Economic Theory
https://doi.org/10.1007/s00199-023-01491-3

RESEARCH ARTICLE

Sorting of trustees: the good and the bad stay in the game

Eberhard Feess[1] · **Florian Kerzenmacher**[2]

Received: 9 June 2022 / Accepted: 10 March 2023

Abstract
We extend the theoretical and experimental analysis of endogenous sorting in social dilemma games to decisions of trustees in trust games. Trustees first decide about the amount they send back if the trustor sends the money and then learn that they can exit the game for a payoff that is identical to the trustor's endowment. We develop a behavioral model where trustors and trustees have reciprocal preferences, and hence put positive weight on the other player's payoff if they perceive their behavior as kind. Our model yields two possible constellations: Only trustees with high reciprocity participate, or all types except those with intermediate reciprocity participate. Our data lend strong support for the second pattern, as we observe a U-shaped relation between the trustees' participation rate and the amount they return. Trustors are hence left with an extreme pool of participants where they are either matched with particularly selfish or generous trustees.

Keywords Trust game · Self-selection in games · Sorting · Reciprocity · Altruism

JEL Classification C92 · D02 · D63 · D71

We are grateful to Frauke v. Bieberstein, Jan Feld, Stefen Lippert, Gerd Muehlheusser, James Tremewan, and seminar participants at the Universities of Auckland and Wellington for helpful remarks. Financial support from the Austrian Science Fund (FWF, SFB F63) and a Faculty Research Grant from Victoria School of Business and Governance are gratefully acknowledged.

✉ Florian Kerzenmacher
florian.kerzenmacher@uibk.ac.at

Eberhard Feess
eberhard.feess@vuw.ac.nz

[1] Victoria University of Wellington, Pipitea Campus, Lambton Quay, Wellington, New Zealand

[2] Department of Economics, University of Innsbruck, Universitätsstr. 15, 6020 Innsbruck, Austria

Published online: 06 April 2023

1 Introduction

Offering people the opportunity to opt out of social dilemma games is likely to be Janus-faced. On the one hand, voluntary participation may attract people with high social preferences, thereby increasing cooperation and efficiency. On the other hand, people with low social preferences may take part in order to take advantage of cooperators. It is therefore not surprising that the experimental literature on the impact of voluntary participation (endogenous sorting) yields mixed results (see the overview and meta analysis by Guido et al. 2019, and the literature review in Sect. 2). In this paper, we extend the theoretical and experimental analysis of endogenous sorting in social dilemma games to the impact of exit options for trustees in trust games.

Analyzing exit options of trustees is complementary to the existing experimental literature in two respects: First, exit options in trust games have so far only been investigated for trustors, but not for trustees. The main purpose of analyzing exit options for trustors is to distinguish between altruism and reciprocity as possible motives of trustors to return money, since reciprocity can only be a motive when trustees are not forced to send the money. Conversely, analyzing the impact of exit options for trustees is important due to its potential impact on cooperation, and thereby on efficiency: If trustees with high social preferences stay in the game more (less) often, and if this is anticipated by trustors, then self-selection increases (decreases) efficiency. Second, analyzing self-selection of trustors complements the experimental literature, which so far focuses on dictator games and VCM-mechanisms: Decisions in dictator games have purely distributional consequences, but are neutral from an efficiency perspective. In VCM-mechanisms, the decisions of all players affect both efficiency and distribution, and players are (most often) symmetric. By contrast, efficiency in trust games depends only on the behavior of trustors, while distribution depends only on the behavior of trustees.

In addition to complementing the experimental literature on sorting, our more general contribution is twofold: First, we develop a behavioral model that suggests an explanation for why the literature finds mixed results on the impact of sorting. In our model, we assume that trustors and trustees have reciprocal preferences, that is, they put positive weight on their counterpart's payoff if and only if they perceive their behavior as kind (in a sense that is specified in the model). An important feature of the model is that the payoff of trustors enters the trustees' utility function both when they participate and when they exit, and this gives rise to two kinds of equilibria. In the first equilibrium, only trustees with high degrees of reciprocity participate. In the second equilibrium, trustees with particularly low and high degrees of reciprocity participate, while those with intermediate degrees of reciprocity exit. In this equilibrium, endogenous sorting leads to a U-shaped pattern between the amount trustees return in case of participation and the participation rate. In a nutshell, the reason that both of these two patterns are possible is that the trustees' social preferences matter for the participation *and* for the amount they return. If reciprocity increases, this may then lead to a higher incentive to participate (in order to return high amounts), or to a higher incentive to exit (if the amount the trustor gets in this case dominates the trustee's decision).

Our second more general contribution is that our treatments allow us to decompose the impacts of the own social preferences and the beliefs about the behavior of the

player one is matched with on the decision to participate. We achieve this by assigning trustees randomly to either the no-information treatment N or the information treatment I. The only difference between these two treatments is that, in treatment I, we inform trustees about the percentage of trustors who send the money. The belief about the percentage of trustors sending the money, which may well be (positively) correlated with the own social preferences, hence matters in treatment N but not in treatment I. Our results show that accounting for the impact of beliefs is crucial for understanding our experimental results.

Our experimental design is as follows: In addition to their payment for participating, trustors in our experiment, carried out on Amazon MTurk with 1,760 participants, received an endowment they could either keep or send to the trustee they were paired with. In the latter case, the amount was tripled. Trustees were first asked how much they return if the trustor sends the money, and only then learned that they can exit the game for an amount identical to the trustor's endowment. This order of moves allows us to collect the amounts for all trustees, which is required for analyzing how the exit option affects the pool trustors can be matched with. If a trustee takes part and the trustor does not send the money, then the trustee keeps only the fixed payment for participation. As a consequence, the incentive of trustees to stay in the game does not only depend on their social preferences, but also on their beliefs about how many trustors send the money. This is why distinguishing between treatments N and I is important.

In treatment N, where trustees are not informed about the average behavior of trustors, we ask them, after their actual decisions, for their belief about the percentage of trustors who send the money. In simple two-sided t-tests and Wilcoxon tests, trustees who stay in the game in treatment N return significantly higher amounts. However, this does not necessarily mean that trustees participate because of their higher social preferences, as their behavior might also be driven by higher expectations about the percentage of trustors who send the money. Indeed, we find that trustees who return less believe that fewer trustors send the money. This effect is so strong that the difference between the amounts returned by those who participate and those who exit becomes insignificant when we control for the trustees' beliefs about the behavior of trustors.

However, controlling for self-stated beliefs is not sufficient, since participants in social dilemma experiments may (partially) rationalize their selfish behavior. It is hence more instructive to look at treatment I, in which we eliminate the impact of different beliefs by informing trustees about the percentage of trustors who send the money. When we regress the exit decision on the amount in treatment I, we find that the amount has no impact at all ($p = 0.994$). Our data hence strongly reject the hypothesis that only trustees with high social preferences stay in the game.

But this does not mean that we do not find a systematic pattern at all: A closer look at the data supports the hypothesis of a U-shaped pattern between the amount trustees return and the participation rate. When we regress the participation decision on the amount and amount squared, then the amount itself is significantly negative, while amount squared is significantly positive, both at least at the 5%-level. This holds for treatments N and I. Our main results are hence threefold: First, if trustees are not informed about the percentage of trustors who send the money, then sorting by trustees has a positive efficiency effect, as those who stay in the game return significantly

higher amounts.[1] Second, this efficiency effect can be attributed to the correlation of the trustees' degree of reciprocity with their belief about how many trustors send the money. As a consequence, there is no efficiency effect when we ensure that all trustees have the same beliefs. Third, regardless of whether we account for this belief or not, sorting leads to a more extreme pool of trustees.

Section 2 relates to the literature. Section 3 presents the model. Section 4 introduces the experimental design and procedure. Results are presented in Sect. 5. In Sect. 6, we discuss our assumptions with respect to the order of moves and the underlying social preferences. We conclude and point to further research in Sect. 7.

