Experimental Economics (2022) 25:444–482
https://doi.org/10.1007/s10683-021-09719-6

ORIGINAL PAPER

Voluntary redistribution mechanism in asymmetric coordination games

Masaki Aoyagi[1] · **Naoko Nishimura**[2] · **Yoshitaka Okano**[3]

Received: 5 March 2019 / Revised: 5 March 2021 / Accepted: 22 April 2021 /
Published online: 22 June 2021

Abstract

An inequality game is an asymmetric 2×2 coordination game in which player 1 earns a substantially higher payoff than player 2 except in the inefficient Nash equilibrium (NE). The two players may have either common or conflicting interests over the two NE. This paper studies a redistribution scheme which allows the players to voluntarily transfer their payoffs after the play of an inequality game. We find that the redistribution scheme induces positive transfer from player 1 to player 2 in both common- and conflicting- interest games, and is particularly effective in increasing efficient coordination and reducing coordination failures in conflicting-interest games. We explain these findings by considering reciprocity by player 1 in response to the sacrifice made by player 2 in achieving efficient coordination in conflicting-interest games.

Keywords Equity · Efficiency · Transfer · Reciprocity · Sacrifice

Mathematics Subject Classification C72 · D31 · D63

✉ Masaki Aoyagi
aoyagi@iser.osaka-u.ac.jp

Naoko Nishimura
nnaoko@fc.ritsumei.ac.jp

Yoshitaka Okano
oknystk@gmail.com

[1] Osaka University, Suita, Japan

[2] Ritsumeikan University, Kusatsu, Japan

[3] Kansai University, Suita, Japan

1 Introduction

Coordination failures are some of the most important sources of economic inefficiencies. Coordination games have been used extensively to study both theoretically and experimentally the sources and remedies of coordination failures. As observed by Crawford et al. (2008), one instance where severe coordination failures take place is when coordination entails asymmetric payoffs for the players involved. In this paper, we consider a class of 2×2 coordination games with highly asymmetric payoffs between the players, and examine through laboratory experiments if ex post voluntary transfer of payoffs helps eliminate coordination failures and increase efficient coordination.

Formally, an *inequality game* is a 2×2 coordination game with two Nash equilibria (NE) (X, X) and (Y, Y). Player 1 earns a strictly higher payoff than player 2 at every action profile except at (Y, Y), where they earn the same payoff. However, the sum of payoffs at (Y, Y) is substantially lower than that at (X, X), implying a tension between efficiency and equity. The inequality games are further classified into **C**om**M**on-interest (CM) inequality games in which both players' payoffs are higher at (X, X) than at (Y, Y), and **C**on**F**licting-interest (CF) inequality games in which player 1's payoff is higher but player 2's payoff is lower, at (X, X) than at (Y, Y). Examples of CM and CF inequality games are presented in Table 1. The inequality games are also parametrized by the degree of inequality between the two players' payoffs.

The redistribution scheme we propose allows both players to voluntarily transfer their payoffs to the other player after the play of the inequality game. Although such a scheme will have no impact on the outcome of the game under self-interested preferences, our main objective is to analyze its functioning in a laboratory where subjects' motivation may come from other sources than self-interest.

The efficiency-equity trade-off between the two NE in the inequality games can be a fundamental source of coordination failures. The players playing CM and CF games also can have sufficiently different motivations when choosing their actions. In CM games, coordination on the Pareto efficient profile (X, X) will result unless the players are, or expect the other player to be, sufficiently inequality averse. In CF games, on the other hand, coordination on (X, X) would be more difficult since it entails a material sacrifice by player 2 compared with coordination on (Y, Y). With ex post redistribution, this difference in the motivations between the CM and CF games can have a significant impact on the final outcome. In other words, when (X, X) is realized in the CM games, player 1 may interpret it as a result of player 2's self-interested behavior, and may find little reason to reciprocate 2's choice of X with payoff transfer to him. On the other hand, if (X, X) is realized in CF games, player 1 may interpret it as resulting from 2's *self-sacrifice* to achieve an outcome which benefits 1. Player 1 may hence have incentive to reciprocate this with payoff transfer. Expecting this, however, player 2 may strategically choose X in CF to his own benefit.

In our experiments, each subject is randomly assigned the role of either player 1 or player 2, and is randomly and anonymously matched with a subject who is

assigned the other role. The experiment consists of three parts with the subject role fixed throughout. In the first part, we have a half of the subjects in each role make a dictator decision over the action profiles of each inequality game. In the second part, the subjects play a series of inequality games in a standard way. In the third part, they play the inequality games under the redistribution scheme. Our design choice to have the same set of subjects play the games with no redistribution first and then with redistribution next, and provide in the instructions detailed information on how the payoffs are determined by the actions, is motivated by the importance of having the subjects understand the externalities involved in their decision making in the inequality games and the consequences of possible coordination failures. The within-subject design also allows us to associate the heterogeneity in the subjects' behavior with their preferences and beliefs about the behavior of the other player.

Our results show that the redistribution scheme induces significantly positive transfer by player 1 in both CM and CF games. Positive transfer takes place almost exclusively when player 2 chooses action X which corresponds to the efficient NE preferred by player 1. The size and frequency of transfer is higher in CF games than in CM games, and increasing inequality increases the size of transfer but not the frequency of positive transfer. Comparison of the results with and without the redistribution scheme shows that the scheme induces the efficient NE (X, X) strongly significantly in CF games, but only weakly in CM games. We also find that the scheme increases the sum of the two players' payoffs significantly in CF games but only insignificantly in CM games, and significantly improves equity as measured by the payoff ratio between the two players in both CF and CM games.

Since the introduction of ex post payoff redistribution has no impact on the behavior of self-interested individuals, the observed increase in efficient coordination and positive transfer imply the presence of distributive social preferences and/or reciprocity. We attempt to identify the source of these effects based on some key observations. In particular, we remark that positive transfer by player 1 to player 2 takes place almost exclusively following 2's choice of X. This suggests that player 1 reciprocates player 2's action choice that benefits player 1. Furthermore, the observed difference between CM and CF games suggests that player 1 perceives the level of kindness entailed in 2's choice of X differently in the two games. Specifically, the choice of X by player 2 can result from self-interest in CM, but entails a sacrifice in CF. We suppose that player 1's reciprocity is strengthened by the presence of self-sacrifice by player 2, and postulate a psychological utility function that explicitly accounts for sacrifice. Taking advantage of the within-subject design, we also attempt to identify the subjects' motivations by examining their behavior in different tasks. In particular, we find that the increased choice of action X by the role 2 subjects in the redistribution scheme is likely motivated by self-interest: They choose X in anticipation of the choice of X and positive transfer by role 1.

The key contributions of the present paper are summarized as follows. First, we show that social preferences can have a significant impact on the play of coordination games. The literature on social preferences has largely ignored coordination games perhaps because of the intuitive perception that social preferences will only contribute to an increase in coordination. We present a formal framework to test this

Table 1 Inequality games

	CM					CF			
	X		*Y*			*X*		*Y*	
X	440,	110	60,	50	*X*	320,	80	60,	20
Y	380,	60	100,	100	*Y*	260,	60	100,	100

(*X*, *X*) efficient coordination; (*Y*, *Y*) equitable coordination

intuition and identify the working of social preferences in the presence of a tension between equality and efficiency. Second and relatedly, we identify self-sacrifice as a critical trigger of positive reciprocity. Third, we find that individuals respond differently to increase in inequality, and that the increase in efficient coordination in the presence of redistribution opportunities is brought about by those who are intrinsically concerned about inequality and/or the own payoffs. Finally, our methodological contribution lies in the formulation of a class of coordination games with payoff asymmetry between the two players. Specifically, this class usefully nests both common and conflicting interests games with varying degrees of inequality and allows us to study the effects of these elements while controlling for the confounding effect of risk dominance.

The paper is organized as follows. The next section discusses the related literature. The inequality game and the redistribution scheme are described in Sect. 3, and testable implications of social preferences are presented in Sect. 4. Section 5 describes the experimental design, and Sect. 6 presents the analysis. The heterogeneous motives behind the observed action choices and transfer decisions are discussed in Sect. 7. We conclude with a discussion in Sect. 8.

2 Related literature

Reciprocity-based mechanisms originate in the literature on public good games. Reciprocity in the form of a punishment or disapproval of other players is the focus of early study by Fehr and Gächter (2000) and Masclet et al. (2003).[1] While reciprocity is at the core of our analysis, asymmetry between the players in our model offers a significantly different perspective from that in the symmetric environment in the early literature.

The public good literature also offers extensive research on the possible distortion of behavior associated with inequality among the players: Asymmetry is introduced either in the level of individual return (MPCR - marginal per capita return), or in the level of initial endowment (income) of each individual. The findings are largely

[1] Andreoni et al. (2003) find that the punishment option has a much stronger impact on the proposer's behavior than the reward option in a dictator-like giving game. Fehr and Rockenbach (2003) however show that the intention of imposing a sanction can induce a non-cooperative behavior in the trust game. Houser et al. (2008) examines whether intention to sanction is more important than the mere presence of sanctions.

inconclusive.[2] Combining asymmetry with redistribution in public good games, Dekel et al. (2017) and Gangadharan et al. (2017) present analysis most closely related to the present paper. When players with positive and negative MPCR's interact, and redistribution takes the form of either a punishment or reward, Dekel et al. (2017) observe that communication coupled with a reward increases contribution substantially. When players may ex post reward the others, Gangadharan et al. (2017) also find a positive impact of communication on both earnings and contribution, but show that its impact is significantly weakened in the presence of heterogeneity in MPCR.[3] While the present model shares many features with the papers on public good games with heterogeneity and ex post redistribution, its use of coordination games highlights the role of reciprocity more clearly. Specifically, it is intuitive that player 2's choice of X corresponding to the efficient coordination is a favor given to player 1, and that payoff transfer from 1 to 2 is a direct way of returning the favor.[4,5]

Positive reciprocity is studied most intensively in the trust games. Although several authors have studied the interplay of inequality and reciprocity in the trust games, our framework is different in some critical dimensions.[6] First, the sequential nature of a trust game presents no coordination issue which is the source of strategic uncertainty and a central focus of our study. Second, while a sender's action of trust in a trust game is a clear message requesting reciprocation, the simultaneous action choices in an inequality game are strategic and hence much less straightforward to interpret. Third, players in a trust game have inherently asymmetric roles as a sender and receiver, whereas the players of an inequality game are symmetric except for their payoffs. Fourth, the trust games do not allow us to study the effect of self-sacrifice as is done here.

Turning to the extensive literature on coordination games experiments, the primary focus is on the comparison between payoff dominance and risk dominance as the effective predictor of the outcome of play.[7] The literature on coordination

[2] Buckley and Croson (2006) find that the low-income subjects give a higher percentage of their income to the public good than the high-income subjects, whereas (Hofmeyr et al. 2007) observe no impact of heterogeneity on the contribution. Oxoby and Spraggon (2013) find that heterogeneity significantly lowers contributions.

[3] Other public good experiments with redistribution include (Uler 2011), who studies income redistribution under exogenous tax rates, and Belafoutas et al. (2013), who let the subjects choose the redistribution rate before the contribution decisions.

[4] Among other differences, our design does not involve explicit communication or repetition of the game between the same pair of subjects.

[5] Voluntary redistribution following a real-effort tournament is studied by Erkal et al. (2011), who study payoff transfer between the first-ranked and second-ranked subjects in the context of social preferences. See also (Ohtake et al. 2013). Unlike in the present paper, however, this literature provides no analysis of the action choice in the first stage with or without the redistribution possibility.

[6] The literature on the subject includes (Anderson et al. 2006; Rodriguez-Lara 2018), and Greiner et al. (2012). Inequality in the trust games is measured in terms of the initial endowment unlike in the present framework.

[7] Cooper et al. (1990), Cooper et al. (1992) ,Straub (1995), Van Huyck et al. (1990) ,Goeree and Holt (2005), among others, observe that risk dominance predicts subjects' play better than payoff dominance. Cachon and Camerer (1996) propose loss-avoidance as a selection principle.

games also investigates ways to eliminate coordination failures, and finds mixed evidence on the effectiveness of forward induction and correlated equilibrium recommendations.[8] Importantly, our analysis controls for the effect of risk-dominance by using games that have a constant level of risk dominance. Furthermore, while forward induction or correlated equilibrium recommendations are based on self-regarding preferences, the working of the redistribution mechanism hinges on social preferences.

