Correspondence Bias

Yi Han, Yiming Liu, and George Loewenstein*

June 4, 2020

Abstract

When drawing inferences about a person's enduring characteristics from her actions, correspondence bias is the tendency to overestimate the influence of the person's enduring characteristics and underestimate the influence of transient situational factors. Focusing on incentives as one important situational factor, we build a simple model to formalize correspondence bias, and test predictions of the model in an online experiment. All players first play the dictator game, as the dictator, with an unknown receiver. Next, depending on their experimental condition, players are assigned to play a 'benign' game that encourages cooperation with another player, a 'malign' game that encourages selfish behavior, or both games with different players. Everyone then chooses to receive the dictator givings from one of two players who they may have played the benign or malign game with. Consistent with correspondence bias, subjects are on average willing to pay to receive the dictator givings from a player with whom they played the benign game. We show, further, that experiencing both games oneself, as opposed to playing one and observing the other, reduces the bias, and receiving information about how each of the players behaved in both games, eliminates it.

---

*Han: Department of Economics, University of Pittsburgh, yi.han@pitt.edu; Liu: Humboldt University of Berlin, WZB Berlin Social Science Center, yiming.liu@wzb.eu; Loewenstein: Social and Decision Sciences, Carnegie Mellon University, gl20@andrew.cmu.edu. We thank Kareem Haggag, David Huffman, Lise Vesterlund, Alistair Wilson for helpful comments.

1 Introduction

When drawing inferences about a person's enduring characteristics from their behaviors, the *correspondence bias* (Jones and Harris, 1967; Ross, 1977; Gilbert and Malone, 1995) is the tendency to overestimate the influence of the person's enduring characteristics on decisions they make, and to underestimate the impact of situational factors, such as social pressures. The situational factor of greatest interest to economists is the incentives that an individual faces.

One setting in which incentives matter is in decisions to cooperate or defect in interpersonal interactions. In economic games for which defection is a dominant strategy, prior research has found that people cooperate more when the payoff from mutual cooperation is higher (Charness, Rigotti and Rustichini, 2016), when "punishment" from cooperating unilaterally is smaller, and when the payoff from defecting against a cooperator is lower (Mengel, 2018). Correspondence bias, in such situations, would predict that observers will underestimate the impact of the incentives that players face, and so, in games that incentivize defection, attribute other's uncooperative behaviors to their negative traits such as selfishness.

Perhaps reflecting such an effect, a common belief is that the rich are more selfish than the poor. This belief most likely stems from the observation that they avoid taxes more often than others (Christian, 1994; Alstadsæter, Johannesen and Zucman, 2019; Saez and Zucman, 2019), and break traffic laws more often (Piff, Stancato, Côté, Mendoza-Denton and Keltner, 2012). However, Andreoni, Nikiforakis and Stoop (2017) argue that the differences in behaviors can be fully explained by differences in incentives. The rich have greater incentive and ability to protect their income from taxation, and paying traffic tickets is less painful for them due to the diminishing marginal utility of money. The rich may actually be similarly generous as the poor (Andreoni, Nikiforakis and Stoop, 2017), but correspondence bias will lead people to attribute their behaviors to negative character traits.

Correspondence bias, which was previously overlooked by economists, has important implications for everyday life. Consider two regions (countries, neighborhoods, etc.) or ethnic groups where social norms (or institutions) are starkly different. Person A is from region/group 1 where unethical behaviors are harshly punished and everyone finds it optimal to be trustworthy. Person B is from region/group 2 where legal enforcement is weak and sabotaging others is common. Now a person C needs to choose one of the two to work with or hire. She observes that A behaved more ethically in the past than person B. If she is correspondence-biased, then she may jump to the conclusion that A

is inherently more trustworthy than B, and choose A over B and be willing to incur costs to work with or hire A.