2 Relation to the literature

The literature on sorting in social dilemma games can be classified into games where the experimenter assigns people to different groups (exogenous sorting), and games where subjects can self-select (endogenous sorting). With endogenous sorting, subjects might also choose different partners in different stages, or they can only decide whether to take part or not. Our paper is in the tradition of the latter approach and restricted to a one-shot game. Our design hence neither allows for reputation effects nor for the possibility to punish selfish players by changing the partner or exiting the game. Instead, we restrict attention to the impact of the own social preferences and the belief about the partner's behavior on the participation probability, and our paper seems to be the first that separates the impact of the own social preferences from the belief. The results of our treatment *N* are in line with findings that the own social preferences are positively correlated with the beliefs about the social preferences of others (see e.g. Hauk and Nagel, 2001, Keser and Montmarquette, 2011, and the overview in Guido et al, 2019), which suggests that disentangling preferences and beliefs is informative.

Sorting in social dilemmas has most extensively been analyzed in the prisoners' dilemma (PD) and in dictator games. Contrary to their prediction, Orbell et al. (1984) find that defectors are more likely to opt out of a PD than cooperators. Orbell and Dawes (1993) observe a similar result for their comparison of an involuntary and voluntary PD. In Bohnet and Kübler (2005), subjects can bid to play a PD where cooperation is less costly than in the alternatively offered standard PD. They find that sorting triggered by bidding leads to higher cooperation rates than random assignment, but there are also subjects who bid in order to take advantage of cooperators.

For dictator games, Dana et al. (2006) observe a positive but insignificant correlation between amounts and participation. Broberg et al. (2007) let dictators first decide about the allocation and then perform a standard BDM auction to elicit the dictators' reservation prices for exiting the game. Conversely to Dana et al. (2006), they find that more generous dictators demand a lower amount and hence exit more often, but significant only in some specifications, and only at the 10% level. While dictators in Dana et al. (2006) and in Broberg et al. (2007) first decide on their amounts and then about participation, dictators in Lazear et al. (2012) are first asked if they want to exit

[1] Of course, referring to this as an "efficiency effect" requires that this is anticipated by trustors, as efficiency in the trust game, defined as the two players' joint income, depends only on whether trustors send the money or not. We will discuss this in Sect. 6.

or not, and decide on the allocation of the money only when they participate. Lazear et al. (2012) find that the average amount is significantly lower in the sorting treatment, which is driven by a far higher percentage of dictators who give nothing at all. This is in line with our finding that the most selfish trustees have high participation rates. Considering all three papers, it seems appropriate to say that results are inconclusive.

The literature on voluntary contribution games (VCMs) finds slight evidence for a positive effect of sorting (see the meta study by Guido et al. 2019). Nosenzo and Tufano (2017) compare three treatments with groups of two subjects each: The standard VCM, an unconditional treatment where subjects first decide whether to participate and then choose their contributions, and a conditional treatment where they can opt out after they have learned the partner's contribution. They find no difference between the contribution rates in the standard and the unconditional treatment. As expected, contributions are largest in the conditional treatment, as the exit threat serves as a disciplinary device.

The potentially U-shaped pattern for the relation between generosity and participation in our model is due to the assumption that people care about the utility of other participants when they decide about the amounts *and* when they decide about participation. Another string in the literature shows that sorting may depend on whether people draw a positive utility from e.g. altruism, or a negative utility from e.g. social pressure or selfishness. In their field experiment, DellaVigna et al. (2012) distribute flyers informing households that solicitors will come to ask for donations. The altruism model predicts that the information leads to a higher presence of residents at their homes, but the data is more in line with the social pressure model, which predicts the opposite (see DellaVigna et al. 2016 for a related design and result).[2] The design in Lazear et al. (2012) allows identifying three types: "willing sharers" who share a positive amount and stay in the game (and hence draw a positive utility from giving), "reluctant sharers" who share if forced to play but opt out if they receive the same amount (and hence draw a negative utility from selfishness), and selfish "nonsharers". A substantial percentage of people can be characterized as reluctant sharers. Cain et al. (2014) distinguish between people who "give" and who "give in". People who give in share their money if they are in situations where giving nothing can be seen as inappropriate, but prefer to avoid the situation even if doing so leaves the would-be beneficiary empty-handed.[3]

Exit options in trust games have so far only been considered for trustors. The objective of these papers is figuring out if the behavior of trustees is rather driven by other-regarding preferences (altruism and equity concerns) or reciprocity. This is either done by comparing the original trust game to an involuntary treatment where trustors are forced to send the money (and can hence not signal trust), or by comparing dictator games and trust games. Results are somewhat mixed: McCabe et al. (2003), Cox (2004), and Di Bartolomeo and Papa (2015) find evidence for reciprocity as an independent motive, while the results in Cox and Deck (2005) and Brülhart and

[2] They take one step further by estimating the costs of social pressure for the solicited households from a structural model.

[3] See also Klinowski (2021) and the field experiment by Andreoni et al. (2017) showing that people invest to avoid situations in which they are asked for donations.

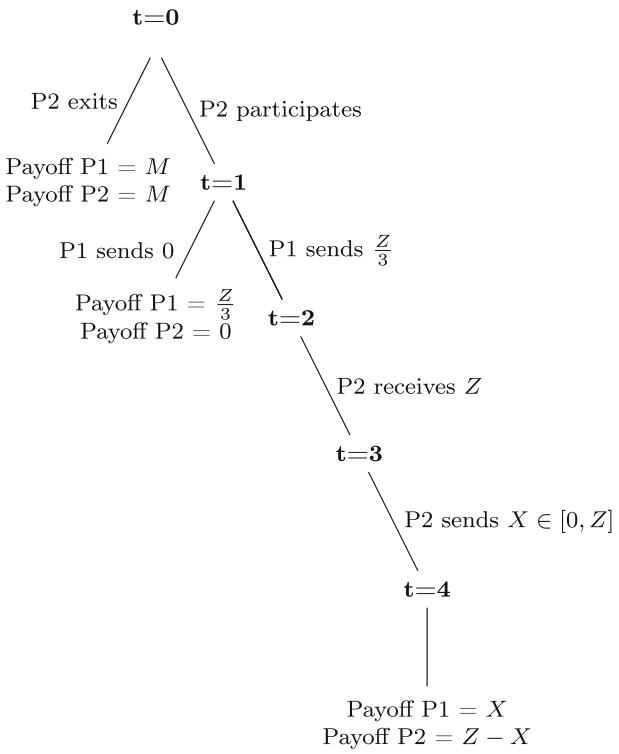


Fig. 1 Time line

Usunier (2012) speak more in favor of altruism. We are not aware of any paper that analyzes sorting of trustees.

3 Theory

3.1 Game structure

We consider a trustor (player 1, P1) and a trustee (player 2, P2). At $t = 0$, P2 decides whether to enter the game or take an outside option. If P2 exits, P1 and P2 both receive an amount $M > 0$ and the game ends. If P2 enters the game, P1 is endowed with an amount $\frac{Z}{3} > 0$ and decides whether to send the full amount $\frac{Z}{3}$ or nothing to P2 at $t = 1$. If P1 does not send the money, P1 keeps their $\frac{Z}{3}$, and P2 receives nothing.

If P1 sends $\frac{Z}{3}$, the amount is tripled, so P2 receives Z at $t = 2$. At $t = 3$, P2 decides on the amount $X \in [0, Z]$ they send back to P1. Payoffs are realized at $t = 4$. We assume $M < \frac{Z}{2}$ to ensure that it is Kaldor-Hicks efficient that P2 enters the game and P1 sends their endowment at $t = 1$. Figure 1 summarizes the timing of events.

3.2 Utilities

For both players, we denote the utility from money as u where $u' > 0$ and $u'' < 0$. To account for reciprocity, we assume that player $i \in \{1, 2\}$ draws positive utility from the other player's payoff iff they perceive their behavior as kind (see below). This utility is denoted r where $r' > 0, r'' < 0$ and $r(0) = 0$. Player i's sensitivity to player j's payoff relative to their own payoff is denoted by θ_i and is private information. The distribution of θ_i is common knowledge and $f(\theta_i)$ ($F(\theta_i)$) denotes the pdf (cdf) with full support on $[0, \infty)$.

