A Online-Appendix

A.1 Deviations in the Stage Game

In a given encounter, how would the belief that j may deviate affect i's willingness to obey the device? Depending on whether i received the message to (a) *claim* or to (b) *concede*, we distinguish two cases: (a) If $(m_i, m_j) = (claim, concede)$, the bourgeois equilibrium calls for $a_i = claim$. In expectations, i prefers to claim as long as j's deviation propensity w^j does not exceed $\overline{w}^j_{claim} = \frac{h}{h+l}$.[43] A player will thus claim her preferred action as long as $(1 - w^j)h$ gives her more utility than $w^j l$. In order to refrain from claiming, a player would need a very high belief w^j about her counterpart's deviation proneness. (b) If $(m_i, m_j) = (concede, claim)$, the bourgeois equilibrium calls for $a_i = concede$. In expectations, i prefers to concede as long as w^j does not exceed $\overline{w}^j_{concede} = \frac{l}{h+l}$.[44] With $0 < l < h$ already a small belief about j's deviation probability could result in i's not conceding. The threshold is the smaller, the larger the difference between l and h.[45]

A.2 Alternative Payoffs for Mutual Conceding

Consider an alternative stage game, where $(concede, concede)$ yields a mutual payoff of l (instead of a payoff of 0 as in G):

		player 2	
		claim	*concede*
player 1	*claim*	$0, 0$	h, l
	concede	l, h	l, l

Figure A1: Stage Game

As in G, there are two pure Nash equilibria $e_1 = (claim, concede)$ and $e_2 = (concede, claim)$. The mixed equilibrium e_{mix} is constituted by playing the action *claim* with $P_{mix}(claim) = \frac{h-l}{h}$ and results in an expected payoff of $E_{mix} = l$, which equals the lower payoff in the pure Nash equilibria.[46]

[43] $E(a_i = m_i | m_i = claim) \geq E(a_i \neq m_i | m_i = claim) \Rightarrow w^j 0 + (1 - w^j)h \geq w^j l + (1 - w^j)0$. For the parameters of the experiment, $h = 10$ and $l = 1$, a player would only refrain from claiming if she believed her counterpart's deviation probability to be larger than $\overline{w}^j_{claim} = \frac{10}{11}$. Note that an inequality averse player would perceive the difference between earning monetary payoff h and l even more starkly, and would thus need an even *higher* belief w^j to refrain from claiming.

[44] $E(a_i = m_i | m_i = concede) \geq E(a_i \neq m_i | m_i = concede) \Rightarrow w^j 0 + (1 - w^j)l \geq w^j h + (1 - w^j)0$. For the parameters of the experiment, $h = 10$ and $l = 1$, a player would refrain from conceding if she believed her counterpart's deviation probability to be larger than $\overline{w}^j_{concede} = \frac{1}{11}$. Note that an inequality averse player would perceive the difference between l and h more pronouncedly and thus would need an even *lower* belief w^j to refrain from conceding.

[45] Note that the tolerance thresholds $\overline{w}^j_{claim}$ and $\overline{w}^j_{concede}$ coincide with the mixing probabilities in the mixed equilibrium of G_ϕ. When a player chooses $a_i = claim$ with $P(claim) = \frac{h}{h+l}$ this is exactly the threshold where the counterpart is indifferent between both actions *claim* and *concede*. Until this threshold, the player who gets favored by the device always wants to follow the recommendation.

[46] The probabilities of the different outcomes in the mixed equilibrium are: $P(claim, claim) = \frac{(h-l)^2}{h^2}$, $P(claim, concede) =$

If $(m_i, m_j) = (claim, concede)$, the bourgeois equilibrium calls for $a_i = claim$. In expectations, i prefers to claim as long as j's deviation propensity w^j does not exceed $\overline{w}^j_{claim} = \frac{h-l}{h}$.[47] For the parameters of the experiment, $h = 10$ and $l = 1$, a player would only refrain from claiming if she believed her counterpart's deviation probability to be larger than $\overline{w}^j_{claim} = \frac{9}{10}$ (instead of $\frac{10}{11}$ in G).

