Belief-dependent Preferences and Reputation: Experimental Analysis of a Repeated Trust Game*

Giuseppe Attanasi (LEM, University of Lille)
Pierpaolo Battigalli (Bocconi University and IGIER, Milan)
Elena Manzoni (University of Milan-Bicocca)
Rosemarie Nagel (ICREA, Universitat Pompeu Fabra, Barcelona GSE)

April 2018

Abstract

We study in a theoretical and experimental setting the interaction between belief-dependent preferences and reputation building in a finitely repeated trust game. We focus mainly on the effect of guilt aversion. In a simple two-types model, we analyze the effect of reputation building in presence of guilt-averse players and derive behavioral predictions. In the experiment, we elicit information on trustees' belief-dependent preferences and disclose it to the paired trustor before the repeated game. Our experimental results show that disclosing information on the trustee's belief-dependent preferences and thus letting players play the repeated trust game in presence of almost complete information leads to higher trust and cooperation than in the corresponding incomplete information game setting. In particular, disclosure of information on preferences of guilt-averse trustees also enhances the trustors' cooperation. Disclosure of information on belief-dependent preferences of reciprocity-concerned trustees, instead, does not lead to higher trust and cooperation. We show that this is theoretically consistent with subjects featuring low reciprocity concerns.

JEL classification: C72; C73; C91.

Keywords: Repeated psychological game; reputation; guilt; reciprocity; almost complete information.

*We thank Chiara Aina, Marco Boretto, Roberto Corrao, Andrea Guido, participants at the at the 9^{th} SEET Workshop in Lecce, the 2^{nd} Workshop on Psychological Game Theory in Norwich, and the seminar participants at University of East Anglia, and Singapore Management University for useful discussions. G. Attanasi gratefully acknowledges financial support by the ERC (grant DU 283953). P. Battigalli gratefully acknowledges financial support by the ERC (grant 324219). R. Nagel gratefully acknowledges financial support by the Barcelona Graduate School of Economics, and the CREA program.

1 Introduction

Why should an agent keep an—implicit or explicit—promise or carry out a threat when this implies a material loss? Traditional game-theoretic models maintain the assumption that players are selfish and provide conditions under which repeated interaction turns such short-run losses into long-run benefits. This can work in two ways:[1] in infinitely repeated games with complete information,[2] failure to comply with an informal, or implicit agreement can trigger the play of a continuation equilibrium that is bad for the deviator; in finitely repeated games where information is incomplete, even if only slightly, deviations can trigger a costly loss of reputation.

But often agents incur material losses to keep promises or carry out threats even when they are not involved in long-run relationships, because they do not only care about their material interests. A huge experimental literature suggests that preferences displaying a concern for the material payoffs of others are important to motivate agents to incur material losses to achieve more equitable outcomes, or to punish selfish behavior (see the survey in Cooper and Kagel 2016). Importantly, experimental evidence suggests that belief-dependent motivations play an important role. For example, the rejection of greedy offers in the Ultimatum Game is positively correlated to how much the responder initially expected to get (see references from Battigalli *et al.* 2015, and Aina *et al.* 2018), and equal sharing in Dictator and Trust Games is positive correlated with how much the chooser believes that the other agent expected to get (see Bellemare *et al.* 2017, 2018; Charness and Dufwenberg 2006; Attanasi *et al.* 2013, and references therein).

In this paper we study both theoretically and experimentally the interaction of reputation and belief-dependent other-regarding preferences in the context of a repeated Trust Game. First, we put forward a theoretical model with role-dependent guilt aversion that highlights how incomplete information on the belief-dependent preferences of the trustee may give rise to reputation building phenomena. We adopt a multi-period model where the trustee experiences guilt (if he defects) at the end of each period (round of the repeated game) and the trustor's disappointment refers to his expectations at the beginning of each period (cf. Battigalli *et al.* 2018). We show that longer cooperative paths arise the higher the trustor's prior belief on the trustee being guilt-averse due to a twofold mechanism: a higher prior on the trustee's guilt sensitivity in our model implies that he is indeed more likely to be a high-guilt type, and therefore more likely to choose the cooperative action. Moreover, a low-guilt type is more likely to cooperate the higher is the trustor's prior belief, due to reputation incentives. As a consequence, the trustor chooses the cooperative action

[1] See Mailath and Samuelson (2006) and the references therein.

[2] Or finitely repeated games where the stage game has multiple equilibria.

more often, the higher is his prior, so that longer cooperative paths are observed. We also argue theoretically that intention-based reciprocity does not give rise to reputation building phenomena if reciprocity concerns are low (see the Appendix).

Second, we analyze a 4-period repeated Trust Minigame experimentally. Building on Attanasi *et al.* (2013), we compare a main treatment in which information on belief-dependent preferences of trustees is disclosed to their matched trustor to a control treatment in which there is no information disclosure. Our results show that disclosing information on the trustee's belief-dependent preferences and thus letting players play the repeated trust game in presence of *almost complete information* leads to higher trust and cooperation than in the corresponding *incomplete information* game setting implemented by the control treatment. In particular, disclosure of guilt-averse preferences also enhances the trustors' cooperation. Disclosure of information on belief-dependent preferences of reciprocity-concerned trustees, instead, does not lead to higher trust and cooperation.

The experimental literature on reputation building in repeated trust games is vast and has addressed several research questions, which are only marginally related to ours. For example, Anderhub *et al.* (2002) study a finitely repeated trust game with incomplete information by explicitly introducing the possibility of a trustee's type (a robot) who always feels obliged to reward trust. Engle-Warnick and Slonim (2004) compare finitely repeated trust games with partner matching with indefinitely repeated ones where in the last round of each interaction the trustor may start a new repeated game with another trustee.

There are few theoretical or experimental studies on the interaction between belief-dependent preferences and reputation building in repeated interactions. We are only aware of Balafoutas (2011), who investigates theoretically the role of guilt aversion for corruption in public administration. Corruption is modeled as the outcome of a game played between a bureaucrat, a lobby, and the public. The three-player game is assumed to be played repeatedly with infinite horizon and a constant continuation probability.

The theoretical and experimental results in our paper contribute to the literature on belief-dependent preferences, suggesting that such preferences matter, and that they should be taken into account when designing experiments on social dilemma games. Charness and Dufwenberg (2006) and follow-up papers suggests that either experimenters elicit and disclose information on the relevant belief-dependent preferences, or the analysis must be an incomplete-information one, a point made forcefully in Attanasi *et al.* (2013). Our paper shows that, if the interaction in the experimental game is repeated, disclosure of information on psychological belief-dependent preferences affects reputation building. Moreover, it suggests that belief-dependent preferences provide an alternative to (or foundation of) the commitment types of standard reputation models.

The rest of the paper is structured as follows. Section 2 presents our theoretical model

with role-dependent guilt. Section 3 describes our experimental design. Section 4 presents our behavioral predictions. Section 5 discusses our experimental results in light of the behavioral predictions. Section 6 concludes. An Appendix presents the theoretical analysis of the case with (role-dependent) intention-based reciprocity. An Online Appendix collects technical details about the experimental instructions.

2 The repeated Trust Minigame

In the paper, we aim to investigate the interaction between belief-dependent preferences and reputation building phenomena. We do so by focusing, as in the experimental design, on the four-period repetition of the Trust Minigame,[3] which—in each period—is played in its simultaneous-move version. The situation of strategic interaction that constitutes the stage game of our problem is the following: Player A ("she") and B ("he") are partners on a project that has thus far yielded a total profit of €2. Player A has to decide whether to *Dissolve* or to *Continue* with the partnership. If player A decides to *Dissolve* the partnership, the contract states that each player receives an equal share of the profit. If player A decides to *Continue* with the partnership, the total profit doubles (€4); however, in that case, player B has the right decide whether to share equally or take the whole surplus. We call *Continue* for player A and *Share* for player B a **cooperative action**. In the simultaneous-move game of Table 1 (the strategic form of the Trust Minigame), player B—before knowing player A's choice—has to state if he would *Take* or *Share* the higher profits.

A/B	*Take*	*Share*
Dissolve	$1, 1$	$1, 1$
Continue	$0, 4$	$2, 2$

Table 1 Strategic form the Trust Minigame.

Players' preferences over outcomes may depend on beliefs. In particular, we focus on the case in which players' preferences over outcomes depend on the co-players' beliefs (Battigalli and Dufwenberg 2007, 2009). The belief-dependent motivations that seem to be relevant in the Trust Minigame are **guilt** and **intention-based reciprocity**. The effect of both belief-dependent motivations in the one-shot Trust Minigame has been analyzed by Battigalli and Dufwenberg (2009), and an experimental and theoretical analysis of the effect of preferences that display both guilt aversion and intention-based reciprocity can be found in Attanasi *et al.* (2013). Moving to a repeated-game setting, however, we focus on guilt aversion only, as

[3] Our theoretical analysis can be extended to any finite repetition of the game, as can be understood from the proof of Proposition 1.

experimental evidence shows that guilt is indeed the dominant psychological motivation in the Trust Minigame (see Attanasi *et al.* 2013). Moreover, we assume the presence of **role-dependent guilt**: Only player B can be affected by guilt, and this is common knowledge. Player A is instead known to be a material payoff maximizer.[4]

We model guilt by adapting the *simple guilt* model of Battigalli and Dufwenberg (2007) to a multi-period setting (cf. Battigalli *et al.* 2018). Player B's guilt depends on his/her guilt sensitivity, θ, and on the expected co-player's disappointment, given her subjective beliefs. In order to analyze the effect of guilt, we need to consider the players' first- and second-order beliefs about behavior. The beliefs that will be relevant for our analysis are player A's first order belief $\alpha = \mathbb{P}_A[S]$, and player B conditional second-order belief $\beta = \mathbb{E}_B[\widetilde{\alpha}|C]$.[5]

Let (m_A, m_B) be the players' monetary payoffs. The **disappointment** of A is the difference, if positive, between A's expected and actual payoffs, that is

$$D(\alpha, m_A) = \max\{0, \mathbb{E}_\alpha[\widetilde{m}_A] - m_A\}.$$

Player A can only be disappointed after (C, T) in which case her disappointment is $D(\alpha, m_A(C,T)) = 2\alpha$. Player B's psychological utility after (C, T) is therefore

$$u_B(m_B, m_A, \alpha, \theta) = m_B(C,T) - \theta D(\alpha, m_A(C,T)) = 4 - 2\theta\alpha,$$

where θ is his guilt parameter, and $\theta = 0$ means player B is selfish, i.e., a material payoff maximizer.[6] We assume that player B's guilt type can take two values, low or high, i.e. $\theta \in \{\theta^L, \theta^H\}$, with $\theta^L = 0$ and $\theta^H > 2$. The stage game with role-dependent guilt is described in Table 2.

A/B	$Take$	$Share$
$Dissolve$	$1, 1$	$1, 1$
$Continue$	$0, 4 - 2\theta\alpha$	$2, 2$

Table 2 Trust Minigame with role-dependent guilt.

Note that B Shares if $4 - 2\theta\beta < 2$, which holds if $\theta_B > 2$ and $\beta \geq 1/2$. With this, the stage game with complete information has a unique equilibrium (D, T) when $\theta = \theta^L$, and two pure strategy equilibria (D, T), (C, S), when $\theta = \theta^H$.

We focus on the repeated game obtained from the 4-period repetition of the psychological

[4] For a detailed analysis of the differences between role-dependent and role-independent guilt in the Trust Game see Attanasi *et al.* (2016).

[5] B's decision depends on his belief conditional on *Continue*, because he is indifferent conditional on *Dissolve*. Even if B does not osserve A's choice before moving, this conditional belief is still well defined as long as B assigns positive subjective probability to *Continue* (see Attanasi *et al.* 2016).

[6] This is a kind of state-dependent utility because B does not observe A's belief α.

stage game described above. We model guilt as experienced at the end of each period, and disappointment refers to player A's expectations at the beginning of each period. An alternative model could consider guilt as experienced at the end of the repeated game, i.e., at the end of period 4, and with A's disappointment being the difference between her total realized payoff and expected total payoff according to her initial belief. The two models describe different strategic interactions. To understand the difference, we refer to the work by Battigalli *et al.* (2018), which makes explicit the difference between periods and stages in dynamic psychological games. The main difference between periods and stages rests in the fact that periods measure the passage of time affecting players' preferences, while stages are merely a representation of moments in which players choose and acquire new information. A period may consist of just one stage, as in our case, or multiple stages. Our interpretation of the Trust Minigame is that each repetition of the stage game constitutes a period. As a consequence, beliefs at the beginning of period t are relevant for the computation of the expected disappointment in period t, and the intertemporal psychological utility is obtained as the sum of the one-period psychological utilities.[7] Indeed, in our experiment we elicit beliefs at the beginning of every round consistently with our interpretation of the model as a 4-period repeated game rather than a 1-period game with 4 stages. As in Battigalli *et al.* (2018), the difference between period and stage is relevant for the determination of A's expectation-based reference points: in a 1-period game the relevant belief is the one A holds at the beginning of the first round, while in a 4-period game the relevant beliefs are those A holds at the beginning of every round.