Our assumptions on P2's reciprocity are as follows[4]: If P1 sends the money, then their payoff enters with full weight, because P2 can only earn money when P1 sends it. If P2 does not enter the game, then they do not know if P1 would have sent the money or not. We then assume that P1's payoff is weighted with the equilibrium probability p_1 that they would have sent the money. P2's utility is thus:

$$U_2 = \begin{cases} u(Z - X^*) + \theta_2 r(X^*) & \text{if P2 participates and P1 sends the money} \\ 0 & \text{if P2 participates and P1 does not send the money} \\ u(M) + p_1 \theta_2 r(M) & \text{if P2 does not participate} \end{cases} \tag{1}$$

For P1, we assume that they care about P2's payoff iff $X \geq \frac{Z}{3}$, that is, if P2 participates and sends an amount at least as high as P1's endowment. P1's utility is[5]:

$$U_1 = \begin{cases} u(X) + \theta_1 r(Z - X) & \text{if P2 participates, P1 sends the money, and } X \geq \frac{Z}{3} \\ u(X) & \text{if P2 participates, P1 sends the money, and } X < \frac{Z}{3} \\ u(\frac{Z}{3}) & \text{if P2 participates but P1 does not send the money} \\ u(M) & \text{if P2 does not participate} \end{cases} \tag{2}$$

3.3 P2's decision about the amount

As solution concept for our dynamic game with private information we apply the Perfect Bayesian Equilibrium. Solving by backward induction, we start with P2's decision at $t = 3$ given that P2 participates and P1 sent the money. P2's utility is

$$U_2(X) = u(Z - X) + \theta_2 r(X). \tag{3}$$

We get

Lemma 1 *The utility-maximizing amount P2 sends back and their utility in case of participation increase with θ_2. Therefore, there exist θ_2^+ and θ_2^{++} such that $X > 0$ iff $\theta_2 > \theta_2^+$ and $X > \frac{Z}{3}$ iff $\theta_2 > \theta_2^{++}$.*

[4] Our specific assumptions on the two players' social preferences are without loss of generality (see Sect. 6 for a discussion of alternative assumptions)

[5] Similar to what we assume for P2's utility, P1 may well care about P2's payoff when P2 participates, but P1 does not send the money. But as P2's payoff is zero in this case, it does not matter.

Proof All proofs are in the Appendix. □

With neoclassical standard preferences (i.e. with $\theta_2 = 0$), P2 returns nothing at $t = 3$. The important part in Lemma 1 is that, while P2's with high social preferences end up with less money, they still have higher utility. To see this, just consider two types of P2, a L -type with low social preferences θ_L who maximizes their utility by sending back X_L, and an H-type with high social preferences who maximizes their utility by sending back $X_H > X_L$. Now suppose that player H sub-optimally sends X_L. Even then, they would have higher utility, as the amount for themselves $Z - X_L$ is the same as for player L, while $\theta_H r(X_L) > \theta_L r(X_L)$. From optimality, it follows that player H's utility is even larger when they send X_H.

3.4 P1's decision to send the money

If P2 has participated, P1 will send the money at $t = 1$ iff their expected utility from the money sent back by P2 (the first addend on the LHS in Inequality (4)) plus their expected utility from reciprocity (the second addend) weakly exceeds their utility from keeping their endowment[6]

$$\begin{aligned}&\int_{\theta^+}^{\infty} u(X(\theta_2))\frac{f(\theta_2)}{1-F(\theta_2^+)}d\theta_2 + \int_{\theta^{++}}^{\infty} \theta_1 r(Z-X(\theta_2))\frac{f(\theta_2)}{1-F(\theta_2^{++})}d\theta_2 \\ &= \mathbb{E}[u(X(\theta_2))|\theta_2 > \theta_2^+] + \theta_1 \mathbb{E}[r(Z-X(\theta_2))|\theta_2 > \theta_2^{++}] \geq u\left(\frac{Z}{3}\right).\end{aligned} \tag{4}$$

We get

Proposition 1 *Suppose* $\mathbb{E}[u(X(\theta_2))|\theta_2 > \theta_2^+] < u(\frac{Z}{3})$. *Then, for any distribution of* θ_2, *there exists a unique* $\widetilde{\theta}_1$ *such that P1 prefers to send (not send) the money iff* $\theta_1 \geq \widetilde{\theta}_1$ $(\theta_1 < \widetilde{\theta}_1)$. $\widetilde{\theta}_1$ *is implicitly defined by* $\mathbb{E}[u(X(\theta_2))|\theta_2 > \theta_2^+] + \theta_1 \mathbb{E}[r(Z-X(\theta_2))|\theta_2 > \theta_2^{++}] = u\left(\frac{Z}{3}\right)$.

For $\mathbb{E}[u(X(\theta_2))|\theta_2 > \theta^+] \geq u(\frac{Z}{3})$, P1 sends the money even without reciprocity $(\theta_1 = 0)$ because their expected utility from their own payoff already (weakly) exceeds the utility from keeping their endowment. For $\mathbb{E}[u(X(\theta_2))|\theta_2 > \theta^+] < u(\frac{Z}{3})$, the incentive to send the money strictly increases with P1's reciprocity parameter θ_1, so $\widetilde{\theta}_1$ is unique. This implies that no type has a profitable deviation by sending the money for $\theta_1 < \widetilde{\theta}_1$, or by keeping the money for $\theta_1 \geq \widetilde{\theta}_1$. Furthermore, $\widetilde{\theta}_1$ decreases if more probability mass is shifted to the right of θ_2's distribution.

[6] In our experiment, P1 does not know if P2 participates. Defining p_2 as the probability that P2 participates, P1 then sends the money iff

$$p_2\left(\mathbb{E}[u(X(\theta_2))|\theta_2 > \theta^+] + \theta_1 \mathbb{E}[r(Z-X(\theta_2))|\theta_2 > \theta^{++}]\right) + (1-p_2)u(M) \geq p_2 u\left(\frac{Z}{3}\right) + (1-p_2)u(M).$$

Again, P1, sends the money iff Inequality 4 holds. The reason why it does not matter if P1 knows whether P2 participates is that their own decision is irrelevant when P2 exits.

3.5 P2's participation decision

P2 enters the game at $t = 0$ iff their expected utility weakly exceeds their utility from the outside option, i.e. iff

$$p_1[u(Z - X^*) + \theta_2 r(X^*)] + (1 - p_1)u(0) \geq u(M) + p_1\theta_2 r(M), \tag{5}$$

where $p_1 = \Pr[\theta_1 \geq \widetilde{\theta}_1] = 1 - F(\widetilde{\theta}_1)$, and where X^* denotes P2's utility-maximizing amount returned in stage 3.

Inequality (5) shows that P2's social preference parameter θ_2 enters both on the LHS and on the RHS, as P1's payoff matters for P2 also when P1 exits the game (though only weighted with p_1). Therefore, the sorting decision is not necessarily monotonic in the reciprocity parameter.[7] Nevertheless, for any θ_2, there exists a unique threshold $\widehat{p}_1$ such that type θ_2 participates iff $p \geq \widehat{p}_1$. For $p_1 = 0$ ($p_1 = 1$) all types of P2 exit (participate). We get Lemma 2.

Lemma 2 *For each type θ_2, there exists a unique $\widehat{p}_1(\theta_2) \in (0, 1)$ such that type θ_2 participates (exits) iff $p_1 \geq \widehat{p}_1(\theta_2)$ ($p < \widehat{p}_1(\theta_2)$). $\widehat{p}_1(\theta_2)$ is inversely U-shaped and has a maximum at some unique $\widehat{\theta}_2$. Furthermore, there exists a unique $\overline{p}_1$ such that for all $p_1 \geq \overline{p}_1$, all types participate.*

The main insight from Lemma 2 is that the probability $\widehat{p}_1(\theta_2)$ such that type θ_2 is indifferent between participation and exit is non-monotonic in θ_2, and inversely U-shaped with a unique maximum at some $\widehat{\theta}_2$. For $\theta_2 < \widehat{\theta}_2$, the critical minimum probability that is required for entry increases with θ_2 (i.e. $\frac{\partial \widehat{p}_1}{\partial \theta_2} > 0$), whereas it decreases in θ_2 for $\theta_2 > \widehat{\theta}_2$. $\overline{p}_1$ is defined as the maximum p_1 for which a type can be indifferent, i.e. $\overline{p}_1 := \widehat{p}_1(\widehat{\theta}_2)$. For $p_1 \geq \overline{p}_1$, all types participate. The global minimum probability is the $\widehat{p}_1$ at the highest possible type and converges to 0, i.e. for $p_1 = 0$, all types prefer to exit.