3 Models of inequality and redistribution

3.1 Inequality games

Formally, an inequality game G is a 2×2 coordination game: Each player i chooses his action x_i from the set $\{0, 1\}$, and their payoff functions are given by

$$\begin{aligned} g_1(x) &= a(1 - x_1)(1 - x_2) + bx_1 + c_1x_2, \\ g_2(x) &= a(1 - x_1)(1 - x_2) + bx_2 + c_2x_1. \end{aligned} \tag{1}$$

For the interpretation of these payoff functions, suppose that each player i chooses whether to allocate his resource to either their private activity ($x_i = 1$) or an activity toward a public project ($x_i = 0$). The private activity generates positive externalities to the other player, whereas the public project results in a success if and only if both players allocate their resources to it. The successful public project is worth a to each player, whereas player i's private activity is worth b to himself and worth c_j to the other player j. When both players engage in private activities, the utility of each player is simply the sum of the benefits from his and the other player's activities.[9] We suppose that the externality benefit that 2's private activity creates for 1 is larger than the externality benefit that 1's private activity creates for 2:

$$a > b > 0 \quad \text{and} \quad c_1 > c_2 > 0.$$

Note that the second condition is the only source of inequality between the two players. Writing X for $x_i = 1$, and Y for $x_i = 0$, we can depict the payoff table as in Table 2.[10]

[8] Cooper et al. (1993) and Evdokimov and Rustichini (2016) find support for forward induction in the battle of the sexes (BOS) game, whereas (Huck and Müller 2005) suggest that the first-mover principle is important rather than forward induction. Among those who take the correlated equilibrium approach, Cason and Sharma (2007) observe a difference in subjects' behavior when they are matched against each other, and against a computer which always follows recommendations. Duffy and Feltovich (2010) find that the recommendations are followed more often when they are payoff-enhancing compared with the NE of the game. Bone et al. (2013) also find that the payoff specification affects subjects' obedience to recommendations. Anbarcı et al. (2018) find a negative impact of payoff asymmetry on obedience.

[9] The experimental instructions use neutral phrasing and express $1 - x_1 = M$ and $x_1 = N$.

[10] For the interpretation of the asymmetric payoffs, consider for example neighboring countries 1 and 2 that have environmental issues between them. $x_i = 1$ corresponds to the reduction of air pollution in country i, and $x_i = 0$ corresponds to the reduction of the pollution of public waters between them. Because of the dominant wind direction, reduction in air pollution in country 1 yields a relatively

Since $a > b > 0$, both (X, X) and (Y, Y) are pure NE. We also assume (i) $c_1 + c_2 > 2(a - b) \Leftrightarrow (X, X)$ uniquely maximizes the sum of payoffs, (ii) $c_1 > b > c_2 \Leftrightarrow g_1(x) > g_2(x)$ for $x \neq (Y, Y)$, and (iii) $2b > a \Leftrightarrow (X, X)$ is *risk dominant.* It follows from (ii) that (Y, Y) is the only profile in which the two players earn the same payoff. We further focus on the following subclasses of inequality games: An inequality game has **C**om**M**on-interest (CM) if $b + c_1 > b + c_2 > a$, and has **C**on**F**licting-interest (CF) if $b + c_1 > a > b + c_2$. In other words, if an inequality game has CM, then both players 1 and 2 prefer the NE (X, X) to the NE (Y, Y) (in terms of material payoffs), whereas if it has CF, then player 1 prefers (X, X) to (Y, Y) and player 2 prefers (Y, Y) to (X, X).

In our experiments, we set $a = 100$ and $b = 60$ and choose six combinations of c_1 and c_2 as in Table 3. This results in three CM inequality games denoted CM2, CM4 and CM6, and three CF inequality games denoted CF2, CF4 and CF6. The suffix represents the degree of inequality between the players and is equal to the payoff ratio at (X, X):

$$\frac{g_1(X,X)}{g_2(X,X)} = \frac{b + c_1}{b + c_2} = \text{k in CF}k \text{ and CM}k.$$

Since $g_1(X, X) - g_2(X, X) = c_1 - c_2$, within each class of games, the larger is k, the larger is the payoff difference at (X, X).[11] The resulting payoff tables are depicted in Tables 4 and 5. Note that all CM games are the same in terms of player 2's payoffs, and so are all CF games. Furthermore, since a and b are held constant in all games, so is the risk dominance level of (X, X).

3.2 Voluntary redistribution

Let u_i denote player i's final material payoff after the possible redistribution of their payoffs. Task 1 (**T1**) is the *baseline scheme* in which no redistribution takes place after the play of the inequality game G. In T1, the players' final payoffs equal their payoffs from G: $u_i = g_i$. Task 2 (**T2**), on the other hand, is the *redistribution scheme* in which the players may give part or all of their payoffs to the other player after publicly observing the outcome of the inequality game. If player i gives $t_i \in [0, g_i]$ payoff points to player j $(i \neq j)$, then i's final (material) payoff is given by $(t = (t_1, t_2))$

Footnote 10 (continued)
small benefit to country 2, but reduction in air pollution in country 2 yields a larger benefit to country 1 than it does to country 2 itself. On the other hand, water pollution cannot be reduced without the joint effort from the two countries. As another example, consider two workers who must allocate their effort between a production line and product development. Worker 1 is inexperienced whereas worker 2 is experienced. Product development requires joint effort from both workers. On the other hand, effort in the production line by either worker yields benefits to both of them with the spillover from the experienced worker to the inexperienced worker large and the spillover in the other direction small.

[11] Inequality may be perceived in terms of the payoff difference rather than the payoff ratio. Our econometric analysis treats k as a dummy variable, and analyzes CM and CF games separately when examining the effect of k. Since the payoff difference also increases with the payoff ratio in each class of games, interpretation of inequality in terms of the payoff difference or payoff ratio is immaterial.

Table 2 Inequality game: $c_1 > c_2$

P1 \ P2	X		Y	
X	$b + c_1$	$b + c_2$	b	c_2
Y	c_1	b	a	a

$$u_i(x,t) = g_i(x) - t_i + t_j \quad \text{for } i = 1,2, j \neq i. \tag{2}$$

Task 0 (**T0**) is the *dictator scheme* in which the final allocation is determined by only one of the players. Specifically, one player in each pair makes a choice among four payoff pairs that correspond to the four cells of the payoff table.

4 Equilibrium under reciprocity

Player i's strategy $x = (x_1, x_2) \in \{X, Y\}^2$ in T0 is the choice of an action profile, whereas his strategy in T1 is $x_i \in \{X, Y\}$. Player i's strategy in T2 is a pair (x_i, σ_i), where $x_i \in \{X, Y\}$ is the action choice and $\sigma_i : \{X, Y\}^2 \to \boldsymbol{R}_+$ is the transfer function that determines transfer to the other player j for each realization of the action profile. The subgame perfect equilibrium (SPE) $(x, \sigma) = (x_i, \sigma_i)_{i=1,2}$ is defined in the standard manner.

The players have *self-interest* preferences if their utilities equal their material payoffs (2): $U_i \equiv u_i$ for $i = 1, 2$. Under self-interest preferences, no redistribution takes place in any SPE of T2 (*i.e.,* $\sigma_i(\cdot) \equiv 0$ for $i = 1, 2$), and x is consistent with an SPE of T2 if and only if it is a NE of G.

We say that the players have *reciprocity preferences* if they reward the other player through positive transfer for the favor given to them in the play of the inequality game. Specifically, we suppose that the reciprocity preferences are given by

$$U_i(x,t) = u_i(x,t) + \gamma_i(x) \log\, u_j(x,t), \tag{3}$$

where for $0 < \mu_i \leq \nu_i$ $(i = 1, 2)$,

$$\gamma_i(x) = \begin{cases} 0 & \text{if } g_i(x) \leq a, \\ \mu_i & \text{if } g_i(x) > a \text{ and } g_j(x) \geq a, \\ \nu_i & \text{if } g_i(x) > a \text{ and } g_j(x) < a. \end{cases}$$

The second term of U_i represents player i's reciprocity concerns, and $\gamma_i(x)$ is the reciprocity weight that measures how kind j is toward i through his action choice in G. Specifically, player i takes $g_i(Y, Y) = a$ as the reference point, and considers j to be kind when j's alternative action choice $x_j = X$ raises i's payoff above a. That is, player i places a strictly positive weight $\gamma_i(x)$ on j's material payoff if and only if

Table 3 Parameter specifications

	CF2	CF4	CF6	CM2	CM4	CM6
c_1	100	260	420	160	380	600
c_2	20	20	20	50	50	50

Table 4 CM inequality games

(a) CM2					(b) CM4					(c) CM6				
	X		Y			X		Y			X		Y	
X	220,	110	60,	50	X	440,	110	60,	50	X	660,	110	60,	50
Y	160,	60	100,	100	Y	380,	60	100,	100	Y	600,	60	100,	100

Table 5 CF inequality games

(a) CF2					(b) CF4					(c) CF6				
	X		Y			X		Y			X		Y	
X	160,	80	60,	20	X	320,	80	60,	20	X	480,	80	60,	20
Y	100,	60	100,	100	Y	260,	60	100,	100	Y	420,	60	100,	100

$g_i(x) > a$.[12] If j's choice of X not only raises i's payoff above a but also lowers j's own payoff from a, then i regards it as the sacrifice made by j in raising i's payoff, and rewards j even more strongly by placing a higher weight on j's material payoff. In the CM inequality games, for example, $\gamma_1(X, X) = \mu_1$ and $\gamma_2(X, X) = \mu_2$ since both players are better off at (X, X) than at (Y, Y). On the other hand, in the CF inequality games, $\gamma_1(X, X) = \nu_1$ and $\gamma_2(X, X) = 0$ since player 1 is better off and player 2 is worse off at (X, X) than at (Y, Y).[13] The following proposition holds for the SPE of T2 under reciprocal preferences.

[12] In order to obtain an interior solution in the optimal transfer choice, we use the log transformation of j's material payoff in the definition of U_i. Any concave transformation yields a qualitatively similar conclusion.

[13] The formulation of reciprocity in (3) closely corresponds to those in the literature based on the psychological game approach. In Rabin's formulation (Rabin 1993) of reciprocity in simultaneous-move games, for example, player i's preferences are given by: $U_i(x_i) = u_i(x_i) + \tilde{f}_j \hat{f}_i(x_i)$, where u_i is i's material payoff, $\tilde{f}_j$ represents i's belief about how kind j is, and $\hat{f}_i(x_i)$ is j's payoff when i chooses action x_i given his belief about j's action. In T2, player j's kindness (in G) is revealed through his action choice in G, and hence i's beliefs need to play no role in the redistribution stage. See (Dufwenberg and Kirchsteiger 2004). One important departure from the literature is our assumption that the reciprocity weight depends on whether or not there is a sacrifice by the other player. Another preference specification that induces positive transfers is guilt-aversion as formulated by Charness and Dufwenberg (2006): A player feels guilty for not making a transfer when the other player expects it. In the present game, however, we find it difficult to specify plausible expectations.

Proposition 1 (SPE transfer under reciprocity) *Suppose that the players' preferences are given by* (3). *σ is an SPE transfer in T2 if and only if for $x \in \{X, Y\}^2$, $i = 1, 2$, $j \neq i$,*

$$\sigma_i(x) = \begin{cases} \max\{\gamma_i(x) - g_j(x), 0\} & \text{if } \gamma_1(x) + \gamma_2(x) < g_1(x) + g_2(x), \\ \min\{\gamma_i(x), g_i(x)\} & \text{if } \gamma_1(x) + \gamma_2(x) > g_1(x) + g_2(x). \end{cases} \tag{4}$$

The SPE transfer $(\sigma_1(x), \sigma_2(x)) \equiv (t_1, t_2)$ in (4) is illustrated in Fig. 1. In what follows, we restrict attention to the case where the reciprocity weights γ_i are not too large so that neither player transfers his entire payoff g_i.[14] When the parameters are in such a range, Fig. 1 shows that at most one player makes a positive transfer, and which player does so depends on the relative magnitude of the reciprocity weights. In particular, as long as those weights are similar in size, it is player 1 who makes positive transfer given that his payoff g_1 is much larger than 2's payoff g_2. Our hypotheses under the reciprocity preferences (3) are as follows.[15]

Hypothesis 1 *Under the reciprocity preferences* (3),

a. *(X, X) is played more often in T2 than in T1.*
b. *Player 1 makes positive transfer only if player 2 has chosen X, and is more likely to make positive transfer at (X, X) in CF than in CM.*
c. *The action choice as well as transfer at (X, X) are both unaffected by the degree k of inequality.*

Hypotheses 1a and 1b are our main hypotheses. The reciprocity preferences as defined in (3) generate behavior different from self-interest only in T2, and no difference is expected either in T0 or T1. As competing hypotheses, we consider the distributional social preferences as follows. The players have *inefficiency aversion* (IEA) preferences if they are concerned about the efficiency of an outcome as measured by the sum of their material payoffs. Specifically, we suppose that

$$U_i(x, t) = u_i(x, t) + \kappa_i\{u_1(x, t) + u_2(x, t)\} \quad \text{for } i = 1, 2, \tag{5}$$

where κ_i represents the degree of the inefficiency concerns relative to the own material payoff. The players have *inequality aversion* (IQA) preferences if they dislike inequality in their material payoffs. Specifically, we suppose that

$$U_i(x, t) = u_i(x, t) - \lambda_i\,|u_i(x, t) - u_j(x, t)| \quad \text{for } i = 1, 2, j \neq i, \tag{6}$$

[14] Specifically, we assume that $\gamma_1 + \gamma_2 < g_1 + g_2$ as in the first line of (4). At $x = (X, X)$, this is equivalent to $\mu_1 + \mu_2 \leq 2b + c_1 + c_2$ in CM and $\nu_1 \leq 2b + c_1 + c_2$ in CF. At $x = (Y, X)$, the equivalent condition is $\nu_1 \leq b + c_1$ in both CM and CF.

