A surprising hot-cold reciprocation gap

Riccardo Ghidoni
University of Bologna, Department of Economics
Tilburg University, CentER, Department of Economics
e-mail: riccardo.ghidoni@unibo.it

Sigrid Suetens
Tilburg University, CentER, Department of Economics
e-mail: s.suetens@tilburguniversity.edu

Jierui Yang
Tilburg University, CentER, Department of Economics
e-mail: j.yang_3@tilburguniversity.edu

June 7, 2024

Abstract

Experiments are conducted to investigate the effect of the decision mode on reciprocation in a one-shot trust game. Trustees either decided directly whether to reciprocate after observing their partner's choice or according to a contingent response plan made before observing their partner's choice. The main finding is that trustees were more likely to reciprocate under contingent decision making than under direct decision making. This reciprocation gap was not present when trust decisions were the outcome of a lottery, thus not made by trustors, which suggests that reciprocation choices must be the outcome of a commitment to reciprocate.

JEL Codes: C70, C90, D70

Keywords: trust game, reciprocity, direct decision mode, contingent decision mode, experiment, psychological game, strategy method

Acknowledgments: We acknowledge financial support from CentERlab at Tilburg University. We thank seminar audiences at Tilburg University, Purdue University, CREED in the University of Amsterdam and IMT Lucca and participants of the CIMEO Workshop at Sapienza University of Rome, ESA 2021 Global Around-the-Clock Virtual Conference, ESA 2022 European Meeting in Bologna and ESA 2023 World Meeting in Lyon for insightful discussions and comments. See https://osf.io/2856n and https://osf.io/tx6yg for the pre-registrations of the two experiments.

1 Introduction

The decision whether or not to reciprocate someone else's trust can be approached with either a direct or contingent decision mode. In the case of the former, players decide to reciprocate upon observing trust, that is, in a *hot* way, whereas in the case of the latter they use a *cold* plan of action, in which the decision is contingent on being trusted, and is made before observing the partner's choice. Given that reciprocity is at the core of multiple economic and social interactions, such as trade, the employer-employee relationship, and lending,[1] it is important to understand the factors underlying it. In an era of emerging technologies like autonomous machines which are programmed to react on the basis of a contingent plan, it is particularly important to understand the role of contingent decision making in reciprocity.

Hot and cold decision modes differ in two important ways, which may generate opposite behavioral effects. First, the hot mode has more scope for an emotional response (e.g. Loewenstein, 2000). Individuals who are pessimistic about trust may be positively surprised when someone indeed trusts them, prompting them to reciprocate. Such response is not present in cold decision making for the simple reason that individuals do not know whether they are being trusted when deciding on a plan of action.[2] Second, in the case of individuals who are inclined to reciprocate trust, the hot decision mode requires more self-control to overcome the temptation to choose the higher monetary payoff associated with a breach of trust. The cold decision mode, in contrast, facilitates commitment to reciprocate. Since there is a chance that one's reciprocation choice is not implemented, namely if the trustor decided not to trust them, the temptation to choose the high monetary payoff is less salient than in the hot decision mode, making reciprocation more likely.[3]

We report the results of two controlled experiments designed to investigate the effect of the decision mode on reciprocation in a trust game and to shed light on the underlying mechanism. The first experiment ($N = 452$) involved a one-shot binary trust game played multiple times in which a first mover chooses whether or not to trust a second mover who in turn

[1]See, for example, Akerlof (1982), Brown et al. (2004), Fehr et al. (1993), Greif (1993), Kehoe and Levine (1993), Kreps (1990) and Thomas and Worrall (1990).

[2]Individuals can also have emotions with regard to their partner's expected choice, but given that surprise is by definition unexpected, these emotions are not equivalent to surprise induced by the actual choices of others. Surprise may lead individuals to update their cognitive representations of others, and to adjust their actions accordingly (Reisenzein et al., 2019).

[3]This intuition is captured in Chen and Schonger (2022) who show that a weak sense of "duty" to reciprocate can be sufficient to make reciprocation optimal for trustees. Moreover, Akdeniz and van Veelen (2021); Akdeniz and van Veelen (2023) show that commitment to reciprocate is key for understanding human morality.

decides whether to reciprocate or defect.[4] The reciprocation choice was either elicited in a hot way, with second movers deciding whether to reciprocate after learning whether the first mover had trusted them, or in a cold way, in which the choice to reciprocate was made before learning whether the first mover had actually trusted them. Furthermore, to induce exogenous variation in the extent of positive surprise within the hot treatment, we manipulated the likelihood of experiencing this emotion. To do so, we divided the pool of first movers into two groups: one with a trust rate of 69 percent and one with a trust rate of 28 percent. Second movers were then matched exclusively with first movers from either the low-trust group or the high-trust group and were informed about the trust rate of the group they were matched with.

By providing second movers with information about the trust rate, we introduced exogenous variation in the likelihood of the second mover being surprised if the matched first mover trusts her, where a higher likelihood is expected in the case of a low trust rate (which was confirmed in a manipulation check).[5] The design assumes a utility function that includes an emotional-psychological component triggered by the experience of being trusted, reflecting the potential for positive surprise (see also Geanakoplos et al., 1989; Ruffle, 1999).[6] This component can only exist in the hot decision mode and its salience can be expected to increase with the second mover's *ex ante* pessimism regarding the first mover's decision to trust them.[7] Thus, the key hypothesis is that the hot-cold reciprocation gap will depend positively on the rate.

Surprisingly, the results of the first experiment contradicted our predictions: the hot reciprocation rate was in fact *lower* than the cold reciprocation rate ($P < 0.001$ across first rounds).[8] The effect was observed regardless of the trust rate: the reciprocation gap was 19 percentage points when the trust rate was high and 13 percentage points when it was low. Moreover, within each decision mode, reciprocation rates were similar across high and low trust rates condition. Thus, the reciprocation gap was not wider in the low trust rate condition as ex-

[4]This experiment was pre-registered at https://osf.io/2856n.

[5]The degree of unexpectedness is deemed to be a crucial driver of the intensity of surprise in the social psychology literature (e.g., Meyer et al., 1997; Reisenzein et al., 2019).

[6]Allowing for disappointment due to a lack of trust, as in Ruffle (1999), is inconsequential in a binary trust game since the second mover has no choice to make if the first mover did not trust her.

[7]The approach is equivalent to that of Aina et al. (2020) who allow for anger in an ultimatum game, though it differs from that of Khalmetski et al. (2015) who assumes people gain non-monetary utility from surprising others in a reciprocation game.

[8]In the experiment, second movers were matched with a different first mover in each round. The matching was organized such that all second movers encountered trust from their partner in the first round, thereby generating the cleanest possible comparison between the hot and cold modes.

pected. The explanations that the result was due to second movers perceiving cold decisions to be more difficult than hot ones or to differential second-order beliefs were ruled out. This leads to the conclusion that positive surprise has no strong behavioral consequences on the trust relationship: although second movers did indeed experience surprise, they did not act on it.

How then are the findings to be understood? We conjecture that the results were due to the presence of the aforementioned commitment mechanism. Thus, individuals attach utility to a commitment to reciprocate which will lead to a lower reciprocation rate under hot decision making than under cold decision making. The explanation is that under hot decision making, a stronger sense of self-control or feeling of duty to reciprocate is needed for a second mover to overcome the temptation to choose the high monetary payoff than under cold decision making. If the second mover plans to reciprocate in the cold decision mode, there is a chance that her choice will not be realized (if the first mover does not trust her). Thus, a weaker preference to commit to reciprocate is enough for a second mover to resist the temptation to breach the first mover's trust.

Alternative mechanisms to explain the observed reciprocation gap include commitment to pro-social behavior *per se* and costly errors. The former operates similarly to commitment to reciprocate, and therefore should lead to a similar hot-cold reciprocation gap: a stronger sense of self-control or a duty to act pro-socially is necessary for the second mover to resist the temptation to choose the higher payoff in the hot decision mode. The costly errors mechanism is based on the idea that self-interested second movers calculate their expected payoffs for each possible action (see, e.g., Mckelvey and Palfrey, 1998) and has been used to rationalize the hot-cold gap in an experimental cheap talk game (Minozzi and Woon, 2019). The logic is that uncertainty about the implementation of the second mover's reciprocation choice, which characterizes cold decision making, reduces the gap in monetary utility between reciprocating and defecting relative to the case of hot decision making, in which no such uncertainty exists. Essentially, the magnitude of error required to justify reciprocation as the optimal strategy is smaller under hot decision making than under cold decision making, leading to the observed reciprocation gap.

The second experiment ($N = 400$) was designed to test the hypothesis that commitment to reciprocate was the key mechanism behind the observed reciprocation gap in the first experiment rather than the two alternative mechanisms.[9] To do so, the possibility that individuals

[9] This experiment was pre-registered at https://osf.io/tx6yg.

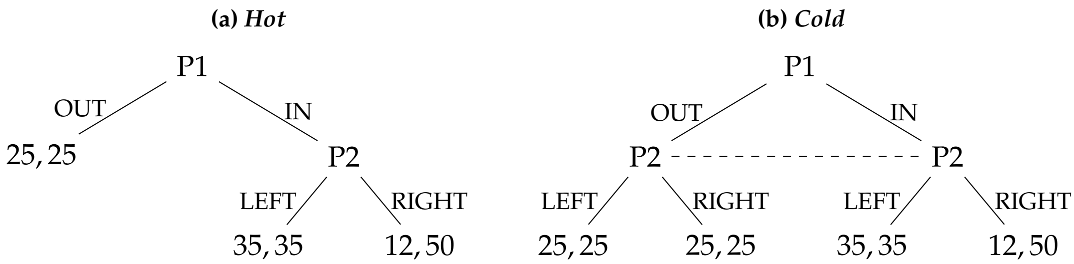
could commit to reciprocate was eliminated by design without removing the treatment variation in the decision mode and in the level of the trust rate. Hence, a lottery was introduced to determine the trust action of the "first mover". This removed the possibility for second movers to commit to reciprocate while leaving intact the possibility of committing to acting pro-socially or making costly errors. Since no behavioral gap was observed in the second experiment, we conclude that commitment to reciprocate is the most plausible mechanism behind the reciprocation gap observed in the first experiment.

The paper is part of the literature on the effect of the decision mode on choice in trust games or other similar games. The results reported in this literature are ambiguous, as shown in the survey by Brandts and Charness (2011) and the meta-study by Johnson and Mislin (2011) of standard, continuous trust game experiments. For example, Casari and Cason (2009) and Cox and Hall (2010) found that trustees send back more money to the first mover while controlling for the amount sent by the first mover, while Solnick (2007) reported the opposite effect (though qualitatively small) and Fong et al. (2007) reported no effect. Neither were hot-cold gaps observed in sequential prisoner's dilemma games (Brandts and Charness, 2000; Reuben and Suetens, 2012) nor in the context of conditional contributions in public goods games (Fischbacher et al., 2012). Our findings are closest to those of Garcia-Pola (2020) who found that players in hot centipede games were more inclined to stop earlier than those in cold centipede games, which is in a sense behaviorally similar to a lower inclination to reciprocate.

Our paper contributes to this literature by theoretically and empirically identifying the mechanisms that underlay the hot-cold reciprocation gap in a trust game, with particular focus on positive surprise and commitment to reciprocate. We employ an experimental design that maximizes the likelihood that positive surprise will influence the decision to reciprocate. In the simple one-shot binary game we employ, the effect of surprise, if present, can only act in one direction, while in continuous trust games (in which a given amount sent by the first mover might induce very optimistic second movers to be disappointed), it might induce the second mover to send back less than they would have under cold decision making.

The remainder of the paper proceeds as follows: Section 2 describes the trust game experiment and report its findings. Section 3 discusses the possible mechanisms that can explain the hot-cold reciprocation gap. Section 4 describes the follow-up experiment and Section 5 concludes.

Figure 1: Binary trust games



Notes: The left-hand (right-hand) panel describes the trust game if choices are elicited using *Hot* (*Cold*) elicitation. P1 (P2) stands for the first mover (second mover). The payoff for the first mover (second mover) is shown on the left (right). The trust choice of the first mover is IN and the reciprocation choice of the second mover is LEFT.

2 Experiment 1

2.1 Theoretical background

Consider a two-player binary trust game in which both players earn 25 if the first mover does not trust the second mover; both earn 35 if the first mover does trust the second mover and the second mover reciprocates; and the first mover earns 12 and the second mover earns 50 if the first mover trusts the second mover and the second mover defects. Assume that the choices of the second mover are elicited using either hot decision making (the *Hot* game) or cold decision making (the *Cold* game), as shown in Figure 1. In the *Hot* game, the second mover makes her choice after observing whether the first mover trusted her (i.e., chose IN), whereas in the *Cold* game the second mover makes a choice, conditional on the first mover trusting her. Thus her optimal choice depends on her expectation about the first mover's choice in the former game but not in the latter game. We focus on the best-reply behavior of the second mover at the point where she makes the decision whether or not to reciprocate (i.e. she chooses LEFT or RIGHT). We denote by p the *ex ante* probability with which the second mover expects the first mover to trust her.

To derive testable predictions, we first consider the best reply of a payoff maximizer and then introduce our psychological typology.[10] If the second mover is a payoff-maximizer, then the payoffs shown in Figure 1 represent her utility. In *Hot*, the second mover's best reply to the first mover's trust is to defect irrespective of p. In contrast, her best reply in *Cold* depends on p: If $p > 0$, then the expected payoff for defecting is strictly higher than the expected payoff for reciprocating, such that the best reply is to defect. If $p = 0$, then the expected payoffs of

[10]We base our theoretical predictions on an analysis of best-reply behavior by second movers given that there is no room in the experiment for convergence to an equilibrium. Equilibrium calculations appear in Section B in the Appendix.

defecting and reciprocating are the same, such that the second mover is indifferent between the two actions.[11]

We analyze next the trust game as a psychological game by allowing the second mover to gain additional psychological utility when reciprocating if she is surprised by the trust shown by the first mover (reminiscent of the assumption made by Geanakoplos et al. (1989) according to which a player in a repeated prisoner's dilemma becomes more sympathetic toward her partner if the latter unexpectedly cooperated earlier in the game). The second mover can only obtain this psychological utility in *Hot* because it is only in that case that the second mover observes the first mover's actual choice before making a choice herself. We assume that the extent to which the second mover cares about the first mover's payoff depends on her surprise sensitivity denoted by $\theta \geq 0$ and the surprise level denoted by $S(p)$.[12] The function S is decreasing in p, such that the lower the expectation that the first mover will trust her, the more surprised the second mover is when the first mover does. To focus on the effect of surprise, we assume that utilities are equal to monetary payoffs, apart from the surprise component. Hence, the utility of the second mover is given by:

$$U_2 = \pi_2 + \theta S(p)\pi_1. \tag{1}$$

It can easily be shown that in *Hot* the second mover is predicted to reciprocate if $\theta > \hat{\theta} \equiv \frac{50-35}{S(p)(35-12)}$ with $S(p) > 0$. In contrast, in *Cold* the second mover is predicted to act like a payoff maximizer because there is no possibility of surprise, i.e. $S(p) = 0$. If she believes with a strictly positive probability that the first mover will trust her, then she will defect, and if she believes that the first mover will not trust her, then she is indifferent.[13]

In summary, if at least some second movers are motivated by positive surprise and as long as $p > 0$, then there are two key predictions.[14] First, the reciprocation rate will be (weakly) higher in the *Hot* game than in the *Cold* game for a given p. Second, the reciprocation gap between the *Hot* game and the *Cold* game is expected to narrow as p increases.

