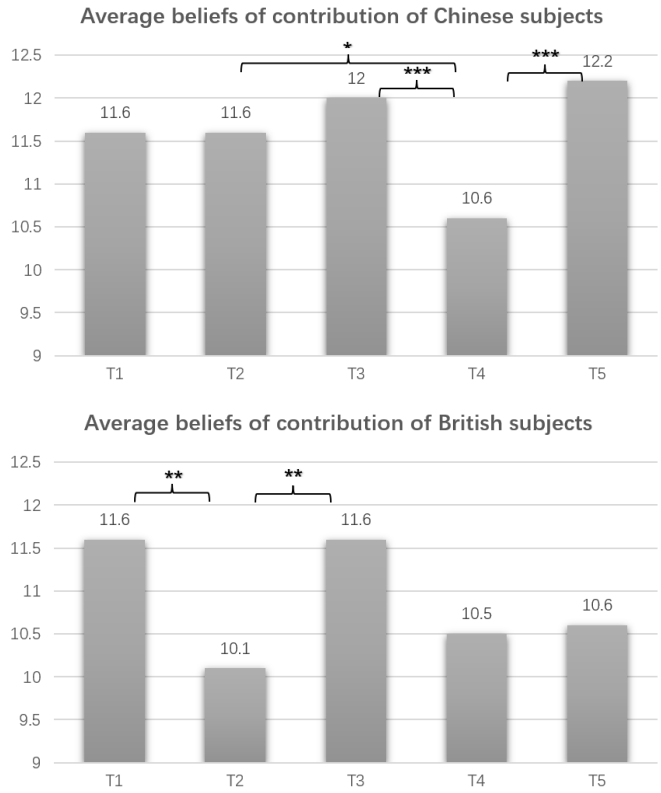

Figure 6: Beliefs of contributions of Chinese and British subjects.

are listed below.

(a) *Chinese subject pool*

T2 > T4: When the signal of normative expectations is low, $s = 9$, the presence of the punishment option (relative to its absence) increases the expectation of the partner's contribution.

T5 > T4: In the absence of a punishment option, subjects increase their expectation of the partner's contribution if the signal of normative expectations is high, $s = 12$, relative to when it is low, $s = 9$.

T3 > T4: When the signal of normative expectations is high, $s = 12$, and a punishment mechanism is available, expected contributions from partners are higher relative to a low signal of normative expectation, $s = 9$, and absence of punishment.

(b) *British subject pool*

T1 > T2: In the presence of the punishment option, a low signal of normative expectations, $s = 9$, decreases expectations of partner's contribution relative to the complete absence of a signal of normative expectations (which is true of T1).

T3 > T2: In the presence of the punishment option, a high signal of normative expectations, $s = 12$, increases the expectations of partner's contribution, relative to a low signal, $s = 9$.

Table 4: Determinants of the beliefs of the partner's contribution. Superscript *** denotes statistical significance at 1%.

	Dependent variable: $E_i g_j$
punishment	0.35
	[0.352]
norm	0.98***
	[0.354]
China	1.90***
	[0.466]
age	0.01
	[0.073]
male	2.16***
	[0.567]
experience	-0.13
	[0.399]
business	-0.37
	[0.409]
China×male	-2.17***
	[0.779]
constant	9.21***
	[1.599]
F-stat	5.62***
Adjusted R^2	0.06
No. Obs.	371

Table 4 reports the results of a robust OLS regression to find the determinants of the beliefs of the partner's contribution. We only use data for treatments T2, T3, T4, and T5, since these four treatments reveal the signals of normative expectations to the subjects. The dependent

variable is the subject's expectation of the partner's contribution. An increase in the signal of normative expectations, s, significantly increases the beliefs of contributions (coefficient of norm is 0.98***). There are important gender differences of a cultural nature. We highlight two calculations of expectations that are also closely reflected in the actual contributions of the relevant subjects in Section 7.2.

1. Chinese males expected significantly lower contribution from the partners as compared to Chinese females ($2.16 - 2.17 = -0.01$). British males expected greater contributions from partners as compared to British females (2.16).

2. Chinese males expected relatively less contribution from the partners as compared to British males ($-2.17 + 1.90 = -0.27$). Chinese females expected more contribution from partners as compared to British females (1.90).

8 Emotions and Punishments

8.1 Relation between emotions and punishments

Table 5 shows the rounded percentages and the proportion of subjects who self-report various negative emotions at the end of Stage 1, the contribution stage. Note that subjects can self-report feeling a range of emotions at the same time. For instance, a subject self-reporting frustration may also report being angry/indignant/dissatisfied. The first row of Table 5 considers subjects who discovered, at the end of Stage 1, that their partners contributed below their expectations ($E_i g_j - g_j > 0$). Following Battigalli et al. (2019), these subjects should be frustrated. However, it is worth exploring if frustration (and the other negative emotions) also arise when $E_i g_j - g_j \leq 0$, but the partner falls short of (i) one's own contributions ($g_j < g_i$), (ii) social norms ($g_j < s$), or (iii) established standards of reciprocity ($g_j < y/2$).[54]

Table 5: Consistency between the self-reported emotions and the relevant economic theory.

Self-reported Emotion		Frustration	Anger	Indignation	Dissatisfaction
$E_i g_j - g_j > 0$		64% (156/243)	74% (145/196)	72% (137/189)	62% (168/271)
$E_i g_j - g_j \leq 0$	$g_j < g_i$	7% (18/87)	8% (15/51)	7% (14/52)	7% (20/103)
	$g_j < y/2$	2% (5/87)	2% (3/51)	2% (3/52)	2% (6/103)
	$g_j < s$	13% (31/87)	9% (17/51)	8% (15/52)	13% (36/103)

In Table 5, a total of 243 subjects self-report being frustrated. Of these, 156/243, or 64%, satisfy the condition $E_i g_j - g_j > 0$. Thus, $243 - 156 = 87$ subjects are frustrated despite $E_i g_j - g_j \leq 0$. The remaining entries in the first column show how these 87 subjects are split into three subcategories. Of these 87 subjects, 18 satisfy $g_j < g_i$ (this is 7% of 243), 5 satisfy $g_j < y/2$, and 31 satisfy $g_j < s$. The remaining entries in the table can be read in a similar manner. Since the subcategories in the second row in Table 5 are not necessarily mutually

[54]The specific inequality $g_j < y/2$ arises from negative sequential conditional reciprocity in Stage 2 if we use the model of Dufwenberg and Kirchsteiger (2004). Details of the derivations are given in the supplementary section.

exclusive,[55] one cannot ascribe them any relative weights. However, it is clear from Table 5 that the main driver of the negative emotions, including frustration and anger, is the condition $E_i g_j - g_j > 0$, which is consistent with frustration aversion, as defined in Battigalli et al. (2019). Furthermore, if partners fall below expectations, a range of other negative emotions might be experienced, such as indignation and dissatisfaction, which was not highlighted before in the relevant theory.

Table 6: Relation between the self-reported emotions and punishment. Superscripts ** and *** denote statistical significance at 5% and 1%, respectively.

Emotion	China	UK
frustration	0.25***	0.31***
anger	0.35***	0.40***
indignation	0.31***	0.34***
elation	-0.16**	-0.20**
satisfaction	-0.19**	-0.18**
dissatisfaction	0.35***	0.25***

Table 5 is not designed to capture the intensity of the emotions. Table 6 exploits the measured intensity of emotions on a 7 point Likert scale. It reports the non-parametric Spearman correlation coefficient between emotions experienced at the end of Stage 1 and the Stage 2 choice of punishment. All the correlations are highly significant for both, Chinese subjects and UK subjects. While frustration, anger, indignation, and dissatisfaction are positively correlated with the choice of punishment; elation and satisfaction are negatively correlated with the choice of punishment.[56] The relevant negative emotions are positively associated with punishment, while the positive emotions are negatively associated with punishment (all correlations are significant). This indicates that (i) our self-reported emotions data is reasonably reliable, and (ii) emotions experienced at the beginning of Stage 2 may reliably predict the punishments chosen in Stage 2 because they temporally precede the choice of punishments.

Do those who are more frustrated, or more angry, also choose higher punishments, in conformity with the transmission mechanism proposed in our paper? Table 7 reports robust OLS regressions using the data from the treatments T1, T2, and T3, where punishments are available; the dependent variable is the subject-specific punishment choice. The results are reported in Table 7. We find that frustration and anger positively affected the choices of punishment (since they are highly correlated we do not include them in the same model). Thus, more frustrated and angry subjects punished more, confirming the result in Proposition 1a(ii). Additionally, higher contributions of the partner (g_j) in Stage 1 significantly reduced the choice of of punishment, which is also consistent with the frustration-aversion hypothesis.[57]

[55] For instance, for any subject it might be simultaneously true that $g_j < y/2$ and $g_j < s$.

[56] Since we described *shame* as "a painful feeling of humiliation or distress caused by the self-realization of socially inappropriate behavior on your part alone which has nothing to do with your partner's decision", we calculated the Spearman correlation coefficients of shame and the choices of contribution, rather than punishment (China: -0.21***; UK: -0.28***).

[57] We also tried several interaction terms in the models in the Table 7, but they were not statistically significant.

Table 7: Punishment and self-reported frustration/anger. Superscripts *, **, and *** denote statistical significance at 10%, 5% and 1%, respectively.

	Model 1	Model 2
frustration	0.16***	
	[0.057]	
anger		0.29***
		[0.067]
g_j	-0.08***	-0.04*
	[0.021]	[0.022]
China	0.41*	0.26
	[0.243]	[0.233]
age	0.01	0.03
	[0.044]	[0.044]
male	-0.33	-0.42**
	[0.209]	[0.203]
experience	-0.27	-0.39*
	[0.233]	[0.220]
business	0.18	-0.01
	[0.217]	[0.212]
constant	1.61	1.04
	[1.054]	[1.045]
F-stat	8.52***	10.17***
Adjusted R^2	0.14	0.18
No. Obs.	290	290

8.2 Changes in emotions after punishment

In Figure 7, we report the data on self-reported emotions of subjects before they punish and after they punish. For each emotion, the left histogram represents the average strength of emotions that subjects felt at the end of Stage 1, just after they had observed the contribution of the partner. The right histogram represents the average strength of emotions felt by subjects at the end of Stage 2, just after they had made the punishment decision, but before they knew about how much they had been punished by their partner. Only the data from the treatments that had a punishment mechanism (T1, T2, and T3) are used here.

For Chinese and British subjects, the strength of all emotions declined after punishing the partner. All changes are significant for Chinese subjects. For British subjects, the decline in shame, elation, satisfaction, and dissatisfaction were statistically significant. The decline in the strength of frustration, anger, and indignation after punishing the partner, is consistent with the *venting hypothesis* (Dickinson and Masclet, 2015). However, it is puzzling that the strength of elation and satisfaction, which are positive emotions, also declined significantly. Table 8 reports the results of an ordered Probit model (one-tail) to try to solve the puzzle, which is a new result in the literature.