We build a simple model to formalize the idea of correspondence bias. In our model, an individual chooses between two players after observing their actions, and the goal is to choose the one who is more likely to be the Good type. One player plays the *benign game* in which both the Good type and the Bad type choose to cooperate, while the other player plays the *malign game* in which the Good type cooperates and the Bad type defects. Borrowing the idea of *cursed equilibrium* (Eyster and Rabin, 2005), we model correspondence bias as the tendency to underestimate the correlation between actions and the game structure when interpreting information obtained about others' play. Consequently, the biased individual tends to over-interpret cooperation in the benign game as a signal that a player is the Good type and under-interpret cooperation in the malign game as a signal of the Good type. Even when the game an individual plays is completely determined by chance, we predict that a correspondence-biased individual is willing to incur a cost to choose the benign-game player.

There are several challenges to empirically identifying correspondence bias. Imagine an experiment in which we randomly assign half of the subjects to play the benign game that incentivizes everyone to cooperate, and the other half to play the malign game that incentivizes some people to defect. We then let them choose between a benign-game player and a malign-game player to play a follow-up game together. Our model predicts that, on average, people are more willing to pick the former. The first challenge is to distinguish between correspondence bias and Bayesian updating. Choosing the benign-game player can be consistent with Bayesian updating, as someone who chooses to cooperate in the benign game can rationally be expected to be more prosocial than someone who chooses to defect in the malign game. Second, reciprocity can also motivate choosing the benign game player; subjects may want to reciprocate the benign-game player's cooperative behavior in the follow-up game. Third, if one believes that people become more (less) prosocial under good (bad) institutions, as shown in (Peysakhovich and Rand, 2015; Cason, Lau and Mui, 2019), then it can make sense to choose the individual who played the benign game.

We seek to rule out these three potential confounds using a three-stage experimental design. In the first stage, all subjects make a decision as the dictator in the dictator game. In the second stage, they are randomly matched into groups of four to play the benign game and the malign game. Both games are 2x2 complete-information games with a strictly dominant strategy for both players. The malign

game is the classic prisoner's dilemma game in which the dominant strategy is to defect, while the benign game is the Harmony Game (Dal Bó, Dal Bó and Eyster, 2018) in which the dominant strategy is to cooperate. At the end of the second stage, subjects are able to see the actions of one or more players, and to obtain information about the payoff structure of the games they played. Based on this information, in the third stage, they choose which of two players to receive the dictator givings from, and we use a multiple price list to elicit their willingness to pay (WTP) for their preferred player.

We address the Bayesian updating confound by randomly assigning players to the two games. This randomization ensures that the benign-game players and malign-game players are equally likely to be the Good type *ex ante*. The Martingale property of Bayesian beliefs then implies that the expected posterior beliefs are the same; a Bayesian model predicts that the individual will be in expectation indifferent between receiving the dictator offerings of the two players. However, our model predicts that a correspondence-biased individual will be (in expectation) willing to pay a positive amount to be matched with the benign-game player.

Our design avoids the possibility that reciprocity could drive the results by using a dictator game in which there are no actions that the receiver can take; thus, there is no way to reciprocate the benign-game player's cooperation in the follow-up game.

Finally, we avoid the potential for participation in the benign or malign game to shape the player's prosociality by sequencing the dictator decision so it occurs before stage 2. Even if individuals become more prosocial after playing the benign game, the dictator decision will have already been made in Stage 1, and cannot be altered by the game.

To better understand correspondence bias, and to explore potential methods to reduce it, we utilize a 4-Treatment design. The treatments differ in how many games subjects play, and how much information they receive.

In Treatment 1, subjects only play one game and are completely unaware of the other game. In Stage 3 they choose whether to obtain the dictator givings of the person they played either the benign or malign game with or those from a randomly chosen other player.

In Treatment 2, subjects still only play one game as in Treatment 1, but those who played the benign (malign) game also learn about the action of a malign-game (benign-game) player at the end of Stage 2. In this treatment they are informed about the payoffs of the game played by the other player, as well as the other player's action, but they do not experience the game themselves. In Stage

3 they choose between the player they played with in Stage 2 and the player who they only received information about. In Treatment 3, subjects actually play each of the games with two different partners in Stage 2. In Stage 3, they then choose whether to obtain the dictator givings of their Stage 2 benign- or malign-game partner.