[11]In a subgame perfect Nash equilibrium, the probability that the second mover reciprocates cannot exceed 0.57 because that would tilt the first mover towards trust, which would make it optimal for the second mover to defect.

[12]Notice that allowing for heterogeneous surprise sensitivities across players does not change the basic theoretical predictions.

[13]A trivial extension of the utility function that leads to a positive reciprocation rate also in *Cold* is to allow for other-regarding preferences in the utility function (e.g., Fehr and Schmidt, 1999; Charness and Rabin, 2002). Yet this extension would not affect the comparative static predictions. To focus on the effect of surprise, we assume that utilities are equal to monetary payoffs, apart from the surprise component.

[14]The assumption of $p > 0$ holds in our experiment since second movers are informed about the (strictly positive) trust rate in the population of first movers they are matched with. We discuss this design feature in Sections 2.2 and 2.3.

2.2 Design and hypotheses

Participants were randomly allocated between first or second movers using an asynchronous design. In the first stage of the experiment, the first movers were asked whether they trusted a randomly matched second mover in a trust game with payoffs as in Figure 1. In the second stage, the second movers were informed about the trust rate in the population of first movers their partners were drawn from and then made their choices. Second movers were randomly assigned to one of four treatment conditions in a between-subjects design. The conditions varied in terms of the decision mode (*Hot* or *Cold*) and the induced prior belief about the trust rate (*Low* or *High*).

We created exogenous variation in the prior belief of trust p in order to induce exogenous variation in the scope for a positive surprise. To implement a *Low* or *High* prior belief of trust, the pool of first-mover participants was divided into two groups: a group with a low trust rate (0.28) and a group with a high trust rate (0.69).[15] Second movers were matched with first movers from the low-trust group *or* the high-trust group. We conjectured that a higher trust rate would be associated with a more optimistic belief about trust, which in turn would lead to a lower level of surprise when encountering a trusting first mover in *Hot*.

After making 20 reciprocation choices with different first movers, second movers completed a brief survey which included a question on the surprise they experienced. The response served as a manipulation check of the variation in trust rate within *Hot*. Specifically, we asked them to rate their level of surprise (not surprised, somewhat surprises, or surprised) at the trust shown by the first mover in games where trust was chosen. In *Hot*, the number of participants in the *Low* condition who reported feeling surprised or somewhat surprised was 13.13 percentage points higher than in the *High* condition ($P = 0.031$).[16] This finding suggests that the induced surprise worked as in the expected direction.

The hypotheses follow directly from the predictions of the surprise model discussed in Section 2.1:

Surprise hypothesis (a) The reciprocation rate is (weakly) higher in *Hot* than in *Cold* in both *Low* and *High*. (b) The difference in the reciprocation rate between *Hot* and *Cold* is (weakly)

[15] Before running the actual experiment, we ran pilot sessions to calibrate the payoff parameters; payoffs for trust and reciprocation in these sessions were 42 and 37, respectively (instead of 35). The aim of this (pre-registered) procedure was to obtain an aggregate trust rate not too different from 40 percent, which would allow us to split the first movers into two subgroups with sufficiently different trust rates (see also Section 2.3).

[16] The reported P-value refers to a Wald test conducted on estimates from linear regressions that include a treatment indicator as an independent variable. Full details on the estimation results are in Table C.4 in the Appendix.

larger in *Low* than in *High*.

2.3 Procedures

The experiment was programmed using Qualtrics and run on Prolific, in compliance with the EU General Data Protection Regulation. The instructions can be found in Sections A.1 and A.2 in the Appendix. Participation was restricted to students from the UK, the most represented nationality on Prolific. To ensure comprehension of the decision making task, participants were required to complete control questions before proceeding. They could make up to six attempts to correctly answer all the questions, and those who failed were excluded.

The asynchronous experiment consisted of two phases, with first-mover decisions elicited in phase 1 and second-mover decisions in phase 2. Phase 1 took place on December 7$^{\text{th}}$ and 8$^{\text{th}}$, 2020, during which 400 participants played as first movers.[17] Each participant in the role of first mover made a single trust decision, unaware of which treatment group their matched second mover belonged to. On average, first movers completed the experiment in 3 minutes and received an average payment of 2.7 pounds, which included a 0.4 pound participation fee. The conversion rate was 15 tokens = 1 pound.

Phase 2 was conducted on December 11$^{\text{th}}$ and 14$^{\text{th}}$, 2020, during which 452 participants played the role of second movers. They were randomly assigned among four treatment conditions.[18] After correctly answering a series of control questions, the second movers played 20 rounds of the game, each time with a different first mover. As mentioned, we informed the second movers of the trust rate in the population of first movers they would be matched with.[19] At the end of each round, the second movers in *Cold* received feedback on the choice made by the first mover and those in all treatments received feedback on their payoffs. To ensure a clear identification of the surprise effect and the same number of observations in *Hot* and *Cold* (and thus sufficient power to reject the null), we arranged the games such that in the first round the second movers all faced a first mover who decided to trust them. Moreover, we also ensured that the order of trust decisions faced by second movers across the 20 rounds was similar between the treatments.[20] Thus, repeating the game for several rounds was instrumen-

[17]We recruited 412 first movers, but 12 of them dropped out during the experiment.

[18]We recruited 611 second movers; however, 159 of them dropped out during the experiment. We show in Section 2.4 that attrition did not generate selection problems.

[19]Recall that we formed two groups of 200 first movers, one with a low trust rate of 0.28 and the other with a high trust rate of 0.69, in order to generate the *Low* and *High* treatment conditions.

[20]We created 20 alternative randomly generated sequences of trust decisions. Thus, roughly five second movers in each treatment encountered the same sequence of trust decisions.

tal to obtaining a substantial difference in trust rate between *High* and *Low* while ensuring that second movers in all treatments encountered trust in the first round.

After the second movers had made their reciprocation choices, they were asked to fill out a short survey that included socio-demographic background (gender, age and education), social preferences, perceived complexity of the experiment, and the level of surprise experienced. At the end of the experiment, one round was selected to pay the second mover and her matched first mover.[21] On average, second movers completed the experiment in 14 minutes and received an average payment of 3.6 pounds, which included a 0.6 pound participation fee. The conversion rate was 15 tokens = 1 pound.

2.4 Possible confounders

Selective attrition. We first investigate whether there was selective attrition in phase 2 of the experiment, which might threaten the identification of treatment effects. In total, 611 second movers began the experiment, of which 159 dropped out before completing it. Of the 159 dropouts, 51 left the experiment before reading the instructions or being assigned to a treatment condition. The attrition rate is balanced across treatments, as shown in Table 1 (first row), and therefore it does not appear to be a concern.

Table 1: Balance table–Experiment 1

	Hot High	*Hot Low*	*Cold High*	*Cold Low*	*P*-value
Attrition rate	0.233	0.154	0.190	0.192	0.425
Female	0.723	0.687	0.694	0.693	0.934
Age	24.241	24.443	23.676	24.132	0.885
Education					0.051
High school or below	0.205	0.122	0.081	0.088	
Bachelor	0.571	0.652	0.568	0.684	
Master	0.170	0.148	0.234	0.158	
PhD or above	0.036	0.078	0.099	0.044	
Other	0.018	0	0.018	0.026	
Observations	112	115	111	114	

Notes: The table shows the attrition rate and averages of the socio-demographic data collected in the post-experimental survey. The attrition rate represents the percentage of second movers who dropped out after viewing the instructions, i.e., after being assigned to a treatment. The education category refers to the highest obtained diploma. The *P*-values in the last column refer to likelihood ratio tests conducted on estimates from a probit regression without a constant that include a treatment indicators as independent variables (for attrition and female); a Wald test conducted on the estimates from the linear regressions (for age); and a Chi-square test (for education).

[21] Since 400 first movers participated, 52 second movers remained unmatched and received a fixed payment of 1 pound, a possibility they were made aware of.

Balancing. To assess the balance of socio-demographic characteristics across treatments, we compared the gender, age, and education of the participants who completed the experiment across treatments based on data collected in the post-experimental survey. Table 1 reports the statistics. It can be seen that there are only minor socio-demographic differences across treatments and they are far from statistically significant. The only exception is educational background, such that the share of students with at most a high-school diploma is substantially higher in *Hot High* than in the other three treatments. Thus, in estimating treatment effects on reciprocation, controlling for educational background will be crucial as a robustness check.[22]

Second-order beliefs. Psychological mechanisms based on second-order beliefs such as guilt aversion and intention-based reciprocity have been shown to be important drivers of reciprocation. Specifically, the more optimistic a guilt-averse second mover thinks the first mover is about reciprocation, the more likely she will be to reciprocate trust (Charness and Dufwenberg, 2006). In contrast, intention-based reciprocity predicts that the more optimistic a second mover thinks the first mover is about reciprocation, the *less* inclined she will be to reciprocate. The intuition is that trust from an optimistic first mover is perceived as less impressive than trust from a pessimistic first mover (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004).[23] Since it is conceivable that the second-order beliefs of second movers are positively affected by the (communicated) trust rate, these beliefs most likely differ between *Low* and *High* conditions, which may in turn affect reciprocation behavior through guilt aversion or intention-based reciprocity. Although this effect does not *per se* invalidate our identification strategy for testing the surprise hypotheses, a crucial identifying assumption is that second-order beliefs do not differ between *Hot* and *Cold* in both *Low* and *High*.

To assess the validity of this assumption, we elicited second-order beliefs. To do so, we elicited first-order beliefs from first movers about the overall reciprocation rate in the experiment after they had made their trust decision. First movers whose guess was within a 5% bound from the actual reciprocation rate earned 20 additional tokens. We then elicited the second-order beliefs of second movers by asking them (before each round) to guess the average (first-order) belief of all first movers who chose to trust the second mover. To minimize the possible influence of the belief elicitation on decisions, second movers did not receive feedback about the accuracy of their beliefs after each round. They received 20 additional tokens if their

[22]The estimated treatment effects on reciprocation while controlling for socio-demographic variables appear in Table C.1 in the Appendix.

[23]The same effect is predicted by the model of surprising gifts in Khalmetski et al. (2015).

belief in a randomly selected round (different from the one used to determine the payment for their reciprocation choice) was within a 5% bound from the average first-order belief of first movers who chose to trust the second mover.

The findings indicate that second-order beliefs did not differ significantly between *Hot* and *Cold* when controlling for the aggregate trust rate. Second movers in *Low* guessed on average that first movers expected a reciprocation rate of 43.53 percent in *Hot* and 41.04 percent in *Cold* ($P = 0.357$), whereas in *High* they expected 56.51 percent and 53.03 percent, respectively ($P = 0.154$).[24]

Cognitive complexity. Participants might have found cold decision making to be cognitively more complicated than hot decision making. Hence, differences in the reciprocation rate between *Hot* and *Cold* might reflect differences in complexity. To study possible treatment effects on complexity, we asked participants to evaluate the complexity of their decision making process on a 10-point Likert scale in the post-experimental survey. As another proxy for complexity, we used the number of attempts (ranging from 1 to 6) needed by second movers to correctly answer the control questions, after reading the instructions but before beginning the reciprocation task.

There are no indications in the data that decision making was perceived as more complex in *Cold* than in *Hot*. Self-reported complexity was on average equal to 4.702 points in *Cold* and 4.995 points in *Hot* ($P = 0.138$), and the number of attempts to solve the control questions was even significantly lower in *Cold* than in *Hot* (on average, 1.871 versus 2.092, $P = 0.030$; see Table C.5 and Table C.6 in the Appendix for full details).

2.5 Results

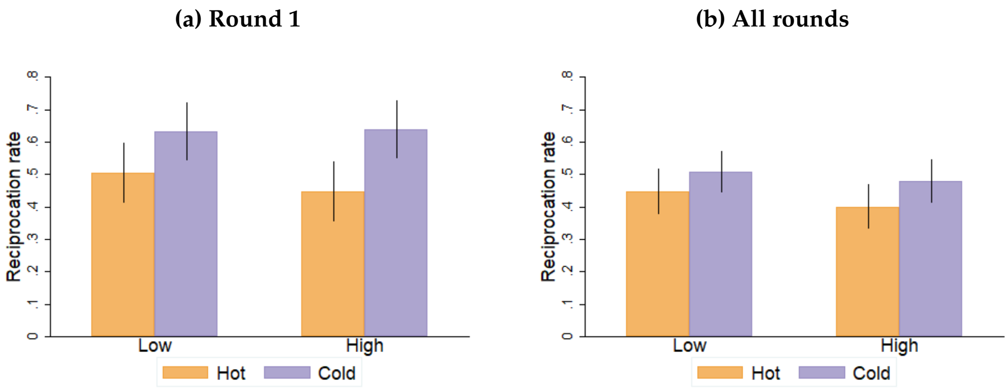
In the following analysis, we differentiate between reciprocation rates observed in the first round and those observed in all the rounds. The first-round data, which are unaffected by prior experiences, offer the cleanest test of our hypotheses, while the data for all rounds reflect the influence of past interactions.[25] Figure 2 shows reciprocation rate by treatment for the first round and for all rounds. Notably, the first-round reciprocation rate is higher in the *Cold* treatment than in the *Hot* treatment, conditional on *Low* or *High*, thus contradicting the surprise

[24] The reported P-values refer to Wald tests conducted on estimates from linear regressions that include treatment indicators as independent variables across the first rounds. Full details of the estimation results based for the first round and for all rounds appear in Table C.3 in the Appendix.

[25] Recall that in the first round, all second movers were matched with a trusting first mover.

hypothesis. Overall, the same effects are visible in the case of all rounds, albeit qualitatively smaller.

Figure 2: Reciprocation rates in Experiment 1



Notes: The graph shows the reciprocation rate (the percentage of times second movers chose LEFT) in Experiment 1 and 95% confidence intervals by treatment. Panel (a) is based on the first round and panel (b) on all 20 rounds. Confidence intervals are based on estimations from probit regression run on treatment dummies with standard errors clustered at the individual level (panel (b) only).