In Table 8, the dependent variables of the ordered probit models are the subjects' self-reports of satisfaction and elation after punishing the partner, but before knowing the partner's choice of punishments. The categories are the emotional intensities from 0-7 on a Likert scale (with

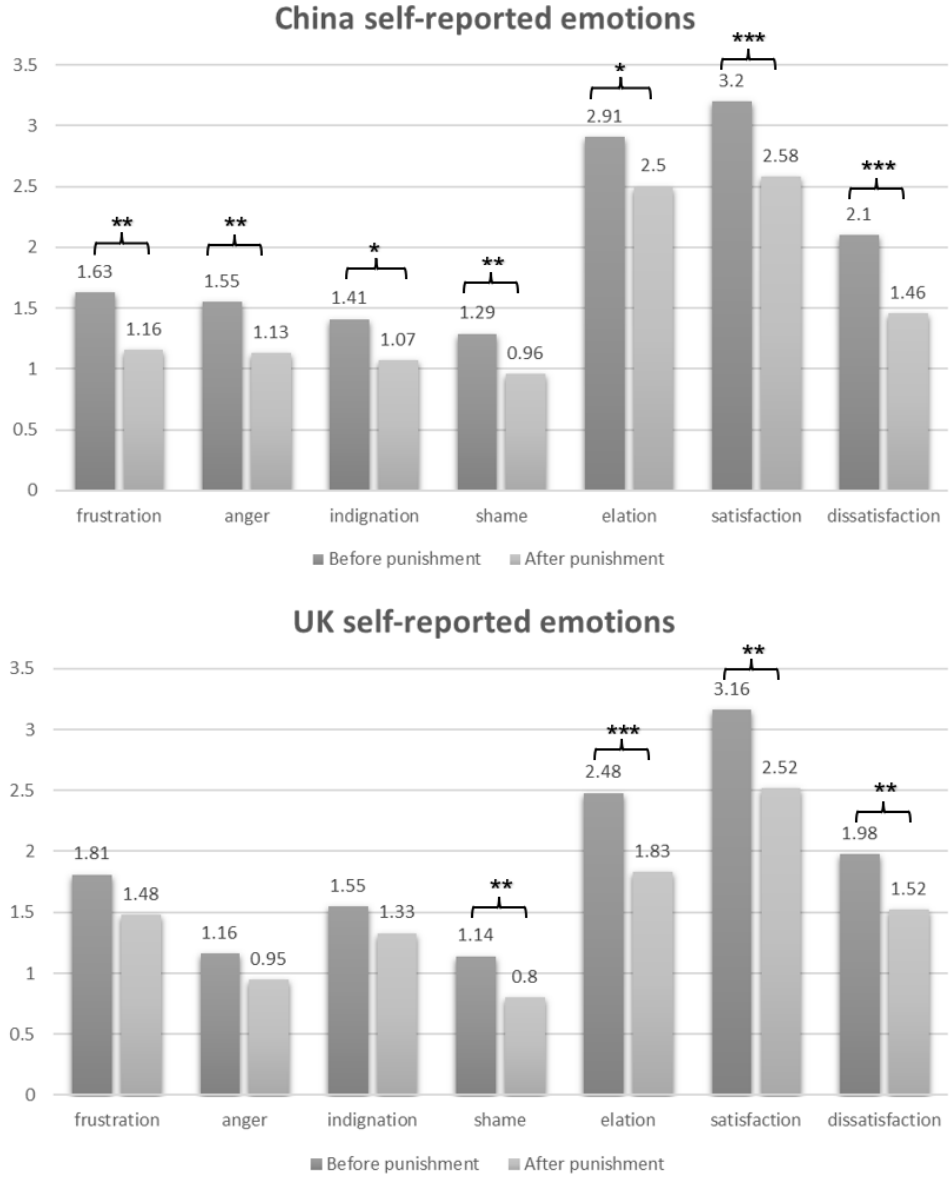

Figure 7: Self-reported change in emotions after punishment for Chinese and British subjects. Superscripts *, **, and *** denote statistical significance at 10%, 5% and 1%, respectively.

0 representing no emotion and 7 representing the highest intensity of the emotion). μ_1, μ_2, ... , μ_7 are the cut points of the ordered probit models. Results are shown for Chinese subjects (CN) and subjects from the United Kingdom (UK).

Table 8: Why did satisfaction decline after punishing? Results of ordered probit models. Superscripts *, **, denote the statistical significance at 10% and 5%, respectively. Standard errors in brackets.

Emotion	Satisfaction		Elation	
Sample	CN	UK	CN	UK
Model	1	2	3	4
p_i	-0.11**	-0.00	-0.05	-0.08
	[0.052]	[0.064]	[0.052]	[0.067]
$E_i p_j$	-0.01	-0.14**	-0.06	-0.09*
	[0.054]	[0.062]	[0.054]	[0.064]
μ_1	-1.00	-1.03	-0.89	-0.68
	[0.143]	[0.164]	[0.140]	[0.156]
μ_2	-0.51	-0.54	-0.40	-0.15
	[0.133]	[0.153]	[0.131]	[0.150]
μ_3	0.03	-0.124	-0.03	0.19
	[0.129]	[0.150]	[0.129]	[0.150]
μ_4	0.39	0.31	0.32	0.67
	[0.131]	[0.151]	[0.130]	[0.157]
μ_5	0.56	0.60	0.68	0.98
	[0.133]	[0.156]	[0.136]	[0.170]
μ_6	0.89	0.89	0.87	1.67
	[0.143]	[0.165]	[0.143]	[0.237]
μ_7	1.18	1.56	1.15	2.24
	[0.157]	[0.220]	[0.157]	[0.380]
Log likelihood	-318.42	-245.15	-318.67	-219.08

From Table 8, one's own choice of punishment (p_i) and the expectation of the partner's punishment choice ($E_i p_j$) reduced the emotions of satisfaction and elation. The reason is that punishing others is costly, and the expectation of being punished is aversive, either in monetary or non-monetary terms, or both. Both factors are likely to reduce satisfaction. However, there are cultural differences. The reduction in *satisfaction* from costly punishment is statistically significant for Chinese subjects but not the UK subjects. However, the aversion from expected punishment is statistically significant in explaining the reduction in *elation* for UK subjects, but not the Chinese subjects. This is supported by the marginal effects calculated from Table 8 and these effects are strongest for those who experience the most intense emotions; the details on the marginal effects can be found in the supplementary section.

9 Conclusions

We combine the literatures on social norms and belief-based modeling of emotions, to explore the foundations of human contributions and punishments in social dilemma situations. We use a modified voluntary contributions mechanism (VCMP) that eliminates dynamic economic

linkages between the two stages of a VCMP but contains dynamic psychological linkages. Our theoretical framework uses belief hierarchies and models in psychological game theory, which we use to derive the relevant predictions. Our solution concept of a psychological best response is motivated by the evidence.

We show that social norms and emotions are important determinants of contributions and punishments, which is consistent with the predictions of our model. Social norms influence the expectations of players about the actions of others in a culture-specific manner, and play a mediating role in determining the expectations, and the contributions of players. More frustrated players punish more, in line with our hypothesis on dynamic psychological linkages. Frustration and anger are also triggered by players falling below the social norms (due to the mediating role of social norms) or below one's own contributions (due to the false consensus effect). There are important cultural and gender differences between Chinese and UK subjects in their (i) contributions, (ii) punishments, and (iii) beliefs. The actions of players are consistent with their hetergenous beliefs, and support our use of a psychological best response to beliefs. We also find evidence supporting the venting role of emotions, and the decline in positive emotions after costly punishment, such as elation and satisfaction, which differ in important ways between Chinese and UK subjects.

While the extensive predictions of our theoretical model are largely supported by the data, there are also important findings that call for a richer model. For instance, the actions of players are influenced by the contemporaneous expected actions of their partners (e.g., higher expected contributions (resp. punishments) by the partner, triggers greater contributions (resp. punishments). This can be explained using models of contemporaneous reciprocity in Rabin (1993); indeed this has been done for public goods game already (Dhami et al., 2019). Our results also show that falling behind the expectations of the partners triggers not just frustration and anger, but also a range of other negative emotions. These findings suggest an even richer theory of emotions, within the ambit of psychological game theory, than the ones that have been developed so far. Yet, our relatively parsimonious model successfully explains most of our data.

Acknowledgements

We are grateful to the National Natural Science Foundation of China (72003100), and China Postdoctoral Science Foundation (2020M670616) for the funding of this research in the UK and in China.

Appendix: Proofs

Proof of Lemma 1: Integrating the LHS by parts, we get

$$\int_{g'=g_i}^{y} \left(g' - g_i\right) dH_i^2(g' \mid s) = \left|\left(g' - g_i\right) H_i^2(g' \mid s)\right|_{g_i}^{y} - \int_{g'=g_i}^{y} H_i^2(g' \mid s) dg'.$$

Simplifying the RHS, we get the desired result. ■

Proof of Proposition 1: Differentiating (4.3) again, we get

$$\frac{d^2U_{i2}(p_i \mid E_ip_j)}{dp_i^2} = \kappa^2 v_i''(y - \kappa p_i - E_ip_j) < 0;\ i = 1, 2. \tag{9.1}$$

Thus, U_{i2} is a strictly concave, twice continuously differentiable, function of p_i and it is defined over the compact interval $p_i \in [0, \overline{p}]$. Hence, given E_ip_j, a unique solution, $p_i^* \in [0, \overline{p}]$, exists; this satisfies Definition 1(i). Suppose that $\lambda_i > 0$. Using (9.1), setting the RHS of (4.3) equal to zero and implicitly differentiating, we get the following comparative static results.

(ai) $\frac{\partial p_i^*}{\partial s} = \left(-\partial^2 U_{i2}/\partial p_i^2\right)^{-1}\left(\lambda_i \frac{\partial E_i(g_j|s)}{\partial s}\right) \geq 0$. [Using Assumption 2]

(aii) $\frac{\partial p_i^*}{\partial [E_i(g_j|s) - g_j]^+} = \left(-\partial^2 U_{i2}/\partial p_i^2\right)^{-1}\lambda_i > 0.$

(aiii) $\frac{\partial p_i^*}{\partial \lambda_i} = \left(-\partial^2 U_{i2}/\partial p_i^2\right)^{-1}[E_i(g_j \mid s) - g_j]^+ > 0.$

(bi) $\frac{\partial p_i^*}{\partial g_j} = \left(-\partial^2 U_{i2}/\partial p_i^2\right)^{-1}(-\lambda_i) < 0.$

(bii) $\frac{\partial p_i^*}{\partial E_ip_j} = \left(-\partial^2 U_{i2}/\partial p_i^2\right)^{-1}(\kappa v_i''(y - \kappa p_i - E_ip_j)) < 0.$ ■

Proof of Proposition 2:

(ai) Differentiating (4.7) again, we get

$$\frac{\partial U_{i1}^2}{\partial g_i^2} = v_i''(y - g_i) - \mu_i \frac{\partial H_i^2(g_i \mid s)}{\partial g_i} + \delta \frac{\partial^2 V_{i2}}{\partial g_i^2}. \tag{9.2}$$

The first two terms on the RHS of (9.2) are negative. Assumption 4 guarantees that the last term is non-positive. It follows that U_{i1} is a strictly concave function of g_i defined over the compact interval $[0, y]$, hence a unique optimal value, g_i^*, exists.

(aii) Using the implicit function theorem, we find

$$\frac{\partial g_i^*}{\partial s} = \left(-\frac{\partial^2 U_{i1}}{\partial g_i^2}\right)^{-1}\left(-\mu_i \frac{\partial H_i^2(g_i^* \mid s)}{\partial s} + \delta \frac{\partial}{\partial s}\left(\frac{\partial V_{i2}}{\partial g_i}\right)\right).$$

Using Assumption 3, $\frac{\partial H_i^2(g_i^*|s)}{\partial s} \leq 0$, and from Assumption 4(iii), $\frac{\partial}{\partial s}\left(\frac{\partial V_{i2}}{\partial g_i}\right) \geq 0$, hence $\frac{\partial g_i^*}{\partial s} \geq 0$.

(b) In the absence of a punishment option, we have a one-stage game with contributions only. Hence, the last term in (4.7), which gives a positive marginal effect on contributions (Assumption 4), is zero. This reduces the positive marginal effect of Stage 1 contributions, reducing them.

(c) Follows from part (aii) by setting $\delta = 0$ in (4.7). ■

References

[1] Aina, C., Battigalli, P., Gamba, A. (2020). Frustration and anger in the Ultimatum Game: An experiment. Games and Economic Behavior 122: 150–167.

[2] Bartke, S., Bosworth, S. J., Snower, D. J., & Chierchia, G. (2019). Motives and comprehension in a public goods game with induced emotions. Theory and Decision, 86(2), 205–238.

[3] Bartling, B., Weber, R. A., and Yao, L. (2015). Do Markets Erode Social Responsibility? The Quarterly Journal of Economics 130(1): 219-266.

[4] Battigalli, P. and Dufwenberg, M. (2022). Belief-Dependent Motivations and Psychological Game Theory. Journal of Economic Literature 60(3):833–82.

[5] Battigalli, P. and Dufwenberg, M. (2009). Dynamic Psychological Games. Journal of Economic Theory, 144(1): 1–35.