2.1 Guilt aversion and reputation building: a model

Let us now focus on the interaction between the repeated structure of the game and the incomplete information on the psychological type, which may give rise to reputation building phenomena. In the experiment we deal with treatments in which subjects playing in role A receive information on their co-players' psychological type, and other treatments in which no information is disclosed to them. However, even when information on B's psychological type is disclosed, the situation is arguably one of *approximately* complete information rather than complete information. The literature on reputation models, which started from Kreps *et al.* (1982), Kreps and Wilson (1982) and Milgrom and Roberts (1982), proves that in a repeated game the presence of a slight uncertainty over the opponent's type may dramatically change the set of equilibrium outcomes, for example enhancing cooperation in games where it cannot be sustained under complete information.

[7] In the language of Battigalli *et al.* (2015), we adopt a "slow-play" model: **slow play** occurs in a multi-period game where each period comprises only one stage, **fast play** occurs in a multistage one-period game.

We build a model of repeated interaction based on the stage game of Table 2. In order to analyze reputation building, we assume that the guilt type of B, θ, is his private information. Player A holds a prior belief on B's type, $\mu_1 = \mathbb{P}\left[\theta = \theta^H | h_\emptyset\right]$, which is common knowledge. Varying how extreme such prior belief is allows us to compare situations in which there is almost complete information on B's type (μ_1 close to either zero or one) to situations in which there is genuine incomplete information on his psychological type (intermediate values of μ_1).

Let a_t denote the realized action profile in period t. Proposition 1 describes a continuum of equilibria in which reputation building phenomena may arise, depending on A's prior on B's guilt aversion, μ_1. The equilibria of Proposition 1 display the traditional structure of reputation models. First of all, a high-guilt B optimally Shares in each period, regardless of the previous history and reputation. This is in itself a relevant result, as it is obtained with a fully rational high-guilt B, and not with the assumption of a commitment type. A first observation that can be drawn from this analysis is therefore that belief-dependent preferences may provide an alternative to (or be the foundation of) some of the commitment types that are assumed in the standard reputation literature.

A low-guilt B always Takes in the last period, as this is weakly dominant in the stage game. In earlier periods, he either Shares, or randomizes, or Takes depending on whether his reputation in the period is higher, equal, or lower than a threshold, which is increasing over time. For example, starting from a high initial reputation one typical situation is that the low-guilt B Shares in earlier periods and then enters a mixed equilibrium path in which he Shares with the probability that makes his reputation equal to the threshold in the subsequent period. As soon as he Takes once, he is recognized as low-guilt and the equilibrium moves on a (D, T) path. Finally, A Continues for high values μ_t, Dissolves for low values, and randomizes for intermediate ones.

Proposition 1 *Let $\mu_t = \mathbb{P}[\theta = \theta^H | h_{t-1}]$ be A's belief about B's type at the beginning of period t, $\gamma_t = \mathbb{P}[C | h_{t-1}]$ be A's behavioral strategy at time t, and $\sigma_t = \mathbb{P}[S | h_{t-1}, \theta^L]$ be a low-guilt B's behavioral strategy at time t. The following is a continuum of sequential equilibria of the 4-period repetition of the psychological game in Table 2: High-guilt B Shares in every period. Low-guilt B's behavioral strategy is*

$$\sigma_t = \begin{cases} 0, & \text{if } t = 4, \text{ or } \mu_t = 0. \\ 1, & \text{if } t < 4 \text{ and } \mu_t \geq \frac{1}{2^{4-t}}, \\ \frac{(2^{4-t}-1)\mu_t}{1-\mu_t}, & \text{if } t < 4 \text{ and } \mu_t \in \left(0, \frac{1}{2^{4-t}}\right). \end{cases}$$

Player A's behavioral strategy is

$$\gamma_t = \begin{cases} 0, & \text{if } \mu_t < \frac{1}{2^{5-t}}, \\ \gamma \in (0,1), & \text{if } t = 1 \text{ and } \mu_t = \frac{1}{2^{5-t}}, \\ \frac{2}{3}\gamma_{t-1}, & \text{if } t > 1 \text{ and } \mu_t = \frac{1}{2^{5-t}}, \\ 1, & \text{if } \mu_t > \frac{1}{2^{5-t}}. \end{cases}$$

A's belief on B's guilt type is

$$\mu_t = \begin{cases} \max\left\{\frac{1}{2^{5-t}}, \mu_{t-1}\right\}, & \text{if } a_{t-1} = (\cdot, S) \text{ and } \mu_{t-1} > 0, \\ 0, & \text{if } a_{t-1} = (\cdot, T) \text{ or } \mu_{t-1} = 0. \end{cases}$$

Proof. We prove the proposition by backward induction.

Period $t = 4$ (last period). A high-guilt player B Shares: A Continues only if $\alpha_4 \geq \frac{1}{2}$. Hence, if the probability that player A Continues is positive, then B's second-order belief is $\beta_4 \geq \frac{1}{2}$. In this case, B finds it optimal to Share. If the probability that player A Continues is 0, we assume that the out-of-equilibrium conditional belief is nonetheless $\beta_4 \geq \frac{1}{2}$.[8] A low-guilt player B Takes: it is a (weakly) dominant action for him in the last period. Given that B Shares only when he is high-guilt, we have that A's first order belief is $\alpha_4 = \mu_4$. Player A Dissolves if $\alpha_4 < \frac{1}{2}$, Continues if $\alpha_4 > \frac{1}{2}$ and mixes if $\alpha_4 = \mu_4 = \frac{1}{2}$.

Period $t < 4$. A high-guilt player B Shares. If there is a positive probability that player A Continues, $\beta_t \geq \frac{1}{2}$. The expected utility from *Share* is:

$$\begin{aligned} \mathbb{E}\left[u_B | S, \theta^H, h_{t-1}\right] &= \sum_{\tau=t}^{4} \left(2\gamma_\tau + (1 - \gamma_\tau)\right) \\ &= \sum_{\tau=t}^{4} \gamma_\tau + 4 - (t-1) \\ &= \sum_{\tau=t}^{4} \gamma_\tau + 5 - t. \end{aligned}$$

The expected utility from *Take* is:

$$\begin{aligned} \mathbb{E}\left[u_B | T, \theta^H, h_{t-1}\right] &= \gamma_t \left(4 - 2\theta^H \beta_t\right) + (1 - \gamma_t) + (4 - t) \\ &= 3\gamma_t - 2\theta^H \beta_t \gamma_t + 5 - t. \end{aligned}$$

[8] This is an arbitrary assumption which is consistent with equilibrium analysis. We note, however, that in the last period forward induction implies $\beta_4 \geq \frac{1}{2}$ even when the probability that player A Continues is 0.

Therefore, Sharing is optimal when

$$\sum_{\tau=t}^{4}\gamma_\tau + 5 - t \geq 3\gamma_t - 2\theta^H\beta_t\gamma_t + 5 - t,$$

that is when

$$\theta^H \geq \frac{1}{\beta_t} - \frac{\sum_{\tau=t+1}^{4}\gamma_\tau}{2\gamma_t\beta_t}.$$

Given that $\frac{1}{\beta_t} - \frac{\sum_{\tau=t+1}^{4}\gamma_\tau}{2\gamma_t\beta_t} < 2$, a high-guilt B finds optimal to Share. Notice that the threshold for Sharing is increasing over time: reputation concerns fade out as the end of the game is near.

A low-guilt B's expected utility if he Takes is

$$\begin{aligned}\mathbb{E}\left[u_B|T,\theta^L,h_{t-1}\right] &= 4\gamma_t + (1-\gamma_t) + 4 - t \\ &= 3\gamma_t + 5 - t.\end{aligned}$$

When B Takes, his reputation drops to zero in the following period ($\mu_{t+1} = 0$) and, as a consequence the equilibrium moves to a (D,T) path, and B's payoff is 1 in each of the $4-t$ remaining periods.

A low-guilt B's expected utility if he Shares is

$$\begin{aligned}\mathbb{E}\left[u_B|S,\theta^L,h_{t-1}\right] &= 2\gamma_t + (1-\gamma_t) + 4\gamma_{t+1} + \left(1-\gamma_{t+1}\right) + 4 - (t+1) \\ &= \gamma_t + 3\gamma_{t+1} + 5 - t,\end{aligned}$$

When computing the expected payoff from Sharing in period t we take into account that the play is going to be on the mixed equilibrium path, in which B is indifferent between Taking and Sharing. Hence, we can compute the expected payoff assuming that B takes in period $t+1$. Under this assumption, the expected utility is given by the expected utility from Sharing in period t ($2\gamma_t + (1-\gamma_t)$), the expected utility from Taking in period $t+1$ ($4\gamma_{t+1} + \left(1-\gamma_{t+1}\right)$), and the expected utility from the (D,T) path in each of the $4-(t+1)$ remaining periods.

Hence, a low-guilt player B Shares if

$$\gamma_t + 3\gamma_{t+1} + 5 - t \geq 3\gamma_t + 5 - t$$

that is if $\gamma_{t+1} \geq \frac{2}{3}\gamma_t$.

If a low guilt B Shares with probability σ_t, his reputation after Sharing is

$$\begin{aligned}\mu_{t+1} &= \mathbb{P}_{\mu_t}[\theta^H | a_t = (\cdot, S)] = \frac{\mathbb{P}_{\mu_t}[a_t = (\cdot, S)|\theta^H]\mathbb{P}_{\mu_t}[\theta^H]}{\mathbb{P}_{\mu_t}[a_t = (\cdot, S)]} \\ &= \frac{\mu_t}{\mu_t + (1-\mu_t)\,\sigma_t}.\end{aligned}$$

In every period t, player A Continues only if $\alpha_t \geq \frac{1}{2}$, given that in the prescribed equilibrium her action does not change her information nor her future payoffs, and she mixes when $\alpha_t = \frac{1}{2}$. A's first-order belief depends on the belief on B's type and B's strategy as follows:

$$\alpha_t = \mu_t + (1-\mu_t)\sigma_t,$$

so that in every period there is a value of μ_t for which A mixes between C and D. In period t, the mixing probability of a low-type B, σ_t is such to induce a μ_{t+1} that implies $\alpha_{t+1} = \frac{1}{2}$. To ease notation, let $r_{t+1} = \frac{1}{\mu_{t+1}}$ denote the inverse of the reputation value μ_{t+1} that yields $\alpha_{t+1} = \frac{1}{2}$. With this, B mixes with probability σ_t such that

$$\mu_{t+1} = \frac{\mu_t}{\mu_t + (1-\mu_t)\sigma_t} = \frac{1}{r_{t+1}},$$

that is

$$\sigma_t = \frac{\mu_t(r_{t+1}-1)}{1-\mu_t}.$$

Player A's first order belief α_t becomes

$$\alpha_t = \mu_t + (1-\mu_t)\sigma_t = \mu_t + (1-\mu_t)\frac{\mu_t(r_{t+1}-1)}{(1-\mu_t)} = r_{t+1}\mu_t.$$

Hence, $\alpha_t = \frac{1}{2}$ when $\mu_t = \frac{1}{2r_{t+1}}$, that is, $r_t = 2r_{t+1}$. Given that $r_4 = 2$, we have that $\alpha_t = \frac{1}{2}$ when $\mu_t = \frac{1}{2^{5-t}}$, and that $\sigma_t = \frac{(2^{4-t}-1)\mu_t}{1-\mu_t}$. ■

The structure of the equilibrium mimics closely the equilibria of the traditional reputation models. There is however a relevant difference that depends on the type of equilibria that we consider. In the equilibrium described by Proposition 1, we have a continuum of mixed equilibria, instead of one mixed equilibrium path. This is due to our assumption that the stage game is simultaneous, together with the fact that we allow players to choose cooperative actions in $t+1$ even if A, by randomizing due to her mixed strategy, chooses D in period t. As a matter of fact, if B Shares in period t, r_{t+1} is still updated correctly and is such that players can be on the mixed equilibrium path in period $t+1$. As the stage game is such that B's action is relevant only if A Continues, also B finds it optimal to remain on the mixed

equilibrium path in $t+1$ if he believes that A will do so.

The existence of a continuum of equilibria distinguishes our model from the traditional reputation models. We also have another kind of equilibrium, more similar to those models, where after A Dissolves for the first time successive play reverts to the non-cooperative profile (D, T).[9] However, in the experiment, several A-subjects who defected revert to the cooperative action after observing that B-subjects Share. Thus, the equilibrium of Proposition 1 allows us to better organized the data (see Section 5.2).