In order to estimate the treatment effects, linear regressions were estimated with the choice to reciprocate as the dependent variable. In a first specification, we test part (a) of the surprise hypothesis, namely that the reciprocation rate is higher in *Hot* than in *Cold* in both *Low* and *High*, by regressing the reciprocation choice on the indicator variables *Hot* and *High*. In a second specification, we test part (b) of the surprise hypothesis, namely whether the difference in reciprocation rates between *Hot* and *Cold* is larger in *Low* than in *High*, by adding an interaction term between *Hot* and *High*. Results are reported in Table 2.[26]

In the non-interacted regressions reported in column (1), the difference in first-round reciprocation rates between *Hot* and *Cold* is large (about 15-16 percentage points) and statistically significant. In the case of all rounds, the difference is smaller but still marginally significant. This result leads us to reject part (a) of the surprise hypothesis. In fact, the effect of the decision mode works in the opposite direction to that hypothesized, such that the reciprocation rate is higher under cold decision making than hot decision making.

A similar conclusion is reached based on the interacted regressions in column (3). The interaction term appears to partially absorb the variation in reciprocation choices, thus reducing the main effect of the decision mode; nevertheless, it does not achieve statistical significance in any specification. Hence, part (b) of the surprise hypothesis is also rejected.

We conclude that although second movers reported being more surprised by their partner's

[26]Results are robust to employing probit regressions (see Table C.2 in the Appendix).

trust in *Hot Low* than in *Hot High*, as can be seen in Table C.4 in the Appendix, this did not translate into a higher reciprocation rate, nor did it generate a higher reciprocation rate in *Hot* than in *Cold*.

Table 2: Treatment effects on reciprocation in Experiment 1

Dep. var.: Reciprocate	(1)	(2)	(3)	(4)
(a) Round 1				
Hot	-0.160 (0.046)***	-0.151 (0.046)***	-0.127 (0.065)*	-0.122 (0.065)*
High	-0.025 (0.046)	-0.016 (0.046)	0.008 (0.066)	0.014 (0.066)
Hot×*High*			-0.066 (0.093)	-0.059 (0.093)
Constant	0.648 (0.040)***	0.451 (0.208)**	0.632 (0.046)***	0.440 (0.209)**
Observations	452	452	452	452
(b) All rounds				
Hot	-0.073 (0.036)**	-0.067 (0.036)*	-0.063 (0.048)	-0.061 (0.048)
High	-0.034 (0.037)	-0.036 (0.036)	-0.029 (0.047)	-0.034 (0.046)
Hot×*High*			-0.016 (0.069)	-0.009 (0.069)
Constant	0.511 (0.029)***	0.404 (0.179)**	0.509 (0.033)***	0.403 (0.179)**
Observations	6684	6684	6684	6684
Controls		✓		✓

Notes: Estimates from linear regressions (standard errors in parentheses). In panel (b), standard errors are clustered at the individual level. In columns (2) and (4), age, gender and education categories are included as controls (see Table C.1 in the Appendix for the estimates). *$P < 0.1$, **$P < 0.05$, ***$P < 0.01$.

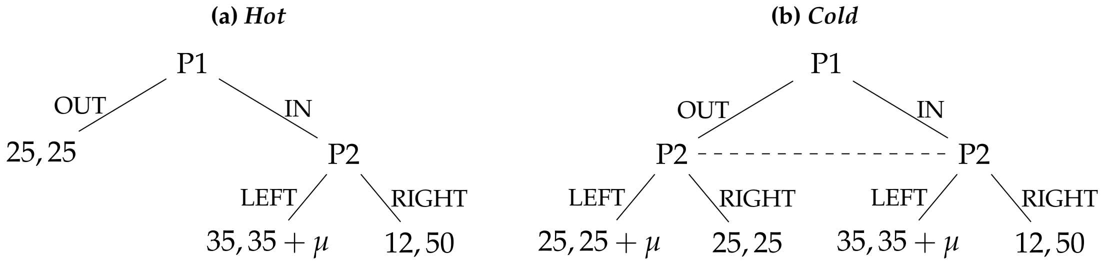
3 Discussion

The observed hot-cold reciprocation gap in Experiment 1 was not due to surprise, second-order beliefs or differences in cognitive complexity. We therefore explore alternative mechanisms to explain this outcome. We conjectured that the gap might be due to a commitment mechanism, such that a sense of duty led individuals to value the act of planning to reciprocate, thus attaching non-monetary utility to this behavior (Chen and Schonger, 2022, 2020).

To model this preference, we introduce a non-monetary utility component, $\mu \geq 0$, into the utilities depicted in Figure 3, in order to reflect the utility from committing to reciprocate.

If p represents the *ex-ante* probability that the first mover chooses IN, then the second mover's expected payoff for choosing LEFT is equal to $(35 + \mu)p + (25 + \mu)(1 - p)$ in *Cold*, whereas that for choosing RIGHT is equal to $50p + 25(1 - p)$. The payoff for LEFT in *Hot* given that the first mover chooses IN is equal to $35 + \mu$, and that of RIGHT is equal to 50. Consequently, the best reply by the second mover in *Cold* is to choose LEFT if $(35 + \mu)p + (25 + \mu)(1 - p) > 50p + 25(1 - p)$ or if $\mu > 15p$, and RIGHT otherwise. In *Hot*, it is optimal for the second mover to choose LEFT if $35 + \mu > 50$, or if $\mu > 15$, and RIGHT otherwise. A second mover with a preference to commit to reciprocate is more likely to reciprocate in the *Cold* game than in the *Hot* game whenever $p < 1$, thus leading to a hot-cold reciprocation gap like the one

Figure 3: Binary trust games with commitment to reciprocate



Notes: The left-hand (right-hand) panel describes the trust game if choices are elicited using *Hot* (*Cold*) elicitation. P1 (P2) stands for the first mover (second mover). The payoff for the first mover (second mover) is shown on the left (right). The trust choice of the first mover is IN and the reciprocation choice of the second mover is LEFT.

observed in Experiment 1.

A second interpretation is that instead of a preference to commit to reciprocate, individuals have a preference to commit to behave pro-socially in general. Preferences would then be as in Figure 3, but can be distinguished from commitment to reciprocate by replacing the first mover by a lottery, as we will argue at the end of this section.

A third interpretation is that individuals make errors when calculating their expected payoffs, with the likelihood of errors being inversely related to their cost (Mckelvey and Palfrey, 1998). Since in *Cold* it is uncertain whether the second mover's choice will be realized, the expected payoffs differ from those in *Hot*. Specifically, the gap between the expected payoff for choosing LEFT and that for choosing RIGHT is higher in *Hot* than in *Cold*, making it more likely that the second mover will choose LEFT when this uncertainty is taken into account. To see this, consider Figure 3 with $\mu = 0$. If p represents the *ex ante* probability of IN, then the second mover's expected payoff for choosing LEFT is $35p + 25(1-p) + \epsilon_{LEFT}$ in *Cold*, whereas that for choosing RIGHT is $50p + 25(1-p) + \epsilon_{RIGHT}$. The payoff for choosing LEFT in *Hot* given IN is $35 + \epsilon_{LEFT}$, while that for choosing RIGHT given IN is $50 + \epsilon_{RIGHT}$. Consequently, the second mover's best reply in *Cold* is to choose LEFT if $35p + 25(1-p) + \epsilon_{LEFT} > 50p + 25(1-p) + \epsilon_{RIGHT}$, or if $\epsilon_{\text{LEFT}} - \epsilon_{\text{RIGHT}} > 15p$, and to choose RIGHT otherwise. In *Hot*, it is optimal for the second mover to choose LEFT if $35 + \epsilon_{LEFT} > 50 + \epsilon_{RIGHT}$, or if $\epsilon_{\text{LEFT}} - \epsilon_{RIGHT} > 15$, and to choose RIGHT otherwise. A second mover who calculates the expected payoffs with random errors will tend to choose LEFT more often in *Cold* than in *Hot* whenever $p < 1$.

To test the hypothesis that the hot-cold reciprocation gap in Experiment 1 stems from commitment to reciprocate rather than one of the other two mechanisms, we designed an additional experiment. Experiment 2 modifies the setup in Experiment 1 by programming the first movers' choices rather than having them make choices themselves. All other aspects of the experiment remain unchanged. Consequently, the only difference between Experiment 1 and Experiment 2 relates to element of reciprocity, which is present in the former case but absent in the latter. If the commitment to pro-social behavior or calculation errors drive the observed outcomes, then we would expect to see a similar hot-cold gap in Experiment 2. Specifically, a second mover who is committed to act pro-socially still experiences a positive non-monetary utility μ_i even if the "first mover" is a passive player. Similarly, the likelihood of the aforementioned calculation errors should remain unchanged, whether the trust choice is made actively by a player or determined by the computer.

4 Experiment 2

4.1 Design and hypotheses

Second movers played a one-shot binary trust game with payoffs as in Figure 1. They were randomly assigned to one of four treatment conditions in a between-subjects design. As in Experiment 1, the conditions varied in terms of the decision mode (*Hot* or *Cold*) and the probability of IN (*High* or *Low*). Second movers in *Hot* were informed of whether the pre-determined action was IN or OUT before making their choices. In contrast, second movers in *Cold* submitted their choice conditional on IN. The probability of IN was equal to 69% in *High* and 28% in *Low*, corresponding to the trust rates in *High* and *Low* of Experiment 1. Second movers were informed about these probabilities and knew that the matched "first movers" did not make the choices themselves.[27] After second movers had made their choices, they were asked to fill out the same short survey as in Experiment 1.

We hypothesize that commitment to reciprocate was the main mechanism underlying the hot-cold reciprocation gap observed in Experiment 1, as stated in the following hypothesis:

Commitment to reciprocate hypothesis The LEFT rate does not differ between *Hot* and *Cold*, conditional on *Low* or *High*.

If participants in Experiment 2 were not more likely to choose LEFT in *Cold* than in *Hot*, we would conclude that commitment to reciprocate is the main mechanism behind the first set of results. If participants were again more likely to choose LEFT in *Cold* than in *Hot*, we would conclude that commitment to pro-social behavior or costly mistakes were the key mechanism.

[27]We employed the same sequences of IN choices as in Experiment 1.

4.2 Procedures

Experiment 2 was conducted between December 11th and 13th, 2023, following closely the procedures of Experiment 1. The instructions can be found in Section A.3 in the Appendix. The main procedural deviation from Experiment 1 involved conducting only one phase, during which second movers made their decisions. Participants in the role of "first mover" were informed that they would not need to make choices and could nonetheless earn some money.[28]

We recruited 400 participants as second movers and randomly assigned them to one of the four treatment conditions.[29] After correctly answering a series of control questions, the second movers played 20 rounds of the game, each time with a different first mover. In each round, the aforementioned process determined whether the outcome was IN or OUT and second movers were informed about the probabilities involved. At the end of each round, the second movers received feedback on the outcomes (in *Cold*) and on the payoffs obtained (in all treatments). As in Experiment 1, we arranged the trust games so that the second movers all encountered IN in the first round. Moreover, we also ensured that the order of IN decisions encountered by second movers across the 20 rounds was similar between the *Hot* and *Cold* treatments.[30] At the end of the experiment, one round was randomly selected to pay the second mover and the first mover with whom she was matched with in that round. Second movers took about 8.5 minutes to complete the experiment, with an average payment of 2.7 pounds (including a participation fee of 0.6 pounds). The conversion rate was 15 tokens = 1 pound.

4.3 Possible confounders

Selective attrition. In total, 529 second movers began the experiment and 129 dropped out before completing it. Among the dropouts, 53 left the experiment before even reading the instructions and being assigned to a treatment. In general, the attrition rate is balanced across treatments, as shown in Table 3 (first row). Thus, attrition does not appear to have generated selection problems.

[28] First movers were contacted on Prolific after the data were collected from the second movers.

[29] In total 529 second movers were recruited; however, 129 of them dropped out during the experiment. See Section 2.4 for a discussion of attrition.

[30] We did not elicit second-order beliefs from second movers given that the "first movers" did not actively make a choice.

Table 3: Balance table–Experiment 2

	Hot High	*Hot Low*	*Cold High*	*Cold Low*	*P*-value
Attrition rate	0.167	0.133	0.203	0.136	0.422
Female	0.500	0.544	0.560	0.680	0.114
Age	28.590	29.837	29.680	28.304	0.633
Education					0.229
High school or below	0.170	0.096	0.117	0.176	
Bachelor	0.550	0.548	0.606	0.549	
Master	0.220	0.250	0.170	0.235	
PhD or above	0.040	0.077	0.106	0.039	
Other	0.020	0.029	0	0	
Observations	100	104	94	102	

Notes: The table shows the attrition rate and averages of the socio-demographic data collected in the post-experimental survey. The attrition rate represents the percentage of second movers who dropped out after viewing the instructions, i.e., after being assigned to a treatment. The education category refers to the highest obtained diploma. The *P*-values in the last column refer to likelihood ratio tests conducted on estimates from a probit regression without a constant that include a treatment indicators as independent variables (for attrition and female); a Wald test conducted on the estimates from the linear regressions (for age); and a Chi-square test (for education).

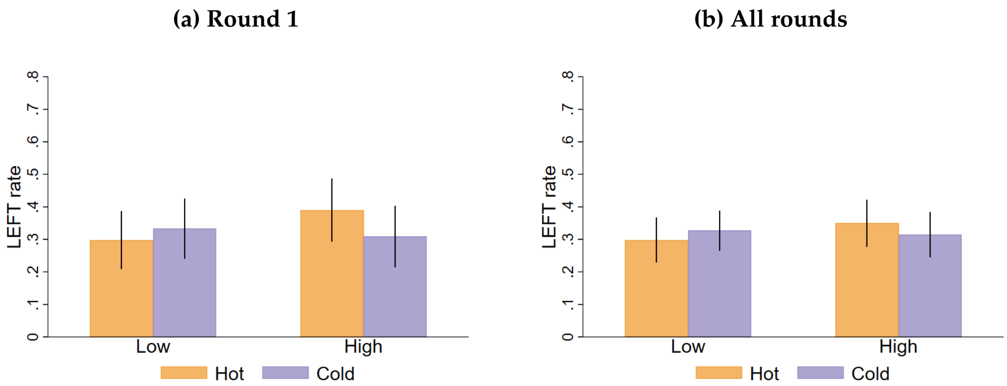
Balancing. In Table 3, we compare the gender, age and education of the participants who completed Experiment 2 across treatments. It can be seen that the differences are small and mostly statistically insignificant.