The setup of Treatment 4 is the same as in Treatment 3, with the exception that subjects are also informed of both of their partners' actions in the other game (benign for the malign-game partner and malign for the benign-game partner) that each of their partners plays with someone else.

Our results show, first, that correspondence bias exists and influences Stage 3 decisions. We measure the impact of correspondence bias through the *benign premium* – the extra amount a subject, in stage 3, is willing to pay for the dictator game givings of a benign-game player compared to the dictator game givings of a malign-game player, which the players had decided upon in Stage 1. While the rational Bayesian model predicts the benign premium to be 0, we find that the benign premium is 11.67 cents on average in Treatment 2, the baseline treatment, which is significantly different from 0 at the 1% level. To receive the dictator game givings of the player who is randomly assigned to the benign game, subjects are on average willing to give up 6% of the $2.00 divided by the dictator (which is the largest possible difference between the two potential dictators), or 12% of the $1.00 (half of the 'pie' is the typical modal amount given in the dictator game; only 11 out of 817 subjects, or 1 percent of subjects, in our experiment gave more than $1.00).

Second, and consistent with the predictions of our model, we find that correspondence bias is driven by both overestimation of the prosociality of the benign-game player and by underestimation of the prosociality of the malign-game player. In Treatment 1, when choosing between a stranger and a benign/malign-game player, subjects are willing to pay more for dictator givings of a benign-game player compared to a stranger, and willing to pay more for dictator givings of a stranger compared to a malign-game player.

We also test two potential methods to reduce or even eliminate correspondence bias. First, we test whether experiencing instead of observing the games can help to reduce the bias. Given that they are likely to behave differently in the two games, themselves, experiencing both games in Treatment 3, as opposed to only learning about it in Treatment 2, should help people to understand that actions are likely to be game-contingent, and they should take the games into account when inferring from the actions. Consistent with such an effect, we find that the benign premium in Treatment 3 is smaller

than that in Treatment 2, although it is still significantly greater than 0 (at the 1% level), suggesting that experiencing both games is not enough to eliminate the bias.

Second, with Treatment 4, we investigate the effect on reducing the bias of providing counterfactual information. In Treatment 4, as subjects know both players' actions in both games, they should be even more aware of the game-contingent nature of play, which should further reduce the bias. Supporting this prediction, we find that providing counterfactual information reduces the benign premium to 2 cents, which is not significantly different from 0 and is significantly smaller than that in treatments 2 and 3.

The research that this study is most closely related to is Haggag, Pope, Bryant-Lees and Bos's (2019) investigation of "Attribution Bias in Consumer Choice." In their study, people underweight the impact of a transitory state, such as hunger, on the utility of consuming a good, and misattribute it to the enduring characteristic of the good. The current research builds on their contribution by showing that attribution bias exists not only when it comes to evaluating consumption experiences, but also in evaluating people. Agents in Haggag, Pope, Bryant-Lees and Bos's (2019) model do not fully appreciate the fact that their preferences are state-dependent; similarly, agents in our work fail to fully recognize that actions of other people are game-dependent. While Haggag, Pope, Bryant-Lees and Bos (2019) has important implications for individual decision making, our analysis shows that attribution bias also plays a vital role in economic interactions. We also build on their work by exploring potential debiasing methods. Following up on earlier research showing that people respond more strongly to games that they actually play as opposed to those that they observe (Simonsohn, Karlsson, Loewenstein and Ariely, 2008), we show that *experiencing* both games, as opposed to playing one and observing the other, reduces attribution bias, and that providing "counterfactual" information eliminates the bias. Our results provide an explanation for the finding of Haggag, Pope, Bryant-Lees and Bos (2019) that the extent of past experiences can attenuate the attribution bias in consumption choice.

Correspondence bias, previously designated the "fundamental attribution bias," has been intensively studied by psychologists since the 1960s (Jones and Harris, 1967; Ross, 1977; Gilbert and Malone, 1995). In a typical study, subjects listen to a speech arguing in favor of or against an opinion, are informed that the speakers' positions are randomly assigned by the experimenter, and are asked to rate the attitudes of the speaker towards that opinion. The repeatedly-replicated finding is that, de-

spite being informed about the random assignment to position, subjects still rate speakers who speak in favor of the opinion as more supportive of it.