Comparative statics. Let us now draw some conclusions on the type of comparative statics implied by our model that we expect to find in the experimental data if reputation building occurs.

First, we focus on the case in which B is truly a low-guilt player. B's reputation incentives are increasing with μ_1. As a matter of fact, the probability that B chooses the cooperative action in the first period is weakly increasing in A's prior belief, μ_1.

Moreover, also the equilibrium path is going to depend on A's belief. The higher μ_1, the higher the number of periods in which B is going to cooperate before hitting the mixed strategy path. Figure 1 shows the periods in which B Shares (pink region), Takes (green), or mixes (violet) depending on the value of μ_1. As the figure shows, the likelihood of observing longer cooperation is increasing with μ_1. However, the **fully cooperative path**, defined as the path on which cooperation occurs in every period t, can never be observed when B is a low-guilt type, as he will never cooperate in the last period.

Let us now consider the case in which B is high-guilt with probability μ_1. In this case, observing the **cooperative path up to** t—that is, cooperation up to a certain period—is more likely the higher is μ_1. This happens in a twofold manner: first, there is a higher probability of B being truly high-guilt, in which case he'll choose the cooperative action in every period. Second, even if B is the low-guilt type, reputation building is more profitable for him, and as a consequence he will keep choosing the cooperative action for a larger number of periods. As a consequence, A best responds choosing *Continue* for a longer time, so that longer paths of cooperation will be observed for higher μ_1. Moreover, the higher μ_1, the higher the probability of a fully cooperative path.

[9] In the equilibrium where (D, T) is played in every period after player A defects, A's mixing probability is pinned down precisely by the backward-induction calculation, so that we have only one equilibrium of this kind.

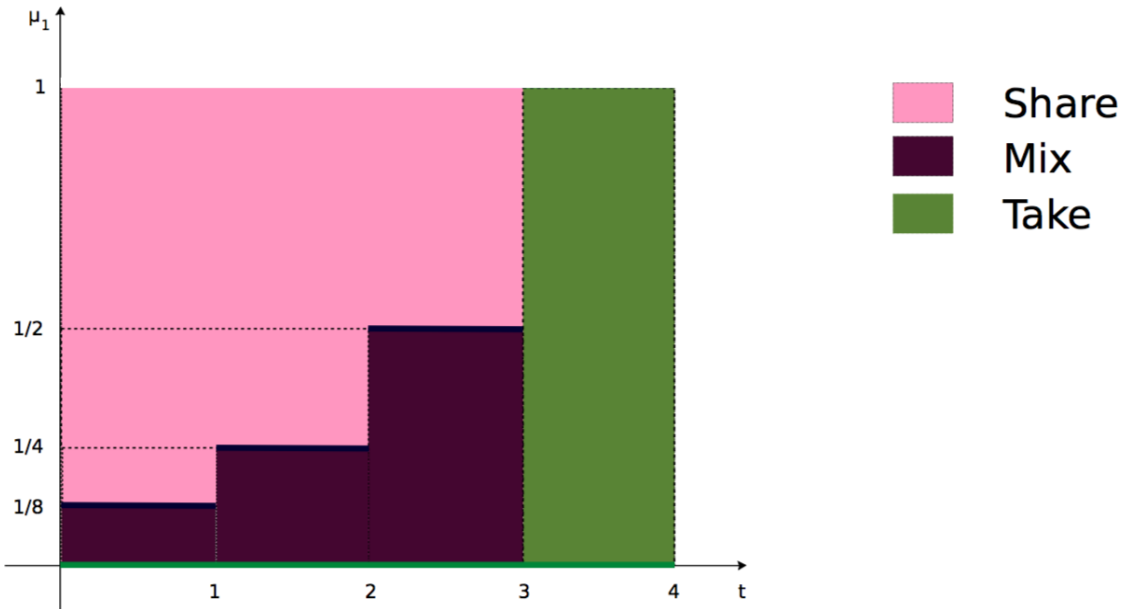

Figure 1: Low guilt B

3 The experiment

3.1 Procedures

Participants were first and second-year undergraduate students in Economics at Bocconi University of Milan. The sessions were conducted in a computerized classroom and subjects were seated at spaced intervals. The experiment was programmed and implemented using the z-Tree software (Fischbacher 2007).

We held 16 sessions with 20 participants per session, hence 320 subjects in total. Each person could only participate in one of these sessions. The majority of these sessions were conducted in the same time span of the experimental sessions of Attanasi *et al.* (2013), with none of their subjects participating in our experiment and vice versa.

Average earnings were €15.48, including a €5 show-up fee (minimum and maximum earnings were respectively €5 and €17); the average duration of a session was 65 minutes, including instructions and payment.

3.2 Design

The design is an extension of the experiment in Attanasi *et al.* (2013). The stage game, namely the Trust Minigame, is the one of Table 1 in Section 2. In this simultaneous-move game (the strategic form of the Trust Minigame), player B has to state if he would (entirely) *Take* or (equally) *Share* the higher profits *before* knowing player A's choice, hence also in

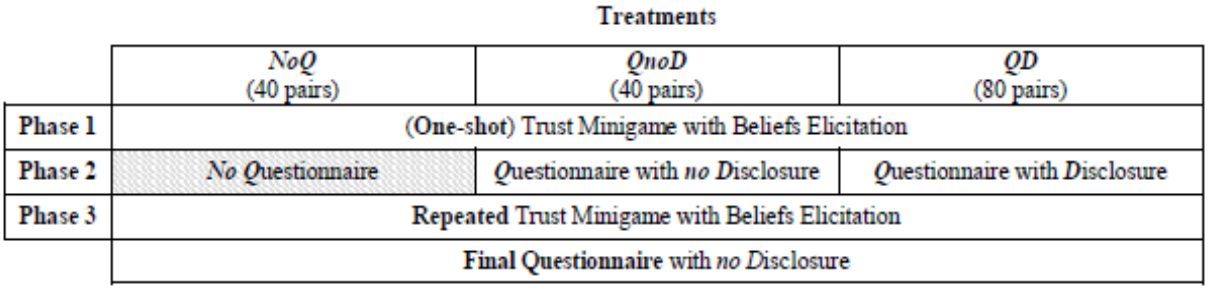

Figure 2: Summary of the design.

the case where A chooses *Dissolve*.

The experimental design is made of three phases and three treatments, explained in detail in Figure 2 (for the experimental instructions see the Online Appendix). The difference among treatments is in phase 2, depending on whether subjects playing in role B are asked to fill in a questionnaire and whether such answers are disclosed. We refer to these treatments, to be explained in detail below, as *No Q*uestionnaire (*NoQ*), *Q*uestionnaire *no D*isclosure (*QnoD*) and *Q*uestionnaire *D*isclosure (*QD*). We run 4 sessions for *NoQ* and for *QnoD* (80 subjects each) and 8 sessions for *QD* (160 subjects).

At the beginning of an experimental session, each of the 20 participants, or subjects, is randomly assigned with equal probability to role A (A-subject) or role B (B-subject) of the Trust Minigame. This determines 10 A-B pairs in each session. Each subject maintains the same role until the end of the session.

Participants are told that the experiment is made of three phases. Instructions of each new phase are given and read aloud only prior to that phase. Our design differs from the one of Attanasi *et al.* (2013) because of a repeated rather than a one-shot Trust Minigame in phase 3.[10]

We now describe in detail the three phases of the experimental design.

Phase 1 Phase 1—the same for all treatments—consists of a random matching between A-subjects and B-subjects, and two subsequent decision tasks:

Belief-elicitation. With regard to the Trust Minigame of Table 1: Each A-subject is asked to guess the percentage of B-subjects in her session who will choose *Share* (A's *initial first-order belief*). Each B-subject is asked to guess the answer of his co-paired A about the

[10]For technical comments on some important features of the experimental design and motivations for specific design choices, see Section 2.3 of Attanasi *et al.* (2013).

percentage of B-subjects who will choose *Share* (B's *unconditional second-order belief*), and to guess the choice—*Dissolve* or *Continue*—that his co-paired A will make (a feature of B's *first-order belief*).

Choice. Within each pair, player A and player B simultaneously make their choice in the Trust Minigame of Table 1. In particular, player A selects *Dissolve* or *Continue* and player B selects *Take* or *Share*.

At the end of phase 1, subjects do not receive any information feedback about the two decision tasks. Indeed, at the beginning of this phase, they are informed that the gains in the belief-elicitation task and in the Trust Minigame will be communicated at the end of the experiment.

Phase 2 In *NoQ*, subjects proceed directly to phase 3.

In *QnoD* and *QD*, subjects are randomly re-matched to form other 10 pairs. B-subjects are asked to fill in the questionnaire of Table 3. In particular, each B-subject is asked to consider the following *hypothetical* situation: His new co-player A has chosen *Continue* and he, B, has chosen *Take*, thereby earning €4 and leaving A with €0. Given this, B has the possibility—if he wishes—to give part of this amount back to A. He is allowed to condition his payback on the new co-player's guess of the percentage of B-subjects choosing *Share*.

Since there are 10 B-subjects, A has 11 possible guesses about how many B-subjects choose *Share* (0%, 10%, ..., 100%), as shown in Table 3. Hence, each B-subject is asked to fill in each of the 11 rows of Table 3 with a value between €0.00 and €4.00. To check for framing effects, half of the sessions of each treatment show the first column of Table 3 in reverse order, with 100% on the first row and 0% on the last row.

B-subjects fill in the questionnaire first on a sheet of paper and then copy the answers on the questionnaire shown on their computer screen. A-subjects read and listen to the instructions of phase 2. Among the subjects in a session of *QnoD* and *QD*, it is made public information that neither the responding B-subject nor anyone else will receive any payment for the answers she gives in the questionnaire of Table 3 (hypothetical payback scheme). Furthermore, in *QnoD* it is public information that B's filled-in questionnaire *will not be* disclosed to anyone.

On the other hand, in *QD* it is public information that B's filled-in questionnaire *will be* disclosed to a randomly-chosen A-subject. Actually, this subject is the one randomly matched with B at the beginning of phase 2. At the end of this phase, the matched B's filled-in questionnaire appears on A's screen, and the latter is invited to copy it on a sheet of paper. At this stage, subjects do not know yet that in phase 3 they are going to play again

the Trust Minigame, with the same matching of phase 2.

A's possible guesses of *Share*	Your payback (in €)
0%	between 0.00 and 4.00
10%	between 0.00 and 4.00
20%	between 0.00 and 4.00
30%	between 0.00 and 4.00
40%	between 0.00 and 4.00
50%	between 0.00 and 4.00
60%	between 0.00 and 4.00
70%	between 0.00 and 4.00
80%	between 0.00 and 4.00
90%	between 0.00 and 4.00
100%	between 0.00 and 4.00

Table 3 Questionnaire (Hypothetical Payback Scheme) in phase 2.

Phase 3 Phase 3—the same for all treatments—consists, with respect to phase 1, of a different random matching (absolute-stranger matching design), and of two decision tasks.

In *NoQ*, subjects are randomly re-matched to form other 10 pairs; in *QnoD* and *QD*, each A-subject is matched with the same B-subject as in phase 2.

The two decision tasks are an extension of those of phase 1 (and of phase 3 of the experiment in Attanasi *et al.* 2013). Indeed, in phase 3 subjects play the Trust Minigame in Table 1 *repeatedly* for four rounds *within the same pair* (from now on, Repeated Trust Minigame). Therefore, subjects go through the two decision tasks of phase 1—belief elicitation and choice—in each round 1-4 of the Repeated Trust Minigame. This leads to the elicitation of 4 first-order beliefs and 4 choices for player A, and of 4 unconditional second-order beliefs, 4 first-order beliefs, and 4 choices for player B.[11]

At the end of each round, each subject within a pair is told the choice of the co-player in that round. Since in each round subjects play the simultaneous-move game of Table 1, A's choice is told to B at the end of the round also in the case where B's choice was payoff-irrelevant, i.e., A chose *Dissolve* in that round.

Subjects do not receive any information feedback about the belief-elicitation tasks. Indeed, at the beginning of this phase, they are informed that the gains in the belief-elicitation task will be communicated at the end of the experiment.

[11] The elicitation of beliefs at the beginning of every round is consistent with our interpretation of the model as a 4-period repeated game rather than a 1-period game with 4 stages. See the discussion in the first part of Section 2.

In *QnoD* and *QD*, each *B*-subject can keep his previously filled-in paper questionnaire with him for the duration of this phase. Additionally, in this phase of *QD*, *A* can keep the matched *B*'s filled-in questionnaire (previously copied on a sheet of paper) with her. At the beginning of phase 3 of *QD*, it is made public information that, in each pair, *B*'s filled-in questionnaire disclosed at the end of phase 2 corresponds to the matched *B*-subject of phase 3 of *QD*. At the end of phase 3, in *QD* and *QnoD* all filled-in questionnaires are collected by the experimenter.