4.4 Results

Figure 4 shows the percentage of LEFT choices in the first round and for all rounds by treatment and Table 4 reports the corresponding regression results. It can clearly be seen that the hot-cold reciprocation gap that existed in Experiment 1 has disappeared, such that the percentage of LEFT choices in Experiment 2 is similar across all treatments. This finding is confirmed by the regressions, which show that of the treatment effects or interactions are statistically significant.[31] These results lead us to conclude that neither commitment to pro-social behavior nor costly mistakes were the key mechanism behind the first set of results, and that commitment to reciprocate is indeed the main mechanism.

[31] Results are robust to employing probit regressions (see Table C.2 in the Appendix).

Figure 4: Percentage of LEFT choices in Experiment 2



Notes: The graph shows the percentage of times second movers chose LEFT in Experiment 2 and 95% confidence intervals by treatment. Panel (a) is based on the first round and panel (b) on all 20 rounds. Confidence intervals are based on predictions from a probit regression run on treatment dummies with standard errors clustered at the individual level (panel (b) only).

Table 4: Treatment effects on the percentage of LEFT choices in Experiment 2

Dep. var.: Choose LEFT	(1)	(2)	(3)	(4)
(a) Round 1				
Hot	0.021 (0.047)	0.030 (0.048)	-0.035 (0.066)	-0.021 (0.067)
High	0.035 (0.047)	0.045 (0.047)	-0.025 (0.068)	-0.008 (0.068)
Hot×*High*			0.117 (0.094)	0.103 (0.095)
Constant	0.305 (0.041)***	0.161 (0.231)	0.333 (0.047)***	0.192 (0.233)
Observations	400	400	400	400
(b) All rounds				
Hot	0.012 (0.037)	0.016 (0.036)	-0.029 (0.047)	-0.014 (0.048)
High	0.007 (0.036)	0.018 (0.036)	-0.012 (0.047)	0.004 (0.047)
Hot×*High*			0.063 (0.069)	0.047 (0.069)
Constant	0.318 (0.028)***	0.284 (0.183)	0.326 (0.031)***	0.294 (0.179)
Observations	5881	5881	5881	5881
Controls		✓		✓

Notes: Estimates from linear regressions (standard errors in parentheses). In panel (b), standard errors are clustered at the individual level. In columns (2) and (4), age, gender and education categories are included as controls (see Table C.1 in the Appendix for the estimates). *$P < 0.1$, **$P < 0.05$, ***$P < 0.01$.

Finally, we combine the data from Experiment 1 and Experiment 2 to test whether the *Hot* treatment effect is significantly different between the two experiments.[32] Given that the samples differ in terms of gender and age distribution—the share of female participants is significantly higher in Experiment 1 than in Experiment 2 and the age of the participants in Experiment 2 is significantly higher than in Experiment 1—we controlled for those variables.[33] As can be seen in Table 5, the reciprocation rate is significantly lower in Experiment 2 than

[32]Note that this analysis was not pre-registered.
[33]See Table C.7 in the Appendix for a balance table comparing the samples of the two experiments.

in Experiment 1, consistent with evidence reported in the literature that reciprocation is distinct from other-regarding (outcome-based) preferences (Cox, 2004). More important in our context is that the hot-cold reciprocation gap is statistically significant in Experiment 1 but not in Experiment 2. This is seen in the difference-in-differences effect, as captured by the coefficient of *Hot*×*Experiment 1*, which is statistically significant at the 1 percent level in the first round and at the 5 percent level in all rounds, whereas the main effect of *Hot* is not significant statistically.[34]

Table 5: Difference-in-differences effect of the hot decision mode

Dep. var.: Choose LEFT	**Round 1**		**All rounds**	
	(1)	(2)	(3)	(4)
Hot	0.022 (0.048)	0.028 (0.048)	0.013 (0.036)	0.017 (0.036)
Experiment 1	0.314 (0.047)***	0.315 (0.048)***	0.173 (0.033)***	0.177 (0.034)***
Hot×*Experiment 1*	-0.181 (0.066)***	-0.184 (0.066)***	-0.093 (0.051)*	-0.095 (0.051)*
Constant	0.321 (0.034)***	0.166 (0.156)	0.321 (0.023)***	0.285 (0.139)**
Controls		✓		✓
Observations	852	852	12565	12565

Notes: Estimates from linear regressions (standard errors in parentheses). In columns (2) and (4), age, gender and education categories are included as controls. In columns (3) and (4), standard errors are clustered at the individual level. *$P < 0.1$, **$P < 0.05$, ***$P < 0.01$.

5 Conclusion

Understanding the conditions under which people reciprocate another person's trust has been a recurring theme in the literature. An experiment was conducted to compare the reciprocation rates in a one-shot binary trust game between hot and cold decision modes, using a design that can shed light on the effect of being positively surprised by a partner's trust. We find that the reciprocation rate is lower in the case of hot decision making than in the case of cold decision making, and that the reciprocation gap does not depend on the likelihood of encountering trust. Both findings are in contrast to what one would expect if surprise were an important behavioral mechanism.

The findings cannot be rationalized by allowing for pro-social preferences or costly errors, nor by differences between hot and cold decision making in second-order beliefs or the level of the game's complexity. Rather, they are consistent with the presence of a commitment or sense of duty to reciprocate, which makes reciprocation a best reply under weaker conditions in the case of cold decision making than in the case of hot decision making. The higher recip-

[34] Results are robust to employing probit regressions (see Table C.9 in the Appendix).

rocation rate under cold decision making is consistent with a non-consequentialist preference for reciprocation whereby reciprocating provides additional non-monetary utility. This additional utility is present when directly reciprocating trust, or when committing to reciprocation in a cold plan, even if it turns out that the trustor has not chosen to trust, and can explain the different outcomes of hot and cold decision modes (Chen and Schonger, 2020).

The results are well-aligned with de Melo et al. (2019) who asked participants in a prisoner's dilemma experiment to program an automaton (defect or cooperate), which would act on their behalf against another participant. Programming the strategy generated a higher cooperation rate than direct play with other participants. We show that this effect extends to reciprocation, suggesting that the difference in the cooperation rate between direct and programmed decision making in that experiment was not just driven by a difference in the partner's beliefs regarding cooperation but by a preference to reciprocate expected cooperation.

The findings help to understand the mixed evidence in the literature on negative reciprocation, in which the second mover responds to a possibly unfair choice by the first mover (see Brandts and Charness, 2011, for an overview). A number of studies find that second movers in a hot decision mode punish unfair choices more severely than in a cold decision mode or are more likely to reject them in an ultimatum game (Aina et al., 2020; Brandts and Charness, 2003; Brosig et al., 2003; Falk et al., 2005; Oxoby and McLeish, 2004). In contrast, a meta-analysis shows that ultimatum games with hot responses have a lower rejection rate (Oosterbeek et al., 2004). The mixed evidence can be understood by viewing emotional responses and commitment to reciprocate as mechanisms leading to opposite effects. A hot response in this context may be motivated by anger or frustration and is therefore associated with a stronger reaction to unfairness than a cold response (Aina et al., 2020), whereas a commitment to (negative) reciprocation may induce individuals to respond more strongly to unfair choices under cold decision making. However, identifying the conditions in which each of the mechanisms is dominant has yet to be investigated.

Finally, our paper also sheds light on how an individual's *state of mind* may influence her willingness to reciprocate. While some individuals may be more prone to cooperative gestures, others may tend to rely more heavily on behavioral rules developed before the interaction (Bilancini et al., 2020). It is difficult, however, to identify an individual's internal decision mode and causally estimate its effect. The advantage of an experiment is that the decision-making mode is induced, thus facilitating causal inference. The experiment we carried out therefore helps to understand how an individuals' propensity for hot versus cold decision-making modes influence their behavior.

References

Aina, C., P. Battigalli, and A. Gamba (2020). Frustration and anger in the ultimatum game: An experiment. *Games and Economic Behavior 122*, 150–167.

Akdeniz, A. and M. van Veelen (2021). The evolution of morality and the role of commitment. *Evolutionary Human Sciences 3*, e41.

Akdeniz, A. and M. van Veelen (2023). Evolution and the ultimatum game. *Games and Economic Behavior 142*, 570–612.

Akerlof, G. (1982). Labor contracts as partial gift exchange. *Quarterly Journal of Economics 97*(4), 543–569.

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

Bilancini, E., L. Boncinelli, V. Capraro, and T. Celadin (2020). The effect of time pressure and motivated delay on cooperation and social norms in the online one-shot public goods game. *Working paper*.

Brandts, J. and G. Charness (2000). Hot vs. cold: Sequential responses and preference stability in experimental games. *Experimental Economics 2*(3), 227–238.

Brandts, J. and G. Charness (2003). Truth or consequences: An experiment. *Management Science 49*(1), 116–130.

Brandts, J. and G. Charness (2011). The strategy versus the direct-response method: A first survey of experimental comparisons. *Experimental Economics* (14), 375–398.

Brosig, J., J. Weimann, and C.-L. Yang (2003). The hot versus cold effect in a simple bargaining experiment. *Experimental Economics 6*, 75–90.

Brown, M., A. Falk, and E. Fehr (2004). Relational contracts and the nature of market interactions. *Econometrica 72*, 747–780.

Casari, M. and T. N. Cason (2009). The strategy method lowers measured trustworthy behavior. *Economics Letters* (103), 157–159.

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

Charness, G. and M. Rabin (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics 117*(3), 817–869.

Chen, D. L. and M. Schonger (2020). A theory of experiments: Invariance of equilibrium to the strategy method and implications. *TSE Working Paper No. 16-724*.

Chen, D. L. and M. Schonger (2022). Social preferences or sacred values? Theory and evidence of deontological motivations. *Science Advances 8*(19), eabb3925.

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior 46*, 260–281.

Cox, J. C. and D. T. Hall (2010). Trust with private and common property: Effects of stronger property right entitlements. *Games* (1), 527–550.

de Melo, C. M., S. Marsella, and J. Gratch (2019). Human cooperation when acting through autonomous machines. *Proceedings of the National Academy of Sciences 116*(9), 3482–3487.

Dufwenberg, M. and G. Kirchsteiger (2004). A theory of sequential reciprocity. *Games and Economic Behavior 47*(2), 268–298.

Falk, A., E. Fehr, and U. Fischbacher (2005). Driving forces behind informal sanctions. *Econometrica 73*, 2017–2030.

Fehr, E., G. Kirchsteiger, and A. Riedl (1993). Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics 108*, 1912–1917.

Fehr, E. and K. M. Schmidt (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics 114*(3), 817–868.

Fischbacher, U., S. Gächter, and S. Quercia (2012). The behavioral validity of the strategy method in public good experiments. *Journal of Economic Psychology 33*, 897–913.

Fong, Y.-f., C.-Y. Huang, and T. Offerman (2007). Guilt driven reciprocity in a psychological signaling game. *University of Amsterdam*.

Garcia-Pola, B., I. N. . K. J. (2020). Hot versus cold behavior in centipede games. *Journal of the Economic Science Association 6*, 226–238.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989). Psychological games and sequential rationality. *Games and Economic Behavior 1*(1), 60–79.

Greif, A. (1993). Contract enforceability and economic institutions in early trade: The Maghribi traders' coalition. *American Economic Review 83*(3), 525–548.

Johnson, N. D. and A. A. Mislin (2011). Trust games: A meta-analysis. *Journal of Economic Psychology* (32), 865–889.

Kehoe, T. J. and D. K. Levine (1993). Debt-constrained asset markets. *Review of Economic Studies 60*(4), 865–888.

Khalmetski, K., A. Ockenfels, and P. Werner (2015). Surprising gifts: Theory and laboratory evidence. *Journal of Economic Theory 159*, 163–208.

Kreps, D. M. (1990). Corporate culture and economic theory. In J. Alt and K. Shepsle (Eds.), *Perspectives on positive political economy*. Cambridge University Press.

Loewenstein, G. (2000). Emotions in economic theory and economic behavior. *American Economic Review 90*(2), 426–432.

Mckelvey, R. D. and T. R. Palfrey (1998, Jun). Quantal response equilibria for extensive form games. *Experimental Economics 1*(1), 9–41.

Meyer, W.-U., R. Reisenzein, and A. Schützwohl (1997). Toward a process analysis of emotions: The case of surprise. *Motivation and Emotion 21*(3), 251–274.

Minozzi, W. and J. Woon (2019). The limited value of a second opinion: Competition and exaggeration in experimental cheap talk games. *Games and Economic Behavior 117*, 144–162.

Oosterbeek, H., R. Sloof, and G. van de Kuilen (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics 7*, 171–188.

Oxoby, R. and K. McLeish (2004). Sequential decision and strategy method vectors in ultimatum bargaining: evidence on the strength of other-regarding behavior. *Economics Letters 84*, 399–405.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review 83*(5), 1281–1302.

Reisenzein, R., G. Horstmann, and A. Schützwohl (2019). The cognitive-evolutionary model of surprise: A review of the evidence. *Topics in Cognitive Science 11*(1), 50–74.

Reuben, E. and S. Suetens (2012). Revisiting strategic versus non-strategic cooperation. *Experimental Economics 15*, 244–43.

Ruffle, B. J. (1999). Gift giving with emotions. *Journal of Economic Behavior & Organization 39*, 399–420.

Solnick, S. J. (2007). Cash and alternate methods of accounting in an experimental game. *Journal of Economic Behavior & Organization 62*(2), 316–321.

Thomas, J. and T. Worrall (1990). Income fluctuation and asymmetric information: An example of a repeated principal-agent problem. *Journal of Economic Theory 51*(2), 367–390.

Online appendix

A Instructions

A.1 Instructions in phase 1 of Experiment 1

Consent Form

1. I understand that my participation in this study is voluntary and that I am free to withdraw my participation at any point. I understand that if I withdraw early, I do not receive a participation fee.
2. I understand that the amount of money I earn depends on my choices and the choices of other participants.
3. I understand that only anonymous data are collected from me and that none of my choices or responses can be linked to my personal data, in accordance with the European General Data Protection Regulation. The software used to achieve this purpose is Qualtrics, which is used by academic institutions worldwide.
4. I understand that the results of this study may be published in a scientific journal and/or presented at meetings or academic conferences.
5. I give permission to publish my anonymous data open access.
6. I confirm that I have read and understood the above and freely consent to participate in this study. I have been given adequate time to consider my participation.

o I CONSENT AND WISH TO PROCEED

o I DO NOT CONSENT AND WISH TO LEAVE THE STUDY

Welcome

Welcome to the study.