If $(m_i, m_j) = (concede, claim)$, the bourgeois equilibrium calls for $a_i = concede$. In expectations, i prefers to concede as long as w^j does not exceed $\overline{w}^j_{concede} = \frac{l}{h}$. [48] For the parameters of the experiment, $h = 10$ and $l = 1$, a player would refrain from conceding if she believed her counterpart's deviation probability to be larger than $\overline{w}^j_{concede} = \frac{1}{10}$ (instead of $\frac{1}{11}$ in G).

In sum, the thresholds are very similar to G and with $0 < l < h$ already a small belief about j's deviation probability could result in i's refusal to concede. In the supergame, while E_{θ_i} is not affected by the parameter change, $E_{mix} = l$ is slightly higher than in G and equals the expected payoff of rank N in the bourgeois equilibrium.

$\frac{hl-l^2}{h^2}$, $P(concede, claim) = \frac{hl-l^2}{h^2}$ and $P(concede, concede) = \frac{l^2}{h^2}$.

[47] $E(a_i = m_i | m_i = claim) \geq E(a_i \neq m_i | m_i = claim) \Rightarrow w^j 0 + (1 - w^j)h \geq w^j l + (1 - w^j)l$.

[48] $E(a_i = m_i | m_i = concede) \geq E(a_i \neq m_i | m_i = concede) \Rightarrow w^j l + (1 - w^j)l \geq w^j h + (1 - w^j)0$.

A.3 Instructions

*Note: The text below shows the instructons of the **no-T** treatment, on which all other treatments build. The additional text in red was only included in instructions for the **T-direct** treatment. The additional text in blue was only included in instructions for the **T-pool** treatment. The additional text in purple was only included in instructions for the **T-admin** treatment. Instructions displayed here are a translation into English.[49] Original instructions where in German and are available from the authors upon request.*

Welcome to our experiment!

If you read the following instructions carefully, you can earn a substantial sum of money, depending on your decisions. It is therefore very important that you read these instructions carefully.

Absolutely no communication with the other participants is allowed during the experiment. Anyone disobeying this rule will be excluded from the experiment and all payments. Should you have any questions, please raise your hand. We will then come to you.

During the experiment, we will speak not of Euros, but of points. Your entire income will therefore initially be calculated in points. The total number of points accumulated by you during the experiment will be paid out to you in Euros at the end, at a rate of:

25 points = 1 Euro.

At the end of the experiment, you will be paid, in cash, the number of points you will have earned during the experiment. In addition to this sum, you will receive payment of 4 Euro for showing up at this experiment.

The experiment consists of at least 50 periods.

After period 50, a draw will decide in each period whether there shall be a further period. With a probability of 75%, there will be a period 51. Should there be a period 51, there will be a period 52 as well, once again with a probability of 75%, etc.

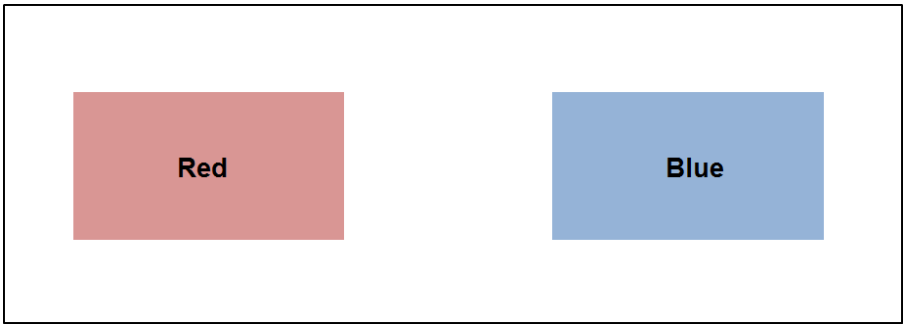
At the beginning of the experiment, participants will be randomly divided into groups of six. Apart from you, your group will therefore be made up of another 5 members. The constellation of your group of six will remain unchanged throughout the entire experiment.