For $p_1 \in (0, \overline{p}_1)$, the inversely U-shaped pattern implies existence of a region in which there exist two types θ_L and θ_H with $\widehat{p}_1(\theta_L) = \widehat{p}_1(\theta_H)$. This region is defined by $p_1 \in [\underline{p}_1, \overline{p}_1)$, where $\underline{p}_1 := \widehat{p}_1(0) = \frac{u(M)}{u(Z)}$. For $p_1 < \underline{p}_1$, there exists a unique type who is indifferent. We summarize:

Proposition 2 *For $p_1 \in [\underline{p}_1, \overline{p}_1)$, there exist two unique types $\underline{\theta}$ and $\overline{\theta} > \underline{\theta}$ such that all types $\theta_2 \leq \underline{\theta}$ and $\theta_2 \geq \overline{\theta}$ prefer to participate and all types $\theta_2 \in (\underline{\theta}, \overline{\theta})$ prefer to exit (U-shaped pattern). For $p_1 \in (0, \underline{p}_1)$, there exists only $\overline{\theta}$, i.e. all types $\theta_2 \geq \overline{\theta}$ ($\theta_2 < \overline{\theta}$) participate (exit).*

For $p_1 \in [\underline{p}_1, \overline{p}_1)$, types with low ($\theta_2 \leq \underline{\theta}$) and high ($\theta_2 \geq \overline{\theta}$) reciprocity participate, while types in-between exit. If the probability p_1 that P1 sends the money is sufficiently high, low reciprocity types benefit from participation by taking advantage of P1. High reciprocity types also participate, as they gain a high utility from returning money. For intermediate types, the outside option is more attractive, as they gain a balanced utility from the own and P1's payoff.

[7] If P1's payoff mattered only in case they send the money, then Lemma 1 would imply that P2's incentive to participate increases monotonically with θ_2.

For $p_1 \in (0, \underline{p}_1)$, however, there is only one indifferent type, and hence just one switching point with regards to the reciprocity parameter θ_2. In this case, only types with high social preferences participate.

The reason is that, if the participation rate of P1 is low, then the *expected* (reciprocal) benefit from returning money to P1 is lower for low θ_2-types (who return only small amounts) compared to the (reciprocal) benefit from P1's exit option. On the other hand, for high θ_2-types (who return relatively large amounts), the (reciprocal) benefit from returning money to P1 can be larger then the utility from the own payoff in case of exit. Therefore, types θ_2 who put low (high) weight on P1's payoff exit (participate). Summing up, there are two possible constellations – either only highly reciprocal trustees participate, or trustees with low *and* with high reciprocity participate, while trustees in-between exit.

3.6 Equilibrium

The equilibrium is characterized by the conditions in Propositions 1 and 2. The threshold for P1's reciprocity parameter such that they send the money iff $\theta_1 \geq \widetilde{\theta}_1$ depends on the conditional distribution of the reciprocity parameter for those P2 who enter the game. For any such distribution given, the threshold $\widetilde{\theta}_1$ for the indifferent type is uniquely defined by the condition stated in Proposition 1, and no type has an incentive to deviate.

For P2, all that matters is the probability that P1 enters the game, and hence the initial type distribution of the reciprocity parameter θ_1.[8] For any distribution of θ_1 given, the types θ_2 who enter the game are characterized by Proposition 2. The amount P2 sends back is as characterized in Lemma 1. By definition of the thresholds in Propositions 1 and 2, neither type has a profitable deviation. We get

Proposition 3 *There exists a Perfect Bayesian Equilibrium in which P1's decision to send the money is as characterized in Proposition 1, and P2's decision to enter is as characterized in Proposition 2.*

In our model, we assumed that the only difference to neoclassical standard preferences is reciprocity, that is, players put positive weight on the other player's payoff if and only if their behavior is perceived as kind. While our assumptions seem reasonable, they are ad hoc in the sense that they are not derived from underlying utility functions. In Sect. 6, we therefore show that our Propositions remain qualitatively the same when we assume other kinds of social preferences. First, we substitute reciprocity by the simplest kind of other-regarding preferences where players always put positive weight on the other player's payoff. Second, we combine reciprocity with Fehr-Schmidt preferences (inequity aversion).[9]

[8] The conditional distribution of θ_1 for those P1 who enter does not matter.

[9] See Cox et al. (2007) for a more general approach on reciprocity and fairness that derives the weights on other players' payoffs endogenously from the underlying utility functions.

4 Experimental design and procedure

Subjects were randomly paired in groups of two and assigned to the role of a trustor or a trustee. In addition to their fixed payment of \$0.50 for their participation, trustors received \$0.50. Trustors were informed that they could either keep this amount or give it to the subject they were paired with. They were told that the game ends if they keep the money, and that the overall amount of \$1 would be transferred to their account. They learned that the amount of \$0.50 was tripled to \$1.50 if they send it. We then informed them that the trustee could send any part of the \$1.50 back. The trustor's bonus was then explained with numerical examples.

So far, our setting is just a standard trust game. Then, however, trustors were informed that the trustee they are matched with has two options: They can either play the game as just described or get a certain bonus of \$0.50 instead. If the trustee chooses the certain bonus of \$0.50, then the game ends and the trustor just keeps their additional payment of \$0.50 as a bonus; regardless of whether they decided to send the money or not. Finally, we asked comprehension questions to ensure that participants understood everything accurately.[10]

The instructions for trustees were similar to those for trustors. Trustees were not informed about whether trustors actually sent the money before they decided about the amounts they wanted to send back. Instead, we asked trustees: "If this game is played and the other worker in your group decides to give you the money, how much do you want to give back?" Only after they had made their decision, we informed trustees about their opt-out possibility. We phrased this as follows: "However, you have two options: You can either play the game as described before or get a certain bonus of \$0.50 instead. If you choose the certain bonus of \$0.50, then the game ends and the other worker just keeps their additional amount of \$0.50 as a bonus. If you decide to play the game, we will take your previous decision on how much money to send back in case the other worker has given the money to you. Recall that your bonus is \$0 if the other worker does not send you the money." This order of information allows us to elicit the amounts for all trustees, regardless of whether they eventually participate or not.

While all trustors received the same instructions, trustees were assigned to two different treatments. The first group was assigned to the *no-information treatment* N and got no additional information. In this treatment, we asked trustees, after they had made their decisions, for their belief on the percentage of trustors who would send the money. We did so because the exit decision is likely to depend not only on social preferences, but also on a trustee's assumption on how many trustors participate. Asking for the belief, however, does not fully account for this, as trustees may rationalize their decision ex post.[11] We therefore assigned the second group of trustees to the *information treatment* I, in which we informed them about the percentage of trustors who sent the money in treatment N. We provided this information after the trustees' decision on the amount they want to send back, but before their exit decision.

[10] See Sect. 6 for a discussion of the differences between our model and the experiment.

[11] We could not have avoided this endogeneity problem by asking first for the belief, as this would have contaminated our main variable of interest.

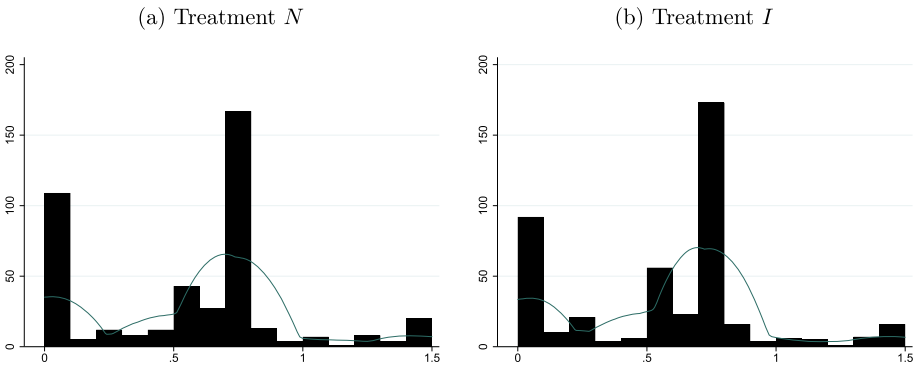


Fig. 2 Distribution of amounts. *Notes:* Amounts are displayed on the horizontal axes, numbers of observations on the vertical axes. Both histograms use a bin width of \$0.10. The lines display kernel density estimates (Epanechnikov)

Treatment I hence takes care of the problem that social preferences may initially be correlated with the beliefs about the behavior of trustors.