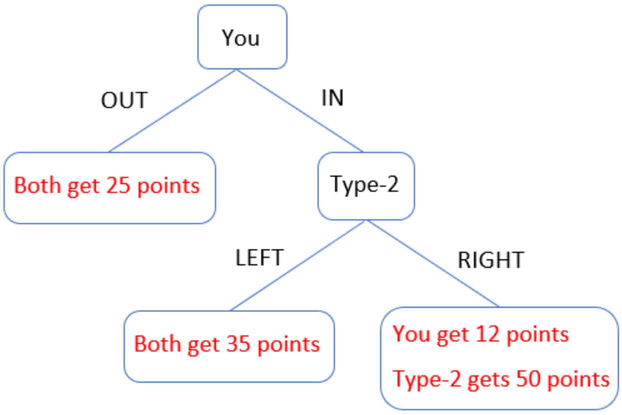
Please read the instructions carefully.

The amount of money you earn on top of the participation fee of 0.4 Pounds depends on decisions made by you and other participants. These earnings will be expressed in points, with points converting to pounds in the following way: 15 points = 1 Pound.

Your earnings will be transferred to you after the study has ended, at the latest within one month.

Your identity will not be revealed to other participants, and others' identities will not be revealed to you.

Please click "Next" to continue.



Instructions

There are two types of participants in this study: type-1 and type-2.

Type-1 participants make choices in a first phase of the study and type-2 participants make choices in a second phase, which will take place after the first phase has ended.

This is phase 1 of the study and you are a type-1 participant.

You will be paired with a type-2 participant and will be asked to make one choice (described on the next screen).

Please click "Next" to continue.

Instructions (continued)

As a type-1 participant, you are asked to choose between two options: IN and OUT. The type-2 participant you will be paired with will also be asked to choose between two options: LEFT and RIGHT.

The figure below shows how earnings (in points) depend on choices.

If you choose OUT, then both of you earn 25 points.

If you choose IN, the paired type-2 participant will be asked to choose between two options: LEFT and RIGHT. If the paired type-2 participant chooses LEFT, then both of you earn 35 points. If the paired type-2 participant chooses RIGHT, then you earn 12 points and the paired type-2 participant earns 50 points.

Because it is not certain that the number of participants recruited in both phases matches perfectly, there is a tiny chance that you will not be paired with any type-2 participant. If this occurs, you get a bonus payment of 1 Pound.

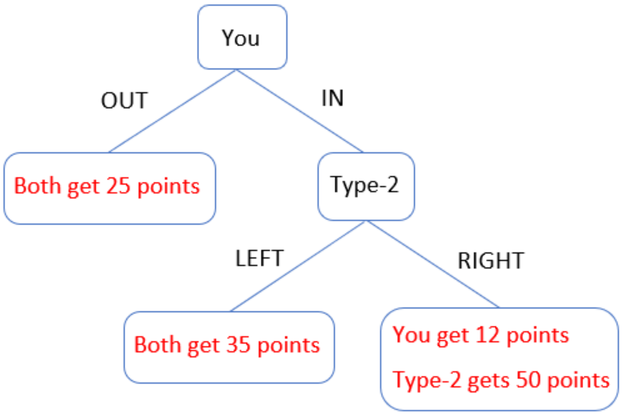
All payments will be administered after phase 2 has ended for all participants in this study.

Please click "Next" to continue to a quiz that tests your understanding of the instructions.

Quiz

Before you can enter your choice, a short quiz is ran to check whether you understand the instructions. You can only proceed if you answer all questions correctly. You have six chances to answer the questions.

Remember that you are a type-1 participant. The figure with earnings is included as a reminder.



1. How many points do you earn if you choose IN and the paired type-2 participant chooses LEFT?

2. How many points does the paired type-2 participant earn if you choose OUT?

3. How many points does the paired type-2 participant earn if you choose IN and the paired type-2 participant chooses RIGHT?

4. How many points do you earn if you choose IN and the paired type-2 participant chooses RIGHT?

Please click "Next" to continue.

Start

You have answered all quiz questions correctly. Please click "Next" to start.

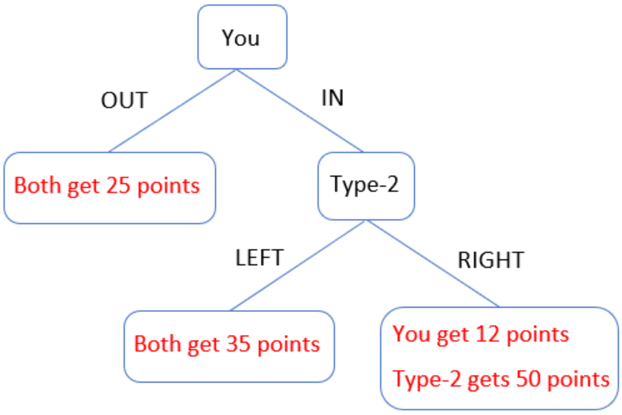
Enter Choices

You can enter your choice by clicking below on IN or OUT. The figure with earnings is included below as a reminder.

I choose:

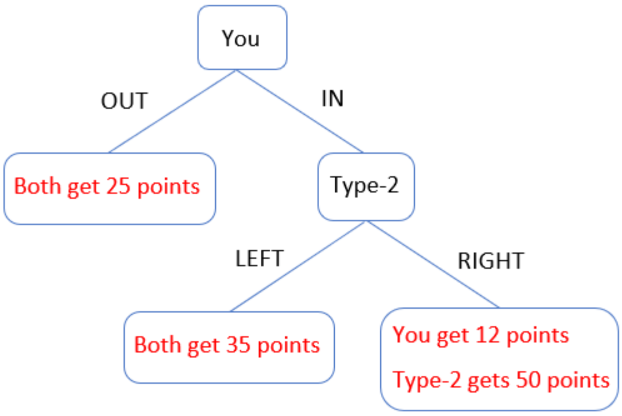
o OUT

o IN



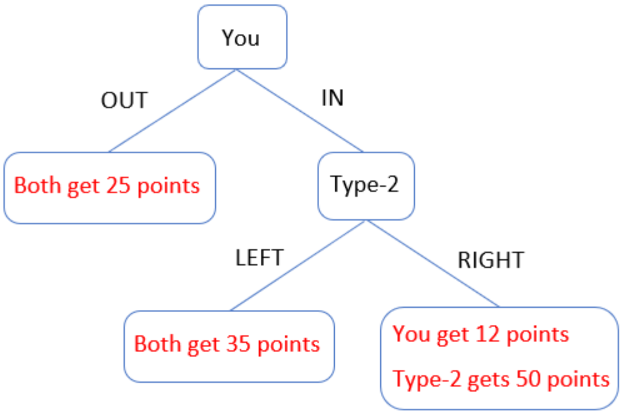
Guess the result (for the first movers who have chosen IN)

What percentage of type-2 participants do you think will choose LEFT? If your guess differs less than 5% from the actual percentage of type-2 participants who choose LEFT, you will get 20 points extra. The figure with earnings is included below as a reminder.
Please move the slider to the point that you think is the actual percentage
[Slider]



Guess the result (for the first movers who have chosen OUT)

Some of the other type-1 participants have chosen IN. What percentage of type-2 participants do you think will choose LEFT, in this case? If your guess differs less than 5% from the actual percentage of type-2 participants who choose LEFT, you will get 20 points extra. The figure with earnings is included below as a reminder.
Please move the slider to the point that you think is the actual percentage.
[Slider]



Payment information

You will be informed about the choice made by the paired type-2 participant and your earnings in the choice task and the guessing task after phase 2 of the study has ended. Your earnings will be transferred through Prolific.

Please make sure you click "Submit" to complete the study. Otherwise, your data will not be recorded.

A.2 Instructions in phase 2 of Experiment 1

Consent Form

1. I understand that my participation in this study is voluntary and that I am free to withdraw my participation at any point. I understand that if I withdraw early, I do not receive a participation fee.

2. I understand that the amount of money I earn depends on my choices and the choices of other participants.

3. I understand that only anonymous data are collected from me and that none of my choices or responses can be linked to my personal data, in accordance with the European General Data Protection Regulation. The software used to achieve this purpose is Qualtrics, which is used by academic institutions worldwide.

4. I understand that the results of this study may be published in a scientific journal and/or presented at meetings or academic conferences.

5. I give permission to publish my anonymous data open access.

6. I confirm that I have read and understood the above and freely consent to participate in this study. I have been given adequate time to consider my participation.

o I CONSENT AND WISH TO PROCEED

o I DO NOT CONSENT AND WISH TO LEAVE THE STUDY

Welcome

Welcome to the study. Please read the instructions carefully.

The amount of money you earn on top of the participation fee of 0.6 Pounds depends on decisions made by you and other participants. These earnings will be expressed in points, with points converting to pounds in the following way: 15 points = 1 Pound.

Your earnings will be transferred to you after the study has ended, at the latest within one month.

Your identity will not be revealed to other participants, and others' identities will not be revealed to you.

Please click "Next" to continue.

Instructions

There are two types of participants in this study: type-1 and type-2. Type-1 participants make choices in a first phase of the study and type-2 participants make choices in a second phase, which takes place after the first phase has ended.

This is phase 2 of the study and you are a type-2 participant.

Phase 2 consists of 20 rounds. In each round, there are two tasks: a choice task and a guessing task.

Please click "Next" to continue.

Instructions choice task (*Hot*)

In phase 1 of the study, type-1 participants have been asked to choose between two options: IN and OUT. As a type-2 participant, you will be asked to choose between two options: LEFT and RIGHT.

The figure below shows how earnings (in points) depend on choices.

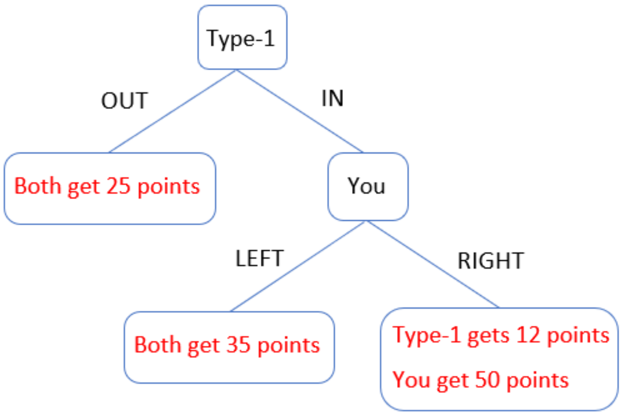
If the paired type-1 participant chooses OUT, then both of you earn 25 points.

If the paired type-1 participant chooses IN, then you are asked to choose between two options: LEFT and RIGHT. If you choose LEFT, then both of you earn 35 points. If you choose RIGHT, then you earn 50 points and the paired type-1 participant earns 12 points.

The percentage of type-1 participants who chose IN in phase 1 of the study is equal to 69%, and the percentage who chose OUT is equal to 31%.

Before you make your choice, you receive information about the choice of the paired type-1 participant.

There will be 20 rounds of the choice task. In each round, you are paired with a different type-1 participant.



Once all 20 rounds have been finished, one of these rounds is selected for payment.

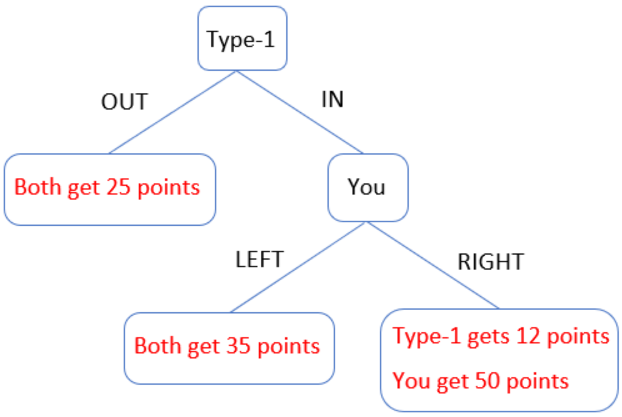
Because it is not certain that the number of participants recruited in both phases matches perfectly, there is a tiny chance that you will not be paired with any type-1 participant. If this occurs, you get a bonus payment of 1 Pound.

Please click "Next" to continue.

Instructions choice task (*Cold*)

In phase 1 of the study, type-1 participants have been asked to choose between two options: IN and OUT. As a type-2 participant, you will be asked to choose between two options: LEFT and RIGHT.

The figure below shows how earnings (in points) depend on choices.



If the paired type-1 participant chooses OUT, then both of you earn 25 points.

If the paired type-1 participant chooses IN, then you are asked to choose between two options: LEFT and RIGHT. If you choose LEFT, then both of you earn 35 points. If you choose RIGHT, then you earn 50 points and the paired type-1 participant earns 12 points.

The percentage of type-1 participants who chose IN in phase 1 of the study is equal to 69%, and the percentage who chose OUT is equal to 31%.

You will be asked whether you choose LEFT or RIGHT supposing that the paired type-1 participant chose IN. You have no information about the choice of the paired type-1 participant before you make your choice. You receive information about his/her choice after you made your choice.

There will be 20 rounds of the choice task. In each round you are paired with a different type-1 participant.

Once all 20 rounds have been finished, one of these rounds is selected for payment.

Because it is not certain that the number of participants recruited in both phases matches perfectly, there is a tiny chance that you will not be paired with any type-1 participant. If this occurs, you get a bonus payment of 1 Pound.

Please click "Next" to continue.

Instructions guessing task

In phase 1 of the study, type-1 participants who chose IN were asked to guess the percentage of type-2 participants who would choose LEFT. In each of the 20 rounds, before the choice task, you will be asked to state what you think is the average guess of these type-1 participants.

At the end of the study, one round will be selected to pay the guessing task. The round selected for the guessing task payment is different from the round selected for the choice task payment.

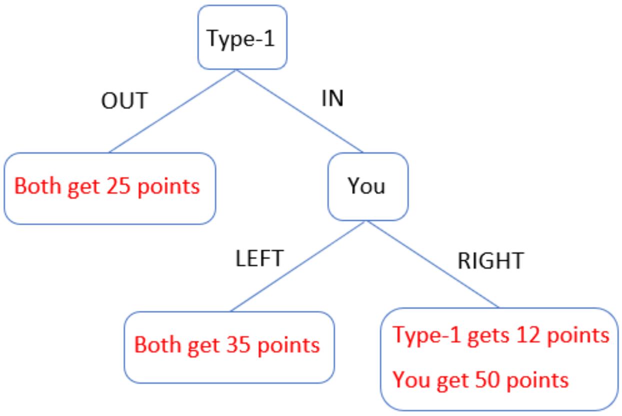
If your guess in the selected round differs less than 5% from the average guess of type-1 participants who chose IN, you earn 20 points extra.

Please click "Next" to continue to a quiz that tests your understanding of the instructions.

Quiz

Before you can enter your choice, a short quiz is ran to check whether you understand the instructions. You can only proceed if you answer all questions correctly. You have six chances to answer the questions.