[15] See the proof of Proposition 1 in Appendix A.4 for the exact descriptions.

where $\lambda_i > 0$ represents the degree of the inequality concerns relative to the own material payoff.[16] The implications of these preferences are as follows.[17]

Hypothesis 2 *Under the IEA preferences* (5),

a. *The action profiles are the same under T1 and T2.*
b. *No transfer takes place in T2.*
c. *(X, X) is played more often as the degree k of inequality increases.*

Hypothesis 3 *Under the IQA preferences* (6),[18]

a. *(X, X) is played more often in T2 than in T1.*
b. *Player 1 makes positive transfer except at (Y, Y).*
c. *As the degree k of inequality increases, (Y, Y) is played more often in T1. The size of transfer in T2 is larger when k is larger, or in CM than in CF.*

5 Experimental design

The experiments were conducted at the Experimental Economics Laboratory at the ISER, Osaka University, with the subjects recruited from undergraduate and graduate students of Osaka University of various majors. There were six sessions with a total of 124 subjects (four sessions of 20 subjects and two sessions of 22 subjects). No subject attended more than one session. The subjects in each session were divided randomly into two groups of the same size with the first group of subjects assigned the role of player 1, and the second group assigned the role of player 2. The player roles stay the same throughout the session. The role assignment is done privately on the PC screen in front of each subject. The instruction presents the payoff formula (1), and provides its illustration by means of numerical examples and graphs.[19], [20]

[16] For simplicity, this formulation defines inequality in terms of the payoff difference. A definition based on the payoff ratio is possible.

[17] See Appendix A.4 for the exact descriptions.

[18] This hypothesis assumes that a player is sufficiently inequality averse: $\lambda_1 > \frac{1}{2}$. If $\lambda_1 < \frac{1}{2}$, no transfer takes place and the behavioral predictions are the same in T1 and T2.

[19] The program was coded using z-tree (Fischbacher 2007). The formula was included in the instructions for T0 and T1, and the graphs were included in the instructions for T1 which involved strategic interactions for the first time. The instructions for T2 didn't include them to avoid redundancy, but instead stated that the payoffs were determined by the same way as in T1. We also had five additional sessions with no payoff formula in the instructions but otherwise identical. See Sect. 8 for some discussion and Appendix A.3 for the analysis of these sessions. Every session had one additional task T3 which is not discussed in this paper: T3 involves the pre-play communication stage in which the two players simultaneously agree or disagree to have the voluntary post-play redistribution stage. The redistribution stage follows the game with agreement from at least one player. [Communication here is hence unlike the free-form communication in Dekel et al. (2017) and Gangadharan et al. (2017)]. T3 was given at the end of each session and didn't influence the subjects' behavior in the preceding tasks.

[20] See Appendix A.5 for an English translation of the instructions. The average time spent on tasks T0-T2 in six sessions is 89 minutes including time spent on instructions (20 min at the start, and 10

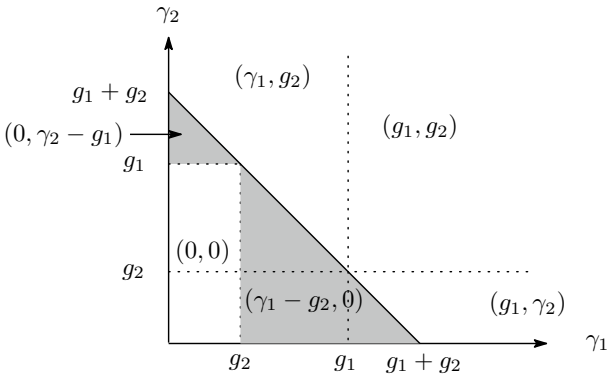

Fig. 1 SPE transfer $t = (t_1, t_2)$ as a function of the reciprocity weights (γ_1, γ_2)

The inclusion of the payoff formula is intended to help the subjects understand the source of inequality between the two roles, and also the externalities involved in their decision making. The payoff matrix is shown also on the PC screen in front of each subject. At the end of each session, the earning of a subject is computed from the sum of his/her payoff points during the session with the conversion rate of 1 payoff point to JPY1.3.[21] The average earnings are JPY9946.1 for the role 1 subjects and JPY3122.8 for the role 2 subjects.[22] The subjects were also given a record sheet in which they describe their action and transfer choices as well as the reason behind those choices.

The experiments adopt the within-subject design and every session is divided into three *task blocks* that correspond to T0–T2 described in the previous section. Each task block in turn consists of six *rounds* that correspond to the six inequality games CF2-CF6 and CM2-CM6. In all four sessions, the ordering of the task blocks is fixed and given by

$$\text{T0} \rightarrow \text{T1} \rightarrow \text{T2}.$$

As mentioned in the Introduction, the fixed task order was adopted so that the subjects would become fully aware of the externalities involved in their decision making through the standard play of the inequality games in T1.[23] The six games appear in random order in the six rounds of each task block. In T0, a half of the role 1

Footnote 20 (continued)
minutes between tasks). The subjects were given three minutes before each task to self-check their comprehension and ask questions.

[21] Azrieli et al. (2018) show that paying in all rounds may distort behavior when there exists complementarity across decisions in different rounds. Given the random matching design, however, we expect no such complementarity. See also the discussion in Sect. 8.

[22] These translate to about US$90-127 for role 1 and US$28-40 for role 2 according to the exchange rates at the time of the experiments.

[23] This design choice is also based on the four pilot sessions which rotated task orders. See Footnote 45.

subjects and a half of the role 2 subjects are randomly chosen to make a choice.[24] After every round, each subject observes his and the other player's action choices on their PC screen, and then is anonymously rematched to another subject of the opposite role in the stranger format. Since ten or eleven pairs are formed in each session consisting of 18 rounds, the probability that the same pair of subjects are matched is positive. One concern may be that it generates stronger incentive to coordinate on the efficient action profile than in the one-shot environment. The literature on repeated prisoners' dilemma however finds that cooperation rates are not high under random matching (Duffy and Ochs 2009). Furthermore, since the same format is used for T1 and T2, the marginal impact of such an incentive on the effectiveness of T2, if any, would be smaller.[25]

6 Analysis

6.1 Dictator decisions

We begin by examining the outcome of the dictator decision task (T0). Figure 2 shows the frequency of each of four choices for CM and CF games, where $A = (X, X)$, $B = (X, Y)$, $C = (Y, X)$, and $D = (Y, Y)$. As seen, the subjects' choices are almost entirely limited to (X, X) and (Y, Y), which are NE of the game.[26] Additionally, the role 1 subjects choose (X, X) more than 94% of the time, whereas the choice of the role 2 subjects is approximately reversed depending on whether the game is CM or CF. In CM, they choose (X, X) more than 90% of the time, whereas in most cases of CF, they choose (Y, Y) 80% of the time.[27] The inequality dummy k is mostly insignificant in both CM and CF.[28]

Role 1's choice of (X, X) is consistent with self-interest and IEA, whereas role 2's choice of (X, X) in CM and (Y, Y) in CF is consistent with self-interest and IQA.

Observation 1 (Dictator decision) *The behavior of the role 1 subjects in T0 is mostly consistent with self-interest and IEA. The behavior of the role 2 subjects in T0 is mostly consistent with self-interest and IQA.*[29]

[24] Once chosen, they make choices in all six rounds of the T0 block (against different opponents).

[25] There were also no comments in the subjects' record sheets that would indicate that the subjects attempted to influence future interactions through their action choice.

[26] The hypothesis that the four outcomes are randomly chosen with equal probabilities is rejected ($p = 0.01$).

[27] A multinomial logit regression over the four outcomes is no more informative than the descriptive statistics because of the skewed distribution of the choice data.

[28] The unique exception is the choice by the role 2 subjects who choose (Y, Y) less often in CF4 than in CF2 (Mann-Whitney U-test, $p = 0.1$).

[29] Further analysis of role 2's choice in CF-T0 is given in Sect. 7.

6.2 Action choice

Figure 3 shows the frequency of each action (X and Y) by each subject role. As seen, there is a significant difference in the subjects' action choice in T1 and T2, and the difference is more prominent in CF: Going from CM-T1 to CM-T2, the choice of X increases by 6 percentage points for role 1 (90% → 96%) and by 4 percentage points for role 2 (90% → 94%). On the other hand, going from CF-T1 to CF-T2, the choice of X increases by 16 percentage points for role 1 (68% → 84%), and 10 percentage points for role 2 (73% → 83%). It also shows that both roles choose Y more often in CF than in CM. These observations are confirmed by the random effects logit regressions in Table 11 in Appendix A.1, where the dependent variable equals one if a subject chooses Y. Models (4) and (6) in Table 11, which include inequality dummies k_4 and k_6, show that increasing inequality has different effects in CM and CF: While higher inequality overall has a positive impact on the choice of Y in CM, higher inequality has little to no impact in CF. In CM where the increasing inequality increases the choice of Y, this effect is independent of the subject role (model (5)). The subject role has no significant impact on the action choice in CM and CF (models (2) and (3)). This is in sharp contrast with our observation in CF-T0, where the dominant choice is $A = (X, X)$ for role 1 and $D = (Y, Y)$ for role 2.[30] The observation on the action choice can be summarized as follows:

Observation 2 (Action choice)

1. *In CF, T2 raises the choice of X compared with T1.*
2. *Higher inequality increases the choice of Y in CM.*

Observation 2.1 supports our main reciprocity Hypothesis 1a as well as the IQA Hypothesis 3a but not the IEA Hypothesis 2a. The effect of inequality in Observation 2.2 on CM is also consistent with the IQA Hypothesis 3c, but not with the IEA Hypothesis 1c or the reciprocity Hypothesis 2c.

6.3 Coordination

Table 6 describes the realized distribution of four action profiles in T0 through T2. It shows that the redistribution scheme induces efficient coordination particularly effectively in CF: Going from T1 to T2, the efficient coordination (X, X) increases by 9% percentage points (81% → 90%) in CM, but by 22 percentage points in CF (48% → 70%). Furthermore, the redistribution scheme also reduces coordination failures much more substantially in CF: Coordination failures (X, Y) and (Y, X) decrease by

[30] In addition to some role 2 subjects who switch from D in T0 to X in CF-T1 or CF-T2, there are also some role 1 subjects who switch from $A = (X, X)$ in T0 to Y in CF-T1 or CF-T2. See more analysis on the behavior of individual subjects in Sect. 7, which also discusses the possible mechanism behind Observation 2.2.