After all decisions had been made, trustors and trustees were asked for their gender, age, and attitude towards risk on a scale from 0 (not willing to take risks at all) to 10 (highly willing to take risks). We pre-registered the experiment on November 30, 2020 at AEA RCT Registry (AEARCTR-0006836). The experiment was carried out on Amazon MTurk between December 2020 and February 2021. 880 subjects participated as trustors and 880 as trustees; 440 each in the no-information and the information treatment. On average, subjects in the role of trustors earned \$1.03. Trustees earned an average \$1.06. On average, it took participants a bit more than four minutes to complete the experiment which translates into an hourly compensation of more than \$15.

5 Results

Recall from the design section that we asked trustees about the amount they return if trustors send the money. We hence have data for all 440 trustees in each of the two treatments; the *information treatment I* in which trustees are informed about the percentage of trustors who sent the money, and the *no-information treatment N* in which they did not get this information. 42% of trustees exit in *Treatment N* compared to only 35% in *Treatment I*. The difference in the exit frequencies is significant with a p-value of 0.051 in a χ^2-test. As the incentive to exit decreases in the belief about the percentage of trustors who send the money, this suggests that trustees, on average, underestimated the trustors' level of trust. In both treatments, slightly less than 25% returned nothing. Out of those fully selfish trustees, 43% exit in treatment *N* compared to 31% in treatment *I*.

Figures 2 and 3 display two kinds of information for each of the two treatments. The diagrams on the left (right) refer to the no-information (the information) treatment.

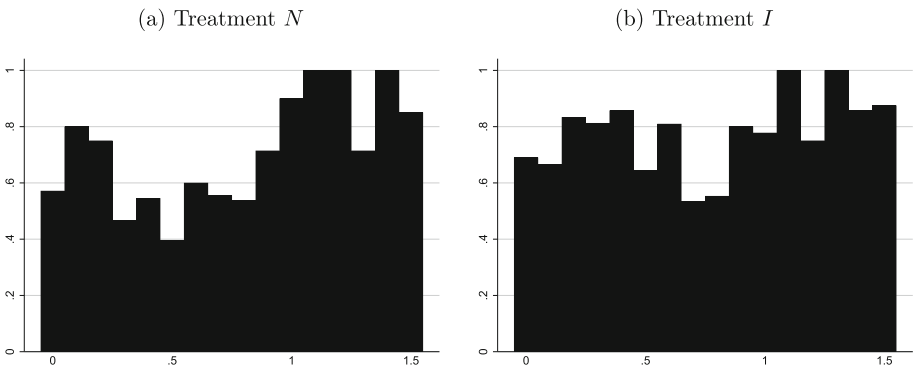


Fig. 3 Share of participation for given amount. *Notes:* Amounts are displayed on the horizontal axes, the share of trustees who stay in the game on the vertical axes

Figure 2 shows the number of trustees who decided to return the amounts shown on the horizontal axis before they were informed about their exit option. For both treatments, there are two peaks, one at zero (with 109 and 92 trustees returning less than \$0.10 in treatment N and I, respectively) and one around \$0.75 (with 167 and 173 trustees, respectively), which leads to an equal split of the money between trustors and trustees. As all numbers above \$0.75 are negligibly low, our data display a U-shaped pattern for both treatments: most trustees are either fully selfish (peak below \$0.10) or rather generous (peak at \$0.75). The lower peak at around \$0.50 suggests that there are also quite some trustees who feel that they should leave trustors with the amount they would have received when they had not sent the money in the first place.

Figure 3 shows the percentages of trustees who stayed in the game for the amounts they wanted to return. The exact picture differs somehow between the two treatments, but there is a U-shaped pattern for both of them.

Before we get back to these U-shaped patterns and turn to our main result, we first consider whether the amounts trustees want to return differ, on average, between those who participated and those who exited. Without control variables, trustees who stay in the game return more in treatment N, significant with $p = 0.045$ in a two-sided t-test and with $p = 0.096$ in a Wilcoxon test. Conversely, the amount and the participation decision are not related in treatment I ($p = 0.812$ in a two-sided t-test, and $p = 0.608$ in a Wilcoxon test).

Table 1 presents results, separated by the two treatments, from a linear probability model on participation.[12] The dependent variable is a dummy that takes the value "1" if a trustee participated. In Treatment N, the amount is significant at the 5%-level without controls, and also when we control only for gender and age (models N1 and N2). As expected, those who are less risk averse stay in the game more often because exiting the game yields a certain payoff (model N3). Adding the degree of risk aversion renders the amount insignificant because those with higher risk aversion return lower

[12] Results are qualitatively the same for probit or logit models.

Table 1 Linear regression results on participation

	Treatment N				Treatment I	
	N1	N2	N3	N4	I1	I2
Amount	0.118**	0.117**	0.064	0.030	−0.014	0.000
	(0.033)	(0.036)	(0.233)	(0.581)	(0.801)	(0.994)
Male		−0.000	−0.054	−0.052		−0.126***
		(0.999)	(0.242)	(0.248)		(0.004)
Age		0.000	0.002	0.001		−0.001
		(0.830)	(0.399)	(0.455)		(0.407)
Risk			0.067***	0.063***		0.067***
			(0.000)	(0.000)		(0.000)
Belief				−0.004***		
				(0.000)		
Observations	440	440	440	440	440	440

Columns (1) to (4) consider treatment *N* and columns (5) and (6) consider treatment *I*. *Amount* is the money trustees send back to trustors. *Male* is a dummy variable that takes the value 1 if trustees selected male as gender and 0 otherwise. *Age* is the trustee's age. *Risk* is the degree of risk attitude on a 10 point scale (from highly risk-averse to highly risk-seeking). *Belief* is the estimation of the percentage of trustors who keep the money for themselves. p-values are reported in parentheses, and *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively (robust standard errors)

amounts *and* exit more often.[13] A similar argument holds when we finally add the belief in model N4.[14] The regression analysis hence reveals that the overall benefit from sorting from the trustors' point of view can be attributed to the fact that the degree of risk aversion and the belief on the percentage of trustors who send the money are both correlated with the amount *and* with the participation decision.

Controlling for beliefs by a simple question runs the risk that subjects might rationalize their own behavior. It is hence more reliable to consider treatment *I* in the next two columns, which reinforces the result that sorting has no impact on the average amount that trustees return when we eliminate the impact of beliefs by informing trustees about the percentage or trustors who send the money (see the p-value of 0.994 for the amount in regression I2). Attitude towards risk is still highly significant in the expected direction, and males exit more often.

We now get back to Fig. 3, which suggests that the relation between the amount and the participation rate is U-shaped. The only difference between the linear regression in Table 2 to the one in Table 1 is that we now add amount squared to account for the non-linearity. Again, the first four (last two) columns refer to treatment *N* (treatment *I*). The results confirm the U-shaped relation displayed in figure 3: In all specifications, the amount itself is significantly negative and amount squared significantly positive. This holds for treatments *N* and *I*. We use the Akaike information criterion (AIC) to

[13] The correlation between the degree of risk aversion and the participation probability is 0.35, and the correlation between the degree of risk aversion and the amount is 0.12 in treatment *N*.

[14] The correlation between the belief on the percentage of trustors who keep the money for themselves and the participation probability is −0.21, and the correlation between this belief and the amount is −0.18 in treatment *N*.

Table 2 Linear regression results on participation including squared amounts

	Treatment N				Treatment I	
	N1	N2	N3	N4	I1	I2
Amount	−0.296**	−0.311**	−0.231*	−0.286**	−0.393***	−0.289**
	(0.035)	(0.030)	(0.086)	(0.033)	(0.006)	(0.036)
Amount^2	0.352***	0.364***	0.252***	0.270***	0.327***	0.249**
	(0.000)	(0.000)	(0.008)	(0.004)	(0.001)	(0.011)
Male		−0.023	−0.068	−0.067		−0.132***
		(0.640)	(0.142)	(0.138)		(0.003)
Age		0.001	0.002	0.002		−0.001
		(0.744)	(0.369)	(0.419)		(0.520)
Risk			0.064***	0.060***		0.065***
			(0.000)	(0.000)		(0.000)
Belief				−0.004***		
				(0.000)		
Observations	440	440	440	440	440	440

Columns (1) to (4) consider treatment *N* and columns (5) and (6) consider treatment *I*. The description of variables is the same as in Table 1. p-values are reported in parentheses, and *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively (robust standard errors)

compare the goodness of fit of the linear and the quadratic model. For this measure, lower values translate into a higher quality of the model in terms of goodness of fit and simplicity. The AIC for regressions N4 (I2) in Tables 1 and 2 are 567 (545) and 563 (541), respectively, indicating that the quadratic model is a much better fit than the linear one. Our main result is hence that sorting by trustees makes the remaining population more extreme: Trustors will be either matched with particularly selfish or generous trustees.