Remember that you are a type-2 participant. The figure with earnings is included as a reminder.



1. How many points do you earn if the paired type-1 participant chooses IN and you choose LEFT?

2. How many points does the paired type-1 participant earn if the paired type-1 participant chooses OUT?

3. How many points do you earn if the paired type-1 participant chooses IN and you choose RIGHT?

4. How many points does the paired type-1 participant earn if the paired type-1 participant chooses IN and you choose RIGHT?

5. You are paired with the same type-1 participant for all rounds.

6. In each round, you know what the paired type-1 participant chose before you make your choice.

7. What is the percentage of type-1 participants who chose IN?

Please click “Next” to continue.

Start

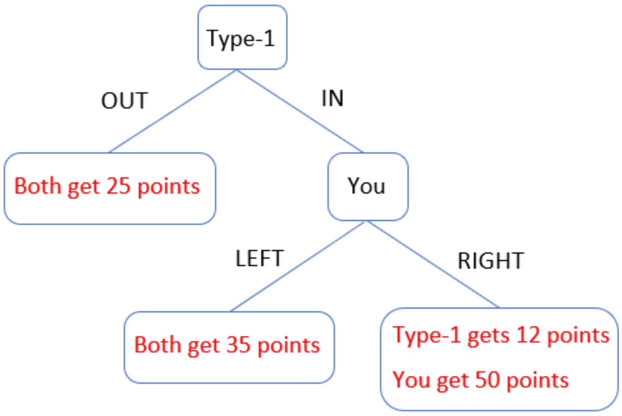
You have answered all quiz questions correctly. Please click “Next” to start.

Round xx – Guessing task

In phase 1 of the study, type-1 participants who chose IN were asked to guess the percentage of type-2 participants who would choose LEFT. The figure with earnings is included below as a reminder.

Please move the slider to the point that you think is the average guess of type-1 participants who chose IN.

[Slider]

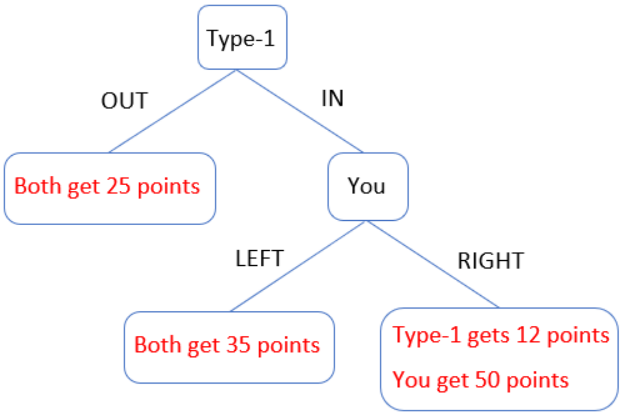


Round xx – Choice task (*Hot*)

The paired type-1 participant in round xx chose OUT, you don't need to make any choices. Remember that the percentage of type-1 participants who chose IN is equal to 69%, and the percentage who chose OUT is equal to 31%.

Round xx – Choice task (*Hot*)

The paired type-1 participant in round xx has chosen IN. You can enter your choice by selecting LEFT or RIGHT. The figure with earnings is included below as a reminder.
Remember that the percentage of type-1 participants who chose IN is equal to 69%, and the percentage who chose OUT is equal to 31%.
I choose:
o RIGHT
o LEFT



Round xx – Choice task (*Cold*)

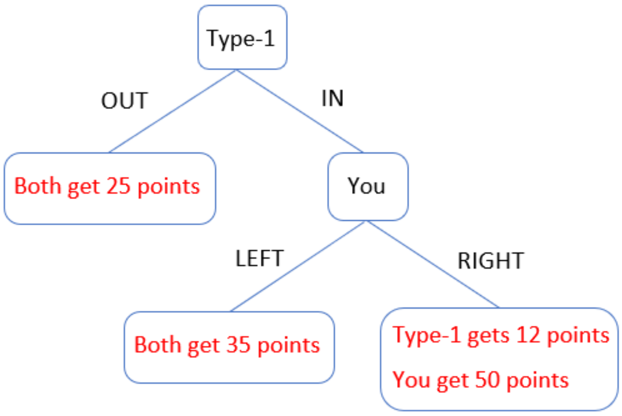
You can enter your choice by selecting LEFT or RIGHT. The figure with earnings is included below as a reminder.

Remember that the percentage of type-1 participants who chose IN is equal to 69%, and the percentage who chose OUT is equal to 31%.

I choose:

o RIGHT

o LEFT



Round xx – Result

The paired type-1 participant in round xx chose xx. You chose xx.

The paired type-1 participant earns xx points and you earn xx points in this round.

Payment information

At the end of the study, you will be informed about the selected round for the choice task payment. Choice made by the paired type-1 participant in that round and your earnings in the choice task.

You will also be informed about the selected round for the guessing task payment. The actual average guess of type-1 participants who chose IN and your earnings in the guessing task.

Your earnings will be transferred through Prolific.

Please click "Next" to fill in the end of survey questionnaire.

Questionnaire

Please fill in the end of study questionnaire.

1. Did you think about what the paired type-1 participant would choose before you made your own choice?

o Yes

o A little

o No

2. Surprise Were you surprised by the choice of the paired type-1 participant when he/she chose IN?
o Yes
o A little
o No

3. What do you think of the complexity of this study on a scale from 0 to 10?

4. How well does the following statement describe you as a person? When someone does me a favor I am willing to return it. Please indicate your answer on a scale from 0 to 10.

5. Please think about what you would do in the following situation. You are in an area you are not familiar with, and you realize that you lost your way. You ask a stranger for directions. The stranger offers to take you to your destination. Helping you costs the stranger about 20 Pounds. However, the stranger says he or she does not want any money from you. You have 6 presents with you. The cheapest present costs 5 Pounds, the most expensive one costs 30 Pounds. Do you give one of the presents to the stranger as a ”thank-you”-gift? If so, which present do you give to the stranger?
o No present
o The present worth 5 Pounds
o The present worth 10 Pounds
o The present worth 15 Pounds
o The present worth 20 Pounds
o The present worth 25 Pounds
o The present worth 30 Pounds

6. What is your age?

7. What is your gender?
o Female
o Male
o Other

o I would rather not say

8. Your educational background
o High school or below
o Bachelor student
o Master student
o PhD student or above
o Other

9. Comment Do you have any comments to this study?
Please make sure you click "Submit" to complete the study. Otherwise, your data will not be recorded.

A.3 Instructions in phase 2 of Experiment 2

Consent Form

1. I understand that my participation in this study is voluntary and that I am free to withdraw my participation at any point. I understand that if I withdraw early, I do not receive a participation fee.
2. I understand that the amount of money I earn depends on my choices.
3. I understand that only anonymous data are collected from me and that none of my choices or responses can be linked to my personal data, in accordance with the European General Data Protection Regulation. The software used to achieve this purpose is Qualtrics, which is used by academic institutions worldwide.
4. I understand that the results of this study may be published in a scientific journal and/or presented at meetings or academic conferences.
5. I give permission to publish my anonymous data open access.
6. I confirm that I have read and understood the above and freely consent to participate in this study. I have been given adequate time to consider my participation.
o I CONSENT AND WISH TO PROCEED
o I DO NOT CONSENT AND WISH TO LEAVE THE STUDY

Welcome

Welcome to the study.
Please read the instructions carefully.

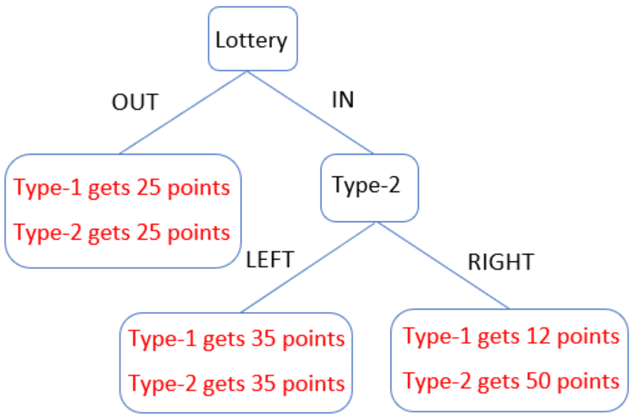
The amount of money you earn on top of the participation fee of 0.6 Pounds depends on decisions made by you. These earnings will be expressed in points, with points converting to pounds in the following way: 15 points = 1 Pound.
Your earnings will be transferred to you after the study has ended, at the latest within one month.
Your identity will not be revealed to other participants, and others' identities will not be revealed to you.
Please click "Next" to continue.

Instructions

There are two types of participants in this study: type-1 and type-2. Type-1 participants don't make choices, while type-2 participants make choices. You are a type-2 participant. The study consists of 20 rounds. In each round, you are paired with a different type-1 participant and are asked to make a choice. Please click "Next" to continue.

Instructions choice task (*Hot*)

In each round, one of the following two options is randomly drawn in a lottery: IN or OUT. As a type-2 participant, you will be asked to choose between two options: LEFT and RIGHT
The figure below shows how your earnings and the earnings of the paired type-1 participant (in points) depend on the outcome of the lottery and your choice.



If the outcome of the lottery draw is OUT, then both you and the paired type-1 participant earn 25 points.
If the outcome of the lottery draw is IN, then you are asked to choose between two options: LEFT and RIGHT. If you choose LEFT, then both you and the paired type-1 participant earn 35 points. If you choose RIGHT, then you earn 50 points and the paired type-1 participant earns 12 points.

The probability that IN is drawn is equal to 69%, and the probability that OUT is drawn is equal to 31%.

Before you make your choice, you receive information about the outcome of the lottery draw.

There will be 20 rounds of this task. In each round, you are paired with a different type-1 participant.

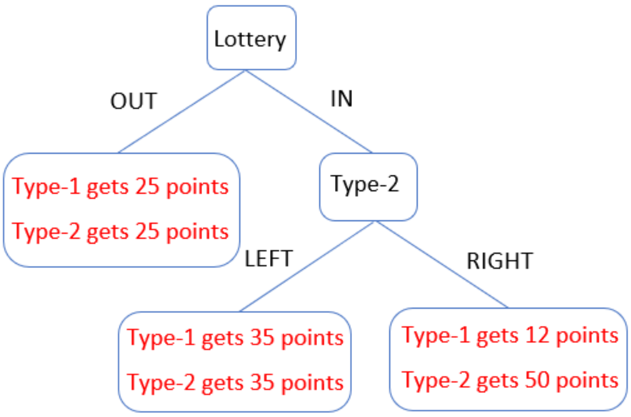
Once all 20 rounds have been finished, one of these rounds is selected for payment.

Please click "Next" to continue to a quiz that tests your understanding of the instructions.

Instructions choice task (*Cold*)

In each round, one of the following two options is randomly drawn in a lottery: IN or OUT. As a type-2 participant, you will be asked to choose between two options: LEFT and RIGHT

The figure below shows how your earnings and the earnings of the paired type-1 participant (in points) depend on the outcome of the lottery and your choice.



If the outcome of the lottery draw is OUT, then both you and the paired type-1 participant earn 25 points.

If the outcome of the lottery draw is IN, then you are asked to choose between two options: LEFT and RIGHT. If you choose LEFT, then both you and the paired type-1 participant earn 35 points. If you choose RIGHT, then you earn 50 points and the paired type-1 participant earns 12 points.

The probability that IN is drawn is equal to 69%, and the probability that OUT is drawn is equal to 31%.

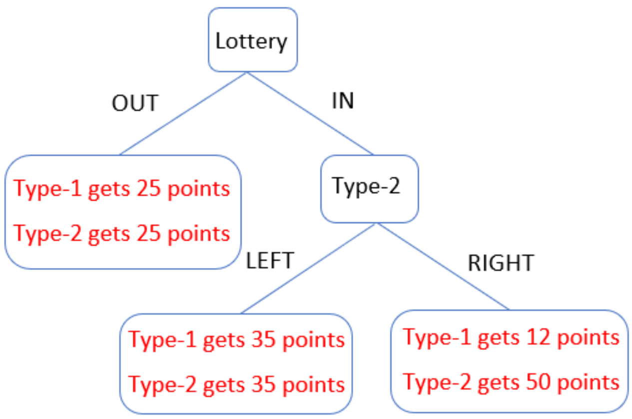
You will be asked whether you choose LEFT or RIGHT supposing that the outcome of the lottery draw is IN. You have no information about the outcome before you make your choice. You receive information about it after you made your choice.

There will be 20 rounds of this task. In each round, you are paired with a different type-1 participant.

Once all 20 rounds have been finished, one of these rounds is selected for payment.
Please click "Next" to continue to a quiz that tests your understanding of the instructions.

Quiz

Before you can enter your choice, a short quiz is ran to check whether you understand the instructions. You can only proceed if you answer all questions correctly. You have six chances to answer the questions. It is your sixth chance.
Remember that you are a type-2 participant. The figure with earnings is included as a reminder.



1. How many points do you earn if the outcome of the lottery draw is IN and you choose LEFT?
2. How many points does the paired type-1 participant earn if the outcome of the lottery draw is OUT?
3. How many points do you earn if the outcome of the lottery draw is IN and you choose RIGHT?
4. How many points does the paired type-1 participant earn if the outcome of the lottery draw is IN and you choose RIGHT?
5. You are paired with the same type-1 participant for all rounds.
6. In each round, you know the outcome of the lottery draw before you make your choice.
7. What is the probability that IN is drawn in the lottery in a certain round?

Please click "Next" to continue.

Start

You have answered all quiz questions correctly. Please click "Next" to start.

Round xx – Choice task (*Hot*)

The outcome of the lottery draw is OUT, so you don't need to make a choice.
Remember that IN is drawn with probability 69% and OUT is drawn with probability 31%.

Round xx – Choice task (*Hot*)

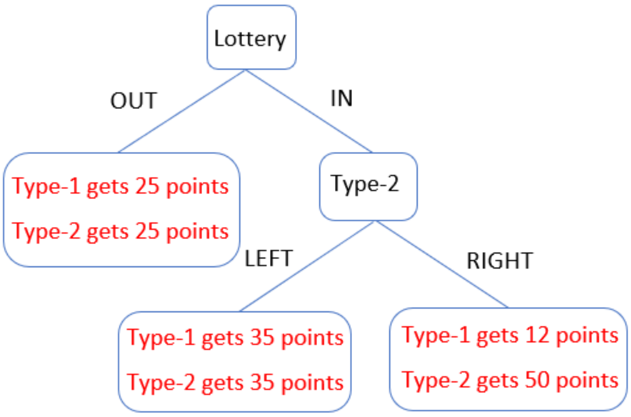
The outcome of the lottery draw is IN. You can enter your choice by selecting LEFT or RIGHT. The figure with earnings is included below as a reminder.