[6] Battigalli, P. and Dufwenberg, M., Smith, A. (2019). Frustration, aggression, and anger in leader-follower games. Games and Economic Behavior 117:15–39.

[7] Bellemare, C., Sebald, A. and Strobel, M. (2011). Measuring the Willingness to Pay to Avoid Guilt: Estimation Using Equilibrium and Stated Belief Models. Journal of Applied Econometrics 26(3): 437--453.

[8] Ben-Shakhar, G., Bornstein, G., Hopfensitz, A., and van Winden, F. (2007). Reciprocity and emotions in bargaining using physiological and self-report measures. Journal of Economic Psychology, 28(3): 314–323.

[9] Bicchieri, C. (2006). The Grammar of Society: The Nature and Dynamics of Social Norms. Cambridge University Press: Cambridge.

[10] Bicchieri, C., (2017). Norms in the Wild: How to Diagnose, Measure, and Change Social Norms. Oxford University Press.

[11] Bicchieri, C. and Xiao, E. (2009). Do The Right Thing: But Only if Others Do So. Journal of Behavioral Decision Making, 22(2): 191–208.

[12] Bosman, R., and van Winden, F. (2002). Emotional hazard in a power-to-take experiment. Economic Journal, 112(476): 147–169.

[13] Bosman, R., Sutter, M., and van Winden, F. (2005). The impact of real effort and emotions in the power-to-take game. Journal of Economic Psychology, 26(3): 407–429.

[14] Bowles, S., and Gintis, H. (2011). A cooperative species: Human reciprocity and its evolution. Princeton University Press: Princeton.

[15] Camerer, C. F. (2003). Behavioral game theory: Experiments in strategic interaction. Russell Sage Foundation.

[16] Carlsmith, K. M., Darley, J. M., Robinson, P. H. (2002) Why do we punish? Deterrence and just deserts as motives for punishment. Journal of Personality and Social Psychology 83: 284–299.

[17] Carpenter, J., Matthews, P., (2012). Norm enforcement: anger, indignation or reciprocity. J. Eur. Econ. Assoc.10, 555–572.

[18] Cubitt, R.P., Drouvelis, M., Gächter, S., (2011). Framing and free riding: emotional responses and punishment in social dilemma games. Experimental Economics 14, 254–272.

[19] d'Adda, G., Dufwenberg, M., Passarelli, F., Tabellini, G. (2020) Social norms with private values: Theory and experiments. Games and Economic Behavior 124: 288-304

[20] Dal Bó, Pedro, and Guillaume R. Fréchette. (2018). On the Determinants of Cooperation in Infinitely Repeated Games: A Survey. Journal of Economic Literature, 56 (1): 60-114.

[21] de Quervain, D. J. F., U. Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder, A. Buck, and E. Fehr. (2004). The neural basis of altruistic punishment. Science 305:1254-8.

[22] Dhami, S. (2019). The Foundations of Behavioral Economic Analysis. Volume II: Other-Regarding Preferences, Oxford University Press: Oxford.

[23] Dhami, S. (2020). The Foundations of Behavioral Economic Analysis. Volume IV: Behavioral Game Theory, Oxford University Press: Oxford.

[24] Dhami, S. (2020). The Foundations of Behavioral Economic Analysis. Volume V: Bounded Rationality, Oxford University Press: Oxford.

[25] Dhami, S. (2019). The Foundations of Behavioral Economic Analysis. Volume VII: Further Topics in Behavioral Economics, Oxford University Press: Oxford.

[26] Dhami, S., Arshad, J., and al-Nowaihi, A. (2022). Microfinance Contracts: Theory and Evidence. Journal of Development Economics. 158: 102921.

[27] Dhami, S, and Sunstein, C. R. (2022) Bounded rationality: Judgement, heuristics, and public policy. MIT Press: Massachussetts.

[28] Dhami, S., Wei, M. and al-Nowaihi, A. (2019). Public Goods Games and Psychological Utility: Theory and Evidence. Journal of Economic Behavior and Organization. 167: 361-390.

[29] Dhami, S., Wei, M. and al-Nowaihi, A. (2023). Classical and Beliefs-Based Models of Gift Exchange: Theory and Evidence. Forthcoming in Games and Economic Behavior. https://doi.org/10.1016/j.geb.2022.12.008.

[30] Dickinson, D.L.,Masclet, D., (2015). Emotion venting and punishment in public good experiments. Journal of Public Economics 122, 55–67.

[31] Drouvelis, M., & Grosskopf, B. (2016). The effects of induced emotions on pro-social behaviour. Journal of Public Economics, 134, 1–8.

[32] Dufwenberg, M., and Kirchsteiger, G. (2004). A Theory of Sequential Reciprocity. Games and Economic Behavior. 47(2): 268-98.

[33] Ellingsen, T., Johannesson, M., Tjøtta, S. and Torsvik, G. (2010). Testing Guilt Aversion. Games and Economic Behavior, 68(1): 95–107.

[34] Elster, J. (2011). Norms, in P. Bearman and P. Hedström, eds, The Oxford Handbook of Analytical Sociology. Oxford University Press, Oxford, UK, pp. 195–217.

[35] Eyster, E. (2019). Errors in strategic reasoning. In Bernheim, B. D., DellaVigna, S., and Laibson, D. (eds.) Handbook of Behavioral Economics: Applications and Foundations, Volume 2, pages 187–259.

[36] Fehr, E. & Schurtenberger, I. (2018a). Normative foundations of human cooperation. Nature Human Behaviour 2: 458–468.

[37] Fehr, E. & Schurtenberger, I. (2018b) The Dynamics of Norm Formation and Norm Decay Working Paper Department of Economics, University of Zurich.

[38] Fehr, E. and Gächter, S. (1999). Cooperation and Punishment in Public Goods Experiments. Working Paper No. 10. University of Zurich.

[39] Fehr, E. and Gächter, S. (2000). Cooperation and Punishment in Public Goods Experiments. American Economic Review, 90(4): 980-994.

[40] Fehr, E., and Schmidt, K. M. (1999). A Theory of fairness, competition, and cooperation. The Quarterly Journal of Economics. 114 (3): 817-68.

[41] Fessler, D. (2004). Shame in Two Cultures: Implications for Evolutionary Approaches. Journal of Cognition and Culture, 4(2): 207–262.

[42] Fessler, D. M. T. and Haley, K. J. (2003). The strategy of affect: emotions in human cooperation. In P. Hammerstein (ed.) Genetic and cultural evolution of cooperation, pp. 7-36. Cambridge: MIT Press.

[43] Frank, R.H., (1988). Passions Within Reason: The Strategic Role of the Emotions. W.W. Norton & Co., Chicago.

[44] Gächter, S., Herrmann, B. (2009). Reciprocity, culture, and human cooperation: previous insights and a new cross-cultural experiment. Philosophical Transactions of the Royal Society B: Biological Science 364: 791-806

[45] Geanakoplos, J., Pearce, D. and Stacchetti, E. (1989). Psychological Games and Sequential Rationality. Games and Economic Behavior 1(1): 60–79.

[46] Gintis, H. (2017). Individuality and Entanglement: The moral and material bases of social life. Princeton University Press: Princeton.

[47] Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences 33(2-3): 61-83.

[48] Herrmann, B., Thöni, C., Gächter, S. (2008). Antisocial punishment across societies. Science, 319, 1362–1367.

[49] Hirshleifer, J., (1987). On the emotions as guarantors of threats and promises. In: Dupre, John (Ed.), The Latest on the Best: Essays on Evolution & Optimality. MIT Press, Cambridge, pp.307–326.

[50] Hofstede, G. (1980). Culture's Consequences: International Differences in Work-Related Values. Beverly Hills, CA: SAGE.

[51] Hopfensitz, A., and Reuben, E. (2009). The importance of emotions for the effectiveness of social punishment. Economic Journal, 119(540): 1534–1559.

[52] Isen, A.M., (1987). Positive affect, cognitive processes, and social behavior. In: Berkowitz, L. (Ed.), Advances in experimental social psychology. Academic Press, San Diego, CA.

[53] Joffily, M., Masclet, D., Noussair, C. N., Villeval, M. C. (2013). Emotions, sanctions and cooperation. Southern Economic Journal, 80(4): 1002–1027.

[54] Klimecki, O. M., Sander, D., and Vuilleumier, P. (2019) Distinct Brain Areas involved in Anger versus Punishment during Social Interactions. Nature Scientific Reports Vol. 8. Article number 10556.

[55] Khalmetski, K., Ockenfels, A. and Werner, P. (2015). Surprising Gifts: Theory and Laboratory Evidence. Journal of Economic Theory, 159: 163–208.

[56] Luria, G., Cnaan, R., Boehm, A. (2015). National culture and prosocial behaviors: results from 66 countries. Nonprofit Voluntary Sector Quarterly 44, 1041–1065.

[57] Martí-Vilar, M., Serrano-Pastor, L., Sala, F. G. (2019). Emotional, cultural and cognitive variables of prosocial behaviour. Current Psychology 38, 912–919.

[58] Masclet, D., and M. C. Villeval. (2008). Punishment, inequality, and welfare: A public good experiment. Social Choice Welfare 31:475–502.

[59] Mauersberger, F. and Nagel, R. (2018). Levels of Reasoning in Keynesian Beauty Contests: A Generative Framework, in Handbook of Computational Economics. Vol. 4, Elsevier, pp. 541--634.

[60] Ostrom, E. (2000). Collective action and the evolution of social norms. Journal of Economic Perspectives 14, 137–158.

[61] Ostrom, Elinor; Walker, James and Gardner, Roy. (1992) Covenants With and Without a Sword: Self-Governance is Possible. American Political Science Review 86(2): 404–17.

[62] Ostrom, E. (1990). Governing the Commons: The Evolution of Institutions for Collective Action. New York: Cambridge University Press.

[63] Persson, E. (2018). Testing the impact of frustration and anger when responsibility is low. Journal of Economic Behavior & Organization 145: 435–448.

[64] Puurtinen, M., Mappes, T. (2009). Between-group competition and human cooperation. Proceedings: Biological Sciences 276(1655): 355–360.

[65] Rabin M. (1993). Incorporating fairness into game theory and economics. The American economic review. 83(5): 1281-1302.

[66] Sanfey, A. G., J. K. Rilling, J. A. Aronson, L. E. Nystrom, and J. D. Cohen. (2003). The neural basis of economic decision-making in the Ultimatum Game. Science 300: 1755-1758.

[67] Trivers, R. L. (1971). The evolution of reciprocal altruism. Qarterly Review of Biology 46, 35-57.

[68] Vale, G. L. and Brosnan, S. F. (2017). Inequity aversion. In J. Vonk, and T.K. Shackelford (eds.), Encyclopedia of Animal Cognition and Behavior, Springer, Cham. https://doi.org/10.1007/978-3-319-47829-6_1084-1.

[69] Weber, T. O., Beranek, B., Gächter, S. (2021). The behavioural mechanisms of voluntary cooperation in WEIRD and non-WEIRD societies. No. 2021-03. CeDEx Discussion Paper Series.

[70] Walker, J., and Halloran, M. (2004). Rewards and Sanctions and the Provision of Public Goods in One-shot Settings. Experimental Economics 7(3): 235-247.

[71] Xiao, E. & Houser, D. (2005). Emotion expression in human punishment behavior. Proceedings of the National Academy of Sciences USA 102, 7398–7401.

10 Supplementary Section: Theoretical Model

10.1 Assumption 2 is implied by first order stochastic dominance

Consider the assumption of first order stochastic dominance.

Assumption 5. *(First order stochastic dominance) Consider the following assumption on the conditional beliefs of the players.*

$$\frac{\partial F_i^1(g_j \mid s)}{\partial s} \leq 0 \text{ for all } g_j \in (0, y)\,,\ s \in S.$$

From Assumption 5, it follows that, for any signal of normative expectations $s \in S$, a higher value of s induces first order stochastic dominance in $F_i^1(g_j \mid s)$. In other words, a higher signal, s, makes it more likely, in the minds of a player, that the partner will make higher contributions. Assumption 5 is likely to hold if players assign a non-zero probability that their partner has some proclivity to follow social norms. Assumption 5 is intuitive and it is a relatively weak assumption. It does not require us to specify exactly how much higher is the expected contribution, just that player j is 'more likely' to make a higher contribution.