[49] We thank Brian Cooper from the MPI for Collective Goods for the translation.

Also at the beginning of the experiment, the computer will name the participants of each group of six, assigning to each a randomly drawn letter (a, b, c, d, e, or f). Each participant in the group is equally likely to receive a particular letter (a, b, c, d, e, f). Each letter is distributed once in each group of six.

In each period, you will interact with exactly one of the other participants from your group. The computer will randomly determine at the beginning of each period who that other player is. The other 4 participants in your group of six will each also be randomly matched with another participant from the group. In total, there will hence be three parallel encounters in your group in each period.

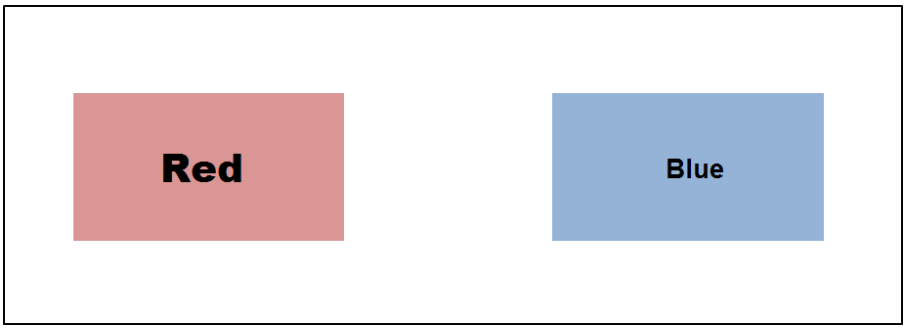
In each period, your task is to choose one of two decision fields:



How many points you earn in a period depends on your decision as well as on the decision of the participant with whom you are interacting.

- If you choose “Red” and the other participant chooses “Blue”, you will earn 10 points, and the other participant will earn 1 point.
- If you choose “Blue” and the other participant chooses “Red”, you will earn 1 point, and the other participant will earn 10 points.
- If both participants choose “Red”, you will both earn 0 points.
- If both participants choose “Blue”, you will both earn 0 points.

In each period, one of these fields will be in bold:



Whenever you see the field "Red" in bold, the other participant sees the field "Blue" in bold, and vice versa. You are free to decide whether you wish to follow the marking or not.

The computer decides on the basis of your letter which field is in bold. Whichever participant's letter comes first in the alphabet sees the field "Red" in bold. If, for example, the computer assigned you the letter c at the beginning of the experiment, and you interact with participant d, e, or f, you will see "Red" in bold. If you interact with participant a or b, however, you will see "Blue" in bold.

For example, if you were assigned the letter a at the start, "Red" will be in bold in all periods. If you are participant f, "Blue" will always be in bold, etc.

Only once both participants have made their decisions will you find out what the other participant has chosen.

At the end of each period, your computer screen will give you an overview of:

- which field you have opted for;
- which field the other participant has chosen;
- the income you and the other participant have each earned in this period;
- how the participants of the other encounters have chosen.

Further, you have the chance to transfer to the other participant any part of your income from the current period. To do this, enter on your screen the number of points you wish to transfer to the other participant, and confirm your entry by clicking "Continue". You are free to decide whether or not you wish to transfer points and, if you do, how many points you wish to donate.

Further, you have the chance to transfer to the other participants any part of your income from the current period. To do this, enter on your screen the number of points you wish to transfer to the

other participants, and confirm your entry by clicking "Continue". You are free to decide whether or not you wish to transfer points and, if you do, how many points you wish to donate.

If you earned 10 points in the current period, your transfer goes into a pool, whose content is payed out equally to all participants who earned 1 point in the current period. If you earned 1 point in the current period, your transfer goes into a pool, whose content is payed out equally to all participants who earned 10 point in the current period.

Subsequently, the computer will automatically transfer some part of your income from the current period to the other participant. The amount transferred can vary from period to period.