6 Discussion

In this section, we discuss two elements of our analysis that deserve some attention. The first concerns alternative assumptions on social preferences, and the second the order of moves in the model and the experiment.

Alternative assumptions on social preferences At the end of the theory section, we mentioned that our Propositions are qualitatively the same when we assume other kinds of social preferences (instead or in addition to reciprocity). Consider first the simplest form of other-regarding preferences where players put *always* positive weight on the other player's payoff. Note first that, regardless of the probability p_1 with which player 1 sends the money, this would not change the amount P2 sends back, as their utility when deciding on the amount is still the same.

Furthermore, P1's expected utility now depends on the unconditional instead of the conditional expectation of P2's payoff, but that does not change our result in

Proposition 1. It only leads to a higher likelihood of P1 sending the money, as a P1 who is indifferent with reciprocal preferences strictly prefers to send the money with other-regarding preferences.

For P2's participation decision, a few things change. Even for $p_1 = 0$, high types θ_2 will prefer participation iff $\frac{Z}{3} > M$, i.e. iff the payoff to P1 is higher than their payoff when P2 exits. For $p_1 = 1$, all types still prefer participation. The incentive to participate still increases in p_1, and a sufficient condition for the existence of a unique $\widehat{p_1}(\theta_2)$ is that $\frac{Z}{3} \leq M$ (as in our experiment). In this case, $\widehat{p_1}(\theta_2)$ is still inversely U-shaped. The main difference to our main model is that there now exists a third equilibrium, in which only low types participate.[15]

Assume next inequity aversion à la Fehr and Schmidt (1999) in addition to reciprocity, and consider the terms that change in this case. Players face a disutility from inequity, with weight α ($\beta \leq \alpha$) in case they (the other player) have a lower payoff. Furthermore, we restrict attention to the case where P2 returns weakly less than 50% (i.e. $X \leq \frac{Z}{2}$), which holds for 80% of our subjects in the experiment. Then, P1's expected utility in case they send the money decreases by $p_2\alpha \int_0^\infty (Z - X(\theta_2)) - X(\theta_2) d\theta_2 = p_2\alpha \int_0^\infty Z - 2X(\theta_2) d\theta_2$. If P1 does not send the money, their utility decreases by $p_2\beta\frac{Z}{3}$, as P2 gets nothing with probability p_2, while P1 keeps $\frac{Z}{3}$. Which of these effects dominates hence depends on the expectation on the amount P2 sends back, and on the difference between the inequity aversion parameters α and β. The overall effect on P1's decision can go either way.

The direct reduction in P2's utility from inequity when they participate is $p_1\beta\,(Z - 2X^*) + (1 - p_1)\,\alpha\frac{Z}{3} > 0$, while there is no impact of inequity aversion when P2 exits (both players then get the same payoff). However, this does not necessarily mean that P2 participates less often than without inequity aversion: We know that P2's incentive to participate increases with p_1, and inequity aversion enhances P1's incentive to send the money if $\alpha \int_0^\infty Z - 2X(\theta_2) d\theta_2 \geq \beta\frac{Z}{3}$. Then, the higher probability that P1 sends the money may well overcompensate the direct effect of inequity aversion. For $p = 0$ ($p = 1$), all types prefer exit (participation). Again, there exist a unique $\widehat{p_1}(\theta_2)$ for which P2 is indifferent and which exhibits an inversely U-shaped pattern. Our Propositions hence still hold.

Relation between model and experiment In the model, trustees (P2s) first decide on participation and then on their amount, while this order is reversed in the experiment. The reasons are as follows: If P2s in the model first decide on their amount, then those who anticipate to exit would be indifferent among all amounts that trigger exit. While this makes the formal analysis more tedious, subjects anticipating to stay in the game still choose their utility-maximizing amount. In the experiment, we need to ensure that all P2s have incentives to report their amount truthfully, which requires to ask them first for their amounts, and to inform them only then about their exit option.

Related, P2 in our experiment does not know whether P1 sends the money when they decide upon the amount they send back. P2's expected utility is then $U_2(X) =$

[15] Note that this equilibrium exists iff $\frac{u(M)}{u(Z)} < \frac{r(M)-r(\frac{Z}{3})}{r(Z)-r(\frac{Z}{3})}$. This is excluded in our experiment because $\frac{Z}{3} = M$.

$p_1 [u(Z - X) + \theta_2 r(X)]$, which, however, yields the same first order condition for the utility-maximizing amount.

In the model, we assume that trustors (P1) observe whether P2 participates or not. In the experiment, we ask them about their amounts under the assumption that P2 participates. This, however, does not make a difference, as P1's decision does not matter anyway if P2 exits the game.

For the reasons mentioned above, we ask P2s in our experiment first upon their amounts and inform them only then about their exit option. By contrast, we have told P1s "If they (trustees) decide to play the game, we will ask them how much money they will send back in case you have given the money to them", because the reversed order seems to be (far) more difficult to understand for P1s. Note that, as discussed above, this does not make any difference for P1s, because their decision is independent of the probability that P2 enters the game.

7 Conclusion

We contribute to sorting in social dilemma games by offering trustees an exit option after they decided about the amount they send back if the trustor sends the money. The key assumption of our behavioral model is that the trustor's payoff enters the trustee's utility when they decide on the amount they send back, and also when they decide upon participation. The model yields two possible constellations: Only trustees with high reciprocity enter, or there is a U-shaped pattern between reciprocity and participation. Our data with 880 trustees support the hypothesis of a U-shaped relation. Sorting thus leads to a more extreme pool that consists of mostly rather selfish or generous trustees. Our experimental finding that the impact of social preferences on the incentive to take part in social dilemma games is non-monotonic seems relevant from an applied perspective (in particular, when the consequences of reciprocal behavior are non-linear). From a conceptual point of view, our result may help to explain why the results on the impact of sorting on cooperation are mixed, and why null results are no exception.

One of the main questions analyzed in the literature is whether sorting increases efficiency, for instance due to the possibility to punish selfish people by changing partners. In our one-shot game, efficiency depends only on how sorting influences the pool of remaining trustees, and whether this is anticipated by trustors. As to the first point, our main results are threefold: (i) the amounts trustees' return are higher with sorting, (ii) this effect, however, disappears when trustees are informed about the percentage of trustors that send the money, and (iii) sorting leads to a more extreme pool of trustees. As to the second point, our design does not allow us to shed light on whether trustors anticipate the impact of sorting on the remaining pool of trustees. This would have required two treatments, one where trustees can opt out and one where they cannot, and to compare the frequencies of trustors who send the money between these two treatments. As efficiency in trust games ultimately depends only on the behavior of trustors, this deserves further research.

A second straightforward question for further research is whether the U-shaped pattern found for our setting extends to other social dilemma games like PDs, dictator

games, and VCMs. If so, this would imply, from an applied perspective, that allowing people to self-select into social dilemma situations is superior to random assignment if and only if having mainly extreme types is better compared to the full population.

Funding Open access funding provided by University of Innsbruck and Medical University of Innsbruck.

Appendix

A Proofs

A.1 Proof of Lemma 1

The utility maximizing amount X^* at $t = 3$ is implicitly given by

$$-u'(Z - X^*) + \theta r'(X^*) = 0, \tag{6}$$

where $X^* = 0$ for $\theta_2 = 0$. From the implicit function theorem, we get

$$\frac{\partial X^*}{\partial \theta_2} = -\frac{r'(X^*)}{u''(Z - X^*) + \theta_2 r''(X^*)} > 0, \tag{7}$$

i.e. the amount increases with the social preference parameter θ_2. θ^+ (θ^{++}) defines the threshold type that sends back a positive amount (an amount that exceeds P1's initial endowment). There exists a type $\theta^{\max}$ who maximizes their utility by returning the full amount to P1, i.e. $X^* = Z$ for $\theta_2 \geq \theta^{\max}$.