Lemma 2. Assumption 5 implies Assumption 2.

Proof of Lemma 2: Suppose Assumption 5 holds. Integrating (2.5) by parts, we get: $E_i(g_j \mid s) = y - \int_{g_j=0}^{y} F_i^1(g_j \mid s) dg_j$ and, hence, $\frac{\partial}{\partial s} E_i(g_j \mid s) = -\int_{g_j=0}^{y} \frac{\partial}{\partial s} F_i^1(g_j \mid s) dg_j \geq 0$, for all $s \in S$. It follows that Assumption 2 holds. ■

For the purposes of our experiments, where we use two different values of s, a testable implication is formalized by Corollary 1, below.

Corollary 1. *(First order stochastic dominance for first order positive beliefs of players): If for $s_1, s_2 \in S$, we have $0 \leq s_1 < s_2$, then $F_i^1(g_j \mid s_2) \leq F_i^1(g_j \mid s_1)$ for $g_j \in (0, y)$.*

10.2 Extension of preferences to include sequential conditional reciprocity

The choice of public goods contributions of the two players, g_1, g_2, are made public at the end of Stage 1. Based on these choices, both players may infer the *kindness intentions* of the partner. Let k_{i1}, $i = 1, 2$ be the 'actual kindness' of player i to player j in Stage 1; $k_{i1} \geq 0$ denotes kindness and $k_{i1} < 0$ denotes unkindness. The kindness intentions of the players are private information. Hence, the actual Stage 1 kindness of player j to player i is k_{j1}, but at the beginning of Stage 2 player i infers this kindness to be $\widehat{k}_{j1}$.

Based on the perceived Stage 1 kindness or unkindness of the partner, players might wish to reciprocate through their punishment choices in Stage 2; thus, reciprocity is both sequential and conditional. We define the reciprocity of player i, which is defined only in Stage 2, once the Stage 1 actions of the partner are observed, as follows

$$R_i = k_{i2}\widehat{k}_{j1}. \tag{10.1}$$

In (10.1), k_{i2} is the Stage 2 *kindness of player i towards player j, as perceived by player i* and $\widehat{k}_{j1}$ is the Stage 1 *kindness of player j to player i, as perceived by player i*. Thus, if player j

is perceived to be kind in Stage 1, $\widehat{k}_{j1} > 0$ (resp. unkind $\widehat{k}_{j1} < 0$), then player i increases own Stage 2 utility by reciprocating with kindness, $k_{i2} > 0$ (resp. unkindness, $k_{i2} < 0$).

Proposition 3. *: Suppose that we follow the Dufwenberg and Kirchsteiger (2004) definition of kindness.*[58] *Then, the Stage 2 reciprocity of player i towards player j is given by*

$$R_i = r\left(g_j - \mu y\right)\left(\left(1-\mu\right)\overline{p} - p_i\right), \tag{10.2}$$

where μ is a parameter related to the perception of player i's entitlement to an equitable payoff.[59]

Proof of Proposition 3: The calculations for reciprocity payoffs are conducted entirely in terms of material payoffs of the players and not the utilities from the material payoffs. The first stage material payoff is given by

$$m_{i1}(g_i, g_j) = (y - g_i) + r\left(g_i + g_j\right). \tag{10.3}$$

while the second stage material payoff is given by

$$m_{i2}(p_i, p_j) = y - \kappa p_i - p_j. \tag{10.4}$$

The computation of k_{j1} (actual Stage 1 kindness of player j to player i) requires the specification of an *equitable material payoff* to player i, m_{i1}^E. Following Dufwenberg and Kirchsteiger (2004), this is defined as

$$m_{i1}^E = \mu \max\left\{m_{i1}(g_i, g_j),\ g_j \in [0, y]\right\} + (1-\mu)\min\left\{m_{i1}(g_i, g_j),\ g_j \in [0, y]\right\}, \mu \in (0,1). \tag{10.5}$$

where $m_{i1}(g_i, g_j)$ is defined in (10.3). The equitable payoff, m_{i1}^E, in (10.5) is a weighted average of the maximum and the minimum payoffs that player j can guarantee player i through the contribution decision, g_j. The parameter μ represents some commonly agreed norms of behavior between the players that allow them to have a 'shared understanding' of what an equitable payoff is (Fehr and Schurtenberger, 2018a). Estimates of μ are not available in the literature, nor is it clear how to experimentally estimate μ, although Dufwenberg and Kirchsteiger (2004) suggest $\mu = \frac{1}{2}$. However, we proceed with the more general case. Given the definition of $m_{i1}(g_i, g_j)$ in (10.3), the highest possible material utility to player i arises when $g_j = y$ and the lowest when $g_j = 0$. Thus, we can rewrite (10.5) as

$$m_{i1}^E = (y - g_i) + r g_i + \mu r y, \tag{10.6}$$

We now define the actual kindness of player j to player i in Stage 1, k_{j1}, as the difference between the material payoff and the equitable payoff of player i, as perceived by player j

$$k_{j1} = m_{i1} - m_{i1}^E, \tag{10.7}$$

From (10.7), player j is kind to player i if through the choice of a contribution, g_j, player i receives expected material payoff greater than the equitable payoff. Otherwise player j is unkind

[58] The kindness functions in Rabin (1993) and Dufwenberg and Kirchsteiger (2004) are related in spirit, although the specifications are slightly different. However, only the latter model considers sequential reciprocity that we are interested in.

[59] For the exact definition of μ, see the proof of Propostion 3.

to player i. Substituting (10.3), (10.6) in (10.7), we get $k_{j1}(g_j) = r(g_j - \mu y)$. At the end of Stage 1, player $i = 1, 2$ directly observes g_j and can compute the kindness of player j using the function $k_{j1}(g_j)$, hence, player i perceives the kindness of player j to be

$$\widehat{k}_{j1} = r(g_j - \mu y), \tag{10.8}$$

where g_j is the actual choice of contribution of player j in Stage 1.

Having computed the Stage 1 kindness of the partner, player i now wishes to reciprocate in Stage 2. In Stage 2, the material payoff of player $i = 1, 2$ is given in (10.4). Both players in Stage 2 simultaneously choose their punishment levels, p_i, p_j. When choosing the punishment level p_i, player i does not observe p_j. The computation of k_{i2} (actual Stage 2 kindness of player i to player j as perceived by player i) requires the specification of a Stage 2 *equitable payoff* to player j, m_{j2}^E, as perceived by player i.

$$m_{j2}^E(p_j) = \mu \max\{m_{j2}(p_i, p_j),\ p_i \in [0, \overline{p}]\} + (1-\mu)\min\{m_{j2}(p_i, p_j),\ p_i \in [0, \overline{p}]\},\ \mu \in (0,1), \tag{10.9}$$

where $m_{j2}(p_i, p_j)$ is defined in (10.4). Note that in (10.9), the equitable payoff depends on the punishment p_j chosen by player j that is not yet observed by player i. From (10.4), it follows that $m_{j2}(p_i, p_j)$ is maximized when $p_i = 0$ and minimized when $p_i = \overline{p}$. Substituting this in (10.9) we get

$$m_{j2}^E(p_j) = [\mu(y - \kappa p_j) + (1-\mu)(y - \kappa p_j - \overline{p})] \tag{10.10}$$

We now define k_{i2} as the difference between the payoff and equitable payoff of player j in Stage 2, as perceived by player i

$$k_{i2} = m_{j2}(p_i, p_j) - m_{j2}^E(p_j), \tag{10.11}$$

Substitute (10.4), (10.10) in (10.11), we get

$$k_{i2} = (1-\mu)\overline{p} - p_i. \tag{10.12}$$

Hence, substituting (10.8), (10.12) in (10.1) we can define the reciprocity term for player i in Stage 2 by

$$R_i = r(g_j - \mu y)((1-\mu)\overline{p} - p_i). \blacksquare \tag{10.13}$$

The RHS in (10.2) is the product of the following two terms. (1) The Stage 1 kindness of player j to player i, as perceived by player i, $\widehat{k}_{j1} = r(g_j - \mu y)$. Thus, player i perceives player j to be kind if player j contributes more than a fraction μ of the Stage 1 endowment. (2) The Stage 2 kindness of player i to player j, as perceived by player i, $k_{i2} = ((1-\mu)\overline{p} - p_i)$. Thus, player i perceives being kind to player j if he/she chooses less than a fraction $1-\mu$ of the maximum punishment $\overline{p}$. In order to operationalize this definition of reciprocity, for the empirical part of our paper, we may use the suggestion of Dufwenberg and Kirchsteiger (2004) to set $\mu = \frac{1}{2}$.

The Stage 2 psychological utility, U_{i2}, of player $i = 1, 2$ in the presence of frustration and reciprocity is given by

$$U_{i2}(p_i, p_j) = u_{i2}(p_i, p_j) - \lambda_i B_i(E_i(g_j \mid s), g_j)(y - \kappa p_j - p_i) + \alpha_i R_i, \tag{10.14}$$

where $\lambda_i \geq 0, \alpha_i \geq 0$. It is then straightforward to prove the following result (the proofs available on request).

Proposition 4. *Denote by p_i^*, the optimal Stage 2 punishment chosen by player i in a psychological best response. Then,*
(i) p_i^* is decreasing in the Stage 1 contributions of the partner relative to the equitable contributions, $g_j - \mu y$.
(ii) p_i^* is decreasing (increasing) in the reciprocity aversion parameter, α_i, if $g_j > \mu y$ ($g_j < \mu y$).

10.3 Further issues with a sequential equilibrium in our model

Following the available evidence, discussed in the paper, we do not use a sequential equilibrium (SE) in our paper. We now offer further observations on some of the requirements for a SE in our model.

At the time of choosing contributions in the beginning of Stage 1, the information set of player i is $I_{i1} = \{s, E_i(g_j \mid s), H_i^2\}$. However, the information set of player i at the time of making the optimal second stage choice of punishment is $I_{i2} = \{s, E_i(g_j \mid s), H_i^2, g_i, g_j, E_i p_j, \lambda_i\}$ and the corresponding information set of player j is $I_{j2} = \{s, E_j(g_i \mid s), H_j^2, g_i, g_j, E_j p_i, \lambda_j\}$. In a SE, the Stage 2 information sets I_{i2} and I_{j2}, jointly determine the Stage 2 indirect utility, V_{i2}, but are unknown in Stage 1. In a SE, we must compute the signs of the derivatives $\frac{\partial V_{i2}}{\partial g_i}$, $\frac{\partial^2 V_{i2}}{\partial g_i^2}$, $\frac{\partial}{\partial s}\left(\frac{\partial V_{i2}}{\partial g_i}\right)$ for player $i = 1, 2$.

This requires player $i = 1, 2$ in Stage 1 to form beliefs about how each player will choose Stage 2 punishments for each possible set of Stage 1 contributions. Hence, player i will need to form third order beliefs about the second order normative beliefs of (the unknown) player j, H_j^2 and its partial derivatives; beliefs about the preference parameters of player j, μ_j, λ_j; and second order beliefs about the Stage 2 beliefs of each player about the expected punishment from the other player, conditional on all possible combinations of Stage 1 contributions. These calculations can be accomplished in theory, but require auxiliary assumptions that are either untestable or untested in our data. Imposing these restrictions is unreasonable in a one-shot experimental game. We are able to explain our data without imposing these restrictions.

11 Supplementary Section: Empirical Results

11.1 Distribution of Punishments for Chinese and UK subjects

Figure 8 compares the treatment-specific distributions of punishments between Chinese and British subjects in treatments T1, T2, T3, where the punishment mechanism is present. We use the same format as in Figure 1, but along the horizontal axis we have deviation from the expected contribution, $E_i g_j - g_j$, which is the main predictive component in the presence of frustration-averse subjects. The results are similar if we replaced $E_i g_j - g_j$ by $g_i - g_j$, or by $s - g_j$. The numbers above the histograms indicate the corresponding number of observations. Note there are no observations for the Chinese subjects in the interval $[-20, -14)$ in T2 and T3; and there are no observations on the British subjects in the interval $(14, 20]$ in T2 and

$[-20, -14)$ in T3. Conditional on players being frustrated, $E_i g_j > g_j$, they punish more in all treatments, the more frustrated they are.