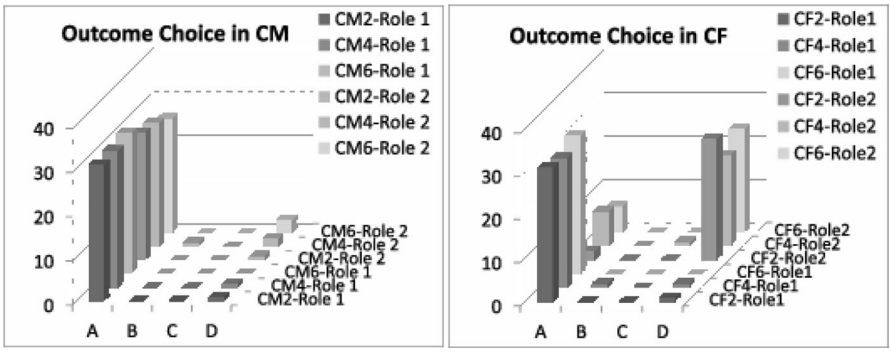

Fig. 2 Outcome distribution in T0

8 percentage points in CM (18% → 10%), whereas they decrease by 18% percentage points in CF (45% → 27%). In fact, the difference in the distributions between T1 and T2 is strongly significant only in CF ($p = 0.00$) and only weakly so in CM ($p = 0.07$). The difference between CM and CF is strongly significant in T0, T1 and T2. These observations are again confirmed by logit regressions of action profiles in Table 12 in Appendix A.1, where the dependent variable is either the efficient coordination ($\mathbf{1}_{\{a=(X,X)\}}$), or the total coordination ($\mathbf{1}_{\{a=(X,X)or(Y,Y)\}}$).[31] As in the case of individual action choices, models (3)-(6) show that the impact of increasing inequality (*i.e.,* signs of the inequality dummies k_4 and k_6) is qualitatively different between CM and CF: higher inequality reduces coordination in CM, but either increases it or has no effect in CF. We summarize our findings as follows:

Observation 3 (Coordination)

1. *In both CM and CF, T2 increases efficient coordination (X, X) and reduces both inefficient coordination (Y, Y) and coordination failures.*
2. *In both T1 and T2, higher inequality decreases efficient and total coordination in CM, but has either positive or no effect on coordination in CF.*

Observation 3.1 is our central finding that supports our main reciprocity hypothesis 1a. It is also consistent with the IQA hypothesis 3a, but contradicts the IEA hypothesis 2a which would imply no difference between T1 and T2. On the other hand, the reciprocity hypothesis 1c is not supported by the negative impact of high inequality on efficient coordination in CM in Observation 3. The negative impact is consistent with the IQA hypothesis 3c in the case of T1, but not with the IEA hypothesis 2c.

[31] A multinomial logit regression is infeasible because of the heavily skewed frequency distribution of action profiles as indicated in Table 6.

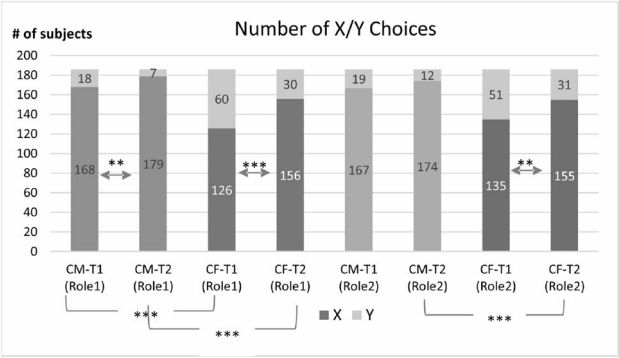

Fig. 3 Action choices in T1 and T2. ** and ***: significant difference at 5% and 1%, respectively, between the respective pair of distributions (χ). Shown in each column are the numbers of each action choice aggregated for $k = 2$, 4, and 6

Table 6 Realization of action profiles

Action profile	CM				CF			
	T0		T1	T2	T0		T1	T2
	Role 1	Role 2			Role 1	Role 2		
(X, X)	94	83	151	167	93	16	89	130
(Y, Y)	2	6	2	0	2	73	14	5
(X, Y)	0	1	17	12	1	0	37	26
(Y, X)	0	0	16	7	0	1	46	25
p-value (χ^2 test):								
T0=T1=T2	0.00	0.00			0.00	0.00		
T1=T2			0.07				0.00	
CM=CF		0.00	0.00	0.00				

The three lines in the bottom report the p-values of the χ^2 tests of the hypothesis that the distributions are the same for T0-T2 (first line), between T1 and T2 (second line), and between CM and CF (third line)

The observed increase in efficient coordination in T2 leads to improvement in efficiency as measured by the sum of the two players' payoffs ($g_1 + g_2 = u_1 + u_2$). T2 raises the average total payoff by 4.7% in CM (490.65 in T1 $\Rightarrow$ 513.92 in T2, $p = 0.28$ t-test), and by 12.8% in CF (301.51 in T1 $\Rightarrow$ 340.00 in T2, $p = 0.03$ t-test).[32] In line with Observation 3.1, T2 has a more substantial impact on efficiency in CF than in CM. Further scrutiny of the subjects' payoffs reveals interesting facts.

[32] As seen in Fig. 4 in Appendix A.2, the cumulative distribution of the total payoff in T2 (dashed-line) approximately first-order dominates that in T1 (solid-line) in both CM and CF, and the relationship is indeed significant in CF ($p = 0.07$, two-sided Kolmogorov-Smirnov test).

Figure 5 in Appendix A.2 depicts the average final payoffs (u_i) of each role in T1 and T2. While redistribution raises role 2's payoff in both CM and CF ($p < 0.01$ in both CM and CF, t-test and Mann-Whitney test), no such effect is observed for role 1 (by either test). Tobit regression analysis in Table 13 in Appendix A.1 confirms that redistribution has a significantly positive impact only on role 2's payoff. In summary, our finding suggests that role 1 transfers away any payoff gain from more efficient coordination achieved in T2.[33]

6.4 Transfer

Table 7 shows the average transfer and the number of occurrences of positive transfers after each action profile. As seen, a dominant share of positive transfer is made by role 1: In total, role 1 makes positive transfers in 31.2% of all occasions in CM (58 times out of 186 occasions), and 43.5% of all occasions in CF (81 times out of 186 occasions). Role 1's average transfer is significantly positive in both CM and CF, implying that they are on average not self-interested. When aggregated over k, 94.8% and 84.0% of all positive transfers by role 1 are observed after the realization of (X, X) in CM and CF, respectively. Role 1's average transfer amount is significantly higher conditional on (X, X) than conditional on (X, Y) in both CM and CF ($p < 0.01$, t-test).

Table 15 in Appendix A.1 presents regressions of absolute and relative transfer amounts as well as the likelihood of positive transfer focusing on role 1 after his own choice of X.[34] It shows that both the average transfer and the likelihood of positive transfer are larger when role 2 chooses X in both CM and CF, and also larger in CF than in CM.[35] Moreover, while the inequality dummy k has a positive impact on absolute transfer (models (4) and (7)), it has no significant impact on the likelihood of positive transfer (models (6) and (9)).

Observation 4 (Size and frequency of transfer)

1. *The average transfer by role 1 is significantly positive.*
2. *Positive transfer by role 1 is more likely after the choice of X by role 2, and both absolute and relative transfer is larger in this case.*
3. *Positive transfer by role 1 is more likely in CF, and the size of transfer is larger in CF.*
4. *Higher inequality increases absolute transfer, but not the likelihood of positive transfer.*

[33] One possible hypothesis behind this observation is that role 1 uses their payoff in T1 as the reference point and transfers away any additional gains in T2 to role 2. However, we find no support for this hypothesis as shown in Table 14 in Appendix A.1, which computes the average payoff of role 1 in T2 conditional on the action profile they experienced in T1.

[34] Relative transfer equals the absolute transfer amount divided by the payoff in the game: $\frac{t_1}{g_1}$.

[35] Figure 6 in Appendix A.2 also shows that CF dominates CM in terms of the cumulative distribution of relative transfer by role 1 ($p < 0.04$, Kolmogorov-Smirnov test).

Observations 4.1-4.3 support the reciprocity hypothesis 1b. Furthermore, the insignificance of the degree k of inequality in Observation 4.4 is consistent with the reciprocity hypothesis 1c. On the other hand, the size of transfer in Observation 4.4 is not implied by our theory of reciprocity. Turning to the competing hypotheses, the positive transfer is consistent with the IQA hypothesis 3b, but not with the IEA hypothesis 2b. Regarding the size of transfer, the effect of k is consistent with the IQA hypothesis 3c. However, the smaller transfer in CM-T2 than in CF-T2 does not support IQA since CM has higher inequality than CF for the same level of k.

Positive transfer by the role 1 subjects improves equity as measured by the ratio of the final payoffs ($\frac{u_1}{u_2}$). Figure 7 in Appendix A.2 shows the average payoff ratio for T1 and T2. For each k, we see that redistribution raises equity in both CM and CF. In fact, the null hypothesis of no difference between T1 and T2 is rejected for all k in CM ($p < 0.05$, t-test) and for $k = 2$ and $k = 4$ in CF ($p < 0.01$, t-test).[36] The strongly significant impact of redistribution on equity in both CM and CF is confirmed by the Tobit regressions of the payoff ratio in Table 16 in Appendix A.1.

7 Heterogeneity across subjects

All taken together, the analysis of the previous section strongly suggests reciprocity as the main driver behind the working of the redistribution scheme. However, there are also indications of distributional social preferences, and our formulation of reciprocity does not capture well the response to the change in the degree of inequality. To further investigate this point, we now look at the possible heterogeneity in motives across subjects.

As observed earlier, some fraction of role 2 subjects choose $A = (X, X)$ in the dictator-decision task CF-T0. There are also subjects who switch from $D = (Y, Y)$ to A as inequality k increases. In other words, we can interpret these role 2 subjects as preferring efficiency to equity as the efficiency gap between A and D widens. We hence say that role 2 subjects are *inefficiency averse* (IEA) if, in CF-T0, they choose A for all k, or switch once from D to A as k increases. On the other hand, those role 2 subjects who choose D for all levels of k, or switch from A to D once as k increases are either self-interested or do not tolerate high inequality at A.[37] We call them *inequality averse* (IQA) type. Out of thirty role 2 subjects who made a choice in CF-T0, twenty are IQA, whereas six are IEA. As seen in Table 8, while type IEA chooses X most of the time in T1 and T2 and doesn't substantially change behavior from T1 to T2, type IQA chooses X less often overall, but increases the choice of X substantially in T2. As far as role 2 is concerned, hence, we can deduce that the increased choice of X in CF-T2 is by type IQA motivated by the reduced concern over inequality at

[36] The reduction in the payoff ratio in T2 is also significant by the Kolmogorov-Smirnov test ($p < 0.03$ for CM and $p < 0.01$ for CF).

[37] Role 2 subjects who choose D for every k can either be self-interested, or dislikes inequality even at $k = 2$. Based on CM-T0, only three subjects are classified as IQA. This implies that the classification here is specific to each class of games.

Table 7 Average transfer $(\bar{t}_1, \bar{t}_2)$ in T2 by game and action profile

	CM2				CM4				CM6			
	X		Y		X		Y		X		Y	
X	8.4	1.2	–	–	24.8	0.8	1.7	–	43.8	1.4	–	–
	$\frac{16}{59}$	$\frac{5}{59}$	$\frac{0}{3}$	$\frac{0}{3}$	$\frac{20}{56}$	$\frac{3}{56}$	$\frac{1}{3}$	$\frac{0}{3}$	$\frac{19}{52}$	$\frac{4}{52}$	$\frac{0}{6}$	$\frac{0}{6}$
Y	–	–	–	–	33.3	0.3	–	–	70.0	–	–	–
	$\frac{0}{0}$	$\frac{0}{0}$	$\frac{0}{0}$	$\frac{0}{0}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{0}{0}$	$\frac{0}{0}$	$\frac{1}{4}$	$\frac{0}{4}$	$\frac{0}{0}$	$\frac{0}{0}$
	CF2				CF4				CF6			
	X		Y		X		Y		X		Y	
X	10.9	0.6	6.7	–	27.9	1.0	–	–	42.4	2.5	0.3	–
	$\frac{17}{39}$	$\frac{2}{39}$	$\frac{2}{9}$	$\frac{0}{9}$	$\frac{24}{44}$	$\frac{3}{44}$	$\frac{0}{9}$	$\frac{0}{9}$	$\frac{27}{47}$	$\frac{5}{47}$	$\frac{1}{8}$	$\frac{0}{8}$
Y	4.6	1.8	–	–	15.8	–	–	–	44.0	0.2	–	–
	$\frac{4}{12}$	$\frac{2}{12}$	$\frac{0}{2}$	$\frac{0}{2}$	$\frac{4}{8}$	$\frac{0}{8}$	$\frac{0}{1}$	$\frac{0}{1}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{0}{2}$	$\frac{0}{2}$

The table lists for each action profile the average transfer amounts ($\bar{t}_1 : 1 \to 2$ and $\bar{t}_2 : 2 \to 1$) in line 1, and (#obs. of positive transfer)/(#obs. of the action profile) (by role 1 and role 2) in line 2

(X, X) and/or the payoff loss at (X, X) (compared with (Y, Y)) in anticipation of the choice of X and positive transfer by role 1.[38]

To see if role 2 is rational in their thinking, Table 17 in Appendix A.1 incorporates the average transfer $\bar{t}_1$ from role 1 to role 2 in T2 into their payoffs in CF. It shows that X is a dominant action for role 2 when $k = 6$, and (X, X) is a payoff dominant equilibrium when $k = 4$. Furthermore, if role 2 expects that role 1's action choice is given by its empirical frequency, X is his uniquely optimal action for every k.[39]

In T1, we also have subjects of both roles whose action choice either is constant for every k, or switches once as k goes up. We call a subject *type TX1* if his choices are either $X \to X \to X$, $Y \to X \to X$, or $Y \to Y \to X$ as k increases from $2 \to 4 \to 6$ in T1. On the other hand, we call a subject *type TY1* if his choices are either $Y \to Y \to Y$, $X \to Y \to Y$, or $X \to X \to Y$ as k increases from $2 \to 4 \to 6$

[38] This inference is consistent with the subjects' comments in their record sheets: As for the choice of X in CF-T2, the most popular reason given by the role 2 subjects is "afraid of the low payoff resulting from my choice of Y and the other player's (role 1) choice of X," followed by "expecting the choice of X by role 1", "wanted to maximize the sum of payoffs", and "role 1 would reciprocate my choice of X with transfer." No role 2 subject describes altruistic motives such as wanting to be kind to role 1. Some of the role 1 subjects, on the other hand, thank role 2 for choosing X, implying the presence of reciprocity motives.