Final questionnaire After phase 3, there is a final questionnaire, which is the same for all treatments (see Table 3), and equal to the one in phase 2 of *QnoD* and *QD*.

In *NoQ*, this is the first time *B*-subjects fill in the questionnaire of Table 3.

In *QnoD* and *QD*, we ask *B*-subjects to fill in the questionnaire of Table 3 on a sheet of paper as in phase 2, knowing that it *will not be* disclosed to anyone; they are allowed to give answers different from those given in phase 2.

Payment Each subject learns the co-player's choice in the Trust Minigame in phase 1, and whether her first-order belief (*A*-subject) or his first- and second-order beliefs (*B*-subject) in phase 1 and in each round of phase 3 were correct.

Each subject is paid the sum of the resulting payoffs in the Trust Minigame in phase 1 and in the Repeated Trust Minigame in phase 3, and is also paid for correct guesses (elicited beliefs). Specifically, €5 are added to the total payoff of *A*-subjects for each correct first-order belief (in phase 1 and in each of the four rounds of phase 3). Similarly, €5 are added to the total payoff of each *B*-subject for every time he guessed correctly both the choice and the first-order belief of the co-player (in phase 1 and in each of the four rounds of phase 3).

4 Experimental Hypotheses

The model analyzed in Section 2 informs our behavioral predictions, i.e., experimental hypotheses. In the setup of Proposition 1, the difference between our treatments can be interpreted as a difference in the prior belief about *B*'s type. More precisely, we expect *A*-subjects to hold high beliefs on *B* being high(low)-guilt when a filled-in questionnaire compatible with a high(low)-guilt sensitivity is disclosed—period 1 of phase 3 of treatment *QD*. On the contrary, we expect *A*-subjects to hold intermediate and more dispersed beliefs when no information is disclosed—period 1 of phase 3 of treatments *NoQ* and *QnoD*. This implies that, comparing the treatment with approximately complete information to the treatments with incomplete information, we consider three kinds of equilibrium path: one with a particularly high μ_1, which corresponds to the case in which *A* is matched with a high-guilt *B* in phase

3 of treatment QD; one with intermediate values of μ_1, which corresponds to phase 3 of treatments NoQ and $QnoD$; and the last one with low values of μ_1, which corresponds to phase 3 of treatment QD for pairs in which B is low-guilt.

The validity of our analysis rests on one main auxiliary assumptions: we assume that eliciting information does not affect subjects' behavior if the information is not disclosed.

H0.i: Subjects in treatments NoQ and $QnoD$ show the same behavior.

A consequence of this hypothesis is that we can pool data from the two treatments that do not disclose information, NoQ and $QnoD$, in what we call the **no information disclosure** (NoQ-$QnoD$) treatment.

We introduce other two auxiliary hypotheses concerning the behavior of subjects with belief-dependent preferences different from guilt aversion. As mentioned in Section 2, there is another major type of belief-dependent preferences that may play a role in the Trust Minigame, intention-based reciprocity (see Dufwenberg and Kirchsteiger 2004, Battigalli and Dufwenberg 2009). The theoretical analysis of Section 2 does not give predictions for reciprocal subjects. However, we show in the Appendix that if reciprocity concerns are mild, reciprocal subjects behave as selfish ones not only in the stage game, but also in the repeated game, and no reputation building arises.[12] Hence, we put forward the following auxiliary hypothesis.

H0.ii Reciprocity concerned B-subjects behave as selfish ones.

We also assume that, in phase 3 of QD, A-subjects matched with a B-subject disclosing reciprocity concerns behave as those matched with a selfish B-subject. Hence, we put forward the following behavioral hypothesis.

H0.iii In treatment QD, A-subjects matched with reciprocity concerned B-subjects behave as if matched with selfish ones.

If H0.ii and H0.iii are verified, the predictions we make for selfish B-subjects and for the corresponding matched A-B pairs will be applicable to reciprocity concerned ones.

[12] We expect that reciprocal subjects may have small reciprocity concerns, as it occurs in Attanasi *et al.* (2013), where none of the subjects classified as reciprocal has an estimated value higher that the relevant treshold for their model. We have replicated Attanasi *et al.* (2013) estimation procedure on our dataset, and also found that none of our reciprocal B-subjects would be predicted to choose *Share* in their model (see Section 5.1 below).

4.1 Hypotheses on *A*-subjects

First we focus on hypotheses that are only related to A-subjects, and, specifically, to their first-order beliefs.

As discussed above, we expect A-subjects to hold more polarized beliefs on their matched B's type when they receive the information on the questionnaire. Recall also that, in the equilibrium of Proposition 1, a high-guilt B Shares more often than a low-guilt one. As a consequence, A-subjects matched with a high-guilt B in the treatment with information disclosure should hold higher first-order beliefs on their partner choosing *Share*. In the experiment, we elicit A's first-order beliefs on the percentage of B-subjects that will choose the cooperative action. These observations together suggest the following hypothesis.

HA1: In *QD*, A-subjects' first-order beliefs are higher if matched with a high-guilt rather than a low-guilt or reciprocal B-subject.

A second prediction on the beliefs of A-subjects comes from the comparison of beliefs across treatments, both in terms of initial beliefs and in terms of their evolution across periods of the repeated game. As argued above, A-subjects in treatment *NoQ-QnoD* hold intermediate beliefs on their co-player's type. This observation, jointly with the equilibrium behavior of Proposition 1, implies that A-subjects hold more intermediate and dispersed initial beliefs on the frequency of B-subjects choosing *Share* in treatment *NoQ-QnoD* than in treatment *QD*. Moreover, in the treatment with information disclosure, there is very little to learn on the opponent's type over time: having looked at B's filled-in questionnaire, A should have a quite precise prior on B's type. Also, low-guilt B-subjects in an environment where μ_1 is low have little reputation incentives.[13] On the contrary, without information disclosure, A learns B's type over time. Finally, behavior of low-guilt B-subjects changes over time as well, due to the incentives to build reputation in early periods. As a consequence, the distribution of first-order beliefs on the probability of B choosing *Share* evolves and polarizes more over time in the treatment without information disclosure.

HA2: A-subjects' first-order beliefs vary more over time in *NoQ-QnoD* than in *QD*. Furthermore, in *NoQ-QnoD* they are more polarized in the last than in the first period; in *QD* they are equally polarized in the last and in the first period.

[13] This depends on the low number N of repetitions: the key insight of Kreps and Wilson (1982) and Kreps *et al.* (1982) is that even if μ_1 is low, there is a sufficiently large N such that reputation matters. Hence the discontinuity: with complete information behavior is independent of N, with approximately complete information behavior depends on N because of reputational concerns, as long as N is large enough.

4.2 Hypotheses on B-subjects

We now describe the hypotheses on beliefs and behavior of B-subjects. First, as discussed after Proposition 1, the equilibrium is such that low-guilt B-subjects display weakly higher reputation building when there is no information disclosure about their psychological type. As a matter of fact, in *NoQ-QnoD*, A holds a higher prior belief on B's type (average prior *vs.* prior concentrated on low-guilt type). Hence, B displays more reputation building, i.e., a higher frequency of *Share* in period 1 in *NoQ-QnoD* than in *QD*. This is summarized below:

HB1: Low-guilt and reciprocal B-subjects display more reputation building in *NoQ-QnoD* than in *QD*.

Moreover, reputation building, which increases the likelihood of observing B-subjects choosing the cooperative action, happens only when subjects are not recognized as being low-guilt. As a consequence, the following hypothesis describes the expected difference between frequencies of Sharing in period 1 of the repeated game and in the one shot Trust Minigame of phase 1. The difference is expected to be positive when reputation building matters, that is, in all cases but the one in which information on B being low-guilt is disclosed.

HB2: In *NoQ-QnoD*, independently of B's type, for any given second-order belief of B sharing is more likely in period 1 of phase 3 than in phase 1. In *QD* this is true only for high-guilt B-subjects.

We finally have an hypothesis on the link between high-guilt B-subjects' second-order beliefs and choices of cooperative action. Our theoretical model was derived under the simplifying assumption that the level of guilt of B-subjects is either zero or extremely high ($\theta^H > 2$). More in general, if we allow guilt levels to be positive but not so high, we expect to observe more cooperation from subjects who hold a higher second-order belief. Moreover, as shown in the proof of Proposition 1, the threshold for sharing is lower the earlier the period, as in earlier periods reputation concerns are added to the underlying psychological motivations for cooperation.

HB3: In both *QD* and *NoQ-QnoD*, sharing is more likely for higher beliefs of high-guilt B-subjects. The threshold is lower for earlier periods.

4.3 Hypotheses on matched A-B pairs

We finally have a set of hypotheses on the frequencies of cooperative paths of matched A-B pairs. These predictions are direct implications of the comparative statics discussed after Proposition 1. The first two, HP1 and HP2, follow from the observation that longer

cooperative paths are more likely the higher is the initial belief on the guilt type of B. The belief on B being high-guilt should be lowest when a low-guilt questionnaire is disclosed, intermediate without disclosure, and highest when a high-guilt questionnaire is disclosed. Thus, we obtain that cooperative paths up to t are more likely without (with) information disclosure when B's type is low-guilt (high-guilt).

HP1: In pairs including low-guilt & reciprocal B-subjects, the frequencies of the cooperative path up to t are higher in *NoQ-QnoD* than in *QD.*

HP2: In pairs including high-guilt B-subjects, the frequencies of the cooperative path up to t are higher in *QD* than in *NoQ-QnoD.*

A last hypothesis follows from the equilibrium behavior of a pair in which A has high beliefs on B being high-guilt, and B is indeed high-guilt. In this case the fully cooperative path is observed in equilibrium.

HP3: In *QD*, pairs including high-guilt B-subjects are on a fully cooperative path.

5 Results

In the next two sub-sections we analyze our experimental data in light of the theoretical model. In Section 5.1 we provide a categorization of B's belief-dependent preferences derived from the answers to the questionnaire of Table 3. We rely on this categorization to classify B-subjects as low-guilt, reciprocity, and high-guilt. With this classification in mind, in Section 5.2 we test the behavioral predictions of Section 4.

5.1 Elicitation of belief-dependent preferences through the filled-in questionnaire

As in Attanasi *et al.* (2013), the experimental elicitation of B's belief-dependent preferences in the Trust Minigame relies on his answers to the questionnaire of Table 3. We call "**payback pattern,**" $p(\alpha)$, the actual answers of a B-subject, with one payback value for each hypothesized α (A's belief about B's action *Share*). Recall that the payback pattern gives 11 observations for B's payback function, i.e., one for each $\alpha \in \{0, 0.1, ..., 1\}$. The left panel of Figure 3 shows B-subjects' average payback pattern, disentangled by treatment.

As the panel suggests, there are no treatment differences: We performed a Kruskal-Wallis test of the equality of distributions of payback values in the three treatments for each one of

the 11 hypothesized α's and found the smallest $P\text{-}value = 0.483$ for $\alpha = 0.1$.[14] Furthermore, recall that in phase 2 of *QnoD* and *QD* treatments *B*-subjects were asked to fill in again the questionnaire at the end of the experiment (cf. Table 2, final questionnaire). This was made to check whether *B*-subjects truthfully revealed their belief-dependent preferences.[15] With very few exceptions (4/40 in *QnoD*, 5/80 in *QD*), *B*-subjects confirmed the payback pattern of phase 2. Therefore, for these two treatments we only refer to the questionnaire in phase 2, while for treatment *NoQ* we rely on the final questionnaire—the only one filled in by *B*-subjects in this treatment. Finally, we checked that in each treatment there is no framing effect on the payback due to the presentation of the 11 lines of the questionnaire in reverse order in half of the experimental sessions of each treatment (Mann-Whitney test, smallest $P\text{-}value = 0.396$ for $\alpha = 0.2$ in *NoQ*). This is confirmed by a similar ratio of increasing over decreasing payback patterns in each order of presentation (χ^2 test, $P\text{-}value = 0.276$).