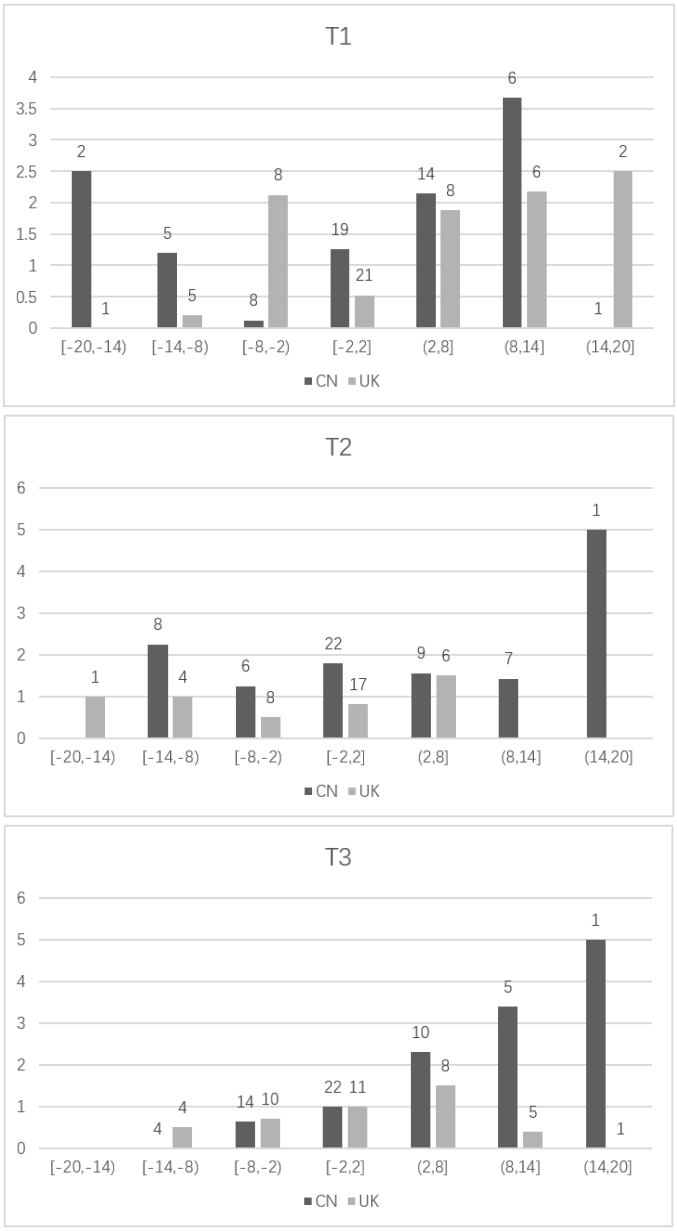

Figure 8: Distribution of punishments in treatments T1, T2, T3. Deviation from the expected contribution, $E_i g_j - g_j$, in bins of various sizes on the horizontal axis.

Table 9 reports the Kolmogorov-Smirnov (KS) test to compare the differences in the China/UK distributions of punishments for each of the three treatments T1, T2, and T3. The high p-values of the KS test show that the distributions of punishments between Chinese and British subjects are *not* significantly different in any treatment, suggesting no significant underlying cultural differences in the overall distributions.

Table 9: KS test of the distributions of punishment.

Treatment	KS Statistic	KS p-value
T1	0.155	0.551
T2	0.255	0.122
T3	0.154	0.644
All	0.149	0.086

11.2 Beliefs on Punishment

Figure 9 shows the histograms of the beliefs of the players on how much they will be punished. The results are shown separately for Chinese and British subjects and for each treatment. The horizontal axis denotes the various levels of punishment $0, 1, ..., 5$ and the vertical axis denotes the fraction of subjects holding those beliefs. The beliefs are dispersed at all levels of punishment in both cultures and the modal belief is zero punishment, although a slightly greater fraction of the Chinese subjects subscribe to the modal belief.

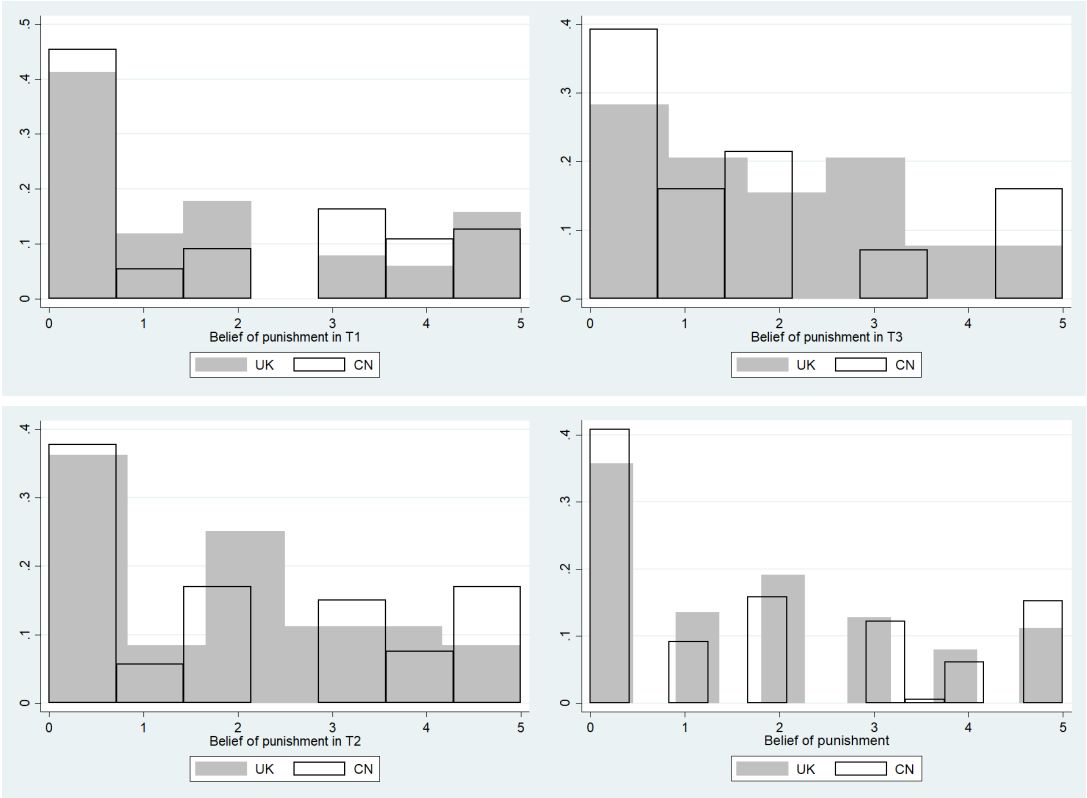

Figure 9: Distributions of the beliefs of punishment.

Table 10 reports the results of the Kolmogorov-Smirnov (KS) test to compare the cultural differences in the distributions of the beliefs of punishment between Chinese and UK subjects in each treatment. However, there are no statistically significant differences in any of the treatments.

Figure 10 shows the average beliefs of the punishment of Chinese and British subjects in the three treatments, in the presence of the punishment mechanism (T1, T2, T3). However,

Table 10: Kolmogorov-Smirnov (KS) test of the distributions of the beliefs of punishment.

Treatment	KS Statistic	KS p-value
T1	0.106	0.928
T2	0.091	0.995
T3	0.127	0.853
All	0.051	0.992

the average beliefs of punishment are not significantly different across cultures or treatments.

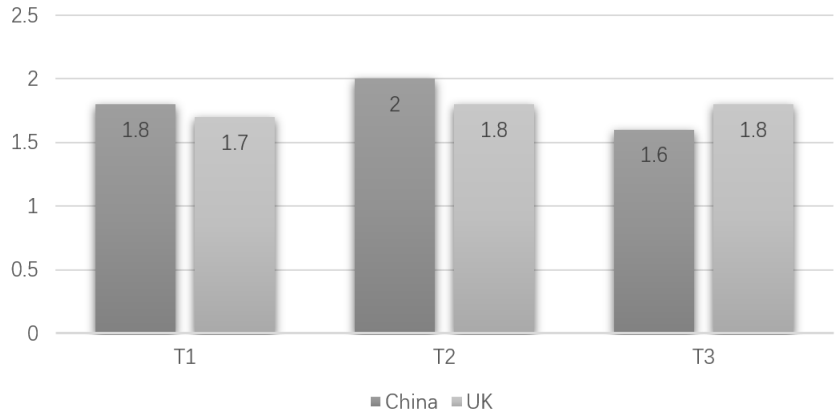

Figure 10: Comparison of the average beliefs on punishment between China and UK.

Table 11 reports robust OLS regression estimates of the determinants of the beliefs of punishment. Since the variables $g_i - g_j$ and $s - g_i$ are highly correlated, each model in Table 11 only uses one of the two variables. Recall that normative expectations are absent in treatment T1 but present in treatments T2 and T3, hence Model 1 does not include the variable $s - g_i$. From Table 11, subjects expected to be punished less, the greater is the difference in contributions from the partner $(g_i - g_j)$. But they expected to be punished more the greater was the shortfall in their contributions relative to the social norm, $(s - g_i)$. Males expected to be punished less.[60]

11.3 Distribution of contributions

Figure 11 plots the histograms of contributions, $g_i \in [0, 20]$, for each of the 5 treatments and for pooled data across all treatments for Chinese and UK subjects. The vertical axis measures the proportion of subjects making a particular level of contribution. The contributions are dispersed, but with a well defined mass around a contribution level of 10 in each treatment.

Table 12 reports the results of a Kolmogorov-Smirnov (KS) test for cultural differences in contributions. The high p-values of the KS test indicate that the distributions of the contributions between Chinese and British subjects are *not* significantly different in each treatment and in the pooled data across all the five treatments.

Table 13 reports the results of a Kolmogorov-Smirnov (KS) test to study the differences in contributions for Chinese and UK subjects. Out of the 4 pairs of contrasts reported in

[60] We tried the regressions with interacting terms, but none was significant. Hence, we did not report those regression results here to save space.

Table 11: Determinants of the beliefs of players about punishment. Superscripts ** and *** denote the statistical significance at 5% and 1%, respectively.

Treatment	Model 1 T1,T2,T3	Model 2 T2,T3
$g_i - g_j$	-0.04** [0.018]	
$s - g_i$		0.10*** [0.025]
China	-0.26 [0.253]	-0.27 [0.293]
age	-0.02 [0.054]	-0.01 [0.050]
male	-0.62*** [0.214]	-0.76*** [0.244]
experience	-0.33 [0.253]	-0.42 [0.282]
business	0.19 [0.213]	0.10 [0.250]
constant	2.74 [1.130]	2.77 [1.071]
F-stat	3.49***	6.02***
Adjusted R^2	0.05	0.11
No. Obs.	290	184

Table 12: KS test of the differences in the contribution distributions between Chinese and British subjects.