[39] For role 1, on the other hand, X cannot be a dominant action for any k. Again, however, if role 1 expects that role 2's action choice is given by its empirical frequency, X is the uniquely optimal action for role 1 for every k.

Table 8 Rate of action X in T1 and T2 by role 2's type in CF-T0

Type	CM-T1			CM-T2		
	$k = 2$	$k = 4$	$k = 6$	$k = 2$	$k = 4$	$k = 6$
IQA	0.90	0.80	0.85	0.90	0.90	0.95
IEA	1.00	1.00	1.00	1.00	1.00	1.00
Type	CF-T1			CF-T2		
	$k = 2$	$k = 4$	$k = 6$	$k = 2$	$k = 4$	$k = 6$
IQA	0.60	0.70	0.50	0.70	0.75	0.70
IEA	0.83	0.83	1.00	0.83	1.00	1.00

in T1.[40] It is worth noting that the two types cover more than 90% of all cases.[41] Table 9 shows the likelihood of action X in T1 and T2 by types TX1 and TY1.[42] As seen, type TX1 of either role chooses X most of the time in both T1 and T2 for each k. On the other hand, the change in the likelihood of X by type TY1 is dramatic. While by definition they never choose X at $k = 6$ in T1, they choose X more than 60% of the time in T2.[43] Table 9 also hints at the possible mechanism behind Observation 2.2 on the effect of inequality on the action choice. As seen, type TY1 in CM-T1 sharply decreases the choice of X as k goes up, whereas TX1 is mostly unresponsive to the change in k. We can interpret the negative impact of k on X in CM-T1 as resulting from the inequality averse response by TY1. In CF-T1, on the other hand, TY1 and TX1 move in the opposite directions as k goes up, offsetting the impact of k.

Observation 5 *Between T1 and T2, the increase in the choice of X by both roles 1 and 2 is mostly due to type TY1 who responds to increased inequality with the choice of Y in T1. In T2, the choice of X by role 2 is motivated by the anticipation of role 1's choice of X and positive transfer.*

[40] The symbols TX and TY signify that the response moves **T**owards **X** and **T**owards **Y**, respectively, as k increases. These types are again specific to each class of games. Type TX1 is hence similar to type IEA in T0 and type TY1 is similar to type IQA in T0. In T1, this interpretation is also consistent with the equilibrium behavior under the IEA and IQA preferences as described in Hypotheses 2 and 3 in Sect. 4. It is important to note that behavior as specified by types TX and TY reflects not only their preferences but also their beliefs over the strategic action choice by the other player.

[41] The number of each type is as follows: In CM-T1, 15 (7 for role 1+8 for role 2) are TY1 and 102 (52 + 50) are TX1. In CF-T1, 32 (15 + 17) are TY1 and 75 (34 + 41) are TX1. Together, 90.3% of the subjects are classified as either type (94.4% in CM-T1, 86.3% in CF-T1).

[42] When types TX2 and TY2 are defined similarly in T2, Table 18 in Appendix A.1 shows the association of the types in T1 and T2. In both CM and CF, nearly all of type TX1 become TX2 whereas approximately two-thirds of TY1 become TX2. In other words, those who respond to higher inequality with the choice of Y in T1 tend to change their behavior and respond to higher inequality with the choice of X in T2.

[43] On the other hand, the tendency of TY1 to decrease the choice of X for a larger k is unchanged in CM-T2.

How does the difference in behavior in T1 translate to the difference in the transfer decisions in T2? For each role 1 subject, let his *reciprocation index r* be defined by

$$r = \frac{\#\text{positive transfers after } (X, X)}{\#\text{occurrences of } (X, X)}.$$

Each role 1 subject experienced (X, X) up to six times, and we call role 1 *strongly reciprocating* (SR) if $r \geq \frac{2}{3}$, *weakly reciprocating* (WR) if $r \in \left[\frac{1}{3}, \frac{2}{3}\right)$, and *non-reciprocating* (NR) if $r < \frac{1}{3}$. Table 19 in Appendix A.1 shows overall downgrading of reciprocity types going from CF-T2 to CM-T2: Nearly half of type SR in CF-T2 become NR in CM-T2, while few type NR in CF-T2 become SR in CM-T2. Table 10 shows the relationship between the type classification in T1 (*i.e.,* TX1 and TY1 for role 1) and reciprocity types in T2. In CF-T2, the distribution of reciprocity types is almost identical between TY1 and TX1 with an even split between SR and NR. On the other hand, in CM-T2, nearly two-thirds of TX1 are NR, while it is much less likely that TY1 becomes NR. This shows that TY1 and TX1 are equally likely to reciprocate role 2's choice of X accompanied by payoff sacrifice, but that TX1 is more likely to ignore 2's choice of X if not accompanied by payoff sacrifice. In other words, TX1 and TY1 are different in the perception of kindness by role 2 that has led to the increase in their own payoff.

Observation 6

1. *The degree of reciprocation is stronger in CF-T2 than in CM-T2 even for the same subject.*
2. *Type TX1 tends to reciprocate role 2's choice of X only when it is accompanied by self-sacrifice.*

Observation 6.1 supports the reciprocity hypothesis 1b. Observation 6.2 suggests that role 1's behavior in T1 can be related to the difference in reciprocity between CM and CF.

8 Conclusion

The analysis of the paper is based on data from the sessions which presented the payoff formula (1) in the instructions. Apart from these sessions, we also ran five sessions in which the instructions did not present the payoff formula.[44] Appendix A.3 reports some analysis that compares the results with and without the formula. Most notably, we observe that the inclusion of the formula had significantly positive impacts on the subjects' action choice both in T1 and T2. In particular,

[44] These sessions were identical to the main sessions otherwise. The five sessions had 20, 20, 20, 24, and 22 subjects with the total of 106 subjects.

Table 9 Rate of action *X* by T1 types

	CM-T1						CF-T1					
	Role 1			Role 2			Role 1			Role 2		
	$k=2$	$k=4$	$k=6$	$k=2$	$k=4$	$k=6$	$k=2$	$k=4$	$k=6$	$k=2$	$k=4$	$k=6$
TY1	0.86	0.57	0.00	0.75	0.50	0.00	0.47	0.20	0.00	0.53	0.35	0.00
TX1	0.94	0.98	1.00	0.98	1.00	1.00	0.79	0.91	1.00	0.78	1.00	1.00
	CM-T2						CF-T2					
	Role 1			Role 2			Role 1			Role 2		
	$k=2$	$k=4$	$k=6$	$k=2$	$k=4$	$k=6$	$k=2$	$k=4$	$k=6$	$k=2$	$k=4$	$k=6$
TY1	1.00	0.71	0.71	0.88	0.75	0.63	0.53	0.67	0.60	0.59	0.71	0.65
TX1	1.00	0.98	0.96	0.98	1.00	0.96	0.88	0.94	0.97	0.90	0.90	0.93

inclusion of the formula increases the frequency of action *X* and increases the frequency of efficient coordination (*X*, *X*). These results suggest that inclusion of the formula raises the awareness of the externalities involved in decision making in the inequality games. We believe that such awareness of externalities is key to the inducement of social preferences including reciprocity.[45]

How would our findings help inform policy making in the possible presence of coordination inefficiencies? First and foremost, our finding suggests the importance of a salient opportunity for ex post reciprocation. While we formulate reciprocation as direct payoff transfer, the interaction may as well take an alternative form as long as it is sufficiently intuitive. Second, as noted above, our analysis suggests the importance of making the parties aware of the externalities involved in their decision making. Third, we note that player 2 in the CF inequality games likely uses self-sacrifice to convey a credible message behind his action choice to player 1, and is confident that player 1 understands this message. The lack of such a message in the CM inequality games has reduced the effectiveness of the redistribution scheme.[46] This observation suggests that the policy should create a channel through which the parties can credibly convey the intention behind their action choice.

One interesting extension of the present work involves elicitation of beliefs before the play of the games. It would be interesting to find out beliefs about the other player's action choice, and the amount of transfer they expect from the other player after the realization of each action profile. Although we have studied the redistribution scheme in the presence of inequality between the players, it is important to check its validity

[45] As mentioned in Footnote 23, we had four pilot sessions with rotated task orders. These sessions also presented no payoff formula in the instructions. Analysis combining data from all sessions without the payoff formula shows that the effect of rotation on the likelihood of coordination is insignificant, but tends to be negative. The effect of T2 on coordination is in line with our main analysis.

[46] In relation to this point, the reason why redistribution alone didn't lead to efficiency in the public good experiments discussed in Sect. 2 may be that the players in those experiments also had difficulty interpreting the intention behind the other's action choice.

Table 10 T1 types and reciprocity types in T2

	CM-T2			CF-T2			
	SR	WR	NR	SR	WR	NR	Other
TY1	5	0	2	5	1	5	4
TX1	12	9	31	16	3	15	0
Other	1	2	0	7	2	4	0

Reciprocity type "other" refers to role 1 who didn't experience (X, X)

in other classes of games. In the BOS game, for example, we would expect that the redistribution scheme as proposed here is valid if the sum of payoffs at one NE is substantially higher than that at the other NE. If, on the other hand, both NE are equally efficient, then some modification to the scheme would be required. In view of the literature, communication may play a critical role in such an environment. Examining the validity of the scheme under various payoff specifications is a topic of future research.