The maximum utility and its derivative w.r.t. θ_2 are

$$\begin{aligned} U(X^*) &= u(Z - X^*) + \theta_2 r(X^*) \\ \frac{\partial U(X^*)}{\partial \theta_2} &= r(X^*) + \frac{\partial X^*}{\partial \theta_2}(\theta_2 r'(X^*) - u'(Z - X^*)) = r(X^*) > 0, \end{aligned}$$

with $\theta_2 r'(X^*) - u'(Z - X^*) = 0$ as FOC for the optimal amount. □

A.2 Proof of Proposition 1

P1's incentive to send the money strictly increases in their type because the derivative of the LHS of inequality (4) w.r.t. θ_1 is $\mathbb{E}[r(Z - X(\theta_2))|\theta_2 > \theta^{++}] > 0$. For $\theta_1 = 0$, it

only matters whether $\mathbb{E}[u(X(\theta_2))|\theta_2 > \theta^+] \geq u(\frac{Z}{3})$ because all θ_1-types then prefer to send the money ($\widetilde{\theta_1} = 0$). For $\theta_1 \to \infty$, P1 sends the money. If P1 expects $X = Z$ (i.e. when the utility from reciprocity disappears), they will still send the money due to the higher own monetary benefit.

For $\mathbb{E}[u(X(\theta_2))|\theta_2 > \theta^+] < u(\frac{Z}{3})$, there exists a unique $\widetilde{\theta_1} > 0$ such that P1 is indifferent:

$$\widetilde{\theta_1} = \frac{u\left(\frac{Z}{3}\right) - \int_{\theta^+}^{\infty} u(X(\theta_2)) \frac{f(\theta_2)}{1-F(\theta^+)} d\theta_2}{\int_{\theta^{++}}^{\infty} r(Z - X(\theta_2)) \frac{f(\theta_2)}{1-F(\theta^{++})} d\theta_2}. \tag{8}$$

$\widetilde{\theta_1}$ strictly decreases with $\mathbb{E}[u(X(\theta_2))|\theta_2 > \theta^+]$ and $\mathbb{E}[r(Z - X(\theta_2))|\theta_2 > \theta^{++}]$. □

A.3 Proof of Lemma 2

For $p_1 = 0$, the LHS of inequality (5) is 0, while the RHS is strictly positive. Hence, P2 exits if they know that P1 does not send the money. To see that the LHS is strictly larger than the RHS for $p_1 = 1$, assume that P2 could (sub-optimally) choose $X^* = \frac{Z}{2}$. This yields a higher monetary utility and a higher utility from reciprocity compared to the outside option. Hence, P2 participates if they know for sure that P1 sends the money.

The LHS and the RHS both increase with p_1, but the LHS at a larger rate since $u(Z - X^*) + \theta_2 r(X^*) > u(M) + \theta_2 r(M) > \theta_2 r(M)$. Therefore, there exists a unique $\widehat{p_1}(\theta_2) \in (0, 1)$ that makes P2 indifferent.

The probability for which a type is indifferent $\widehat{p_1}(\theta_2)$ and its derivative are:

$$\widehat{p_1}(\theta_2) = \frac{u(M)}{u(Z - X^*) + \theta_2 r(X^*) - \theta_2 r(M)}$$
$$\frac{\partial \widehat{p_1}(\theta_2)}{\partial \theta_2} = \frac{[r(M) - r(X^*)]u(M)}{[u(Z - X^*) + \theta_2 r(X^*) - \theta_2 r(M)]^2}$$

The numerator determines the sign of the derivative. It is positive for $X^* < M$. Therefore, there exists a $\widehat{\theta_2}$, where $\widehat{p_1}$ increases (decreases) with θ_2 for $\theta_2 < \widehat{\theta_2}$ ($\theta_2 > \widehat{\theta_2}$). $\widehat{\theta_2}$ is implicitly defined by $X^*(\widehat{\theta_2}) = M$.

Note that $\lim_{\theta_2 \to \infty} \widehat{p_1}(\theta_2) = \lim_{\theta_2 \to \infty} \frac{u(M)}{\theta_2 r(Z) - \theta_2 r(M)} = 0$, and $\lim_{\theta_2 \to 0} \widehat{p_1}(\theta_2) = \frac{u(M)}{u(Z)} =: \underline{p_1}$. Hence, for $p_1 \in (0, \underline{p_1})$ ($p_1 \in [\underline{p_1}, \overline{p_1})$) there exists exactly one (two) indifferent types. For $p_1 \geq \widehat{p_1}(\widehat{\theta_2}) =: \overline{p_1}$ any type θ_2 prefers to participate. □

A.4 Proof of Proposition 2

From the proof of Lemma 2, we know that, for $p_1 \in (0, \overline{p_1}]$ given, there exists at least one θ_2 such that

$$F(p_1, \theta_2) := \widehat{p_1}[u\left(Z - X^*\right) + \theta_2 r\left(X^*\right)] - u(M) - \widehat{p_1}\theta_2 r(M) = 0. \tag{9}$$

All types with $F(p_1, \theta_2) > 0$ ($F(p_1, \theta_2) < 0$) strictly prefer to participate (to exit).

The derivative of $F(p_1, \theta_2)$ w.r.t. θ_2 is

$$\frac{\partial F}{\partial \theta_2} = \widehat{p_1}[r(X^*) - r(M)]. \tag{10}$$

To determine the sign of $\frac{\partial F}{\partial \theta_2}$, we need to distinguish between two cases:

1. $p_1 \in [\underline{p_1}, \overline{p_1})$:
   From the proof of Lemma 2 it follows that there exist two unique types $\underline{\theta_2} < \widehat{\theta_2} < \overline{\theta_2}$ such that $\widehat{p_1}(\underline{\theta_2}) = \widehat{p_1}(\overline{\theta_2})$, i.e. $F(p_1, \underline{\theta_2}) = F(p_1, \overline{\theta_2}) = 0$. At $\underline{\theta_2}$, $X^* < M$, hence $\frac{\partial F}{\partial \theta_2} < 0$. Types $\theta_2 \leq \underline{\theta_2}$ hence participate. At $\overline{\theta_2}$, $X^* > M$, hence $\frac{\partial F}{\partial \theta_2} > 0$. Types $\theta_2 \geq \overline{\theta_2}$ hence participate while types $\underline{\theta_2} < \theta_2 < \overline{\theta_2}$ exit.
2. $p_1 \in (0, \underline{p_1})$
   From the proof of Lemma 2 it follows that there exists a unique type $\overline{\theta_2} > \widehat{\theta_2}$ such that $F(p_1, \overline{\theta_2}) = 0$. At $\overline{\theta_2}$, $X^* > M$, hence $\frac{\partial F}{\partial \theta_2} > 0$. Types $\theta_2 \geq \overline{\theta_2}$ ($\theta_2 < \overline{\theta_2}$) hence participate (exit). □

A.5 Proof of Proposition 3

Assume that $\mathbb{E}[u(X(\theta_2))|\theta_2 > \theta^+] < u(\frac{Z}{3})$. Then, there exists a unique $\widetilde{\theta_1}$ such that P1 sends the money iff $\theta_1 \geq \widetilde{\theta_1}$ according to Proposition 1.

If $1 - F(\widetilde{\theta_1}) < \widehat{p_1}(\widehat{\theta})$, there exist one or two $\widetilde{\theta_2}$ that make P2 indifferent between participation and exit iff $\theta_2 = \widetilde{\theta_2}$. From Proposition 2, we know that there exist types θ_2 who prefer to (not) participate iff $\theta_2 \geq \widetilde{\theta_2}(\widehat{p_1})$ ($\theta_2 < \widetilde{\theta_2}(\widehat{p_1})$).

Hence, there exists a distribution over θ_1 and θ_2 such that some types P1 do (not) send the money and some types P2 do (not) participate. □

B Experimental instructions

Role of trustor

Thank you for participating in this study. Please read the instructions carefully!

You get $0.50 for your participation and have the possibility to earn an additional bonus. We will first ask you for a decision that could influence your bonus. Afterwards, we will ask you to provide some information about yourself.

You will be randomly paired with another worker to form a group of two. The other worker also receives $0.50 for their participation. Both you and the other worker will remain anonymous.

You get an additional amount of $0.50. You can either keep this amount or give it to the other worker you are paired with.