The most common explanation that psychologists offer for the correspondence bias is that, when attempting to make sense of a person's behavior, the characteristics of the person are typically more "salient" than their situation, resulting in an overestimation of the influence of the former, and an underestimation of the latter. We formulate the bias in a different way. We are less focused on the salience of other people's characteristics, but more on assessments of their stability. In our formulation in the following section of the paper, it is people's failure to fully account for the incentive-contingent nature of other's actions that leads them to under-attribute actions to incentives.

The current study augments the existing psychology research on correspondence bias in three ways. First, the standard experimental paradigm for studying correspondence bias in psychology suffers from the potential confound that subjects may believe that the randomly assigned positions can potentially shape the speakers' attitudes. As we discussed, our design rules this out. Second, in an environment that closely mimics real-life interpersonal interactions, our design clearly shows that correspondence bias is welfare-reducing. Third, we provide a suggestion for how to reduce or eliminate correspondence bias by providing counterfactual information.

We also contribute to the literature on people's belief updating relative to Bayesian updating. Previous evidence suggests that people generally infer less from evidence than Bayes' Theorem predicts (Phillips and Edwards, 1966; Edwards, 1968; Möbius, Niederle, Niehaus and Rosenblat, 2014; Ambuehl and Li, 2018). However, as pointed out by Kahneman, this finding is in contrast to the everyday experience that people often jump to conclusions based on little information. We provide another reason, in addition to the Law of Small Numbers (Kahneman and Tversky, 1972) and base-rate neglect (Kahneman and Tversky, 1973), for why people may draw overly extreme conclusions from small samples.[1] In our case, people jump from observations of others' actions in narrow contexts to conclusions about those people's underlying qualities, without paying sufficient attention to the transient incentives they are facing.

The paper proceeds as follows: Section 2 describes a simple model of correspondence bias. Sec-

[1]For more discussion on over-inference, see Benjamin (2019).

tion 3 introduces our experimental design and the predictions it tests. Section 4 presents results, and Section 5 concludes and discusses policy implications.

2 Model

In this section, we build a simple model of correspondence bias that is in a similar spirit to the cursed equilibrium of Eyster and Rabin (2005). In our model, the individual does not fully take into account the fact that other people's actions depend on the incentives they face (or the game they play); they are aware of the distributions of others' actions, but underestimate the correlation between actions and the game structure when they try to interpret those actions.

Consider two games $\tau \in \{b, m\}$, the *benign game* b and the *malign game* m. In each complete information game, there are two actions to take: $\{C, D\}$. There are two types of agent, the Good type G and the Bad type B. Let the probability of being the Good type be $P(t_i = G) = p_0$. In the benign game τ_b, both the Good type and the Bad type choose C; in the malign game τ_m, the Good type chooses C and the Bad type chooses D. Half of the players are assigned to play the benign game, and the other half are assigned to play the malign game.

After observing player i's action in the benign game a_i^b and player j's action in the malign game a_j^m, player k needs to choose between i and j to play a follow-up game. k's payoff in the follow-up game is defined by the type of the partner of her choosing. If we standardize the payoff of choosing type B to 0 and choosing type G to 1, then player k's expected payoff for choosing i to play the follow-up game with is given by

$$U_{ki} = P(t_i = G) - P(t_j = G). \tag{1}$$

Without loss, the payoff for choosing j is standardized to 0.[2]

Define $p(\cdot)$ as the true probability and $\pi(\cdot)$ as a person's potentially biased belief. For a Bayesian agent, as both types choose C in the benign game, the posterior $\pi(t_i = G \mid a_i^b) = p(t_i = G \mid a_i^b)$ is

[2]By formulating the expected payoff in this way, we assume that the decision maker is risk neutral. However, our main results remain unchanged by assuming risk aversion.

equal to the prior, p_0. In the malign game, player j's type is perfectly revealed. If she chooses C, then $\pi(t_j = G \mid a_j^m = C) = 1$; if she chooses D, then $\pi(t_j = G \mid a_j^m = D) = 0$. Therefore, $\pi_D^m < \pi_C^b < \pi_C^m$, where π_a^τ is the Bayesian agent's posterior belief about someone who chooses action a in game τ being the Good type.