Appendix

A.1 Tables

Table 11 Random effects logit regressions of action choice: $y = \mathbf{1}_{\{a_i=Y\}}$

Model	(1) All	(2) CM	(3) CF		(4) CM	(5) CM	(6) CF	(7) CF
t2	-0.58	0.31	− 0.71***	Role	− 0.07	0.60	0.38	− 0.25
	(0.40)	(0.67)	(0.26)		(0.65)	(0.86)	(0.23)	(0.52)
cf	1.86***			k_4	0.92***	1.32**	0.21	− 0.81*
	(0.30)				(0.15)	(0.60)	(0.40)	(0.48)
cf * t2	− 0.23			k_6	1.11***	1.57***	− 0.01	− 0.01
	(0.37)				(0.29)	(0.58)	(0.40)	(0.50)
Role		− 0.06	0.38	Role * k_4		− 0.77		1.90***
		(0.53)	(0.23)			(1.20)		(0.65)
Role * t2		− 0.65	− 0.44	Role * k_6		− 0.87		0.00
		(0.97)	(0.37)			(1.26)		(0.63)
1/round	6.55	22.00***	5.22	1/round	1.29**	1.32***	− 0.12	− 0.12
	− 5.94	(7.34)	(6.62)		(0.51)	(0.46)	(0.41)	(0.44)
Constant	− 3.81***	− 6.05***	− 2.04***	Constant	− 5.10***	− 5.51***	− 1.57***	− 1.31***
	(0.73)	(0.85)	(0.60)		(1.13)	(0.94)	(0.33)	(0.36)
Log-likelihood	− 513.09	− 167.24	− 342.10	Log-likelihood	− 106.22	− 105.93	− 206.11	− 201.55
#obs.	1,488	744	744	#obs.	372	372	372	372
#subjects	124	124	124	#subjects	124	124	124	124

Model (1) combines data from CM and CF whereas models (2)-(7) separate them. Independent variables: cf = 1 if CF, t2 = 1 if T2, role = 1 if role 1, $k_4 = 1$ if $k = 4$, and $k_6 = 1$ if $k = 6$. The variable 1/round equals the inverse of the round number within each task block, and is included given that all other independent variables are dummies. *, ** and ***: significant at 10%, 5% and 1%, respectively. Robust standard errors clustered by session in parentheses

Table 12 Logit regressions of action profiles

Model	(1)	(2)	(3)	(4)	(5)	(6)
	All		CM		CF	
Dep. var.	$\mathbf{1}_{\{(X,X)\}}$	$\mathbf{1}_{\{(X,X) \text{ or} (Y,Y)\}}$	$\mathbf{1}_{\{(X,X)\}}$	$\mathbf{1}_{\{(X,X) \text{ or} (Y,Y)\}}$	$\mathbf{1}_{\{(X,X)\}}$	$\mathbf{1}_{\{(X,X) \text{ or} (Y,Y)\}}$
t2	0.81***	0.68***	1.09	1.06	0.85***	0.80***
	(0.23)	(0.22)	(0.79)	(0.78)	(0.33)	(0.21)
cf	− 1.84***	− 1.45***				
	(0.23)	(0.16)				
cf*t2	0.29	0.15				
	(0.38)	(0.35)				
1/round	− 0.37	− 0.40*	− 0.64*	− 0.45	0.02	− 0.13
	(0.27)	(0.21)	(0.36)	(0.36)	(0.37)	(0.39)
k_4			− 0.90***	− 0.79***	0.00	0.23
			(0.15)	(0.13)	(0.48)	(0.55)
k_6			− 0.98***	− 0.85***	0.38	0.69**
			(0.24)	(0.28)	(0.45)	(0.34)
t2*k_4			0.26	0.10	0.43	0.06
			(0.96)	(0.94)	(0.79)	(0.79)
t2*k_6			− 0.59	− 0.68	0.34	− 0.02
			(0.68)	(0.59)	(0.49)	(0.33)
Constant	1.86***	1.84***	2.96***	2.78***	− 0.24	− 0.02
	(0.20)	(0.17)	(0.31)	(0.32)	(0.34)	(0.18)
#obs	744	744	372	372	372	372
Log likelihood	− 374.40	− 377.83	− 136.58	− 136.81	− 233.66	− 232.96

Models (1) and (2) combine data from CM and CF whereas models (3)-(6) separate them. See Table 11 for the definitions of the independent variables. *, ** and ***: significant at 10%, 5% and 1%, respectively. Robust standard errors clustered by session in parentheses

Table 13 Mixed effects Tobit regressions of final payoffs

Model	Role 1		Role 2	
	CM	CF	CM	CF
	(1)	(2)	(3)	(4)
t2	− 3.12	7.691	10.08***	12.56***
	(6.48)	(5.33)	(2.48)	(1.78)
k_4	174.40***	132.10***	− 5.61***	3.07
	(11.45)	(7.22)	(1.18)	(4.10)
k_6	367.60***	230.10***	− 5.63***	5.86***
	(18.49)	(10.53)	(1.76)	(1.60)
t2 * k_4	12.18	− 6.034	19.68***	11.44
	(10.65)	(16.43)	(4.99)	(7.42)
t2 * k_6	− 15.65	21.00	32.65***	23.03***
	(23.98)	(18.25)	(6.54)	(5.39)
1/round	12.01	0.713	− 0.13	1.75
	(19.92)	(11.45)	(4.60)	(3.72)
Constant	204.80***	115.90***	103.90***	66.93***
	(6.11)	(6.20)	(2.71)	(0.80)
# of obs.	372	372	372	372
Log likelihood	− 2333.37	− 2283.6033	− 1944.21	− 1839.3891

*, ** and ***Significant at 10%, 5% and 1%, respectively. Standard errors clustered by session in parentheses

Table 14 Role 1's payoff in T2 conditional on the action profile in T1

T1	CF2	CF4	CF6	CM2	CM4	CM6
(X,X)	132.429	231.393	384.857	206.855	405.438	560.208
	(40.547)	(106.812)	(136.046)	(34.280)	(85.694)	(166.431)
	28	28	35	55	48	48
	0.001	0.002	0.002	0.0106	0.008	0.001
(X,Y)	134.059	266.375	472.727	166.667	345.143	565.714
	(34.965)	(92.751)	(13.484)	(92.376)	(137.914)	(224.117)
	17	8	11	3	7	7
	0	0.0004	0	0.1835	0.0016	0.001
(Y,X)	99.800	275.286	322.444	214.500	366.667	548.333
	(31.134)	(65.198)	(193.891)	(9.713)	(78.655)	(153.677)
	15	21	9	4	6	6
	0.9805	0.2954	0.1696	–	–	–
(Y,Y)	100.000	220.000	337.143	–	438.000	460.000
	(0.000)	(88.318)	(142.912)	–	–	–
	2	5	7	–	1	1
	–	–	–	–	–	–

For each action profile in T1, the table lists the average payoff in T2 (line 1), standard deviations (line 2), the number of observations (line 3), and *p*-value of the hypothesis: "payoff in T1= payoff in T2" by t-test (line 4). "–" implies insufficient observations

Table 15 Determinants of the size and likelihood of transfer

Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	All			CM			CF		
	Absolute	Relative	Likelihood	Absolute	Relative	Likelihood	Absolute	Relative	Likelihood
2's Y	− 149.14***	− 0.28***	− 2.16***						
	(7.71)	(0.01)	(0.27)						
cf	23.54**	0.08***	1.06***						
	(10.20)	(0.02)	(0.27)						
2's Y * cf	28.21	0.132*	− 0.23						
	(31.12)	(0.08)	(0.67)						
k_4				40.83***	0.04**	0.31	32.53**	0.00	0.33
				(12.88)	(0.02)	(0.29)	(15.68)	(0.03)	(0.42)
k_6				81.93***	0.07***	0.38	51.14***	0.00	0.56
				(13.29)	(0.02)	(0.26)	(13.96)	(0.03)	(0.49)
1/round	23.64	0.0469*	0.81*	46.44**	0.10***	1.48***	23.92	− 0.01	0.16
	(18.79)	(0.03)	(0.48)	(19.81)	(0.03)	(0.55)	(22.00)	(0.05)	(0.79)
Constant	− 48.33***	− 0.10***	− 1.22***	− 134.15***	− 0.20***	− 2.00***	− 54.43**	− 0.03	− 0.62
	(18.26)	(0.04)	(0.37)	(13.34)	(0.02)	(0.38)	(24.99)	(0.05)	(0.70)
#obs.	335	335	335	179	179	179	156	156	156
Log-likelihood	− 808.31	− 17.03	− 157.29	− 379.61	− 23.88	− 87.71	− 432.53	− 11.46	− 88.52

Models (1), (2), (4), (5), (7) and (8) are the mixed effects Tobit regressions of the relative and absolute transfer amounts, whereas (3), (6) and (9) are the random effects probit regressions of the likelihood $\mathbf{1}_{\{t_i>0\}}$ of positive transfer. The variable "2's Y"= 1 if role 2's action is Y. *, ** and ***Significant at 10%, 5% and 1%, respectively. Robust standard errors clustered by session in parentheses

Table 16 Tobit regressions of the payoff ratio

Variables	(1) CM	(2) CM	(3) CM	(4) CF	(5) CF	(6) CF
t2	− 0.51***	− 0.51***	− 0.17***	− 0.47***	− 0.47***	− 0.37***
	(0.15)	(0.16)	(0.05)	(0.15)	(0.11)	(0.06)
k4		1.64***	1.79***		1.51***	1.59***
		(0.18)	(0.16)		(0.12)	(0.18)
k6		3.36***	3.73***		2.80***	2.85***
		(0.24)	(0.30)		(0.13)	(0.05)
1/round	0.88	0.29	0.26	− 0.87	− 0.01	− 0.05
	(0.77)	(0.50)	(0.40)	(0.85)	(0.11)	(0.15)
t2 * k4			− 0.28			− 0.17
			(0.20)			(0.17)
t2 * k6			− 0.73**			− 0.11
			(0.35)			(0.27)
Constant	3.51***	2.09***	1.93***	3.98***	2.21***	2.18***
	(0.09)	(0.28)	(0.12)	(0.34)	(0.06)	(0.08)
Log likelihood	− 794.54	− 681.63	− 679.59	− 768.93	− 686.67	− 686.56
#obs	372	372	372	372	372	372
#subject pairs	62	62	62	62	62	62

*, ** and ***Significant at 10%, 5% and 1%, respectively. Robust standard errors clustered by session in parentheses

Table 17 Payoffs incorporating the average transfer from role 1: $(g_1 - \bar{t}_1, g_2 + \bar{t}_1)$

CF2					CF4					CF6				
	X (51)		Y (11)			X (52)		Y (10)			X (52)		Y (10)	
X(48)	149.1,	90.9	53.3,	26.7	X(53)	292.1,	107.9	60,	20	X(55)	437.6,	122.4	59.7,	20.3
Y(14)	95.4,	64.6	100,	100	Y(9)	244.3,	75.8	100,	100	Y(7)	376,	104	100,	100

#observations in parentheses

Table 18 Types in T1 and T2

	CM						CF					
	Role 1			Role 2			Role 1			Role 2		
	TY2	TX2	Other	TY2	TX2	Other	TY2	TX2	Other	TY2	TX2	Other
TY1	2	4	1	2	4	2	6	9	0	5	9	3
TX1	2	49	1	2	48	0	1	32	1	2	34	5
Other	0	3	0	1	3	0	0	12	1	1	3	0

Table 19 Reciprocity types in CM-T2 and CF-T2

		CM-T2			
		SR	WR	NR	Other
CF-T2	SR	14	3	11	0
	WR	2	2	2	0
	NR	1	5	18	0
	Other	1	1	2	0

Type "other" refers to role 1 who didn't experience (X, X)

A.2 Figures

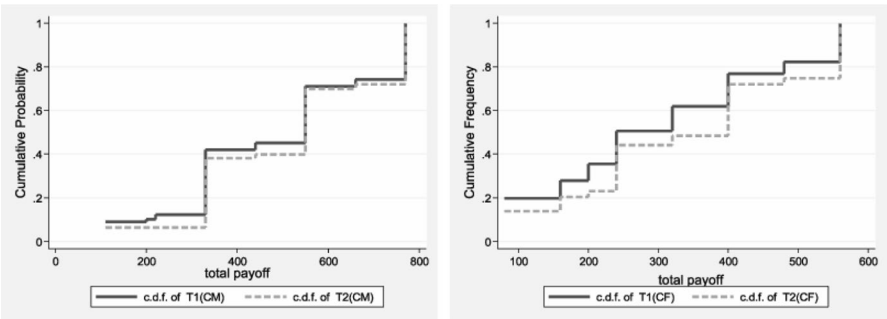

Fig. 4 Cumulative distributions of total payoffs in T1 and T2: CM (left) and CF (right)

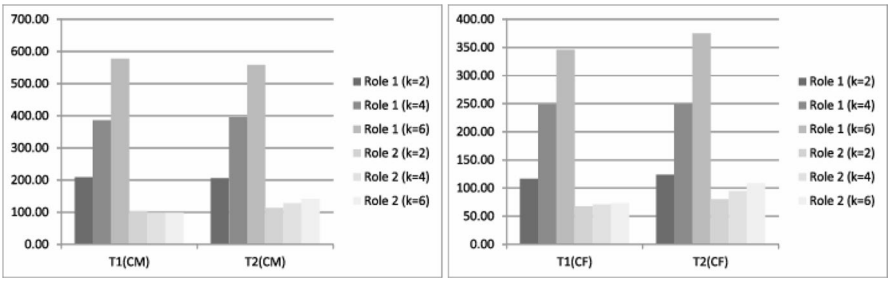

Fig. 5 Final payoffs u_i in CM (left) and CF (right): role 1 (dark) and role 2 (light)