In addition, your computer screen will give you an overview of:

- the number of points transferred by you to the other participant;
- the number of points the other participant has transferred to you;
- the income you have earned in this period, after transfers;
- the number of points transferred by you to the respective pool;
- the number of points payed out to you from the respective pool;
- the income you have earned in this period, after transfers;
- the number of points the computer has automatically transferred to the other participant;
- the number of points the computer has automatically transferred to you from the other participant;
- the income you have earned in this period, after transfers;
- the total income you have made up to now;
- the total income each of the other participants in your group of six has made so far

In the next period, you will once again be randomly matched with one other participant from your group of six, with whom you will then interact.

Do you have any questions? If yes, please raise your hand. We will come to you.

A.4 Post-Experimental Tests

Other-regarding preferences. Other-regarding preferences were measured using the social value orientation (SVO) ring measure (**?**). The ring measure consists of 15 modified dictator games in which players allocate money between themselves and another player. Players are characterized on a continuous type space ranging from competitiveness and individualism to prosociality and altruism. To make the *Social Value Orientation (SVO)* comparable to the other scales, we divide the ring degree by 45, such that 0 describes a perfectly selfish (0 degree) individual, and 1 a perfectly pro-social (45 degree) individual.

Risk. The risk elicitation consisted of 1 general risk question and 6 domain-specific questions. The general question read: "Are you, generally speaking, a person willing to take risks or do you rather try to avoid risks?" The domain-specific questions read: "How would you rate your willingness to take risks in the following domains...(i) when driving a car, (ii) in financial investments, (iii) in leisure and sports, (iv) in your carreer, (v) concerning your health, (vi) when trusting unfamiliar people?". There are 10 answer options ranging from *not at all willing to take risks* to *very willing to take risks*. We use the arithmetic mean over the 7 questions as our measure of an individual's risk attitude. The scale runs from -1 "very risk averse" to 1 "very risk seeking".

Trust. The trust questions read: "Please rate the following three statements: (i) Generally, people can be trusted. (ii) Nowadays you cannot trust anybody. (iii) When dealing with strangers it's better to be careful before trusting them." The answer options are: *fully agree*, *rather agree*, *rather disagree*, *fully disagree*. The composite trust measure is the arithmetic mean of the three question whereby the first is coded negatively, and the other two positively. The scale runs from -1 "very low trust" to 1 "very high trust".

.

A.5 Additional Results

Table A1: Claiming and Conceding over Treatments

		$m = concede$								
		$a = claim$			$a = concede$					
$m = claim$	$a = claim$	nT	51%	1222	**nT**	**42%**	**1011**	**nT**	**93%**	**2233**
		Td	29%	704	**Td**	**63%**	**1516**	**Td**	**93%**	**2220**
		Tp	33%	790	**Tp**	**58%**	**1381**	**Tp**	**90%**	**2171**
		Ta	21%	496	**Ta**	**71%**	**1711**	**Ta**	**92%**	**2207**
	$a = concede$	nT	5%	118	nT	2%	49	nT	7%	167
		Td	4%	95	Td	4%	85	Td	8%	180
		Tp	6%	151	Tp	3%	78	Tp	10%	229
		Ta	3%	73	Ta	5%	120	Ta	8%	193
		nT	56%	1340	**nT**	**44%**	**1060**			
		Td	33%	799	**Td**	**67%**	**1601**			
		Tp	39%	941	**Tp**	**61%**	**1459**			
		Ta	24%	569	**Ta**	**76%**	**1831**			

Relative and absolute frequencies of stage 1 behavior in *no-T* (nT), *T-direct* (Td), *T-pool* (Tp), and *T-admin* (Ta). We report data from periods $1-50$. Per treatment, we thus have 4800 observations (96 subjects $\times$ 50 periods) on the variable $a = \{claim, concede\}$, half of which saw the message $m = claim$ (i.e. red field shown in bold) and $m = concede$ (i.e. blue field shown in bold), respectively. Behavior consistent with the bourgeois equilibrium ($a_i = m_i$) is shown in **bold**.