The correspondence-biased agent knows the distribution of actions across the whole population, but they cannot fully account for their opponent's action based on the game they played. Borrowing from Eyster and Rabin's (2005) *cursed equilibrium*, we define an agent as χ-biased if

$$\tilde{\pi}_a^\tau = \chi(P(b \mid a)\pi_a^b + P(m \mid a)\pi_a^m) + (1-\chi)(\pi_a^\tau), \tag{2}$$

where $\tilde{\pi}_a^\tau$ is the correspondence-biased agent's posterior belief about the player who chooses action a in game τ, and $P(\tau \mid a)$ is the probability of the game being τ given action a. Intuitively, a χ-biased agent only recognizes the action of her opponent but ignores the incentive/game structure she faces with probability χ. In this case, she replaces the actual probability $p_a^\tau = \pi_a^\tau$ of her opponent being the Good type given action a in game τ with the average posterior of her opponent being the Good type given action a across the two games, $P(b \mid a)\pi_a^b + P(m \mid a)\pi_a^m$. If $\chi = 0$, then the χ-biased agent's posterior is the same with the Bayesian agent. As long as $\chi > 0$, we can conclude that

$$\pi_D^m = \tilde{\pi}_D^m < \pi_C^b = p_0 < \tilde{\pi}_C^b \leq \tilde{\pi}_C^m < \pi_C^m, \tag{3}$$

where $\tilde{\pi}_C^b = \tilde{\pi}_C^m$ when $\chi = 1$.

Therefore, the probability of the benign-game player i being the Good type given she chooses C is overestimated: $\pi_C^b < \tilde{\pi}_C^b$. The probability of the malign-game player j being the Good type given he chooses C is underestimated: $\tilde{\pi}_C^m < \pi_C^m$. It is important to note that the χ-biased agent can still make the inference that a_C^m is a stronger signal for the Good type than a_C^b, i.e. $\tilde{\pi}_C^b \leq \tilde{\pi}_C^m$, but they underestimate the difference between the two: $\tilde{\pi}_C^m - \tilde{\pi}_C^b < \pi_C^m - \pi_C^b$.

Correspondence bias implies an over-evaluation of the payoff for choosing the benign-game player. Depending on the malign-game player j's choice, there are two cases: $(a_i^b = C, a_j^m = D)$ and $(a_i^b = C, a_j^m = C)$. In the first case, the Bayesian agent should conclude that the payoff for choosing the benign-game player is equal to

$$U_{ki} = \pi_C^b - \pi_D^m. \tag{4}$$

However, the correspondence-biased agent would believe that the expected payoff of choosing i, $\tilde{U}_{ki}$, is

$$\tilde{U}_{ki} = \tilde{\pi}_C^b - \tilde{\pi}_D^m. \tag{5}$$

As $\tilde{\pi}_C^b > \pi_C^b$ and $\tilde{\pi}_D^m = \pi_D^m$, $\tilde{U}_{ki} > U_{ki}$. Similarly, when $a_i^b = C$ and $a_j^m = C$, as $\tilde{\pi}_C^b > \pi_C^b$ and $\tilde{\pi}_C^m < \pi_C^m$, $\tilde{U}_{ki}$ is also larger than U_{ki}. As $\tilde{U}_{ki}$ is larger than U_{ki} in both cases, we conclude that the perceived payoff for choosing the benign-game player is larger for the correspondence-biased agent.

For a Bayesian agent, the expected benefit of choosing i is 0. The Martingale property of Bayesian updating implies that

$$E[\pi \mid \tau = b] = E[\pi \mid \tau = m] = E[\pi] = p_0. \tag{6}$$

Intuitively, as whether one is assigned to the benign game or the malign game is completely determined by chance, i and j are equally likely to be the Good type *ex ante*. As the expected posterior is equal to the prior, they are equally likely to be the Good type in expectation *ex post*.