Remember that IN is drawn with probability 69%, and OUT is drawn with probability 31%.

I choose:

o RIGHT

o LEFT



Round xx – Choice task (*Cold*)

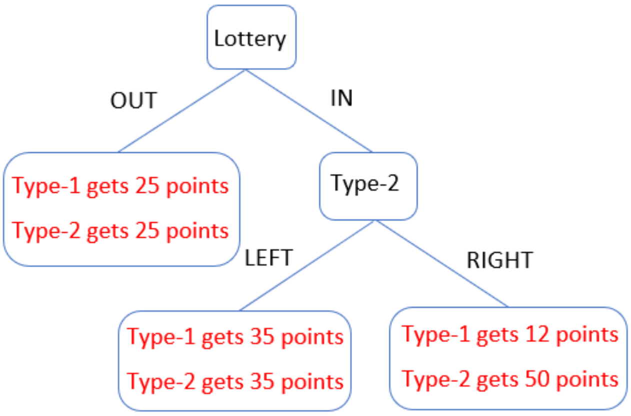
You can enter your choice by selecting LEFT or RIGHT. The figure with earnings is included below as a reminder.

Remember that IN is drawn with probability 69%, and OUT is drawn with probability 31%.

I choose:

o RIGHT

o LEFT



Round xx – Result

he outcome of the lottery draw in round xx was xx. You chose xx.
The paired type-1 participant earns xx points and you earn xx points in this round.

Payment information

At the end of the study, you are informed about the selected round for payment, the outcome of the lottery draw in that round and your final earnings.
Your earnings will be transferred through Prolific.
Please click ”Next ” to fill in the end of survey questionnaire.

B Equilibrium calculations

B.1 Standard theory

We calculate the subgame perfect Nash equilibrium for the *Hot* and *Cold* game under the standard assumption that payoffs represent utilities.

We first focus on pure strategy Nash equilibrium. From Figure 1 we can see that there is a unique pure strategy Nash equilibrium in this game: the first mover does not trust and the second mover defects. In both *Hot* and *Cold* decision making, this equilibrium is subgame perfect. In the *Hot* game, the second mover's best response is to defect in the subgame where the matched first mover trusts. Anticipating that, the best response for the first mover is not to trust. In the *Cold* game, the game consists of only one subgame. Therefore, all Nash equilibria are subgame perfect. The pure strategy Nash equilibrium (Not trust, Defect) is subgame perfect.

Next, we focus on mixed strategy Nash equilibrium. We can calculate the mixed strategy Nash equilibria using Figure 1. There exists an infinite number of subgame perfect equilibria in mixed strategies, where with probability one the first mover does not trust and with a probability lower or equal to 0.57 the second mover reciprocates. In what follows we present the proof of the mixed strategy Nash equilibria.

The first mover doesn't have incentive to deviate to choose trust given that the second mover reciprocates with probability $q \leqslant 0.57$:

$$U(\text{Trust}) = 35 \times q + 12 \times (1 - q) \leqslant U(\text{Not trust}) = 25$$

when $q \leqslant 0.57$.

The second mover doesn't have incentive to deviate to another probability given that the first mover does not trusts. The reason is that the monetary payoff the second mover can get is always 25 not matter what she chooses.

In the *Hot* game, the mixed strategy Nash equilibria is not subgame perfect because in the subgame where the first mover trusts, defect with probability one is the best response. However, in the *Cold* game, the mixed strategy Nash equilibria is subgame perfect because all Nash equilibria are subgame perfect in the *Cold* game where the game consists of only one subgame.

B.2 Surprise model

We calculate the equilibrium predictions if the second mover has a utility function as defined by Equation 1. We denote the payoff that both players get if second mover reciprocates first mover's trust by c, the payoff that the second mover gets if she reciprocates and first mover trusts by t, and the payoff that first mover gets if the second mover defects and she trusts by s. We furthermore assume that the surprise sensitivity is individual-specific and denoted by θ_i for the ith second mover, and is distributed in the population according to a specific distribution.[35] We calculate a intrapersonal equilibrium of the second mover in the *Hot* game along the lines of Aina et al. (2020). The equilibrium prescribes that planned behavior must be consistent with the incentives determined by the plan itself.

We denote the second mover's plan of action, that is, her initial first-order belief about herself reciprocating trust, as q. We calculate probability q^* that the second mover will reciprocate trust in the intrapersonal equilibrium given her surprise sensitivity θ_i.

We define the level of surprise $S(p,q) \geqslant 0$ to be the gap (if positive) between the second mover's actual payoff and her expected payoff: [36]

$$S(p,q) = [(1-q)t + qc - E(p,q)]^+ = (1-p)(1-q)t + (1-p)qc - (1-p)d^{37} \qquad (2)$$

Hence, $S(p,q)$ is assumed to be negatively correlated with p and q. Using Equations 1 and 2, we can calculate the intrapersonal equilibrium in *Hot* for different θ_i value. Consider the binary

[35] The functional form of the distribution of θ_i does not influence the comparative-static predictions of the surprise model.

[36] In Battigalli et al. (2019), instead of using the actual diminished expectation, frustration depends on the difference between the maximum payoff the second mover can get if first mover chooses greedy offer and his expected payoff. The reason why they didn't use the actual diminished expectation for frustration is that this substitution would dramatically impact behavior (see discussion section of their paper). They gave an example of a common-interest game: Ann chooses Out or In; In the former case the game ends with payoffs (1, 1), and in the latter case Bob chooses between (0, 0)and (2, 2). If Bob has high sensitivity parameter for anger, there is a sequential equilibrium where Ann chooses Out and Bob would go for (0, 0). Following In, Bob would be frustrated because he (so-to-say) feels locked-in with his stage-2 planned action. Their formulation of frustration rules this circumstance out. This problem doesn't appear in our surprise model. We also consider different surprise formulations. For example, consider substituting $(1-q)t + qc$ to be t for $S(p,q)$ (Here surprise is the gap between the maximum payoff and the expectation). And consider substituting $(1-q)t + qc$ to be c (Here surprise is the gap between the minimum payoff and the expectation). Both specifications will lead to different intrapersonal equilibria from those calculated based on the first specification. However, all three specifications generate the same predictions about the comparative static of reciprocation rate across treatments and all three specifications will not lead to the problem described above.

[37] Note that $(1-q)t + qc - E(p,q)$ is always positive in our setting because $t > c > d$.

trust game played in *Hot*. The second mover expects to reciprocate trust, i.e., $q = 1$, if

$$\begin{aligned} U_i(Reciprocate) &> U_i(Defect) \\ c + \theta_i S(p,1)c &> t + \theta_i S(p,1)s \\ \theta_i > \hat{\theta}(p,1) &= \frac{t-c}{S(p,1)(c-s)} \end{aligned}$$

Thus, the choice to reciprocate is incentive compatible when θ_i is big enough.

The second mover can be expected to defect if

$$\begin{aligned} U_i(Reciprocate) &< U_i(Defect) \\ c + \theta_i S(p,0)c &> t + \theta_i S(p,0)s \\ \theta_i < \hat{\theta}(p,0) &= \frac{t-c}{S(p,0)(c-s)} \end{aligned}$$

Thus, the plan to defect is incentive compatible when θ_i is small enough.

Note that $\hat{\theta}(p,0) < \hat{\theta}(p,1)$ because $S(p,q)$ is decreasing in q. For intermediate values θ_i, deterministic plans are self-defeating, i.e., a plan to reciprocate induces defection and a plan to defect induces reciprocation via the effect of q on the initially expected payoff and surprise. [38] Thus, the only incentive-compatible plan is a probability of reciprocate $q^* \in (0,1)$ that makes the second mover indifferent. Therefore, for intermediate value $\hat{\theta}(p,0) < \theta_i < \hat{\theta}(p,1)$, the second mover will choose mixed strategy q^* such that:

$$\begin{aligned} U(Defect) &= U(Reciprocate) \\ t + \theta_i S(p,q)s &= c + \theta_i S(p,q)c \\ q^* &= 1 - \frac{1}{\theta_i(1-p)(c-s)} - \frac{d-c}{t-c} \end{aligned}$$

In conclusion, the intrapersonal equilibrium is:

$$\begin{cases} q = 1, & if \theta_i < \hat{\theta}(p,1) \\ q = q^*, & if \theta_i \in [\hat{\theta}(p,1), \hat{\theta}(p,0)] \\ q = 0, & if \theta_i > \hat{\theta}(p,0) \end{cases} \tag{3}$$

We can use the intrapersonal equilibrium to calculate the comparative static of reciprocation

[38]The reason is that $S(p,q)$ is decreasing in q, this means that a plan to reciprocate induces small surprise such that the level of surprise is too small to incentivize the second mover to reciprocate. A plan to defect induces a large surprise such that the second mover is willing to reciprocate. Similar self-defeating property in the calculation of intrapersonal equilibrium can be found in Aina et al. (2020).

rate across treatments. Remember that $S(p,q) = 0$ in *Cold*. In *Cold*, the second mover's best reply depends only on her first-order beliefs p. If $p = 0$, then any probability that the second mover will reciprocate is the best reply since her choice can not influence the final outcome. Whereas if $p > 0$, her best-reply is to defect. Because even when there is a small chance that the first mover will trust, the second mover will defect because her temptation payoff is bigger than her reciprocal payoff. The crux of the surprise model is that for any $p > 0$, there is a threshold $\hat{\theta}$ in *Hot* and no threshold in *Cold*. Therefore, for any distribution θ_i and any $p > 0$, the intrapersonal equilibrium predicts that the reciprocation rate is higher in *Hot* than that in *Cold*.

Remember that $S(p,q)$ is decreasing in p, making $\hat{\theta}(p_{low}, 1) < \hat{\theta}(p_{high}, 1), \hat{\theta}(p_{low}, 0) < \hat{\theta}(p_{high}, 0)$ and $q^*(p_{low}) > q^*(p_{high})$. It means that the thresholds $\hat{\theta}$ when $p = p_{low}$ is lower than the threshold when $p = p_{high}$ and the probability to reciprocate in mix strategy intrapersonal equilibrium q^* when $p = p_{low}$ is higher than that when $p = p_{high}$. The reciprocation rate in *Hot* is higher, the lower p. The reciprocation rate in *Cold* is not affected by p. The comparative static predictions following intrapersonal equilibrium are consistent with the predictions based on the best reply analysis in our main text.

C Supplementary tables

Table C.1: Treatment effects on the percentage of LEFT choices (full)

Dep. var.: choose LEFT	Round 1 (1)	(2)	All rounds (3)	(4)
(a) Experiment 1				
Hot	-0.151 (0.046)***	-0.122 (0.065)*	-0.067 (0.036)*	-0.061 (0.048)
High	-0.016 (0.046)	0.014 (0.066)	-0.036 (0.036)	-0.034 (0.046)
Hot×*High*		-0.059 (0.093)		-0.009 (0.069)
Constant	0.451 (0.208)**	0.440 (0.209)**	0.404 (0.179)**	0.403 (0.179)**
Controls:				
Age	0.004 (0.003)	0.004 (0.003)	0.005 (0.002)**	0.005 (0.002)**
Female	0.097 (0.050)*	0.097 (0.050)*	0.082 (0.043)*	0.082 (0.043)*
High school	-0.135 (0.198)	-0.138 (0.198)	-0.139 (0.164)	-0.139 (0.164)
BSc	0.052 (0.188)	0.046 (0.188)	-0.077 (0.156)	-0.077 (0.156)
MSc	0.006 (0.193)	-0.001 (0.193)	-0.031 (0.158)	-0.031 (0.158)
PhD	0.019 (0.206)	0.007 (0.207)	-0.021 (0.169)	-0.023 (0.170)
Observations	452	452	6684	6684
(b) Experiment 2				
Hot	0.030 (0.048)	-0.021 (0.067)	0.016 (0.036)	-0.014 (0.048)
High	0.045 (0.047)	-0.008 (0.068)	0.018 (0.036)	0.004 (0.047)
Hot×*High*		0.103 (0.095)		0.047 (0.069)
Constant	0.161 (0.231)	0.192 (0.233)	0.284 (0.183)	0.294 (0.179)
Controls:				
Age	0.003 (0.002)	0.003 (0.002)	0.002 (0.002)	0.002 (0.002)
Female	0.106 (0.044)**	0.102 (0.044)**	0.098 (0.033)***	0.096 (0.034)***
High school	-0.045 (0.222)	-0.058 (0.223)	-0.118 (0.177)	-0.123 (0.172)
BSc	-0.021 (0.215)	-0.027 (0.215)	-0.076 (0.173)	-0.079 (0.167)
MSc	0.004 (0.219)	-0.004 (0.219)	-0.103 (0.178)	-0.107 (0.172)
PhD	-0.139 (0.232)	-0.136 (0.232)	-0.223 (0.185)	-0.223 (0.181)
Observations	400	400	5881	5881

Notes: Estimates come from linear regressions (standard errors in parentheses). In column (3) and (4), standard errors are clustered at the individual level. *$P < 0.1$, **$P < 0.05$, ***$P < 0.01$.

Table C.7: Balance table Experiment 1 versus Experiment 2

	Experiment 1	Experiment 2	*P*-value
Attrition rate	0.244	0.260	0.163
Female	0.699	0.552	0
Age	24.126	29.098	0
Education		0.447	0.447
High school or below	0.092	0.106	
Bachelor	0.458	0.425	
Master	0.131	0.166	
PhD or above	0.047	0.049	
Other	0.011	0.009	
Observations	452	400	

Notes: The table shows the attrition rate and averages of the socio-demographic variables collected in the post-experimental surveys. The attrition rate is equal to the percentage of second movers who dropped out of the experiments after the instructions were shown on the screen, i.e. after they were assigned to a treatment. The education category refers to the highest obtained diploma. The *P*-values in the last column refer to likelihood ratio tests conducted on estimates from probit regression without a constant that include treatment indicators as independent variables (for attrition and female), to a Wald test conducted on estimates from linear regressions (for age), or to a Chi-square test (for eduction).