Table A2: Individual Characteristics - Summary Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
SVO	384	.31	.50	-3.20	2.34
Trust	384	-.25	.28	-1	.33
Risk	384	-.12	.32	-1	1
Age	384	23.84	5.42	16	65
Female	384	.57	.49	0	1
Siblings	384	1.58	1.14	0	6

The *Trust* scale runs from -1 "very low trust" to 1 "very high trust". The *Risk* scale runs from -1 "very risk averse" to 1 "very risk seeking". To make the *Social Value Orientation (SVO)* comparable to the other scales, we divide the ring degree (**?**) by 45, such that 0 describes a perfectly selfish (0 degree) individual, and 1 a perfectly pro-social (45 degree) individual. See Appendix A.4 for further detail.

Table A3: Individual Characteristics - Pairwise Correlations

	SVO	*Trust*	*Risk*	*Age*	*Female*	*Siblings*
SVO	1					
Trust	0.124**	1				
Risk	0.108**	0.189***	1			
Age	-0.144***	0.031	0.031	1		
Female	-0.085*	-0.020	-0.239***	0.031	1	
Siblings	0.052	0.088*	-0.020	0.048	0.046	1

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

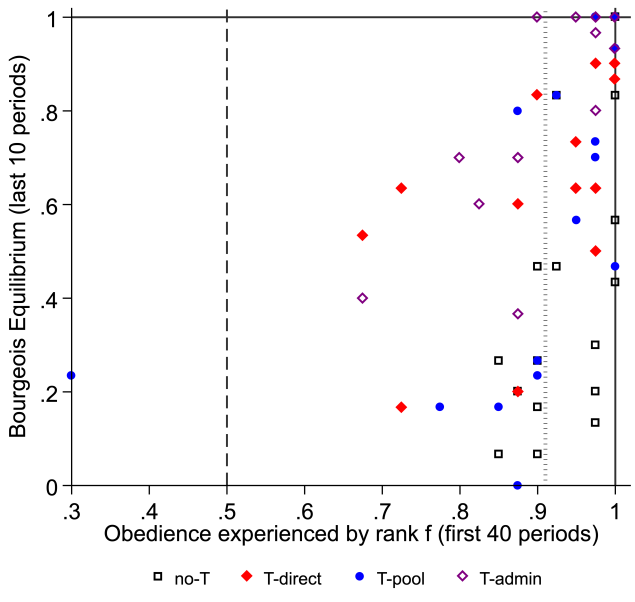


Figure A2: Conditions for Bourgeois Equilibrium, by Groups

Each dot depicts one group. There are 16 groups per treatment. Mean obedience of others experienced by the lowest-ranked player f (in the first 40 periods), and relative frequency of full coordination on the *bourgeois equilibrium* (in the last 10 periods). The dotted line at .91 (dashed line at .5) denotes the minimum obedience necessary for the existence of the *bourgeois equilibrium* in the absence of transfer opportunities (in the presence of transfer opportunities).