Treatment	KS Statistic	KS p-value
T1	0.089	0.985
T2	0.095	0.990
T3	0.109	0.949
T4	0.141	0.781
T5	0.175	0.482
all	0.046	0.968

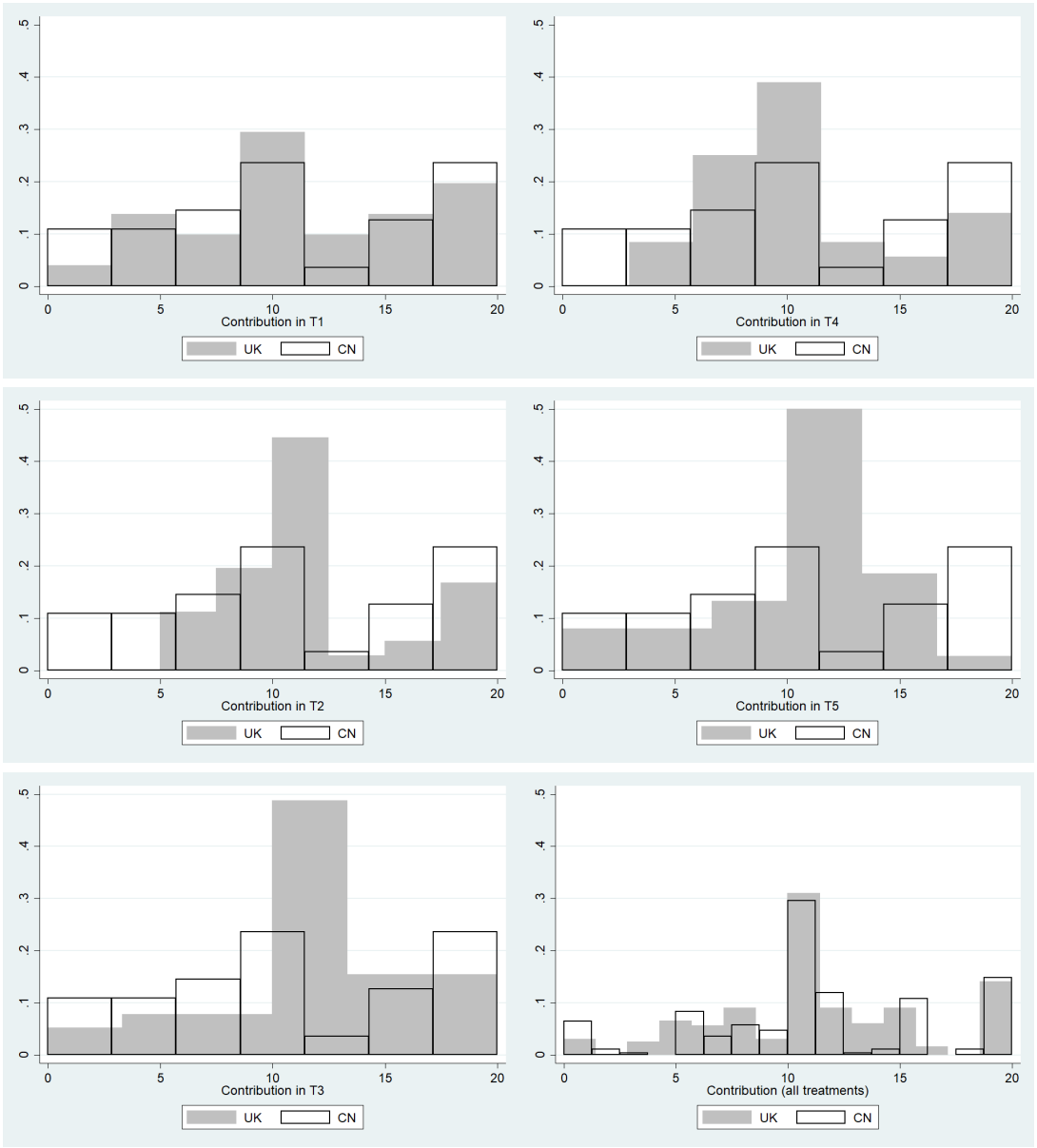

Figure 11: Distributions of contributions in each treatment.

Table 13 (the numbers in the table are p-values of contrasts), the contributions distributions are statistically different (at the 1% level) in only one contrast, T4 vs T5 for Chinese subjects; they are not statistically different in the other three contrasts.

Table 13: p-values of distributional contrasts from punishment and social norms.

Norm				Punishment			
T2 vs T3		T4 vs T5		T2 vs T4		T3 vs T5	
CN	UK	CN	UK	CN	UK	CN	UK
0.126	0.148	0.004	0.677	0.268	0.992	0.917	0.203

11.4 Distributions of beliefs about contributions

Figure 12 compares the distributions of the beliefs of contributions of Chinese and British subjects. We plot the histograms for each treatment. Along the horizontal axis, we measure the 21 different possible levels of contributions ($g_i = 0, 1, ..., 20$), and the vertical axis shows the fraction of subjects choosing each of the 21 contribution levels. In each case, the modal level of contributions is close to 10.

In order to compare the beliefs on contributions across the two cultures, we engage in two types of comparisons: a distributional comparison (and a comparison of the averages, which we undertake in the main body of the paper). The distributional comparison of the beliefs of Chinese and UK subjects is based on the results of the Kolmogorov-Smirnov (KS) test, for each treatment, reported in Table 14. The high p-values of the KS test indicate that the distributions of the beliefs of contributions of Chinese and British subjects are *not* significantly different in any treatment.

Table 14: KS test of the differences in the distributions of beliefs of contribution of Chinese and UK subjects.

Treatment	KS Statistic	KS p-value
T1	0.084	0.992
T2	0.170	0.567
T3	0.161	0.589
T4	0.114	0.939
T5	0.252	0.108
All	0.074	0.549

Table 15 uses a Kolgomorov-Smirnov (KS) test to compare the country-specific beliefs of contributions for selected treatment contrasts that were outlined in Section 7.1. In order to test the effect of the punishment option, we compared the treatments with and without punishment, keeping social norms fixed (T3 vs T5 and T2 vs T4). In order to test for the effect of social norms, we compared the treatments with high and low social norms, keeping fixed the availablity of punishment (T2 vs T3 and T4 vs T5). The results show that the belief distributions of the Chinese subjects in the contrasts T2 vs T3 and T4 vs T5 are significantly different at the 1%

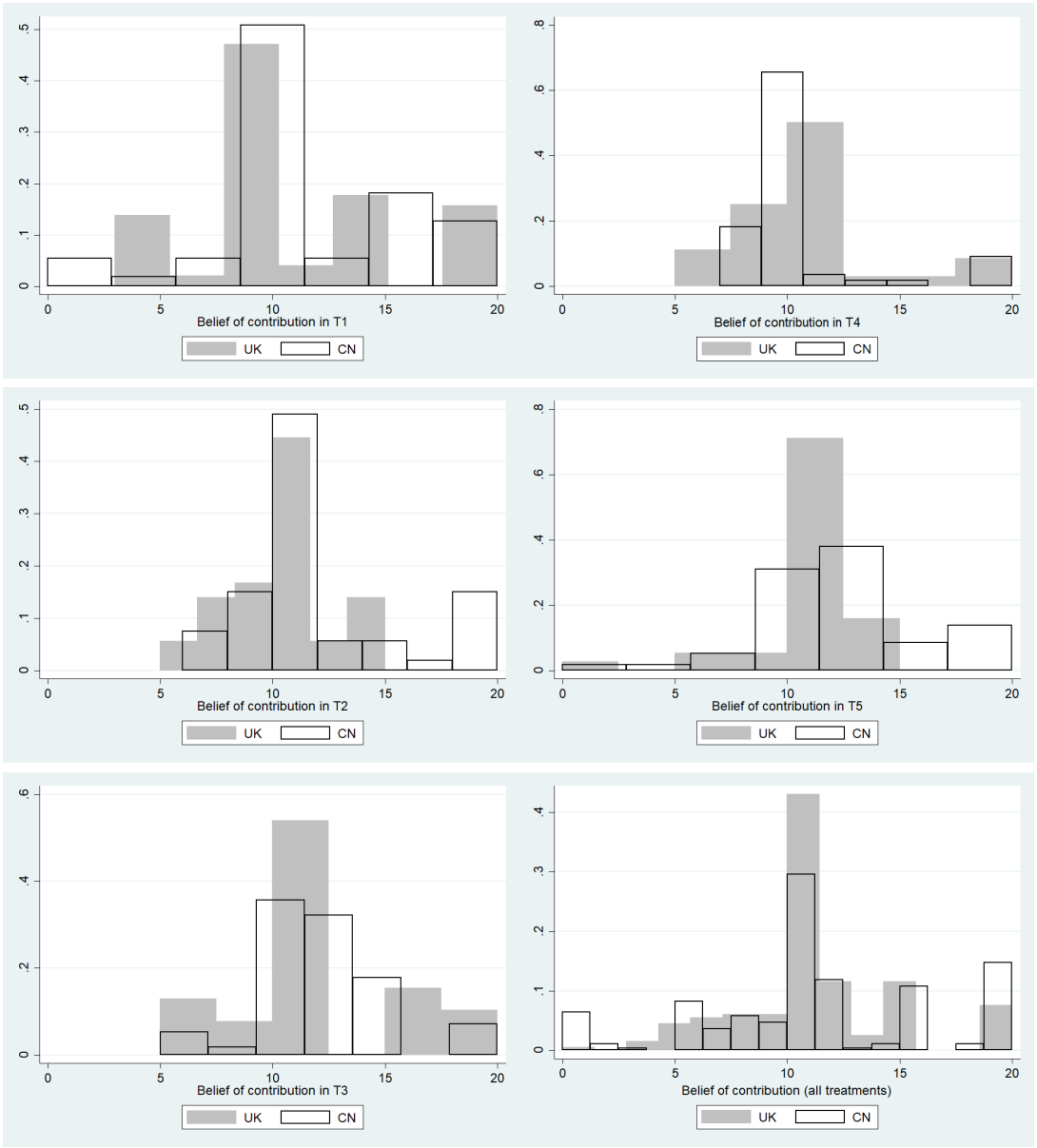

Figure 12: Distributions of the beliefs of contribution.

level. Thus, social norms significantly influence the belief distributions of Chinese subjects. None of the other comparisons in Table 15 are significantly different.

Table 15: p-values of treatment contrasts for punishment and the social norm for the distributions of beliefs of contribution, based on a Kolgomorov-Smirnov test.

Norm				Punishment			
T2 vs T3		T4 vs T5		T2 vs T4		T3 vs T5	
CN	UK	CN	UK	CN	UK	CN	UK
0.001	0.442	0.000	0.244	0.839	0.766	0.917	0.823

11.5 Marginal effects of the Probit model in Table 8

Table 16: Marginal effects of the models in Table 8. Superscripts *, **, denote the statistical significance at 10% and 5%, respectively. Standard error in brackets.

Intensity	Satisfaction (CN) Model 1		Satisfaction (UK) Model 2		Elation (CN) Model 3		Elation (UK) Model 4	
	p_i	$E_i p_j$	p_i	$E_i p_j$	p_i	$E_i p_j$	p_i	$E_i p_j$
0	0.030**	0.004	0.001	0.039**	0.017	0.020	0.027	0.031*
	[0.015]	[0.015]	[0.018]	[0.018]	[0.016]	[0.017]	[0.023]	[0.022]
1	0.009**	0.001	0.000	0.011**	0.004	0.005	0.003	0.003
	[0.005]	[0.005]	[0.005]	[0.006]	[0.004]	[0.004]	[0.003]	[0.003]
2	0.001	0.000	0.000	0.002	0.000	0.000	-0.002	-0.003
	[0.002]	[0.001]	[0.001]	[0.002]	[0.001]	[0.001]	[0.002]	[0.002]
3	-0.005**	-0.001	-0.000	-0.007**	-0.002	-0.003	-0.007	-0.008*
	[0.003]	[0.003]	[0.003]	[0.004]	[0.002]	[0.002]	[0.006]	[0.006]
4	-0.004**	-0.000	-0.000	-0.008**	-0.004	-0.005	-0.005	-0.006*
	[0.002]	[0.002]	[0.004]	[0.004]	[0.004]	[0.004]	[0.005]	[0.005]
5	-0.008**	-0.001	-0.000	-0.009**	-0.002	-0.003	-0.009	-0.011*
	[0.004]	[0.004]	[0.004]	[0.005]	[0.002]	[0.003]	[0.009]	[0.008]
6	-0.007**	-0.001	-0.001	-0.017**	-0.003	-0.004	-0.004	-0.004
	[0.004]	[0.003]	[0.008]	[0.009]	[0.003]	[0.004]	[0.004]	[0.004]
7	-0.017**	-0.002	-0.000	-0.011**	-0.009	-0.010	-0.002	-0.002
	[0.009]	[0.009]	[0.005]	[0.007]	[0.009]	[0.009]	[0.002]	[0.002]

Consider the marginal effects of the ordered probit models in Table 8 that are reported in Table 16. Among those who report a non-zero number on the Likert scale and hence, experience a strictly positive level of emotion after punishment, we get the following results.