If you decide to give the money to the other worker:

- We will triple the amount and the other worker receives $1.50.
- The other worker can then decide how much money they want to send back to you.
- Example:
  - You give the money to the other worker and they receive $1.50.
  - They decide to send $0.75 back.
  - Then, your bonus is $0.75 and the other worker's bonus is $1.50 - $0.75 = $0.75.
  - In addition, each of you gets $0.50 for your participation.

The other worker, however, has two options: They can either play the game as described before or get a certain bonus of $0.50 instead. If they choose the certain bonus of $0.50, then the game ends and you just keep your additional amount of $0.50 as a bonus. If they decide to play the game, we will ask them how much money they will send back in case you have given the money to them.

Please click the following button to continue.

Continue

To see whether we explained everything clearly, we will now ask you to answer the following questions:

- Suppose the other worker decided against playing the game. What is your bonus?
  - ○ $0.50
  - ○ $0.75
- Suppose the other worker decided to take part in the game. You do not give them the money. What is your bonus?
  - ○ $0.50
  - ○ $0.75
- Suppose the other worker decided to take part in the game. You give them the money. They send $0.60 back. What is your bonus?
  - ○ $0.60
  - ○ $0.90
- Suppose the other worker decided to take part in the game. You give them the money. They send nothing back. What is your bonus?
  - ○ $0.50
  - ○ $0

Please click the following button to continue.

Continue

We will now ask you for your decision:

Do you want to give the money to the other worker in your group?
○ Yes
○ No

Please click the following button to continue.

Continue

Finally, some questions about yourself:

- What is your gender?
  ○ Male ○ Female ○ Other ○ Prefer not to answer
- What is your age?
  years
- How willing are you in general to take risks on a scale from 0 (not willing to take risks at all) to 10 (highly willing to take risks)?
  ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Please click the following button to finish.

Continue

Role of trustee

Thank you for participating in this study. Please read the instructions carefully!

You get $0.50 for your participation and have the possibility to earn an additional bonus. We will first ask you for two decisions that could influence your bonus. Afterwards, we will ask you to provide some information about yourself.

You will be randomly paired with another worker to form a group of two. The other worker also receives $0.50 for their participation. Both you and the other worker will remain anonymous.

The other worker you are paired with gets an additional amount of $0.50. They can either keep this amount or give it to you.

If they give the money to you:

- We will triple the amount and you receive $1.50.
- You can then decide how much money you want to send back to the other worker.
- Example:
  - The other worker gives you the money and you receive $1.50.
  - You decide to send $0.75 back.
  - Then, your bonus is $1.50 - $0.75 = $0.75 and the other worker's bonus is $0.75.
  - In addition, each of you gets $0.50 for your participation.

Please click the following button to continue.

Continue

To see whether we explained everything clearly, we will now ask you to answer the following questions:

- Suppose the other worker does not give you the money. What is your bonus?
  - ○ $0
  - ○ $0.50
- Suppose the other worker gives you the money. You send $0.60 back. What is your bonus?
  - ○ $0.90
  - ○ $0.60
- Suppose the other worker gives you the money. You send nothing back. What is your bonus?
  - ○ $0.50
  - ○ $1.50

Please click the following button to continue.

Continue

We will now ask you for your decision:

If this game is played and the other worker in your group decides to give you the money, how much do you want to give back?

$ 0.35

Please click the following button to continue.

Continue

However, you have two options: You can either play the game as described before or get a certain bonus of $0.50 instead. If you choose the certain bonus of $0.50, then the game ends and the other worker just keeps their additional amount of $0.50 as a bonus. If you decide to play the game, we will take your previous decision on how much money to send back in case the other worker has given the money to you. Recall that your bonus is $0 if the other worker does not send you the money.

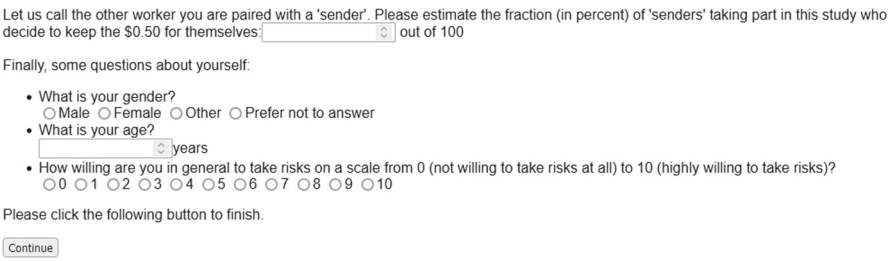
In a similar study, 58% of other workers decided to give the money to workers in your role.

Now, please make a decision:
○ Take part in the game.
○ Do not take part in the game.

Please click the following button to continue.

Continue

Note: The sentence "In a similar study, 58% of other workers decided to give the money to workers in your role." was only included in the information treatment I.



References

Andreoni, J., Rao, J.M., Trachtman, H.: Avoiding the ask: a field experiment on altruism, empathy, and charitable giving. J. Polit. Econ. **125**(3), 625–653 (2017)

Bohnet, I., Kübler, D.: Compensating the cooperators: is sorting in the prisoner's dilemma possible? J. Econ. Behav. Organ. **56**(1), 61–76 (2005)

Broberg, T., Ellingsen, T., Johannesson, M.: Is generosity involuntary? Econ. Lett. **94**(1), 32–37 (2007)

Brülhart, M., Usunier, J.-C.: Does the trust game measure trust? Econ. Lett. **115**(1), 20–23 (2012)

Cain, D.M., Dana, J., Newman, G.E.: Giving versus giving in. Acad. Manag. Ann. **8**(1), 505–533 (2014)

Cox, J.C.: How to identify trust and reciprocity. Games Econom. Behav. **46**(2), 260–281 (2004)

Cox, J.C., Deck, C.A.: On the nature of reciprocal motives. Econ. Inq. **43**(3), 623–635 (2005)

Cox, J.C., Friedman, D., Gjerstad, S.: A tractable model of reciprocity and fairness. Games Econom. Behav. **59**(1), 17–45 (2007)

Dana, J., Cain, D.M., Dawes, R.M.: What you don't know won't hurt me: costly (but quiet) exit in dictator games. Organ. Behav. Hum. Decis. Process. **100**(2), 193–201 (2006)

DellaVigna, S., List, J.A., Malmendier, U.: Testing for altruism and social pressure in charitable giving. Q. J. Econ. **127**(1), 1–56 (2012)

DellaVigna, S., List, J.A., Malmendier, U., Rao, G.: Voting to tell others. Rev. Econ. Stud. **84**(1), 143–181 (2016)

Di Bartolomeo, G., Papa, S.: Does meditation lead to more selfish or pro-social behaviors in a trust game? Technical report, Department of Communication, University of Teramo (2015)

Fehr, E., Schmidt, K.M.: A theory of fairness, competition, and cooperation. Q. J. Econ. **114**(3), 817–868 (1999)

Guido, A., Robbett, A., Romaniuc, R.: Group formation and cooperation in social dilemmas: a survey and meta-analytic evidence. J. Econ. Behav. Organ. **159**, 192–209 (2019)

Hauk, E., Nagel, R.: Choice of partners in multiple two-person prisoner's dilemma games: an experimental study. J. Conflict Resolut. **45**(6), 770–793 (2001)

Keser, C., Montmarquette, C.: Voluntary versus enforced team effort. Games **2**(3), 277–301 (2011)

Klinowski, D.: Reluctant donors and their reactions to social information. Exp. Econ. **24**(2), 515–535 (2021)

Lazear, E.P., Malmendier, U., Weber, R.A.: Sorting in experiments with application to social preferences. Am. Econ. J. Appl. Econ. **4**(1), 136–163 (2012)

McCabe, K.A., Rigdon, M.L., Smith, V.L.: Positive reciprocity and intentions in trust games. J. Econ. Behav. Organ. **52**(2), 267–275 (2003)

Nosenzo, D., Tufano, F.: The effect of voluntary participation on cooperation. J. Econ. Behav. Organ. **142**, 307–319 (2017)

Orbell, J.M., Dawes, R.M.: Social welfare, cooperators' advantage, and the option of not playing the game. Am. Sociol. Rev. **58**(6), 787–800 (1993)

Orbell, J.M., Schwartz-Shea, P., Simmons, R.T.: Do cooperators exit more readily than defectors? Am. Polit. Sci. Rev. **78**, 147–162 (1984)