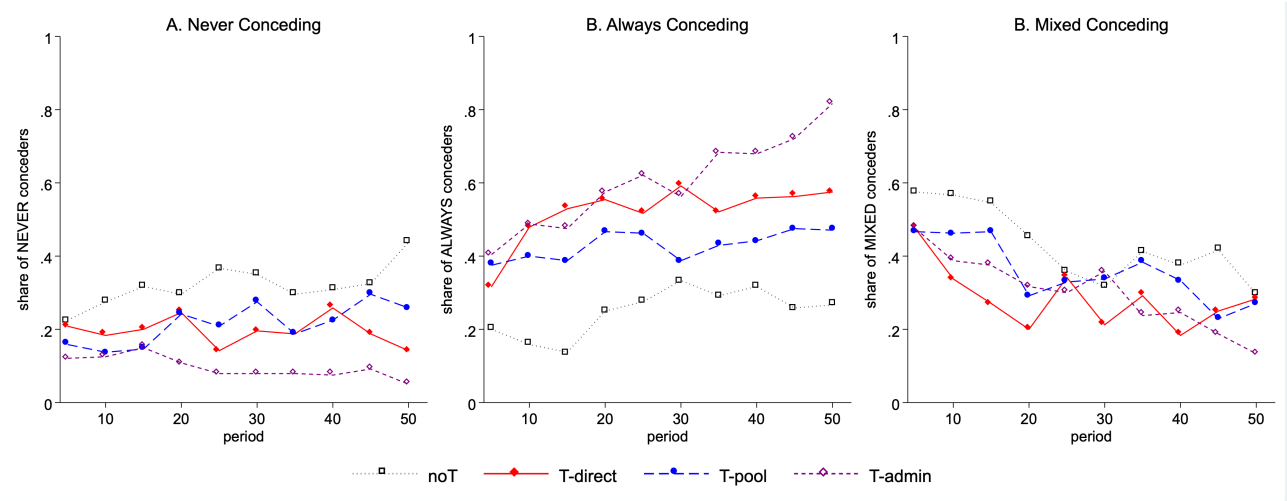


Figure A3: Conceding over Time

Relative frequency of individuals who, within a block of 5 periods, (A) never conceded, (B) always conceded, or (C) sometimes conceded when receiving the message $m = concede$ (i.e. blue field shown in bold). Within each block, individuals are weighted by the number of periods they received $m = concede$.

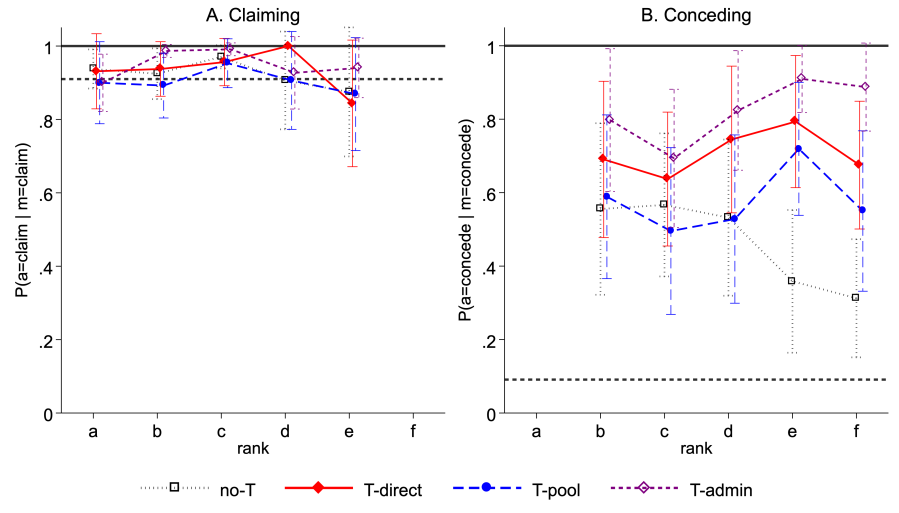


Figure A4: Compliance with the Status Quo (last 10 periods)

Mean relative frequency of complying with the exogenous recommendation when one's message is (A) *claim* (i.e. red field shown in bold) or (B) *concede* (i.e. blue field shown in bold). The 95% confidence intervals capture the between-group variation around the treatment means. In the experiment, the two actions *claim* and *concede* are labeled as *red* and *blue*, respectively.