However, for a correspondence-biased agent, the expected benefit of choosing i is larger than 0. As $\tilde{\pi}_C^b > \pi_C^b = p_0$ for the biased agent, $E[\tilde{\pi} \mid \tau = b] > p_0$. As $\pi_D^m = \tilde{\pi}_D^m$ and $\tilde{\pi}_C^m < \pi_C^m$, $E[\tilde{\pi} \mid \tau = m] < E[\pi \mid \tau = m] = p_0$. Therefore, for a correspondence-biased agent

$$E[\tilde{\pi} \mid \tau = m] < p_0 < E[\tilde{\pi} \mid \tau = b] \tag{7}$$

Definition. We define a correspondence-biased agent's *benign premium* as her expected payoff of choosing the benign-game player over the malign game player, namely $E[\tilde{\pi} \mid \tau = b] - E[\tilde{\pi} \mid \tau = m]$.

We summarize our main results in the following proposition.

Proposition 1. *i) A Bayesian is in expectation indifferent between a benign-game player and a malign-game player to play the follow-up game with, and is expected to pay* $E[\pi \mid \tau = b] - E[\pi \mid \tau = m] = 0$ *for the benign-game player.*

ii) For any $\chi \in (0,1]$*, a* χ*-biased individual's benign premium* $E[\tilde{\pi} \mid \tau = b] - E[\tilde{\pi} \mid \tau = m]$ *is larger than 0.*

iii) For any $\chi \in (0,1]$*, a* χ*-biased individual is willing to pay* $E[\tilde{\pi} \mid \tau = b] - p_0 > 0$ *in expectation for a benign-game player when choosing between the benign-game player and a stranger, and is willing to pay* $p_0 - E[\tilde{\pi} \mid \tau = m] > 0$ *for a stranger when choosing between the stranger and a malign-game player. While a Bayesian is in expectation indifferent between a benign-game player, a malign game*

player, and a stranger.

3 Design

The experiment has three stages. In the first stage, all subjects make a decision as the dictator in the dictator game. In the second stage, they are randomly matched into groups of 4 to play the benign game and the malign game. The benign game was chosen to encourage players to cooperate with their partners, while the malign game was chosen to motivate selfish behavior. Lastly, they are asked, as the receiver, to choose between receiving the dictator givings of two players from the first stage. Our model predicts that there exists a *benign premium*: subjects are, on average, willing to pay to be matched with the benign-game player.

First Stage

The experiment was conducted online, and subjects were recruited through Amazon Mechanical Turk (Mturk). Upon arriving at the study website, each subject was instructed to play a dictator game as the dictator. They divided 200 cents between themselves and a random receiver. As in a standard dictator game, the receiver had no influence over the outcome of the game, and both the receiver and the dictator receive 50 cents of endowment prior to the split decision. Subjects were also informed that, although everyone needed to make the decision, only half of those decisions would be implemented later. At this stage, they had no idea of the existence or nature of the future stages of the experiment or of the identity of the potential random receiver. This dictator decision serves as our measure of each subject's prosociality.

Second Stage

In the second stage, subjects were randomly matched into four-player groups. Everyone was randomly assigned a role. There were four roles in each four-player group. We name them A, B, C and D. Then, the participants played the benign and/or the malign games with individuals in their own group. Depending on the treatment, a subject interacted with one or two individuals at this stage. The two games are defined as follows.

The malign game is a two-player one-shot prisoner's dilemma, see left panel in Table 1. Partic-

ipants in this game must choose between cooperate (C) and defect (D).[3] It is a dominant strategy to choose D for both players. The Nash equilibrium of this game is (D,D), which leads to the payoffs (30,30). Even though D is the dominant strategy, in previous studies not everyone defected, and those who chose to cooperate were more prosocial, as measured by givings in a Dictator Game, than those who choose to defect (Cooper, DeJong, Forsythe and Ross, 1996; Barreda-Tarrazona, Jaramillo-Gutiérrez, Pavan and Sabater-Grande, 2017). In the language of the model, subjects who cooperate in the malign game are *Good* types and subjects who defect in the malign game are *Bad* types.