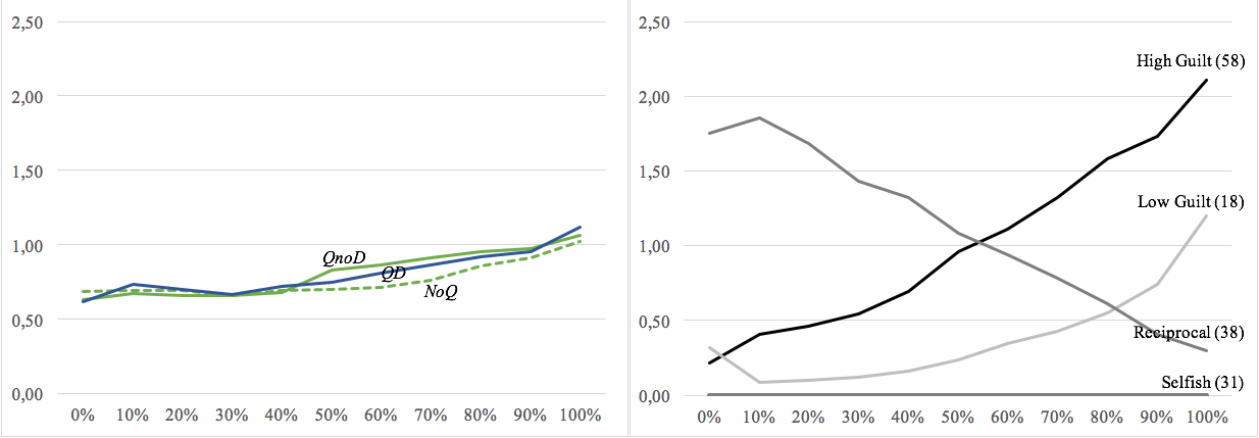

Figure 3 *B*'s average payback pattern, by treatment (left panel) and type (right panel).
The figure reports on the left panel *B*'s average payback pattern in *NoQ* (40 subjects), *QnoD* (40 subjects), and *QD* (80 subjects). On the right panel, it reports the average payback pattern of *B*-subjects according to the predicted shapes of $p(\alpha)$: high-guilt ($p(\alpha)$ increasing at least for $\alpha \geq 0.5$, and $p(1) \geq 2$), low-guilt ($p(\alpha)$ increasing at least for $\alpha \geq 0.5$, and $p(1) < 2$), reciprocal ($p(\alpha)$ decreasing at least for $\alpha \geq 0.5$), and selfish preferences ($p(\alpha) = 0$ for all α); for each average pattern, the intensity of the black color indicates the relative frequency of the corresponding shape in the population of *B*-subjects (reported in parentheses).

The left panel of Figure 3 shows that average payback patterns are increasing. This is the result of the prevalence of subjects whose elicited preferences display guilt aversion.[16]

[14] A Mann-Whitney test with a pairwise comparison between treatments confirms this result (smallest $P\text{-}value = 0.262$ for $\alpha = 0.1$ in *QnoD vs. QD*).

[15] We acknowledge that a tendency of *B*-subjects to provide the same answers in phase 2 and the final questionnaire could be due to a consistency motive (see, e.g., Podsakoff *et al.* 2003).

[16] Besides Attanasi *et al.* (2013), which use the same elicitation method, this result is in line with other

Indeed, according to the theory of belief-dependent preferences (Battigalli and Dufwenberg 2009), the payback pattern $p(\alpha)$ is *increasing* in α if B is **guilt-averse**, and *decreasing* in α if he is **reciprocal** *à la* Dufwenberg and Kirchsteiger (2004). Following this theoretical insight, and looking at the filled-in questionnaires, we can classify our B-subjects according to their elicited psychological type (we find 67/160 guilt-averse and 27/160 reciprocal B-subjects). We refine this classification further by observing that, in the theoretical analysis, B's behavior matters only when A Continues, and she does so only if $\alpha \geq 0.5$. As a consequence, given that what matters is B's behavior for $\alpha \geq 0.5$, we are able to classify some of the non-monotone payback patterns: we consider as guilt-averse also B-subjects with $p(\alpha)$ non-monotone in α but *increasing* for $\alpha \geq 0.5$ (9/160), and as reciprocal those with $p(\alpha)$ non-monotone in α but *decreasing* for $\alpha \geq 0.5$ (11/160). Furthermore, among all guilt-averse B-subjects, we disentangle those with $p(1) \geq 2$ and those with $p(1) < 2$ and call the former **high-guilt** (58/76) and the latter **low-guilt** (18/76), so that we call high-guilt those for whom the cooperative equilibrium exists in the stage game. Finally, we classify as **selfish** the B-subjects with $p(\alpha) = 0$ for each α (31/160). For B-subjects whose payback pattern is not captured by any of the above-described shapes we do not have a clear behavioral prediction, and so we categorize them as **unclassified** (15/160): the majority of them (11/15) have a positive flat payback pattern, which is consistent with inequity aversion.[17]

All this explains the categorization in Table 4 and the right panel of Figure 3. The former reports the distribution of the 160 B-subjects' psychological types across the four possible shapes of the payback pattern $p(\alpha)$ implied by our (belief-dependent) theory-based categorization. The latter reports, for each classified psychological type, their average payback pattern and the corresponding number of B-subjects: the majority of classified B-subjects is guilt-averse (76/145). This fraction is treatment-independent (17/40 in *NoQ*, 19/40 in *QnoD*, 40/80 in *QD*). Table 4 shows no significant difference between the distributions of types in *NoQ* and *QnoD* (χ^2 test, *P-value* $= 0.970$), which allows us to pool elicited types of these two treatments (column *NoQ-QNoD* in Table 4) so as to have the same number of observations without disclosure (*NoQ-QnoD*) and with disclosure (*QD*). Table 4 also shows no significant difference between the distributions of psychological types in *NoQ-QnoD* and *QD* (last two columns of Table 4: χ^2 test, *P-value* $= 0.639$). This is further evidence that the presence or absence of information disclosure does not affect subjects' answers to the questionnaire.[18]

studies eliciting trustees' belief-dependent motivations in the trust game, namely Ederer and Stremitzer (2016), and Bellemare *et al.* (2018). Although using different elicitation methods than ours, all these studies find that the majority of trustees are guilt-averse.

[17] B-subjects' answers to debriefing questions about the interpretation of the filled-in questionnaire seem to confirm the classification. These answers are available from the authors upon request.

[18] Notice that our model-free classification is different from the one in Attanasi *et al.* (2013), where high-guilt *vs.* low-guilt categories are disentangled according to a non-linear least square estimation of guilt

Categories of elicited psychological types	Treatment			
	NoQ	*QnoD*	*NoQ-QnoD*	*QD*
Guilt-averse: High-guilt	14	16	30	28
Guilt-averse: Low-guilt	3	3	6	12
Reciprocal	11	9	20	18
Selfish preferences	9	8	17	14
Unclassified	3	4	7	8
TOTAL	40	40	80	80

Table 4 Classification of B-subjects according to the payback pattern.

The table reports, for each treatment and category of psychological types: the number of B-subjects with payback pattern $p(\alpha)$ in that category. Column *NoQ-QnoD* pools the observations of *NoQ* and *QnoD*.

In the next section we will consider only classified types of Table 4 (145/160), i.e., types for whom our model gives theoretical predictions. In line with these predictions, we will focus on the comparison of behavior of high-guilt *vs.* low-guilt & selfish B-subjects. The latter category includes subjects for whom (*Continue*, *Take*) is not an equilibrium of the stage game. We label this category "**low-guilt**". We also analyze data from reciprocal B-subjects. We show in the Appendix that reciprocal subjects behave as selfish ones if their sensitivity to reciprocity is lower than a given threshold (see auxiliary hypothesis H0.ii). At the beginning of Section 5.2 we check that this hypothesis is verified, together with the auxiliary hypothesis of similar behavior of B-subjects in *NoQ* and *QnoD*, given their psychological type (H0.i).

5.2 Test of the Experimental Hypotheses

Preliminary controls Auxiliary hypothesis H0.i is verified. Considering classified types in Table 4 (37/40 for *NoQ* and 36/40 for *QnoD*), we find no significant difference between *NoQ* and *QnoD* in the behavior of B-subjects both in phase 1 and in all periods of phase 3 (χ^2 test, smallest *P-value* $= 0.407$ in period 1 of phase 3). The same holds if we run the test for the three categories of classified types separately (high-guilt, low-guilt and reciprocal). This is not surprising, given that we show that there is the same distribution of classified

and reciprocity sensitivities, and a (non-parametric) bootstrap estimation of the probability that an elicited psychological type falls into one of the predicted regions of behavior. We have replicated Attanasi *et al.* (2013) classification technique on our dataset, and found no significant difference between the distribution of elicited psychological types in Table 4 across the two classification methods, both over the whole sample of B-subjects (χ^2 test, *P-value* $= 0.850$) and within each treatment (χ^2 test, *P-value* $= 0.843$ for *NoQ-QnoD* and *P-value* $= 0.967$ for *QD*). Estimations of guilt and reciprocity sensitivities from replication of Attanasi *et al.* (2013) are available from the authors upon request.

types between *NoQ* and *QnoD* (see Table 4). We also find no significant difference in the behavior of *A*-subjects both in phase 1 and in the first three periods of phase 3 (χ^2 test, smallest *P-value* = 0.262 in phase 1).[19] Therefore, we pool the data of these two treatments (73 classified types) and from now on we refer to them as the 'no information disclosure' treatment (*NoQ-QnoD*), as opposed to *QD*, the treatment with information disclosure.

Notice that in phase 1 of the two treatments subjects are exposed to the same no-disclosure environment. Therefore, as expected, we find no between-treatment differences in the behavior of *A*-subjects and of *B*-subjects both at the aggregate level (χ^2 test, *P-value* = 0.813 for *A*s and 0.959 for *B*s) and if we disentangle by *B*'s type.[20] Furthermore, we find no within-treatment differences in the behavior of *A*-subjects according to the matched type—which they do not know (χ^2 test: *P-value* = 0.953 for *NoQ-QnoD* and *P-value* = 0.811 for *QD*), and a more cooperative behavior of high-guilt *B*-subjects as compared to both low-guilt *B*-subjects and reciprocal *B*-subjects, independently of the treatment (χ^2 test: in *NoQ-QnoD*, highest *P-value* = 0.062; in *QD*, highest *P-value* = 0.088).

The last control corroborates the auxiliary hypothesis H0.ii. Figure 4 adds support and confirms that H0.ii is verified. In fact, in each treatment we find no significant difference in the behavior of low-guilt and reciprocal *B*-subjects in both phase 1 (χ^2 test: *P-value* = 0.538 for *NoQ-QnoD*, 0.828 for *QD*), and in each period of phase 3 (for *NoQ-QnoD*: highest *P-value* = 0.519 in period 4; for *QD*: highest *P-value* = 0.357 in period 4). Figure 4 reports that the same holds for matched *A*-subjects, thus also the auxiliary hypothesis H0.iii is verified. Therefore, from now on we pool data of the two categories of psychological types (respectively in red and yellow color in Figure 4), and perform the statistical analysis by referring to **low-guilt & reciprocal** *B*-subjects and matched pairs (in orange color in all the next figures).

All of the above also holds for subjects' beliefs.

[19] The significant difference found in period 4 of phase 3 in favor of more trust by *A*-subjects in *QnoD vs. NoQ* (χ^2 test, *P*-value = 0.037) is probably due to a stronger end-game effect in treatment *NoQ*, with only 4/37 *A*-subjects choosing *Continue* (11/36 in *QnoD*) in the last period of the repeated game.

[20] About *A*'s choices, we find a similar fraction of *A*-subjects choosing *Continue* across *NoQ-QnoD* and *QnoD* given the type (χ^2 test: *P*-value = 0.885, 0.717 and 0.825 for *A*s' matched respectively with a high-guilt *B*, a low-guilt *B*, and a reciprocal *B*). About *B*'s choices, we record a similar fraction of *B*-subjects choosing *Share* across the two treatments given the type (χ^2 test: *P*-value = 0.975, 0.559 and 0.791 respectively for a high-guilt *B*, a low-guilt *B*, and a reciprocal *B*).

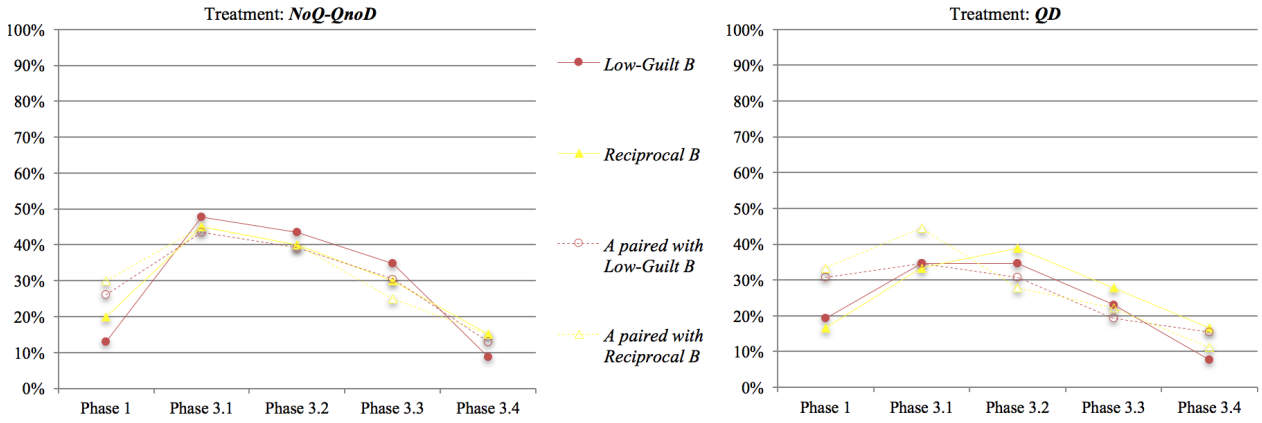

Figure 4 Frequency of B's *Share* and A's *Continue* choices in low-guilt *vs.* recipr. pairs, by treat. Remark: To emphasize differences in frequencies within each subgroup of types, dots of phase 1 are connected to dots of phase 3.1 although game and partner change between phase 1 and phase 3.