Table C.2: Treatment effects on the percentage of LEFT choices (probit)

Dep. var.: choose LEFT	Round 1 (1)	Round 1 (2)	All rounds (3)	All rounds (4)
(a) Experiment 1				
Hot	-0.395 (0.122)***	-0.320 (0.170)*	-0.171 (0.093)*	-0.154 (0.122)
High	-0.041 (0.121)	0.038 (0.174)	-0.093 (0.092)	-0.086 (0.118)
Hot×High		-0.154 (0.243)		-0.026 (0.176)
Constant	-0.137 (0.540)	-0.164 (0.543)	-0.251 (0.448)	-0.254 (0.449)
Controls:				
Age	0.011 (0.008)	0.012 (0.008)	0.013 (0.006)**	0.013 (0.006)**
Female	0.256 (0.132)*	0.259 (0.132)**	0.211 (0.110)*	0.211 (0.110)*
High school	-0.357 (0.516)	-0.368 (0.518)	-0.358 (0.412)	-0.358 (0.412)
BSc	0.134 (0.489)	0.116 (0.492)	-0.192 (0.389)	-0.193 (0.390)
MSc	0.013 (0.502)	-0.009 (0.505)	-0.076 (0.395)	-0.078 (0.396)
PhD	0.045 (0.538)	0.011 (0.542)	-0.051 (0.424)	-0.055 (0.425)
Observations	452	452	6684	6684
(b) Experiment 2				
Hot	0.080 (0.133)	-0.072 (0.187)	0.032 (0.102)	-0.074 (0.136)
High	0.141 (0.132)	-0.019 (0.191)	0.041 (0.102)	-0.009 (0.133)
Hot×High		0.308 (0.265)		0.166 (0.196)
Constant	-0.846 (0.616)	-0.768 (0.626)	-0.341 (0.552)	-0.313 (0.542)
Controls:				
Age	0.009 (0.007)	0.009 (0.007)	0.005 (0.005)	0.005 (0.005)
Female	0.379 (0.135)***	0.374 (0.135)***	0.189 (0.104)*	0.188 (0.104)*
High school	-0.242 (0.598)	-0.268 (0.605)	-0.499 (0.536)	-0.513 (0.525)
BSc	-0.160 (0.577)	-0.169 (0.584)	-0.343 (0.524)	-0.350 (0.512)
MSc	-0.085 (0.588)	-0.101 (0.595)	-0.423 (0.535)	-0.434 (0.523)
PhD	-0.503 (0.633)	-0.489 (0.640)	-0.789 (0.574)	-0.785 (0.563)
Observations	400	400	5881	5881

Notes: Estimates come from probit regressions (standard errors in parentheses). In column (3) and (4), standard errors are clustered at the individual level. *$P < 0.1$, **$P < 0.05$, ***$P < 0.01$.

Table C.3: Treatment effects on second-order beliefs (Experiment 1)

Dep. var.: Second-order beliefs	(1)	(2)	(3)	(4)
(a) Round 1				
Hot	0.030 (0.018)	0.033 (0.018)*	0.025 (0.026)	0.026 (0.026)
High	0.125 (0.018)***	0.122 (0.018)***	0.120 (0.026)***	0.115 (0.026)***
Hot×High			0.010 (0.036)	0.013 (0.037)
Constant	0.408 (0.016)***	0.444 (0.082)***	0.410 (0.018)***	0.446 (0.082)***
Controls:				
Age		-0.001 (0.001)		-0.001 (0.001)
Female		0.048 (0.020)**		0.048 (0.020)**
Highschool		-0.056 (0.078)		-0.055 (0.078)
BSc		-0.049 (0.074)		-0.048 (0.074)
MSc		-0.021 (0.076)		-0.019 (0.076)
PhD		-0.011 (0.081)		-0.008 (0.082)
Observations	452	452	452	452
(b) All rounds				
Hot	0.002 (0.015)	0.005 (0.015)	0.012 (0.022)	0.013 (0.022)
High	0.194 (0.015)***	0.192 (0.015)***	0.204 (0.022)***	0.199 (0.022)***
Hot×High			-0.020 (0.030)	-0.016 (0.030)
Constant	0.371 (0.014)***	0.397 (0.078)***	0.366 (0.016)***	0.394 (0.079)***
Controls:				
Age		0.001 (0.001)		0.001 (0.001)
Female		0.014 (0.017)		0.014 (0.017)
High school		-0.069 (0.071)		-0.070 (0.071)
BSc		-0.066 (0.068)		-0.067 (0.068)
MSc		-0.031 (0.069)		-0.033 (0.069)
PhD		-0.016 (0.074)		-0.019 (0.074)
Observations	9040	9040	9040	9040

Notes: Estimates come from linear regressions (standard errors in parentheses). In panel (b), standard errors are clustered at the individual level. In columns (2) and (4), age, gender and education categories are included as controls. *$P < 0.1$, **$P < 0.05$, ***$P < 0.01$.

Table C.4: Treatment effects on level of surprise

Dep. var.: Surprise (yes and a little = 1)	(1)	(2)	(3)	(4)
(a) Experiment 1				
Hot	0.092 (0.044)**	0.097 (0.045)**	0.090 (0.062)	0.094 (0.062)
High	-0.133 (0.044)***	-0.133 (0.044)***	-0.135 (0.063)**	-0.137 (0.063)**
Hot×*High*			0.004 (0.089)	0.007 (0.089)
Constant	0.675 (0.038)***	1.004 (0.200)***	0.675 (0.044)***	1.005 (0.201)***
Controls:				
Age		-0.007 (0.003)**		-0.007 (0.003)**
Female		-0.066 (0.048)		-0.066 (0.048)
High school		-0.117 (0.190)		-0.117 (0.190)
BSc		-0.105 (0.180)		-0.104 (0.181)
MSc		-0.128 (0.185)		-0.128 (0.185)
PhD		-0.088 (0.198)		-0.087 (0.199)
Observations	452	452	452	452
(a) Experiment 2				
Hot	-0.108 (0.047)**	-0.121 (0.048)**	-0.164 (0.066)**	-0.181 (0.067)***
High	-0.143 (0.047)***	-0.139 (0.048)***	-0.201 (0.068)***	-0.202 (0.069)***
Hot×*High*			0.115 (0.095)	0.122 (0.096)
Constant	0.482 (0.041)***	0.802 (0.233)***	0.510 (0.047)***	0.839 (0.235)***
Controls:				
Age		-0.000 (0.002)		0.000 (0.002)
Female		-0.023 (0.044)		-0.028 (0.044)
High school		-0.358 (0.224)		-0.373 (0.224)*
BSc		-0.295 (0.216)		-0.302 (0.216)
MSc		-0.261 (0.220)		-0.271 (0.220)
PhD		-0.384 (0.234)		-0.380 (0.234)
Observations	400	400	400	400

Notes: Estimates come from linear regressions (standard errors in parentheses). In columns (2) and (4), age, gender and education categories are included as controls. *$P < 0.1$, **$P < 0.05$, ***$P < 0.01$.

Table C.5: Treatment effects on level of complexity

Dep. var.: Complexity level	(1)	(2)	(3)	(4)
(a) Experiment 1				
Hot	0.293 (0.197)	0.226 (0.198)	0.203 (0.277)	0.165 (0.277)
High	0.288 (0.197)	0.315 (0.197)	0.197 (0.280)	0.253 (0.281)
Hot×*High*			0.182 (0.395)	0.125 (0.397)
Constant	4.560 (0.170)***	2.753 (0.888)***	4.605 (0.196)***	2.777 (0.893)***
Controls:				
Age		0.030 (0.013)**		0.030 (0.013)**
Female		0.289 (0.215)		0.288 (0.215)
High school		1.188 (0.844)		1.195 (0.846)
BSc		0.986 (0.803)		0.998 (0.804)
MSc		0.488 (0.823)		0.504 (0.826)
PhD		0.759 (0.880)		0.784 (0.885)
Observations	452	452	452	452
(a) Experiment 2				
Hot	-0.160 (0.225)	-0.153 (0.226)	-0.091 (0.315)	-0.081 (0.317)
High	-0.274 (0.226)	-0.256 (0.224)	-0.202 (0.323)	-0.180 (0.323)
Hot×*High*			-0.142 (0.452)	-0.148 (0.452)
Constant	3.290 (0.194)***	1.998 (1.095)*	3.255 (0.223)***	1.954 (1.105)*
Controls:				
Age		0.035 (0.011)***		0.035 (0.011)***
Female		0.137 (0.207)		0.142 (0.208)
High school		0.644 (1.053)		0.662 (1.056)
BSc		-0.044 (1.017)		-0.036 (1.018)
MSc		0.377 (1.035)		0.389 (1.036)
PhD		0.249 (1.098)		0.243 (1.100)
Observations	400	400	400	400

Notes: Estimates come from linear regressions (standard errors in parentheses). In columns (2) and (4), age, gender and education categories are included as controls. *$P < 0.1$, **$P < 0.05$, ***$P < 0.01$.

Table C.6: Treatment effects on number of attempts

Dep. var.: Number of attempts	(1)	(2)	(3)	(4)
(a) Experiment 1				
Hot	0.221 (0.102)**	0.178 (0.102)*	0.176 (0.143)	0.156 (0.143)
High	0.177 (0.102)*	0.164 (0.102)	0.130 (0.144)	0.141 (0.145)
Hot×*High*			0.093 (0.203)	0.045 (0.204)
Constant	1.784 (0.088)***	1.550 (0.457)***	1.807 (0.101)***	1.558 (0.460)***
Controls:				
Age		0.000 (0.007)		0.000 (0.007)
Female		0.017 (0.111)		0.016 (0.111)
High school		0.667 (0.435)		0.669 (0.435)
BSc		0.206 (0.413)		0.210 (0.414)
MSc		0.131 (0.424)		0.137 (0.425)
PhD		0.126 (0.453)		0.135 (0.456)
Observations	452	452	452	452
(a) Experiment 2				
Hot	0.139 (0.089)	0.142 (0.091)	0.257 (0.124)**	0.282 (0.127)**
High	0.155 (0.089)*	0.154 (0.090)*	0.278 (0.127)**	0.300 (0.129)**
Hot×*High*			-0.242 (0.178)	-0.285 (0.181)
Constant	1.706 (0.077)***	1.191 (0.439)***	1.647 (0.088)***	1.106 (0.442)**
Controls:				
Age		0.002 (0.005)		0.002 (0.005)
Female		-0.046 (0.083)		-0.036 (0.083)
High school		0.579 (0.422)		0.615 (0.422)
BSc		0.494 (0.408)		0.510 (0.407)
MSc		0.515 (0.415)		0.538 (0.414)
PhD		0.206 (0.441)		0.196 (0.440)
Observations	400	400	400	400

Notes: Estimates come from linear regressions (standard errors in parentheses). In columns (2) and (4), age, gender and education categories are included as controls. *$P < 0.1$, **$P < 0.05$, ***$P < 0.01$.

Table C.8: Difference-in-differences effect of hot decision mode (full)

Dep. var.: choose LEFT	**Round 1**		**All rounds**	
	(1)	(2)	(3)	(4)
Hot	0.022 (0.048)	0.028 (0.048)	0.013 (0.036)	0.017 (0.036)
Experiment 1	0.314 (0.047)***	0.315 (0.048)***	0.173 (0.033)***	0.177 (0.034)***
Hot×*Experiment 1*	-0.181 (0.066)***	-0.184 (0.066)***	-0.093 (0.051)*	-0.095 (0.051)*
Constant	0.321 (0.034)***	0.166 (0.156)	0.321 (0.023)***	0.285 (0.139)**
Controls:				
Age		0.004 (0.002)*		0.003 (0.001)**
Female		0.112 (0.035)***		0.073 (0.028)***
High school		-0.104 (0.146)		-0.146 (0.130)
BSc		0.006 (0.140)		-0.089 (0.126)
MSc		-0.007 (0.143)		-0.082 (0.128)
PhD		-0.065 (0.153)		-0.126 (0.136)
Observations	852	852	12565	12565

Notes: Estimates come from linear regressions (standard errors in parentheses). In columns (3) and (4), standard errors are clustered at the individual level. *$P < 0.1$, **$P < 0.05$, ***$P < 0.01$.

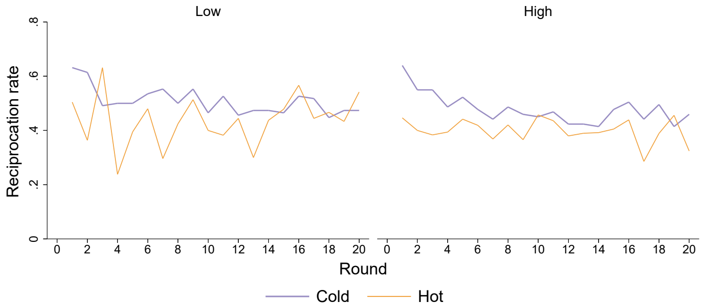
Table C.9: Difference-in-differences effect of hot decision mode (probit)

Dep. var.: choose LEFT	Round 1		All rounds	
	(1)	(2)	(3)	(4)
Hot	0.060 (0.130)	0.080 (0.132)	0.036 (0.100)	0.048 (0.101)
Experiment 1	0.810 (0.126)***	0.831 (0.131)***	0.451 (0.088)***	0.464 (0.092)***
Hot×*Experiment 1*	-0.467 (0.176)***	-0.484 (0.179)***	-0.239 (0.136)*	-0.244 (0.136)*
Constant	-0.464 (0.093)***	-0.894 (0.414)**	-0.465 (0.065)***	-0.578 (0.355)
Controls:				
Age		0.010 (0.005)**		0.008 (0.004)**
Female		0.306 (0.094)***		0.197 (0.076)***
High school		-0.297 (0.390)		-0.387 (0.330)
BSc		0.007 (0.372)		-0.228 (0.319)
MSc		-0.027 (0.381)		-0.211 (0.324)
PhD		-0.187 (0.409)		-0.331 (0.346)
Observations	852	852	12565	12565

Notes: Estimates come from probit regressions (standard errors in parentheses). In columns (3) and (4), standard errors are clustered at the individual level. *$P < 0.1$, **$P < 0.05$, ***$P < 0.01$.

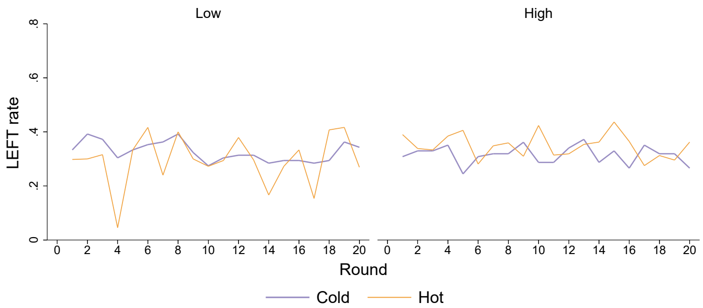
D Supplementary figures

Figure D.1: Dynamics of reciprocation rates across 20 rounds-Experiment 1



Notes: The figure shows the dynamics of reciprocation rates in Experiment 1 across 20 rounds. The unit of observation is a second mover's decision in a round.

Figure D.2: Dynamics of the percentage of LEFT choices across 20 rounds-Experiment 2



Notes: The figure shows the dynamics of the percentage of LEFT choices in Experiment 2 across 20 rounds. The unit of observation is a second mover's decision in a round.