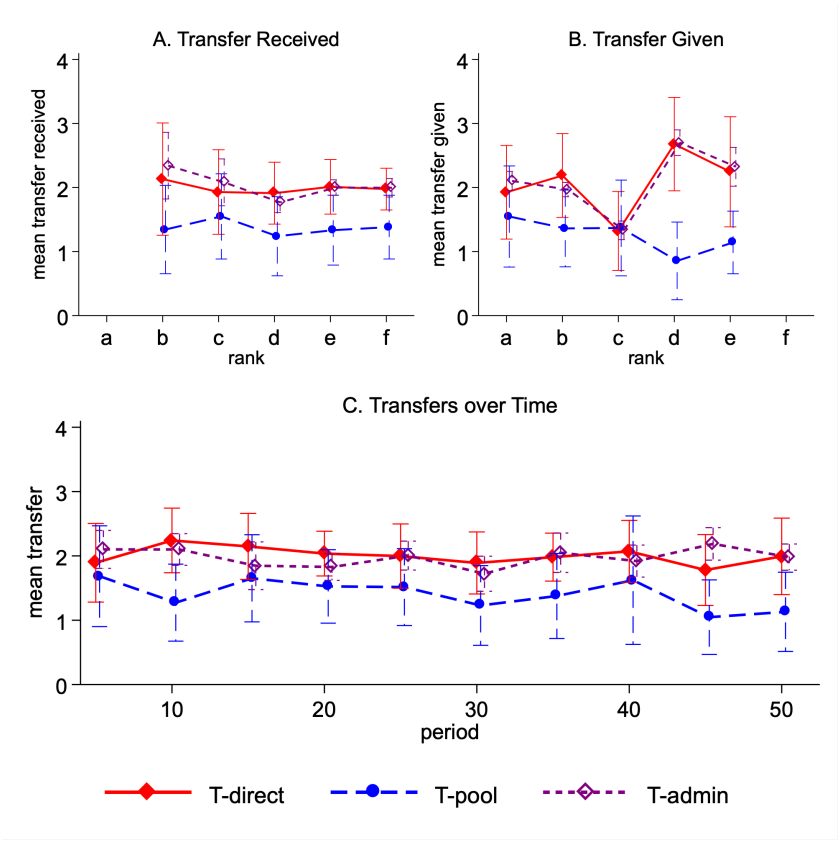


Figure A5: Transfers

Transfers in *T-direct* and *T-pool* were choices of the experimental participants. Transfers in *T-admin* were randomly drawn from the empirical distribution of transfers made by the experimental participants in *T-direct*. Transfer after both players played $a_i = m_i$, averaged first by group and then by treatment. The 95% confidence intervals capture the between-group variation around the treatment means.

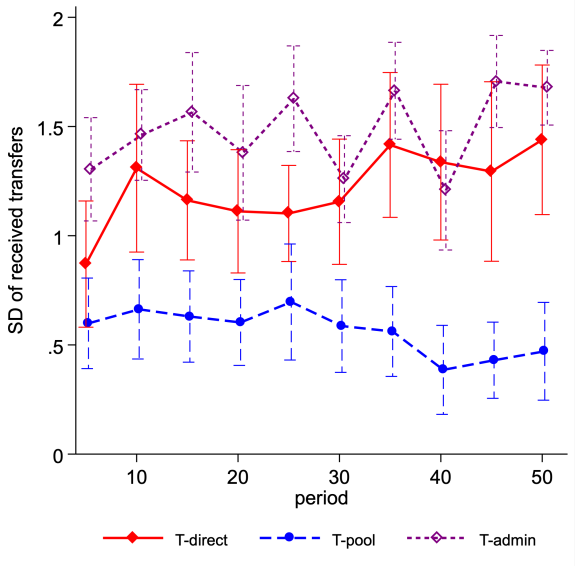


Figure A6: Period-by-period variation of transfers received

Period-by-period standard deviation (SD) of transfers received (after both players played $a_i = m_i$) by an individual within blocks of 5 periods, averaged first by group and then by treatment. The 95% confidence intervals capture the between-group variation around the treatment means of the period-by-period SD.