We report below the results of the tests of the experimental hypotheses derived from our model, following the same order of presentation of Section 4, where these hypotheses are formulated.

HA1: In *QD*, *A*-subjects' first-order beliefs are higher if matched with a high-guilt rather than a low-guilt or reciprocal *B*-subject. Table 5 reports A-subjects' average first-order beliefs (standard errors in parentheses) according to their matched B's type in treatment QD, for phase 1 and for each period of phase 3.

Elicited type	Phase 1	Phase 3.1	Phase 3.2	Phase 3.3	Phase 3.4
High Guilt	27.86 (4.58)	60.36 (4.28)	65.00 (5.45)	66.79 (5.42)	42.86 (6.17)
Low Guilt & Reciprocal	32.05 (4.07)	40.91 (4.20)	40.91 (4.54)	33.64 (3.93)	23.41 (3.75)
P-value (Mann-Whitney)	0.575	0.001	0.001	0.000	0.009

Table 5 A-subjects' average first-order beliefs in QD, by matched B's type, phase and period. Average beliefs in percentages; standard errors in parentheses. Results of Mann-Whitney test on equality of population medians are reported in the last row of the Table.

As shown by the preliminary controls, no significant difference is found for phase 1 across different matched types. As for phase 3, a significantly higher average first-order belief is recorded for A-subjects matched with a high-guilt type *vs.* those matched with a low-guilt

or reciprocal type. As shown in Table 5, this difference is significant at the 1% level (Mann-Whitney test) for each period of the repeated game in *QD*. Therefore, we conclude that **HA1 is verified**.

HA2: *A*-subjects' first-order beliefs vary more over time in *NoQ-QnoD* than in *QD*. Furthermore, in *NoQ-QnoD* they are more polarized in the last than in the first period; in *QD* they are equally polarized in the last and in the first period. Figure 5 presents the distribution of *A*-subjects' coefficients of variation of first-order beliefs over the four periods of the repeated game in phase 3, disentangled by *A*-subjects' first-order belief in the one-shot game of phase 1. Each coefficient of variation is calculated on a within-subject base. The left-hand panel refers to *NoQ-QnoD*; the right-hand panel refers to *QD*. As Figure 5 shows, the distribution of coefficient of variations is significantly higher in *NoQ-QnoD* than in *QD* (Mann-Whitney test, $P\text{-}value = 0.001$). This holds in particular for *A*-subjects whose first-order beliefs in phase 1 are low, i.e., $\alpha < 50\%$ ($P\text{-}value = 0.002$). For these *A*-subjects, (low-guilt or reciprocal) *B*'s strategic reputation building in *NoQ-QnoD* has more room to play a positive effect on their beliefs in the first period of phase 3, thereby making them vary more over the last three periods of the repeated game. Thus, we conclude that **the first part of HA2 is verified**.

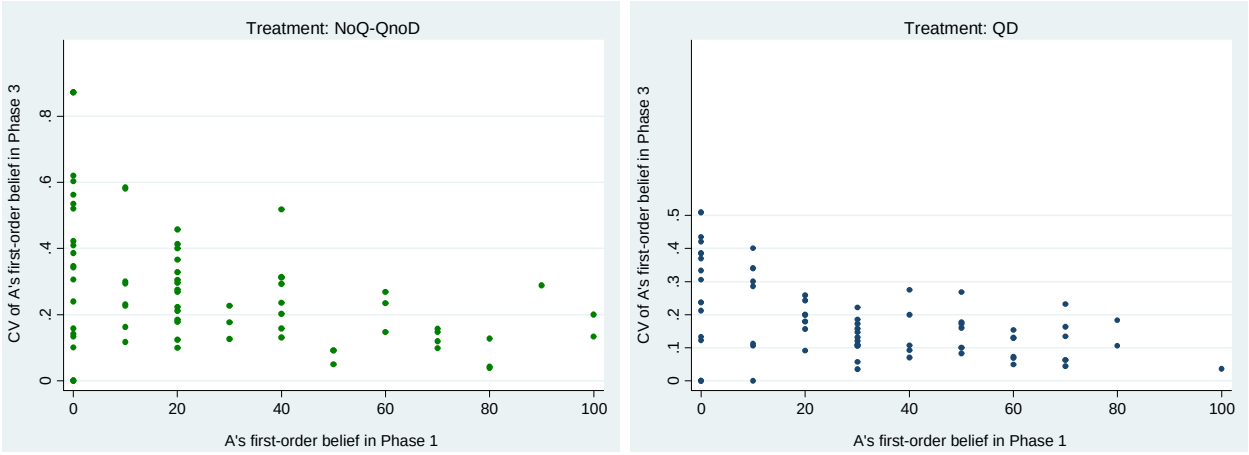

Figure 5 Distribution of within-subject Coefficients of Variation (CV) of *A*'s first-order beliefs across periods of phase 3, by treatment and by *A*'s first-order belief (in %) in phase 1.

Let us now compare, for each treatment, the polarization of *A*-subjects' first-order beliefs in period 4 *vs.* period 1 of phase 3. Figure 6 shows that beliefs are more polarized in period 4 than in period 1 not only in *NoQ-QnoD* (Mann-Whitney: $P\text{-}value = 0.001$) but also in *QD* ($P\text{-}value = 0.000$) where instead our model predicts the same level of polarization. Thus,

we conclude that **the second part of HA2 is verified in *NoQ-QnoD* but not in *QD***.

The latter result might be due to our elicitation method of A-subjects' first-order beliefs α_t across periods of the repeated game ($t = 1, ..., 4$). Recall (see Section 3.2) that A's elicited first-order belief is not only about the matched B, but about all the 10 B-subjects in the session; hence, we should get an elicited α_1 that is less polarized than the true one.

For example, an A who faces a low-guilt B in period 1 of phase 3 of QD is asked how many of the 10 B-subjects in the session (the matched B and the other nine) will Share in that period, and she can rationally presume—despite the disclosed filled-in questionnaire of the matched low-guilt B—that there are some high-guilt B-subjects in the session. As the repeated game with paired matching unfolds, observation of no cooperation by the matched low-guilt B may lead A to decrease α_t across periods. This boosts the frequency of last-period belief $\alpha_4 = 0$ on the cooperative behavior of B-subjects in pairs with a disclosed low-guilt (or reciprocal) B in QD, as one might presume by looking at Figure 6. We come back to this point when discussing the experimental hypotheses on matched pairs.

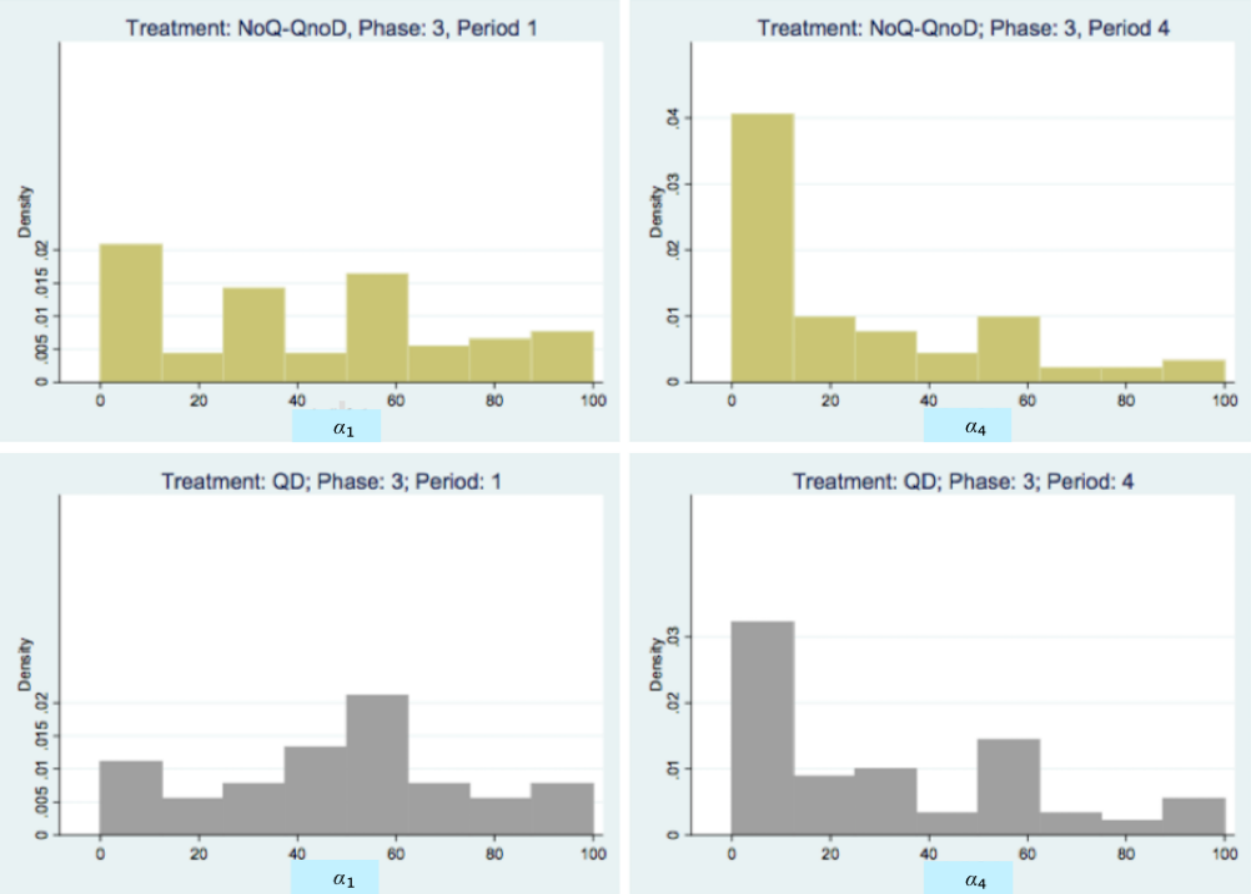

Figure 6 Distribution of A's first-order beliefs in period 1 and in period 4 of phase 3, by treatment.

HB1: Low-guilt and reciprocal *B*-subjects display more reputation building in *NoQ-QnoD* than in *QD*. The frequencies under scrutiny here are represented by the solid orange lines in Figure 7. We verify that the frequency of *Share* of low-guilt & reciprocal *B*-subjects is similar in phase 1 of the two treatments (16% in *NoQ-QnoD vs.* 18% in *QD*; χ^2 test: *P-value* = 0.815). Figure 7 shows a higher percentage of low-guilt & reciprocal *B*-subjects choosing *Share* in the first period of phase 3 of *NoQ-QnoD* than in *QD*: 47% *vs.* 34%, although this difference is not significant (*P-value* = 0.238).

The treatment difference becomes significant if we only consider low-guilt & reciprocal *B*-subjects switching from *Take* in phase 1 to *Share* in period 1 of phase 3, i.e., those for whom reputation building is transparent: 33% in *NoQ-QnoD vs.* 16% in *QD* (χ^2 test, *P-value* = 0.070). Thus, we conclude that **HB1 is verified under some restrictions**.

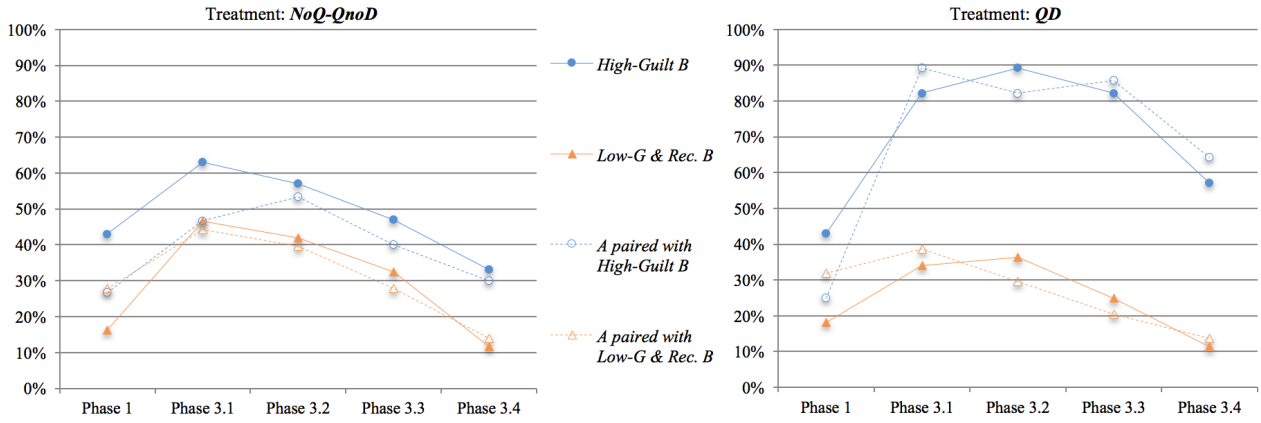

Figure 7 Frequency of *B*'s *Share* and *A*'s *Continue* choices, by *B*'s type and by treatment.