(i) Satisfaction. Consider first the marginal effects of costly punishment, p_i. The average marginal effect for Chinese subjects, across all categories, taking account of only the significant marginal effect is -0.005.[61] Thus, a unit increase in costly punishment reduces satisfaction by 0.5%. However, the satisfaction of those who report the highest intensity of the emotion (7 on the Likert scale) declines by 17%. None of the marginal effects of costly

[61] This is calculated as follows $\frac{1}{6}(0.009 - 0.005 - 0.004 - 0.008 - 0.007 - 0.017)$.

punishment is significant for UK subjects. Now consider the marginal effects of expected punishment by the partner, $E_i p_j$. In this case, none of the marginal effects is significant for Chinese subjects. However, for UK subjects, several of the marginal effects are significant. The average marginal effect, restricting attention to the significant effects, is -0.007.[62] Thus, a unit increase in expected punishment by the partner reduces satisfaction by 0.7%. However, the marginal effect on those who report the highest two intensities of emotions (respectively, 6 and 7 on the Likert scale) is a reduction in satisfaction, by respectively, 17% and 11%.

(ii) Elation: For Chinese subjects, none of the marginal effects are significant with respect to either p_i or $E_i p_j$. For British subjects, none of the marginal effects with respect to p_i is significant. However, the average marginal effect on elation of a unit change in $E_i p_j$, restricting attention to the significant effects, is negative, -0.008.[63]

Among those subjects who experience no emotion (self report of 0 on the Likert scale), an increase in punishment, p_i, increases satisfaction by 3% for Chinese subjects, but this has no effect on UK subjects. On the other hand, an increase in $E_i p_j$ has no effect on the satisfaction reported by Chinese subjects but it increases satisfaction among UK subjects by 3.9%. With respect to elation, the only significant marginal effect for such subjects arises from $E_i p_j$; a unit change in $E_i p_j$ for UK subjects leads to an increase in elation by 3.9%. For these two statistically significant cases, there is no clear pattern of punishments meted out by subjects who self report 0 on the Likert scale.[64]

12 Supplementary section: Experimental Instructions

12.1 General information on the experiment

You are now participating in an economics experiment in which you may be able to earn money depending on your decisions and the decisions of others. During the experiment, you are not allowed to communicate with other participants.

During the experiment, our unit of money will be tokens, hence, all monetary amounts are quoted in terms of *tokens*. Your total income from the experiment, expressed in tokens, will be converted into pound sterling in cash at the end of the experiment at an exchange rate of *1 token = 0.3 pound*. Additionally, you will receive 4 pounds, as a show-up fee for participating in this experiment.

[62] This is calculated as follows $\frac{1}{6}(0.011 - 0.007 - 0.008 - 0.009 - 0.017 - 0.011)$.

[63] This is calculated as $\frac{1}{3}(-0.008 - 0.006 - 0.011)$.

[64] For Chinese subjects whose self-reported satisfaction equals 0 on the Likert scale, 42.9%(= 15/35) chose zero punishment, and the proportions of subjects choosing the punishment levels $0, 1, 2, 3, 4, 5$ are respectively 11.4%(= 4/35), 8.6%(= 3/35), 5.7%(= 2/35), 11.4%(= 4/35), 2.9%(= 1/35), and 22.9%(= 8/35). Similarly, for British subjects whose self-reported satisfaction equals 0, the corresponding proportions are respectively 28.6%(= 8/28), 10.7%(= 3/28), 17.9%(= 5/28), 14.3%(= 4/28), 10.7%(= 3/28), and 17.9%(= 5/28).

12.2 Outline of experiment

We now describe the broad outline of the experiment in brief, to be followed by a more complete description later. You will be paired with the same partner for the entire duration of the experiment. The experiment consists of the following two stages[65].

In **Stage 1** (*Voluntary contributions decision*), you decide on contributions to a joint project with a partner. Before you decide your contributions, you are given the following data taken from a previous similar experiment (you may take this as the behavior/opinion of your social or peer group): (1) The actual behavior of *most* participants in this experiment. (2) *Most* participants' opinion at the end of the experiment about the desired behavior of others.

You and your partner are provided an identical Stage 1 endowment by the experimenter. Using only these endowments, you and your partner will simultaneously, and independently, make contributions to a joint project. You cannot observe or influence each other's choices at the time of making your contribution decisions. The joint project benefits both of you equally, even if your contributions are unequal. At the end of Stage 1, you and your partner can observe each other's actual contributions. At this stage, we shall also elicit the emotions that you are experiencing; these are Frustration, Anger, Indignation, Shame, Elation, Satisfaction, Dissatisfaction.

In **Stage 2** (*Voluntary reduction decision*), you and your partner have observed Stage 1 contributions, and both of you are given identical Stage 2 endowment by the experimenter. Using only the Stage 2 endowment, you and your partner can reduce each other's Stage 2 incomes by paying a cost. At a cost of 1 token, you can reduce your partner's income by 3 tokens. The Stage 2 reduction decisions by you and your partner are made simultaneously, and independently, without observing or influencing each other's choices.

Your income in tokens is calculated separately in each stage and it depends on your decisions and the decisions of your partner. After the experiment, *only one stage* will be randomly chosen to pay you. The identity of your partner stays anonymous to you, and vice-versa.

12.3 Complete description of experiment

We now explain the economic environment in more detail before you make your actual decisions in the experiment. After reading the details, you will need to answer several questions to make sure that the instructions are clear and well understood.

Stage 1: Voluntary Contribution Decision

You and your partner both receive an identical Stage 1 endowment of 20 tokens from the experimenter and this is known to you and to your partner. You can choose any number of tokens, between 0 and 20 tokens, to contribute to a **joint project**. The joint project benefits you and your partner **equally irrespective of the actual individual contributions** that you and your partner make. The remaining tokens, net of your contributions to the public

[65]This instruction applies to the treatment T2 and T3. The treatment T1 only has contribution stage and punishment stage; T2 and T3 reveal the normative expectation (NE) and empirical expectation (EE) signals before the contribution stage and punishment stage; T4 and T5 reveal the NE and EE signals as well but only have the contribution stage.

project belong to you and are yours to keep. Your partner makes an identical decision, i.e., contribute between 0 to 20 tokens to the joint project and keep the rest for themselves. While making your decisions, neither you, nor your partner, observes how much the other contributes to the joint project. Thus, the decisions are made simultaneously and without a chance to communicate with each other. This is the **only decision** you make in Stage 1.

Suppose that you contribute t1 tokens (between 0 and 20 tokens) towards the joint project and keep 20- t1 tokens for yourself. Suppose that your partner contributes t2 tokens (between 0 and 20 tokens) and keeps $20 - t_2$ tokens for themselves. Then total investment in the joint project is denoted by $G = t_1 + t_2$.

The joint project generates a total return that is 160% of the investment G (or 1.6 times G). This return is shared equally between you and the partner, or an 80% return for each of you (or 0.8 times G).

Your **total income in Stage 1** is calculated as follows:

1. Tokens kept for yourself = $20 - t_1$ tokens

2. Return from the project = $0.8 \times G = 0.8 \times (t_1 + t_2)$

Total income is the sum of tokens kept for yourself and the return from the project, or $(20 - t_1) + (0.8 \times G)$.

Similarly, your partner's Stage 1 income is $(20 - t_2) + (0.8 \times G)$. Both of you receive identical incomes from the project but if you contribute different number of tokens, the first part of your income (tokens kept for yourself) is different.

Hypothetical example 1:

$t_1 = 10$ tokens (your contribution to the project). You have kept $20 - 10 = 10$ tokens for yourself;

$t_2 = 15$ tokens (your partner's contribution). S/he has kept $20 - 15 = 5$ tokens for himself/herself.

[Recall that you and your partner choose contributions simultaneously, and independently]

Total investment in the public project is $G = 10 + 15 = 25$ tokens.

You and your partner get an **identical return** $0.8 \times G$ or $0.8 \times 25 = 20$ tokens *despite the fact that the contributions are unequal.*

Your Stage 1 income is:

Tokens kept for yourself $(20 - 10 = 10)$ + project returns $(0.8 \times 25) = 10 + 20 = 30$ tokens.

Your partner's Stage 1 income is:

Tokens kept for himself/herself $(20 - 15 = 5)$ + project returns $(0.8 \times 25) = 5 + 20 = 25$ tokens.

Hypothetical example 2:

$t_1 = 20$ tokens (your contribution to the project). You have kept $20 - 20 = 0$ tokens for yourself;

$t_2 = 10$ tokens (your partner's contribution). S/he has kept $20 - 10 = 10$ tokens for himself/herself.

[Recall that you and your partner choose contributions simultaneously, and independently]

Total investment in the public project is $G = 20 + 10 = 30$ tokens.

You and your partner get an **identical return** $0.8 \times G$ or $0.8 \times 30 = 24$ tokens *despite the fact that the contributions are unequal.*

Your Stage 1 income is:

Tokens kept for yourself $(20 - 20 = 0)$ + project returns $(0.8 \times 30) = 0 + 24 = 24$ tokens.

Your partner's Stage 1 income is:

Tokens kept for himself/herself $(20 - 10 = 10)$ + project returns $(0.8 \times 30) = 10 + 24 = 34$ tokens.

Guess Your Partner's Contribution Decision

Before you make the Stage 1 contribution decision, you are asked to guess how much your partner will contribute to the project out of their endowment of 20 tokens. Your guess *won't* be revealed to any other participant and it remains your private information. At the end of the experiment, the computer will randomly choose one participant whose guess matches his/her partner's actual contribution and give this participant an additional prize of 5 tokens. If nobody guessed correctly, then the computer will randomly choose one participant whose guess is the closest to the partner's actual contribution, and give this participant a prize of 2 tokens.

After you write your guess of the partner's contribution, you will make your Stage 1 contribution decision.

Before you write your guess of the partner's contribution, you are given the following data taken from a previous similar experiment (you may take this as the behavior/opinion of your social or peer group): (1) The actual behavior of *most* participants in this experiment. (2) *Most* participants' opinion at the end of the experiment about the desired behavior of others.

At the end of Stage 1, the following information will be publicly announced within your group: your contributions t1; your partner's contributions t2; the total investment G; your returns from project $0.8 \times G$; and your and your partner's total Stage 1 income.

Your last task in Stage 1 is to reveal the emotions that you are feeling now [Frustration, Anger, Indignation, Shame, Elation, Satisfaction, Dissatisfaction] and the intensity of these emotions.

Stage 2: Voluntary Reduction Decision

At the beginning of Stage 2, you and your partner are given the Stage 2 endowment of 20 tokens each. All the Stage 2 decisions that you make can only use Stage 2 endowment of 20 tokens, and no other source of income. Both you and your partner possess this information.

Your only decision in Stage 2 is to choose to reduce your partner's Stage 2 endowment at some cost to yourself. We call this a *reduction decision.* If you pay a cost of 1 token from your endowment, you can reduce your partner's income by 3 tokens. You can use between 0 to 5 tokens of your endowment to reduce your partner's income. The table below shows the reduction in your partner's Stage 2 income, as you choose to give up 0 to 5 tokens. [You can also use fractions between 0 and 5, such as 1.5, 2.3, e.g., by giving up 2.5 tokens, you can reduce your partner's income by $3 \times 2.5 = 7.5$ tokens.]

Cost paid by you in tokens	0	1	2	3	4	5
Reduction of partner's income in tokens	0	3	6	9	12	15

Your partner, who also has a Stage 2 endowment of 20 tokens, faces an identical choice. S/he can also give up 1 token to reduce your income by 3 tokens. Like you, s/he can choose between 0 to 5 tokens to reduce your endowment.

Both you and your partner make the reduction decision ***simultaneously*** *without knowing the choice of the partner and without influencing each other in any way. The only information you have is the Stage 1 contribution decisions made by you and your partner and your respective incomes.*

Your Stage 2 income is calculated using the following formula:

[Tokens left over after reducing the partner's income] - [the reduction in your endowment due to your partner's decision to reduce your income].

Hypothetical Example 3

You pay a cost of 4 tokens to reduce your partner's endowment by $3 \times 4 = 12$ tokens.

Simultaneously, and unobserved to you when you make your decision to reduce your partner's income, your partner chooses 5 tokens to reduce your income by $3 \times 5 = 15$ tokens.