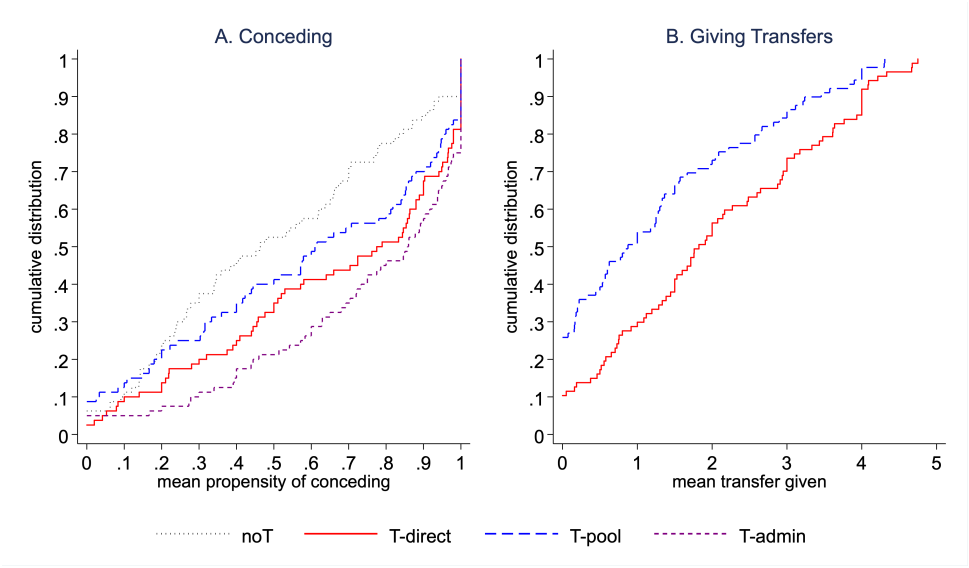


Figure A7: Individual willingness to (A) concede and (B) transfer

Panel A: Propensity to concede (i.e. choose blue when blue is shown in bold), by individual and treatment. Panel B: Mean transfer given, by individual and treatment, after both players played $a_i = m_i$. In the experiment, the two actions *claim* and *concede* are labeled as *red* and *blue*, respectively.

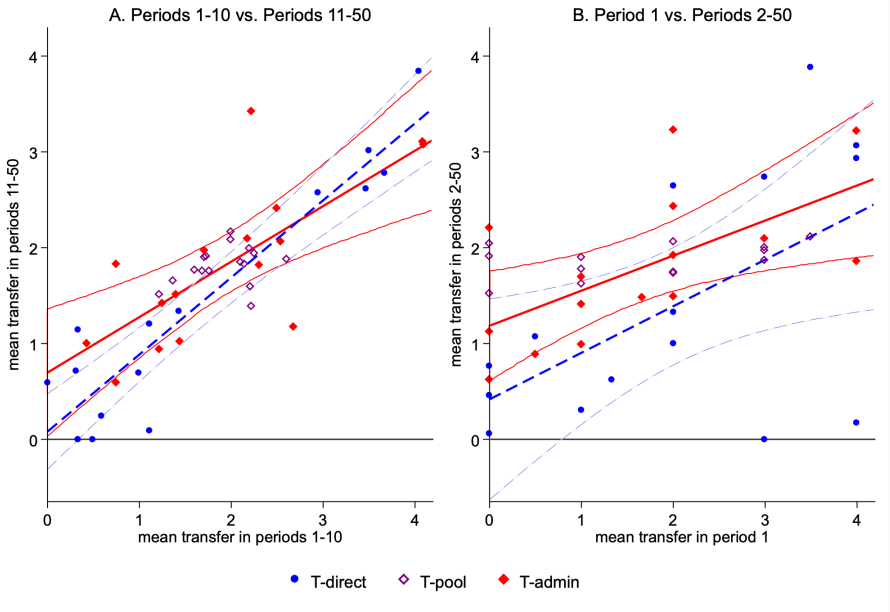


Figure A8: Early vs. Late Transfers of Upper Ranks, by Groups

Each dot depicts one group. There are 16 groups per treatment. Mean transfer given by *upper* ranks ($a - c$) after both players played $a_i = m_i$ in early vs. late periods of the game. Linear regression lines with 95% confidence intervals. Regression lines for *T-admin* not included because they are insignificantly different from zero (p=.308 in Panel A, p=.143 in Panel B).