Remark: To emphasize differences in frequencies within each subgroup of types, dots of phase 1 are connected to dots of phase 3.1 although game and partner change between phase 1 and phase 3.

HB2: In *NoQ-QnoD*, independently of *B*'s type, for any given second-order belief of *B* sharing is more likely in period 1 of phase 3 than in phase 1. In *QD* this is true only for high-guilt *B*-subjects. Figure 7 also offers a check of this hypothesis. We first verify that the distribution of *B*-subjects' second-order beliefs in phase 1 is not significantly different across treatments (Mann-Whitney test: *P-value* = 0.699 for high-guilt; *P-value* = 0.672 for low-guilt & reciprocal), and across types within the same treatment (*P-value* = 0.896 for *NoQ-QnoD*; *P-value* = 0.299 for *QD*).

Let us focus on *NoQ-QnoD* (left panel of Figure 7). We detect a non-significant difference between the frequencies of *Share* in period 1 of phase 3 and in phase 1 for high-guilt *B*s (χ^2 test, *P-value* = 0.120); the difference is significant for low-guilt & reciprocal *B*-subjects (χ^2

test, *P-value* = 0.003). This confirms only partially H.4 for treatment *NoQ-QnoD*.

In *QD*, H4 seems to be confirmed only for high-guilt *B*s, who Share significantly more (at the 1% level) in period 1 of phase 3 than in phase 1 (χ^2 test, *P-value* = 0.003), with the difference being still significant—although only at the 10% level—for low-guilt & reciprocal *B*-subjects (χ^2 test, *P-value* = 0.089), while instead our model predicts no difference for the latter subgroup of *B*-subjects. Therefore, we conclude that **HB2 finds confirmation only for low-guilt & reciprocal types in *NoQ-QnoD* and for high-guilt types in *QD***.

HB3: In both *QD* and *NoQ-QnoD*, sharing is more likely for higher beliefs of high-guilt *B*-subjects. The threshold is lower for earlier periods. Table 6 reports values of the rank-biserial correlation coefficient, **Somers'** *D*, between the *Share* choice and second-order belief of *Share*, for each treatment and each phase-period combination. A high and significant (at the 1% level) positive correlation is found in phase 1 and in each period of phase 3 for *QD*. For *NoQ-QnoD* we find the same result, except for period 1 of phase 3, possibly due to the fact that several *B*-subjects in this treatment chose *Share* in the first period of the repeated game even though they hold low second-order beliefs. Indeed, the reputation-building choice of *Share* for those *B*-subjects was at no (or, at most, low) monetary cost, since—due to perfect monitoring—the matched *A*-subject would be informed about it also in the case she would choose *Dissolve* in period 1. With this, we can state that **the first part of HB3 is verified in both QD and NoQ-QnoD**.

Treatment	Phase 1	Phase 3.1	Phase 3.2	Phase 3.3	Phase 3.4
NoQ-QnoD	0.71***	0.29	0.54***	0.60***	0.84***
QD	0.68***	0.60***	0.95***	0.77***	0.86***

Table 6 Rank correlation between *B*s' choice and belief of *Share*, by treatment, phase and period. Correlation between a dichotomic (choice) and discrete (belief) variable is measured through Somer's *D*; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Now we test the second part of HB3. A lower threshold for second-order beliefs leading to choose *Share* in period t than for those leading to choose *Share* in period $t+1$ of the repeated game empirically gives a distribution of *Share* choices across beliefs of period $t+1$ that stochastically dominates the distribution of period t, for each $t = 1, 2, 3$. Figure 8 reports, for each treatment and for each period of the repeated game, the cumulative distribution of high-guilt *B*-subjects choosing *Share* across the second-order belief they hold.

For *NoQ-QnoD*, the distribution of period 1 is first-order stochastically dominated by the distribution of period 2, but the two distributions are not significantly different (Two-sample Kolmogorov-Smirnov test: *P-value* = 0.490). Furthermore, no significant difference

is detected between the distributions of period 2 and period 3 (*P-value* = 0.812) and between those of period 3 and period 4 (*P-value* = 0.719). Therefore, we conclude that the second part of HB3 is not verified for *NoQ-QnoD*.

For *QD* instead it is easy to see that distribution of period 1 is first-order stochastically dominated by the distributions of any of the other three periods (Two-sample Kolmogorov-Smirnov test: *P-value* = 0.006 for period 1 vs. period 2, 0.006 vs. period 3, 0.000 vs. period 4). Because both the distributions of periods 2 and 3 (*P-value* = 0.966) and the distribution of period 3 and 4 (*P-value* = 0.679) are not significantly different among them, we conclude that **the second part of HB3 is verified only for period 1 *vs.* the following periods of *QD***.

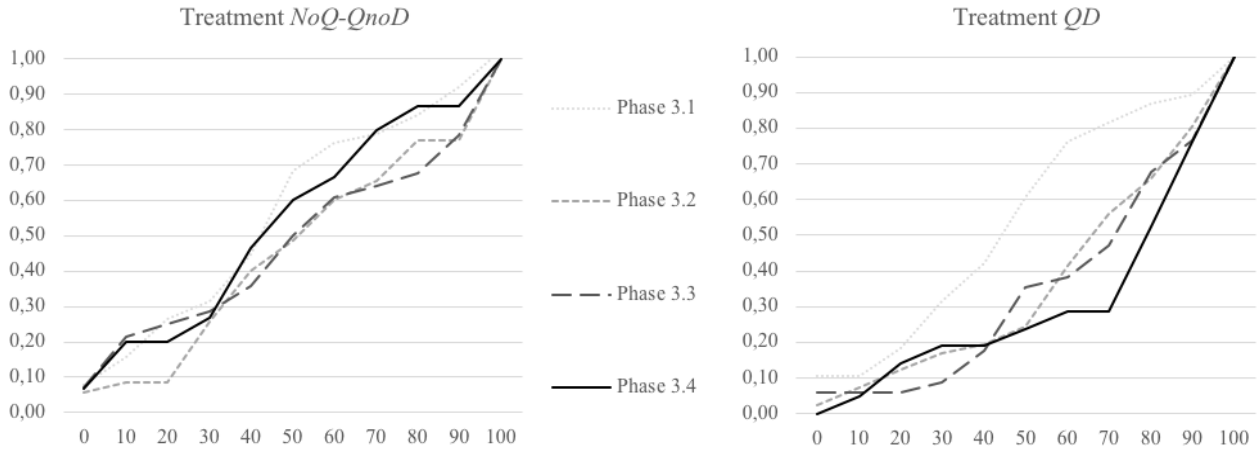

Figure 8 Cumulative distributions of *B*s' *Share* choices across second-order beliefs, by treatment.

HP1: In pairs including low guilt & reciprocal *B*-subjects, the frequencies of the cooperative path up to t are higher in *NoQ-QnoD* than in *QD*. Figure 9 reports, for each treatment, the frequency of (*Continue*, *Share*) choices of matched pairs in period t (dotted line) and the frequency of matched pairs always choosing (*Continue*, *Share*) up to period t (solid line), disentangled by B's type in the pair. Recall that this type is disclosed in *QD* only after phase 1, and never disclosed in *NoQ-QnoD*. The controls in phase 1 work as they should: due to the random matching of *A*-*B* pairs, the frequency of pairs choosing (*Continue*, *Share*) is not significantly different between the two treatments (15% in *NoQ-QnoD vs.* 10% in *QD*: χ^2 test, *P-value* = 0.329). The same holds if we disentangle pairs according to the psychological type of the *B*-subject (high-guilt *vs.* low-guilt or reciprocal).

Figure 9 shows that for each period of phase 3 the frequency of low-guilt & reciprocal *A*-*B* pairs choosing (*Continue*, *Share*) in period t—dotted orange line—is higher in *NoQ-QnoD* than in *QD* for each $t = 1, 2, 3$. However, none of these differences is significant (χ^2 test: lowest *P-value* = 0.176 for $t = 2$), and in period 4 of both treatments we find no pair

choosing (*Continue*, *Share*).

The same holds if, for periods $t > 1$, we consider low-guilt & reciprocal A-B pairs always choosing (*Continue*, *Share*) up to that period—solid orange line (χ^2 test: lowest *P-value* = 0.477 for $t = 2$). In particular, neither in *NoQ-QnoD* nor in *QD* any of the pairs with a low-guilt or reciprocal B-subject choosing (*Continue*, *Share*) in period 1 was able to keep on the (*Continue*, *Share*) path until the end of the repeated game. This is predicted by our model (see the comparative statics after Proposition 1). Despite this, due to the absence of significant between-treatment differences, we have to conclude that **we find only weak support for HP1**.

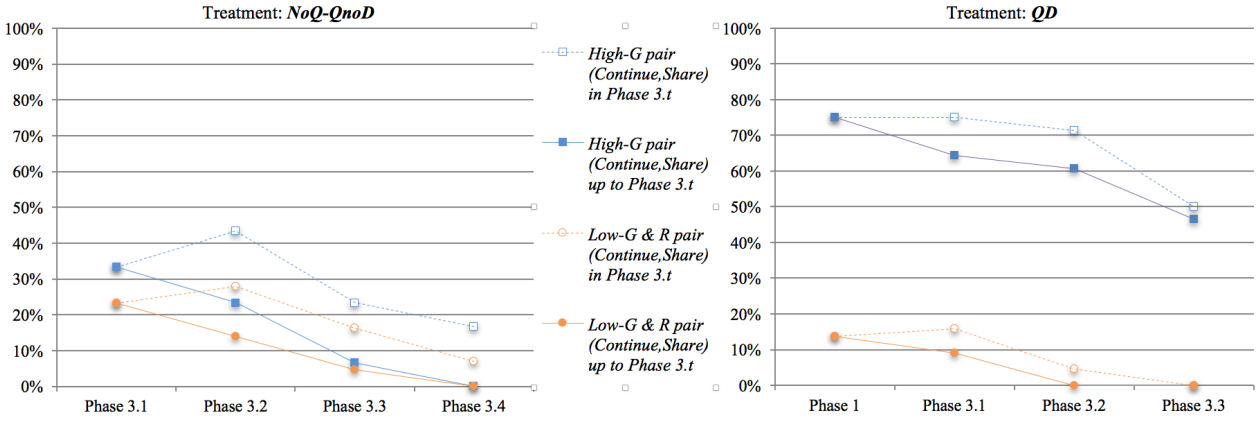

Figure 9 Frequency of A-B pairs' (*Continue*, *Share*) choices in t and path (up to t), by B's type and by treatment.

HP2: In pairs including high-guilt *B*-subjects, the frequencies of the cooperative path up to t are higher in *QD* than in *NoQ-QnoD*. Recall that high-guilt B-subjects should cooperate more than low-guilt ones in the one-shot game with no disclosure (see Attanasi *et al.* 2016). This control in phase 1 works as it should: due to a significantly higher frequency of *Share* by high-guilt subjects (in line with results in Attanasi *et al.* 2013), the frequency of pairs with a high-guilt B choosing (*Continue*, *Share*) is higher than for pairs with a low-guilt or reciprocal B, although this difference is not significant in any of the two treatments (χ^2 test: *P-value* = 0.325 for *NoQ-QnoD*, 0.297 for *QD*).

Figure 9 shows that for each period of phase 3 the frequency of pairs with a high-guilt B choosing (*Continue*, *Share*) in period t—dotted blue line—is significantly higher in *QD* than in *NoQ-QnoD* for each $t = 1, 2, 3, 4$ (χ^2 test: highest *P-value* = 0.014 for $t = 2$). The same holds if, for periods $t > 1$, we consider pairs always choosing (*Continue*, *Share*) up to that period—solid blue line (χ^2 test: highest *P-value* = 0.002 for $t = 2$). In particular, in *QD*, among the 14/28 pairs with a high-guilt B choosing (*Continue*, *Share*) in period 4,

13/14 had also chosen the same action profile in all previous periods of the repeated game. Conversely, in *NoQ-QnoD*, despite 10/30 pairs choosing (*Continue*, *Share*) in period 1, none of these pairs was able to keep on the (*Continue*, *Share*) path until the end of the repeated game. Therefore, we can conclude that **HP2 is verified**.