Your Stage 2 income is: Tokens left over by you $(20 - 4 = 16)$ – reduction in your income by the partner's reduction decision $(3 \times 5 = 15) = 16 - 15 = 1$ token.

Your partner's Stage 2 income is: Tokens left over by your partner $(20 - 5 = 15)$ – reduction in partner's income by your reduction decision $(3 \times 4 = 12) = 15 - 12 = 3$ tokens.

Hypothetical Example 4

You pay a cost of 1 token to reduce your partner's endowment by $3 \times 1 = 3$ tokens.

Simultaneously, and unobserved to you when you make your decision to reduce your partner's income, your partner chooses 3 tokens to reduce your income by $3 \times 3 = 9$ tokens.

Your Stage 2 income is: Tokens left over by you $(20 - 1 = 19)$ – reduction in your income by the partner's reduction decision $(3 \times 3 = 9) = 19 - 9 = 10$ tokens.

Your partner's Stage 2 income is: Tokens left over by your partner $(20 - 3 = 17)$ – reduction in partner's income by your reduction decision $(3 \times 1 = 3) = 17 - 3 = 14$ tokens.

After you make your decision to reduce your partner's income, but before you are informed about your partner's decision to reduce your income, we elicit the emotions that you are feeling [Frustration, Anger, Indignation, Shame, Elation, Satisfaction, Dissatisfaction] and the intensity of these emotions.

Guess Your Partner's Reduction Decision

Before you make the Stage 2 reduction decision, you are asked to guess your partner's reduction decision, i.e., how many tokens will your partner give up to reduce your income. Your guess *won't* be revealed to any other participant and it remains your *private* information. At the end of the experiment, the computer will randomly choose one participant whose guess matches his/her partner's actual reduction decision and give this participant an additional prize of 5 tokens. If nobody guessed correctly, then the computer will randomly choose one participant whose guess is the closest to the partner's actual contribution, and give this participant a prize of 2 tokens.

After you write your guess of the partner's reduction decision, you will make your Stage 2 reduction decision.

Finally, you learn about your partner's decision to reduce your income and your Stage 2 income.

Note: After the experiment, only one stage will be *randomly* chosen by the computer to pay you.

End of Experimental Instructions

Hypothetical practice questions for Stage 1

The following questions are hypothetical and only serve to enhance your understanding.

Question 1. You and your partner contribute 15 tokens each to the project. What is, in tokens,

- your total Stage 1 income?

- your partner's total Stage 1 income?

Question 2. You contribute 14 tokens. Your partner contributes 6 tokens. What is, in tokens,

- your total Stage 1 income?

- your partner's total Stage 1 income?

Hypothetical practice questions for Stage 2

The following questions are hypothetical and only serve to enhance your understanding.

Question 1. You spend 0 tokens to reduce your partner's income. Your partner simultaneously, spends 5 tokens to reduce your income. What is, in tokens,

- your total Stage 2 income?

- your partner's total Stage 2 income?

Question 2. You spend 4 tokens to reduce your partner's income. Your partner simultaneously spends 2 tokens to reduce your income. What is, in tokens,

- your total Stage 2 income?

- your partner's total Stage 2 income?

12.4 Actual Experiment Begins

You are about to start the experiment. You will be randomly paired with a partner whose identity you will never learn (and vice-versa). Your partner is given the same experimental instructions as you are. **Once you complete the decisions and go to the next page, then you cannot go back to the previous page to modify your decisions any more.**

Stage 1 (Voluntary contributions to a joint project)

You are provided with the following data from a previous similar experiment (you may take this as the behavior/opinion of your social or peer group):

(1) Most individuals contributed more than x tokens.

(2) Most individuals said that others who play this experiment, "ought to contribute" at least T tokens, or that it would be "socially desirable" to contribute at least T tokens.

Before you decide on your contributions, you must guess the partner's contributions to the joint project.

What is your best guess of how much your partner is likely to contribute? Please choose any number between 0 and 20 tokens: _____ tokens.

We now ask you to make your only decision in Stage 1, your contribution decision.

What is your contribution to the project? Please choose any number between 0 and 20 tokens: _____ tokens.

This concludes your active decisions in Stage 1. The rest of Stage 1 has two parts.

1. Information about Stage 1 outcomes

Your contribution to the joint project was: t_1 tokens

Your partner's contribution to the joint project was: t_2 tokens

The total investment in the joint project was: $G = t_1 + t_2$ tokens

Your return from the project is: $0.8 \times G$ tokens (your partner gets an identical return)

Your Stage 1 income is: $(20 - t_1) + 0.8G$ tokens.

Your partner's Stage 1 income is: $(20 - t_2) + 0.8G$ tokens

2. Your self-report of the emotions that you are experiencing at this moment.

Tick as many of the emotions that you are experiencing from the list below and then rate the intensity of the emotions on a scale of 1-7. (**Please choose 0 if the emotion you are not experiencing**).

Here is a brief guide to what these emotions mean:

Frustration: The feeling of being upset or annoyed as a result of being unable to change or achieve something.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

Anger: a strong feeling of annoyance, displeasure, or hostility.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

Indignation: anger or annoyance at what is perceived to be unfair treatment relative to what you believe is fair treatment, particularly, in your social group.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

Shame: a painful feeling of humiliation or distress caused by the self-realization of socially inappropriate behavior on your part alone which has nothing to do with your partner's decision.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

Elation: Great happiness and exhilaration.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

Satisfaction: Fulfilment of your wishes, expectations, or needs.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

Dissatisfaction: Non- fulfilment of your wishes, expectations, or needs.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

Stage 2 (Reduction Decisions)

Recall that:

Your Stage 1 contributions were x tokens and your Stage 1 income was x tokens.

Your partner's Stage 1 contributions were x tokens and his/her Stage 1 income was x tokens

.

Your endowment for Stage 2 is 20 tokens. Your partner's Stage 2 endowment is also 20 tokens. Both of you know each other's endowment is 20 tokens.

Guessing your partner's choice to reduce your income

What do you believe is the number of tokens that your partner will give up to reduce your income? Please choose any number between 0 and 5 tokens: _____ tokens.

We now ask you to make your only decision in Stage 2, your reduction decision.

How many tokens do you wish to give up to reduce your partner's income? You know that your partner will be asked to give up 3 times as many tokens. Please choose any number between 0 and 5 tokens: _____ tokens.

This concludes your active decisions in Stage 2. The rest of Stage 2 has two parts.

1. Your self-report of the emotions that you are experiencing at this moment. Tick as many of the emotions that you are experiencing from the list below and then rate the intensity of the emotions on a scale of 1-7. (**Please choose 0 if the emotion you are not experiencing**).

Here is a brief guide to what these emotions mean:

Frustration: The feeling of being upset or annoyed as a result of being unable to change or achieve something.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

Anger: a strong feeling of annoyance, displeasure, or hostility.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

Indignation: anger or annoyance at what is perceived to be unfair treatment relative to what you believe is fair treatment, particularly, in your social group.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

Shame: a painful feeling of humiliation or distress caused by the self-realization of socially inappropriate behavior on your part alone which has nothing to do with your partner's decision.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

Elation: Great happiness and exhilaration.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

Satisfaction: Fulfilment of your wishes, expectations, or needs.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

Dissatisfaction: Non- fulfilment of your wishes, expectations, or needs.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

2. Information about Stage 2 outcomes

You gave up x tokens to reduce your partner's income.

You partner gave up y tokens to reduce your income.

Your Stage 2 income is: $(20 - x - 3y)$ tokens.

Your partner's Stage 2 income is: $(20 - y - 3x)$ tokens

12.5 Post-experimental Questionnaire

In this questionnaire we ask you some questions about yourself. It would really help us to understand the choices you made in the experiment. So please take your time, and please answer as accurately as possible.

1. Age: _____ years old

Gender: (female/male)

Highest qualification: _____

Your year of study: _____

2. Have you participated in similar experiments in the past? (Yes/No)

3. How did you guess your partner's contribution in Stage 1? Tick the choices that apply. If none of the choices applies then pick the last option and provide your own reason.

[ ] I used my own intended contributions and guessed my partner would choose the same.

[ ] I started with my own intended contributions and added some number to it.

[ ] I started with my own intended contributions and subtracted some number from it.

[ ] I chose randomly without giving it much thought.

[ ] None of the above accurately describes my guess. Here is how I chose: _____ .

4. When you were told your partner's contribution at the end of Stage 1, tick the choices that apply.

[ ] I compared my partner's contributions to my own contributions.

[ ] I compared my partner's contributions to my initial guess of how much my partner would contribute.

If you ticked both the choices above, please state their relative importance:

Relative importance of the first option: _____ %

Relative importance of the second option: _____ %

5. Consider your choice of reduction in Stage 2. If you contributed more than your partner in Stage 1, then please answer ONLY Part A below. If you contributed less than the partner in Stage 1, please answer ONLY Part B below. If you contributed equal to the partner in Stage 1, please answer ONLY Part C below. Please do pay particular attention to the italicized text below.

Part A: You contributed MORE than your partner in Stage 1. Tick the choices below that apply to you.

[ ] I chose zero reduction of my partner' income because it *would have reduced my own Stage 2 income.*

[ ] I chose strictly positive reduction of my partner' income because I was *frustrated and angry* that my partner's contributions were *lower than my own contributions.*

[ ] I chose strictly positive reduction of my partner' income because I was *frustrated and angry* that my partner's contributions were *lower than my initial guess of the partner's contributions.*

[ ] I chose strictly positive reduction of my partner' income because my partner's Stage 1 contributions revealed to me that *the partner was unkind to me.* So, I *reciprocated with unkindness* in Stage 2.

[ ] I chose strictly positive reduction of my partner' income because my partner contributed less than the socially appropriate or fair level of contributions.

[ ] I chose strictly positive reduction of my partner' income because my partner contributed less than the socially appropriate or fair level **and that they should be ashamed**.

Part B: You contributed LESS than your partner in Stage 1. Tick the choices below that apply to you.

[ ] I choose zero reduction of my partner' income because it *would have reduced my own Stage 2 income.*

[ ] I choose zero reduction of my partner' income because I felt guilty at having contributed less than my partner.

[ ] I chose strictly positive reduction of my partner' income because I was *frustrated and angry* that my partner's contributions were *lower than my initial guess of the partner's contributions.*

[ ] I chose strictly positive reduction of my partner' income because my partner contributed less than the socially appropriate or fair level of contributions.

[ ] I chose strictly positive reduction of my partner' income because my partner contributed less than the socially appropriate or fair level **and that they should be ashamed**.

Part C: You contributed EQUAL to your partner in Stage 1. Tick the choices below that apply to you.

[ ] I chose zero reduction of my partner' income because it would *have reduced my own Stage 2 income.*

[ ] I choose zero reduction of my partner' income because we contributed equally.

[ ] I chose strictly positive reduction of my partner' income because I was *frustrated and angry* that my partner's contributions were *lower than my initial guess of the partner's contributions.*

[ ] I chose strictly positive reduction of my partner' income because my partner contributed less than the socially appropriate or fair level of contributions.

[ ] I chose strictly positive reduction of my partner' income because my partner contributed less than the socially appropriate or fair level **and that they should be ashamed**.

6. Recall that a contribution of at least T units was rated to be the "socially desirable contribution" by your social group or peers. Tick all options below that apply to you.

[ ] I contributed less than T units, and I feel **no** shame.

[ ] I contributed less than T units and in hindsight I feel **some** shame.

[ ] I contributed T or more units and I feel no specific emotion.

[ ] I contributed T or more units and I feel elated at doing something socially responsible.

7. If your income was reduced by your partner, do you believe that you deserved it? (Yes/No)

8. Suppose that you were asked to play this experiment again. Relative to the contributions that you chose in Stage 1 of this experiment, will your contributions in the new experiment

(tick as appropriate):

[ ] Increase

[ ] Decrease

[ ] Stay the same

End of post-experimental questionnaire

The computer randomly chose the x stage to actually pay you, and you have earned x tokens in this stage. Thus, your total earning in this experiment is x pounds.

This is the end of the experiment. Thank you for your participation.