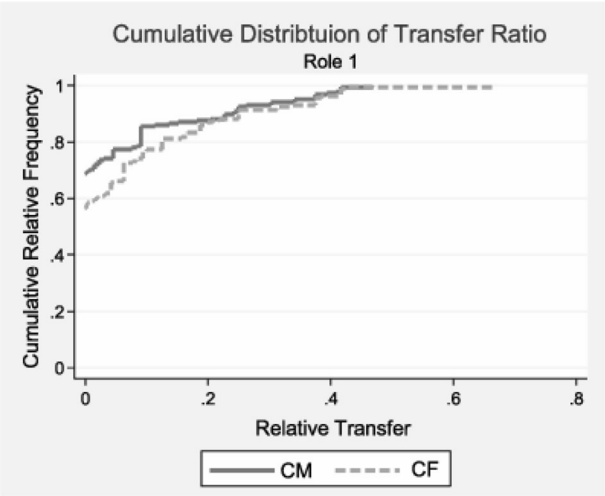

Fig. 6 Cumulative distributions of relative transfer by role 1 subjects

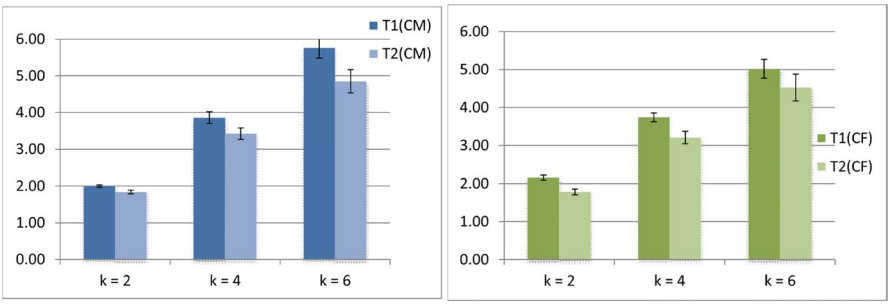

Fig. 7 Final payoff ratios u_1/u_2 in CM (left) and CF (right): T1 (dark) and T2 (light)

A.3 Effect of the Payoff formula in the instructions

This section examines the effects of including the payoff formula (1) in the instructions. There are a total of 106 subjects who participated in five sessions without the payoff formula but with the same task sequence as in the main experiments. Tables 20 and 21 describe the frequency of action Y by each role and the frequency of each action profile, respectively, in T1 and T2 with and without the payoff formula. We observe that role 1 chooses Y less often in every game with the formula, and that role 2 does so in four out of six games (CF2 and CM6). The effect is stronger in T2. In the case of action profiles, the efficient coordination profile (X, X) increases with the formula in every game, whereas the inefficient coordination profile decreases or does not change with the formula in every game. Again, these effects are generally stronger in T2. As seen in logit regressions reported in Tables 22 and 23, many of these changes are significant. In terms of

Table 20 Frequencies of *Y* with and without formula

	Role 1						Role 2					
T1	CF2	CF4	CF6	CM2	CM4	CM6	CF2	CF4	CF6	CM2	CM4	CM6
Without	0.34	0.51	0.43	0.19	0.15	0.17	0.38	0.28	0.38	0.09	0.21	0.21
53	(0.07)	(0.07)	(0.07)	(0.54)	(0.05)	(0.05)	(0.07)	(0.06)	(0.07)	(0.04)	(0.06)	(0.06)
With	0.27	0.42	0.27	0.06	0.11	0.11	0.31	0.21	0.31	0.05	0.13	0.13
62	(0.06)	(0.06)	(0.06)	(0.03)	(0.04)	(0.04)	(0.06)	(0.05)	(0.06)	(0.03)	(0.04)	(0.04)
T2												
Without	0.38	0.26	0.34	0.06	0.13	0.13	0.17	0.17	0.21	0.09	0.09	0.11
53	(0.07)	(0.06)	(0.07)	(0.03)	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)	(0.04)	(0.04)	(0.04)
With	0.23	0.15	0.11	0.00	0.05	0.06	0.18	0.16	0.16	0.05	0.05	0.10
62	(0.05)	(0.04)	(0.04)	(0.00)	(0.03)	(0.03)	(0.05)	(0.05)	(0.04)	(0.03)	(0.03)	(0.04)

Standard errors in parentheses

Table 21 Action profiles with and without formula

		CF2		CF4		CF6	
		Without	With	Without	With	Without	With
T1	(*X*, *X*)	0.42	0.45	0.32	0.45	0.38	0.53
	(*X*, *Y*)	0.25	0.27	0.17	0.13	0.19	0.19
	(*Y*, *X*)	0.21	0.24	0.40	0.34	0.25	0.16
	(*Y*, *Y*)	0.13	0.03	0.11	0.08	0.19	0.11
	Fisher's test	0.365		0.542		0.307	
T2	(*X*, *X*)	0.53	0.63	0.60	0.71	0.55	0.76
	(*X*, *Y*)	0.09	0.15	0.13	0.15	0.11	0.13
	(*Y*, *X*)	0.30	0.19	0.23	0.13	0.25	0.08
	(*Y*, *Y*)	0.08	0.03	0.04	0.02	0.09	0.03
	Fisher's test	0.32		0.459		0.033	
		CM2		CM4		CM6	
T1	(*X*, *X*)	0.77	0.89	0.64	0.77	0.64	0.77
	(*X*, *Y*)	0.04	0.05	0.21	0.11	0.19	0.11
	(*Y*, *X*)	0.13	0.06	0.15	0.10	0.15	0.10
	(*Y*, *Y*)	0.06	0.00	0.00	0.02	0.02	0.02
	Fisher's test	0.145		0.274		0.472	
T2	(*X*, *X*)	0.87	0.95	0.81	0.90	0.79	0.84
	(*X*, *Y*)	0.08	0.05	0.06	0.05	0.08	0.10
	(*Y*, *X*)	0.04	0.00	0.09	0.05	0.09	0.06
	(*Y*, *Y*)	0.02	0.00	0.04	0.00	0.04	0.00
	Fisher's test	0.254		0.318		0.409	

Table 22 Random effects logit regressions of action choice Y with and without formula

Variables	(1) CF role1	(2) CF role1	(3) CM role1	(4) CM role1	(5) CF role2	(6) CF role2	(7) CM role2	(8) CM role2
Formula	− 1.141**	− 0.506	− 1.386***	− 0.855*	− 0.825**	− 0.445	− 1.038	− 0.749
	(0.50)	(0.52)	(0.46)	(0.45)	(0.38)	(0.36)	(0.69)	(0.71)
t2	− 0.057	− 0.042	− 0.0808**	− 0.0699**	− 0.128***	− 0.118***	− 0.0630**	− 0.0564*
	(0.05)	(0.05)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
1/round	0.842***	0.893***	1.185**	1.227**	− 1.062***	− 1.082***	− 0.376	− 0.347
	(0.30)	(0.32)	(0.60)	(0.57)	(0.35)	(0.38)	(0.70)	(0.70)
t2 * formula		− 1.405***		− 1.371***		− 0.804***		− 0.625
		(0.31)		(0.39)		(0.16)		(0.61)
Constant	− 0.937**	− 1.035**	− 3.442***	− 3.546***	− 0.552	− 0.585	− 2.522***	− 2.566***
	(0.46)	(0.47)	(0.63)	(0.64)	(0.36)	(0.37)	(0.61)	(0.64)
Log likelihood	− 351.95	− 342.23	− 178.54	− 175.15	− 332.86	− 329.45	− 198.85	− 197.88
#obs	690	690	690	690	690	690	690	690
#subjects	115	115	115	115	115	115	115	115

The variable formula = 1 for sessions with the formula. *, ** and ***Significant at 10%, 5% and 1%, respectively. Robust standard errors clustered by session in parentheses

Table 23 Logit regressions of action profiles with and without formula

Variables	(1) CF yy	(2) CF yy	(3) CM yy	(4) CM yy	(5) CF xx/yy	(6) CF xx/yy	(7) CM xx/yy	(8) CM xx/yy
Formula	− 1.428***	− 0.950*	− 1.727**	− 0.99	0.417**	0.02	0.918**	0.596
	(0.54)	(0.56)	(0.83)	(0.86)	(0.17)	(0.18)	(0.37)	(0.37)
t2	− 0.09	− 0.08	0.00	0.01	0.04	0.03	0.0766***	0.0695***
	(0.07)	(0.07)	(0.05)	(0.05)	(0.03)	(0.03)	(0.01)	(0.01)
1/round	− 1.563***	− 1.637***	1.405**	1.416**	− 0.431**	− 0.442**	− 0.113	− 0.129
	(0.42)	(0.43)	(0.70)	(0.65)	(0.20)	(0.20)	(0.38)	(0.36)
t2 * formula		− 1.303***		Omitted		0.772***		0.696**
		(0.30)				(0.21)		(0.27)
Constant	− 2.006***	− 2.045***	− 4.784***	− 4.832***	0.340	0.383*	1.395***	1.432***
	(0.57)	(0.58)	(0.55)	(0.56)	(0.22)	(0.22)	(0.51)	(0.52)
Log likelihood	− 166.33	− 163.64	− 51.63	− 50.20	− 456.03	− 450.20	− 298.47	− 296.31
#obs	690	690	690	504	690	690	690	690
#subjects	115	115	115	115	115	115	115	115

*, ** and ***Significant at 10%, 5% and 1%, respectively. Standard errors clustered by session in parentheses

transfer, the inclusion of the payoff formula also has positive impact on the average transfer by role 1 as seen in the Tobit regressions reported in Table 24.

On the other hand, the redistribution scheme increases the choice of X even without the formula: Going from T1 to T2, the rate of X increases by 6.8 percentage points in CM (83% → 89.3% for role 1 and 83% → 90.3% for role (2), and by 13.2 percentage points in CF (57.3% → 67.3% for role 1 and 65.3% → 81.7% for role (2). However, the increase is smaller than the corresponding number with the formula reported in Sect. 6.2.

A.4 Proofs

Proof of Proposition 1 The utility function U_i is concave in the own transfer t_i so that the first-order condition fully characterizes the solution to the maximization problem. In particular, the solution is either at a corner ($t_i = 0$ or $t_i = g_i$) or in the interior ($t_i = \gamma_i(x) - g_j(x) + t_j$). When $\gamma_1(x) + \gamma_2(x) \neq g_1(x) + g_2(x)$, we cannot have both t_1 and t_2 as interior solutions. The first-order condition for t_i against $t_j = 0$ or $t_j = g_j$ then yields the relationship between (γ_1, γ_2) and t_i in (4).

Reasoning for Hypothesis 1 is as follows:

Table 24 Mixed effects Tobit regressions of transfer with and without formula

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
					Conditional on (X, X)			
	CF	CF	CM	CM	CF	CF	CM	CM
Formula	− 6.332	− 31.94*	− 13.3	− 51.41**	− 5.67	− 24.79	− 15.23	− 57.88**
	(7.94)	(18.05)	(15.17)	(25.64)	(6.33)	(17.52)	(13.68)	(24.27)
1/round	9.765	10.08	41.75***	41.87***	7.939	8.215	36.62**	36.74**
	(13.48)	(13.49)	(13.96)	(13.97)	(15.30)	(15.32)	(14.67)	(14.65)
Role	70.99***	50.22***	82.77***	50.53**	80.93***	63.55***	81.71***	45.27**
	(14.68)	(15.36)	(18.61)	(19.64)	(16.04)	(16.76)	(19.98)	(22.07)
Role * formula		41.15**		63.76***		31.21		71.34***
		(20.47)		(23.46)		(23.02)		(25.14)
Constant	− 89.54***	− 77.02***	− 156.2***	− 136.8***	− 83.46***	− 73.09***	− 144.6***	− 122.9***
	(16.97)	(16.65)	(24.65)	(24.58)	(19.21)	(19.90)	(22.29)	(22.71)
	− 1196.78	− 1194.73	− 1026.13	− 1024.40	− 865.89	− 864.96	− 943.98	− 941.82
Observations	690	690	690	690	438	438	596	596
Number of groups	230	230	230	230	200	200	225	225

*, ** and ***Significant at 10%, 5% and 1%, respectively. role = 1 if role 1. Standard errors clustered by session in parentheses

Table 25 Payoff profiles including SPE transfer σ

	CM					CF				
	X		Y			X		Y		
X	$2b+c_1+c_2-\mu_1$,	μ_1	b,	c_2	X	$2b+c_1+c_2-\nu_1$,	ν_1	b,	c_2
Y	$b+c_1-\nu_1$,	ν_1	a,	a	Y	$b+c_1-\nu_1$,	ν_1	a,	a

1a. When $c_1 > a$ and $x = (Y, X)$, $\gamma_1(x) = \nu_1$ and $\gamma_2(x) = 0$ since $g_1(x) = c_1 > a > b = g_2(x)$. It follows that $\sigma_1(x) = \nu_1 - g_2(x)$ and $\sigma_2(x) = 0$ by (4) and hence that the payoff profile including the SPE transfer at $x = (Y, X)$ is given by $(b + c_1 - \nu_1, \nu_1)$. Hence, when $c_1 > a$ and $\nu_1 > a$, then player 2's choice of Y is strictly dominated and (X, X) is the unique SPE action profile. On the other hand, when $c_1 \leq a$ or $\nu_1 \leq a$, (X, X) and (Y, Y) are both SPE action profiles. See Table 25.