HP3: In *QD*, pairs including high-guilt *B*-subjects are on a fully cooperative path. Figure 9—right-hand panel, dotted blue lines—shows that around 70% of *A*-*B* pairs with a high-guilt *B* choose (*Continue*, *Share*) in each of the first three periods of the repeated game and 50% of them choose (*Continue*, *Share*) in the last period. In each period, the fraction of (*Continue*, *Share*) choices is significantly higher than the one (25%) obtained through a random guess over the four possible strategy profiles (χ^2 test: *P-value* < 0.001 for the first three periods, *P-value* $= 0.053$ for the last period). If, for each period t, we only focus on *A*-*B* pairs always choosing (*Continue*, *Share*) up to t, we find that for pairs with a high-guilt *B* this fraction is not significantly different from the previous one in any of the four periods (compare solid and dotted blue lines in the right-hand panel of Figure 9: the smallest *P-value*, 0.383, is found for period 2, χ^2 test). Therefore, almost all high-guilt pairs cooperating in a period $t > 1$ have also cooperated in all previous periods. In particular, over the 21/28 pairs with a high-guilt *B* cooperating in period 1, 13/21 (62%) cooperate until the end of the repeated game.

As a control, we check that pairs with a low-guilt & reciprocal *B* are not on a (*Continue*, *Share*) path. Indeed, Figure 9—right-hand panel, dotted orange lines—shows that the fraction of these pairs choosing (*Continue*, *Share*) in the first two periods is not significantly different from the one of a random guess (χ^2 test: *P-value* $= 0.177$ for $t = 1$, 0.291 for $t = 2$), and in the last two periods the former fraction is significantly lower (χ^2 test: *P-value* $= 0.047$ for $t = 3$, *P-value* < 0.001 for $t = 4$). As highlighted above, none of these pairs chooses (*Continue*, *Share*) in the last period of the repeated game. This control on low-guilt & reciprocal *A*-*B* pairs provides further supports to the fact that **HP3 is verified**.

6 Conclusions

This paper investigates the interaction between belief-dependent preferences and reputation building in a repeated Trust Minigame. We focus on the 4-period repetition of a Trust Minigame, and assume role-dependent guilt: *A* (the trustor) is selfish while *B* (the trustee) can feature a high or low degree of guilt aversion. With this, we analyze in a laboratory experiment the interplay between repetition of the game and information (in)completeness on *B*'s psychological type. We implement two main treatments, where subjects play the 4-period Trust Minigame with partner matching. In the main treatment, subjects playing in

role A receive information on their co-players' psychological type (information disclosure), while in the other one no information is disclosed to them. We derive several theoretical predictions, which can be classified into four main groups.

The first group concerns reputation building without information disclosure, irrespective *of the guilt type*, and can be summarized as follows: B-subjects cooperate more in the first period of the repeated game than in the corresponding one-shot game.

The second group concerns the behavior of specific guilt types, irrespective *of the information setting*: Due to belief-dependent preferences, cooperation is more likely to occur for higher beliefs of high-guilt B-subjects, and the cooperation threshold on these subjects' beliefs is lower for earlier periods.

The third group concerns behavioral differences *across disclosed types*: With information disclosure, A's first-order beliefs is higher if matched with a high-guilt rather than a low-guilt B. Moreover, only high-guilt subjects try to build reputation at the beginning of the repeated game. Finally, pairs with a high-guilt B-subject follow a cooperative path, while those with a low-guilt one do not.

The last group accounts for the majority of our behavioral predictions, namely those concerning the *interplay between the psychological type and its disclosure* within the matched A-B pair. First of all, A's first-order beliefs vary more over time without disclosure, and only in this treatment they are dispersed in the first period and polarized in the last period of the repeated game. Second, low-guilt B-subjects display more reputation building without information disclosure. Third, in pairs with a low-guilt B-subject the frequency of fully cooperative paths is higher without information disclosure, whereas in pairs with a high-guilt B-subject it is higher under information disclosure.

Our theoretical predictions capture well the central tendencies of the data, both across subjects within the same role and across matched pairs, and especially for disclosed high-guilt B-subjects and for low-guilt B-subjects without disclosure. The model predicts particularly well the emergence of a full cooperation path in pairs where a high-guilt B-subject is disclosed: out of 21/28 pairs with a high-guilt B cooperating at the beginning of the repeated game, 13/21 (62%) cooperate until the end of the repeated game, thereby counterbalancing the usual end-game effect detected in finitely repeated games with few periods like ours. Furthermore, subjects featuring low reciprocity concerns behave as low-guilt subjects in each period of the repeated game and under each information setting, as it is also predicted by an enrichment of our model.

We detect two main deviations from the theoretical predictions. First, A-subjects' first-order beliefs are more polarized in the last than in the first period of the treatment with disclosure, while our model predicts the same level of polarization. This might be due to our elicitation method. Indeed, A's elicited first-order belief is not only about the matched

B-subject, but about the whole population of B-subjects in the experiment; hence, these beliefs may reflect in the first period the fact that A only has information about 1/10 B-subjects, and may polarize in later periods according to the behavior of the disclosed type (high guilt or low guilt). Second, low-guilt B-subjects try to build some reputation despite their disclosed type at the beginning of the repeated game. Our informed conjecture is that—as emphasized when introducing our model in Section 2—even when information on B's psychological type is disclosed, this only approximates complete information. Therefore, some learning is still possible, and low-guilt B-subjects may behave as in the treatment with no information disclosure. However, this occurs only in the first period, and with no success: only 10% of such A-B pairs cooperate until the second period, and none of them until the third one.

References

[1] Anderhub, V., Engelmann, D., and W. Guth (2002): "An Experimental Study of the Repeated Trust Game with Incomplete Information," *Journal of Economic Behavior & Organization*, 48, 197–216.

[2] Aina, C., P. Battigalli, and A. Gamba (2018): "Frustration and Anger in the Ultimatum Game: An Experiment," IGIER Working Paper no. 621.

[3] Attanasi G., P. Battigalli and E. Manzoni (2016): "Incomplete Information Models of Guilt Aversion in the Trust Game," *Management Science*, 62, 648–667.

[4] Attanasi G., P. Battigalli and R. Nagel (2013): "Disclosure of Belief-dependent Preferences in a Trust Game," IGIER Working Paper no. 506.

[5] Balafoutas, L. (2011): "Public Beliefs and Corruption in a Repeated Psychological Game," *Journal of Economic Behavior & Organization,* 78, 51–59.

[6] Battigalli, P., and M. Dufwenberg (2007): "Guilt in Games," *American Economic Review, Papers & Proceedings*, 97, 170–176.

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

[8] Battigalli P., R. Corrao, and M. Dufwenberg (2018): "Incorporating Belief-Dependent Motivation in Games," Mimeo.

[9] Battigalli, P., M. Dufwenberg, and A. Smith (2015): "Frustration and Anger in Games," IGIER Working Paper no. 539.

[10] Bellemare, C., A. Sebald, and S. Suetens (2017): "A Note on Testing Guilt Aversion," *Games and Economic Behavior*, 102, 233–239.

[11] Bellemare, C., A. Sebald, and S. Suetens (2018): "Heterogeneous Guilt Aversion and Incentive Effects," *Experimental Economics*, forthcoming.

[12] Charness, G., and M. Dufwenberg (2006): "Promises and Partnership," *Econometrica*, 74, 1579–1601.

[13] Cooper, D.J., and J. Kagel (2016): "Other Regarding Preferences: A Survey of Experimental Results," in J. Kagel and A. Roth (eds.), *The Handbook of Experimental Economics*, Vol. 2, Ch. 4, pp. 217–289, Princeton: Princeton University Press.

[14] Dufwenberg, M., and G. Kirchsteiger (2004): "A Theory of Sequential Reciprocity," *Games and Economic Behavior*, 47, 268–298.

[15] Ederer, F., and A. Stremitzer (2016): "Promises and Expectations," Cowles Foundation Discussion Paper No. 1931.

[16] Engle-Warnick, J., and R.L. Slonim (2004). The Evolution of Strategies in a Repeated Trust Game," *Journal of Economic Behavior & Organization*, 55, 553–573.

[17] Fischbacher, U. (2007): "Z-Tree: Zurich Toolbox for Readymade Economic Experiments," *Experimental Economics*, 10, 171–178.

[18] Kreps, D.M., and R.J. Wilson (1982): "Reputation and Imperfect Information," *Journal of Economic Theory*, 27, 253–279.

[19] Kreps, D.M., P.R. Milgrom, J. Roberts, and R.J. Wilson (1982): "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma," *Journal of Economic Theory*, 27, 245–252.

[20] Mailath, G.J., and L. Samuelson (2006): *Repeated Games and Reputations: Long-Run Relationships*. Oxford, UK: Oxford University Press.

[21] Milgrom, P.R., and J. Roberts (1982): "Predation, Reputation, and Entry Deterrence," *Journal of Economic Theory,* 27, 280–312.

[22] Podsakoff, P.M., MacKenzie, S.B., Lee, J.-Y., and N. P. Podsakoff (2003): "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies," *Journal of Applied Psychology*, 88, 879–903.

Appendix: Reciprocity concerned B-subjects

As mentioned in Section 2 there is a main alternative model of belief-dependent preferences that can be relevant in the Trust Minigame, **intention-based reciprocity**. If we introduced preferences consistent with intention-based reciprocity (see Dufwenberg and Kirchsteiger 2004), B's psychological utility function depends on his monetary payoffs, and on the utility (disutility) of increasing A's payoff if A is kind (unkind) to him. More specifically, B's utility is

$$u_B(m_A, m_B, \alpha) = m_B + R \cdot K(\alpha) \cdot m_A,$$

where $K(\alpha)$ is the kindness of player A, that is the difference between the payoff that B expects to let A have, given B's belief about A's strategy, and A's "equitable" payoff. In the specification of the Trust Minigame that constitutes our stage game, the equitable payoff of B from A's perspective is

$$m_B^e(\alpha) = \frac{1}{2}\mathbb{E}_\alpha[m_B(C,\cdot)] + \frac{1}{2}\mathbb{E}_\alpha[m_B(D,\cdot)] = \frac{4-2\alpha}{2} + \frac{1}{2} = \frac{5}{2} - \alpha.$$

Hence, A's kindness when she Continues is

$$K_A(\alpha_A) = (4-2\alpha) - \left(\frac{5}{2} - 2\alpha\right) = \frac{3}{2} - \alpha_A.$$

Notice that with reciprocity concerns *à la* Dufwenberg and Kirchsteiger (2004), B's willingness to share depends on his perception of A's action as either kind or neutral toward him: The less A expects B to Share, the kinder is her action; therefore, B's willingness to share is decreasing in his second-order belief of *Share*. We exploited this relation between B's second-order beliefs and choices—which is different from the one predicted by guilt-averse preferences—, in Section 5.1, when we classified B-subjects into groups of belief-dependent attitudes according to their filled-in questionnaire.

The one-shot Trust Minigame with reciprocity concerns is therefore

A/B	$Take$	$Share$
$Dissolve$	$1, 1$	$1, 1$
$Continue$	$0, 4$	$2, 2 + R\left(\frac{3}{2} - \alpha_A\right) 2$

For low values of sensitivity to reciprocity concerns, in particular for $R < \frac{2}{3}$, this game has only the non-cooperative equilibrium (D, T).

Proposition 2 shows that, as a consequence, in a model where players can be either selfish or reciprocal, reputation concerns play no role if the reciprocity parameter is low enough. We modify the model of repeated interaction described in Section 2 by changing the set of

possible psychological types of player B. We assume that B's type is now $R_B \in \{0, R^H\}$, with $R^H < \frac{2}{3}$. B's type is still his private information, and we call ρ_1 A's prior belief on B's type, $\rho_1 = \mathbb{P}\left[R_B = R^H | h_\emptyset\right]$, which is common knowledge.

Proposition 2 *The 4-period repeated game described above has a unique sequential equilibrium strategy pair, in which (D, T) is played in every period.*

Proof. We show by induction that the only subgame perfect equilibrium prescribes that the non-cooperative action pair is chosen in every period.

Period $t = 4$. In the stage game, regardless of B's type, there is a unique equilibrium, (D, T). Hence, (D, T) is played in $t = 4$.

Period $t < 4$. Let the equilibrium behavior be (D, T) in every period $\tau > t$. Then, as the actions in period t do not affect future payoffs, the unique equilibrium of the stage game (D, T) is played also in period t. ■

The result of Proposition 2 suggests that we may pool the observation of the low-guilt (and selfish) subjects with those of reciprocal B-subjects if the auxiliary hypotheses H0.ii and H0.iii are verified.