1b. By Fig. 1, at most one player i chooses $\sigma_i(x) > 0$ when $\gamma_1(x) + \gamma_2(x) < g_1(x) + g_2(x)$. If in addition $\gamma_1(x) = \gamma_2(x)$, then $\sigma_2(x) = 0$ while $\sigma_1(x) > 0$ if $\gamma_1(x) > g_2(x)$ and $\sigma_1(x) = 0$ if $\gamma_1(x) < g_2(x)$. Since $\gamma_1(x) > g_2(x)$ implies $g_1(x) > a$, $\sigma_1(x) > 0$ if and only if $x = (X, X)$ or (Y, X). Regarding the comparison between CM-T2 and CF-T2, note that at $x = (X, X)$, player 1 chooses $\sigma_1(x) > 0$ if $\gamma_1(x) = \mu_1 > b + c_2 = g_2(x) > a$ in CM and $\gamma_1(x) = \nu_1 > b + c_2 = g_2(x) < a$ in CF. Hence, if $\mu_1 \leq b + c_2 = 110$ and $\nu_1 > 80$, player 1 chooses $\sigma_1(x) > 0$ at $x = (X, X)$ only in CF.

1c. When $\gamma_1(x) + \gamma(x) < g_1(x) + g_2(x)$, $\sigma_1(x) > 0$ if $\gamma_2(x) > g_2(x)$. Since we specify $g_2(x)$ to be independent of the degree k of inequality in each class of games, $\sigma_1(x)$ is also independent of k. This further implies that the action choice is independent of k as well in each class of games.

□

Equilibrium under distributive social preferences

Let e_i^{T0} denote player i's optimal choice in the dictator task T0, and E^{T1} and E^{T2} denote the set of (pure) NE and SPE action profiles in the inequality game G in tasks T1 and T2, respectively.

(1) Inefficiency aversion

In T0, the optimal action for player 1 is $e_1^{T0} = (X, X)$ regardless of κ_1, and for player 2,

$$e_2^{T0} = \begin{cases} (X,X) \text{ if } \kappa_2 > \frac{a-b-c_2}{2b+c_1+c_2-2a}, \\ (Y,Y) \text{ if } \kappa_2 < \frac{a-b-c_2}{2b+c_1+c_2-2a}. \end{cases}$$

In T1,

$$E^{T1} = \begin{cases} \{(X,X)\} & \text{if } b + c_1 > 2a \text{ and } \kappa_2 > \frac{a-b}{b+c_1-2a}, \\ \{(X,X),(Y,Y)\} & \text{if } b + c_1 \leq 2a, \text{ or if } b + c_1 > 2a \text{ and} \kappa_2 < \frac{a-b}{b+c_1-2a}. \end{cases} \quad (7)$$

Since the threshold $\frac{a-b}{b+c_1-2a}$ decreases as c_1 increases (or $k = \frac{b+c_1}{b+c_2}$ increases), (7) implies that (X, X) is the unique NE for a larger set of κ_2 for a larger k, implying Hypothesis 2c.[47] In T2, the transfer equals zero at any action profile and the set of SPE action profiles is as given in (7): $E^{T2} = E^{T1}$. We hence have Hypotheses 2a and 2b.

(2) Inequality aversion

In T0,

$$e_1^{T0} = \begin{cases} (X,X) \text{ if } \lambda_1 < \frac{b+c_1-a}{c_1-c_2}, \\ (Y,Y) \text{ if } \lambda_1 > \frac{b+c_1-a}{c_1-c_2}, \end{cases}$$

and

$$e_2^{T0} = \begin{cases} (X,X) \text{ if } \lambda_2 < \frac{b+c_2-a}{c_1-c_2}, \\ (Y,Y) \text{ if } \lambda_2 > \frac{b+c_2-a}{c_1-c_2}, \end{cases}$$

In T1,

$$E^{T1} = \begin{cases} \{(Y,Y)\} & \text{if } \lambda_1 > \frac{b}{b-c_2} \text{ or } \lambda_2 > \frac{b}{c_1-b}, \\ \{(X,X),(Y,Y)\} & \text{if } \lambda_1 \leq \frac{b}{b-c_2} \text{ and } \lambda_2 \leq \frac{b}{c_1-b}. \end{cases} \quad (8)$$

Since the threshold $\frac{b}{c_1-b}$ decreases as c_1 increases (or k increases), (Y, Y) is the unique NE for a larger set of λ_2 for a larger k, implying the first part of Hypothesis 3c. In T2, if $\lambda_1 < \frac{1}{2}$, then no transfer takes place in SPE, and the SPE action profile in stage 1 is the same as in T1: $E^{T2} = E^{T1}$. If $\lambda_1 > \frac{1}{2}$, then (x, t) is an SPE if and only if x is a NE of the following game of identical-interest:

P1 \ P2	X		Y	
v	$2b + c_1 + c_2$	$2b + c_1 + c_2$	$b + c_2$	$b + c_2$
Y	$b + c_1$	$b + c_1$	$2a$	$2a$

and the transfer function t in SPE satisfies

$$t_1(x) - t_2(x) = \frac{g_1(x) - g_2(x)}{2} \text{ for every } x.$$

Hence, Hypothesis 3b as well as the second part of Hypothesis 3c hold. Furthermore,

[47] $b + c_1 > 2a$ holds in all but one (CF2) of our parameter specifications. See Table 3.

$$E^{T2} = \begin{cases} \{(X,X),(Y,Y)\} & \text{if } 2a \geq b + c_1, \\ \{(X,X)\} & \text{if } 2a < b + c_1. \end{cases}$$

Except for CF2, $2a < b + c_1$ holds and hence (X, X) is the unique SPE action profile. This along with (8) implies that (X, X) is played more often in T2 than in T1 (Hypothesis 3a).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10683-021-09719-6.

Acknowledgements We are very grateful to the Editor and a referee for the comments that led to a significant improvement of the paper. Financial support from the JSPS (Grant Numbers: 22330061, 23530216, 24330064, 24653048, 15K13006, 15H03328, 16H03597, 16K17088, 15H05728, 20H05631) and the Joint Usage/Research Center at ISER, Osaka University, and the International Joint Research Promotion Program of Osaka University is gratefully acknowledged.

References

Anbarci, N., Feltovich, N., & Gurdal, M. Y. (2018). Payoff inequity reduces the effectiveness of correlated-equilibrium recommendations. *European Economic Review, 108,* 172–190.

Anderson, L. R., Mellor, J. M., & Milyo, J. (2006). Induced heterogeneity in trust experiments. *Experimental Economics, 9*(3), 223–235.

Andreoni, J., Harbaugh, W., & Vesterlund, L. (2003). The carrot or the stick: Rewards, punishments, and cooperation. *American Economic Review, 93*(3), 893–902.

Azrieli, Y., Chambers, C. P., & Healy, P. J. (2018). Incentives in experiments: A theoretical analysis. *Journal of Political Economy, 126*(4), 1472–1503.

Belafoutas, L., Kocher, M. G., Putterman, L., & Sutter, M. (2013). Equality, equity and incentives: An experiment. *European Economic Review, 60,* 32–51.

Bone, J., Drouvelis, M., & Ray, I. (2013). Coordination in 2×2 games by following recommendations from correlated equilibria. Working Paper.

Buckley, E., & Croson, R. (2006). The poor give more: Income and wealth heterogeneity in the voluntary provision of linear public goods. *Journal of Public Economics, 90*(5), 935–955.

Cachon, G. P., & Camerer, C. F. (1996). Loss-avoidance and forward induction in experimental coordination games. *Quarterly Journal of Economics, 111*(1), 165–194.

Cason, T. N., & Sharma, T. (2007). Recommended play and correlated equilibria: An experimental study. *Economic Theory, 33*(1), 11–27.

Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica, 74*(6), 1579–1601.

Cooper, R., DeJong, D. V., Forsythe, R., & Ross, T. W. (1992). Communication in coordination games. *Quarterly Journal of Economics, 107*(2), 739–771.

Cooper, R., DeJong, D. V., Forsythe, R., & Ross, T. W. (1993). Forward induction in the battle-of-the-sexes games. *American Economic Review, 83*(5), 1303–1316.

Cooper, R. W., DeJong, D. V., Forsythe, R., & Ross, T. W. (1990). Selection criteria in coordination games: Some experimental results. *American Economic Review, 80*(1), 218–233.
Crawford, V. P., Gneezy, U., & Rottenstreich, Y. (2008). The power of focal points is limited: Even minute payoff asymmetry may yield large coordination failures. *American Economic Review, 98*(4), 1443–1458.
Dekel, S., Fischer, S., & Zultan, R. (2017). Potential pareto public goods. *Journal of Public Economics, 146,* 87–96.
Duffy, J., & Feltovich, N. (2010). Correlated equilibria, good and bad: An experimental study. *International Economic Review, 51*(3), 701–721.
Duffy, J., & Ochs, J. (2009). Cooperative behavior and the frequency of social interaction. *Games and Economic Behavior, 66*(2), 785–812.
Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior, 47*(2), 268–298.
Erkal, N., Gangadharan, L., & Nikiforakis, N. (2011). Relative earnings and giving in a real-effort experiment. *American Economic Review, 101,* 3330–3348.
Evdokimov, P., & Rustichini, A. (2016). Forward induction: Thinking and behavior. *Journal of Economic Behavior & Organization, 128,* 195–208.
Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review, 90*(4), 980–994.
Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature, 422*(6928), 137–40.
Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics, 10*(2), 171–178.
Gangadharan, L., Nikiforakis, N., & Villeval, M. C. (2017). Normative conflict and the limits of self-governance in heterogeneous populations. *European Economic Review, 100,* 143–156.
Goeree, J. K., & Holt, C. A. (2005). An experimental study of costly coordination. *Games and Economic Behavior, 51,* 349–364.
Greiner, B., Ockenfels, A., & Werner, P. (2012). The dynamic interplay of inequality and trust? An experimental study. *Journal of Economic Behavior & Organization, 81*(2), 355–365.
Hofmeyr, A., Burns, J., & Visser, M. (2007). Income inequality, reciprocity and public good provision: An experimental analysis. *South African Journal of Economics, 75,* 508–520.
Houser, D., Xiao, E., McCabe, K., & Smith, V. (2008). When punishment fails: Research on sanctions, intentions and non-cooperation. *Games and Economic Behavior, 62*(2), 509–532.
Huck, S., & Müller, W. (2005). Burning money and (pseudo) first-mover advantages: An experimental study on forward induction. *Games and Economic Behavior, 51*(1), 109–127.
Masclet, D., Noussair, C., Tucker, S., & Villeval, M.-C. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. *American Economic Review, 93*(1), 366–380.
Ohtake, F., Kinari, Y., Mizutani, N., & Mori, T. (2013). Income, giving, and egalitarianism: A real-effort experiment in Japan. *Journal of Behavioral Economics and Finance, 6,* 81–84 (**(in Japanese)**).
Oxoby, R. J., & Spraggon, J. (2013). A clear and present minority: Heterogeneity in the source of endowments and the provision of public goods. *Economic Inquiry, 51*(4), 2071–2082.
Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review, 83*(5), 1281–1302.
Rodriguez-Lara, I. (2018). No evidence of inequality aversion in the investment game. *PLOS ONE, 13*(10), 1–16.
Straub, P. G. (1995). Risk dominance and coordination failures in static games. *Quarterly Review of Economics and Finance, 35*(4), 339–363.
Uler, N. (2011). Public goods provision, inequality and taxes. *Experimental Economics, 14,* 287–306.
Van Huyck, J. B., Battalio, R. C., & Beil, R. O. (1990). Tacit coordination games, strategic uncertainty, and coordination failure. *American Economic Review, 80*(1), 234–248.