The benign game is the Harmony game as in Dal Bó, Dal Bó and Eyster (2018) (right panel in Table 1). Participants also choose between cooperate (C) and defect (D). However, (C,C) is the dominant strategy equilibrium and Pareto dominates all other strategy profiles in this game. It is easy for the subjects to figure out that they should choose C, and most do so.

To ease understanding, we illustrate the rest of the experimental design in terms of Treatment 2, which we believe to be closest to situations encountered in daily life. We discuss the differences between the treatments at the end of this section. In Treatment 2, subjects play one game, and observe outcomes of the other game. Specifically, players A and B play the benign game, and C and D play the malign game. Even though they only play one game, we also give them the instructions, including payoffs, of the other game so that they can still understand the incentives of the game they are not playing.

After Stage 2 actions were taken, subjects entered the information provision page (Appendix Figure A2). On this page, they learned about their payoffs and the action of their partner in the game they had just played. We also displayed the payoff table of the two games again to minimize confusion and to aid in recall. In addition, in Treatment 2, subjects were informed of the action of one of the two players in the game they did not play. To be more precise, A (B) in the benign game also learns whether D (C) in the malign game chooses to cooperate or not. Similarly, C (D) is also informed of the action of B (A) in the benign game.

Subjects were forced to stay on the information provision page for at least 120 seconds to make

[3]The actions C and D are respectively labeled "Action 1" and "Action 2" in the experiment to ensure a neutral presentation.

sure that they had the time to understand the game structure and make inferences about the types of their partners based on their actions.

Third Stage

In the third stage, every subject had the opportunity to choose whether to receive the dictator transfer from the first stage either from a malign-game player and a benign-game player. In this second condition, one of them was the player they had played with, and the other was the player whose play they only learned about. The two candidates are those two whose actions in Stage 2 were shown to the subject in the information provision phase. For example, A played the benign game with B and observed D's action in the malign game. Then in Stage 3, A chose between B and D's dictator transfers in the first-stage dictator game. Therefore, the only source of information that the subject had to go on when choosing between the two candidates was the action that each had taken in the game they played in Stage 2 (as well as the payoffs for these games).

After reading the instructions for Stage 3 and before making any decisions, subjects were asked three comprehension questions. Only those who answer all three questions correctly could proceed to make their choices in this stage; those who answered at least one question incorrectly were required to re-do all three questions until they get them all right.[4]

After a subject made the choice between the two candidates, we used a multiple price list to elicit her WTP to be matched with the partner of her choosing. The list shown to A if she chooses B over D in the first choice is displayed in Table 2 as an example. In total, subjects made 10 choices, excluding the first one. In each choice, there were two options. D+(x cents) means if in this choice A chooses D, and this is the choice selected at random to count, then she will then get an extra reward of x cents. But if she chooses B, there is no extra reward. The point where A switches from option 1 to option 2 defines A's WTP to get B. One of these 11 choices (including the one between B and D with no extra

[4]Please see Appendix A1 for the comprehension questions. As one cannot proceed to the decision stage of Stage 3 without answering all 3 questions correctly, some subjects dropped out in this stage. Out of 1,008 subjects who signed up for the experiment, 151 of them finished Stage 2 but dropped out in Stage 3. As Stage 3 is not interactive, the dropouts of those subjects have no impact on the use of data from others in the same group. Moreover, there is no significant difference in attrition rate across treatments.

rewards) was randomly selected as the *choice-that-counts*, and the instructions made this clear. We then implement one of the four *choice-that-counts* with a designated matching protocol.

Our matching protocol in stage 3 was designed to deal with a potential confound of correspondence bias. One way to reciprocate the benign-game player is to choose her as the dictator in the stage 3 as dictators are expected to earn more than receivers. The protocol can be divided into 4 steps. In the first step, we randomly chose 1 player from the 4. Let us name her the *chosen player* and assume it was A. Second, both the *chosen player*'s (A) benign- and malign-game partners (B and D) got the dictator role. Therefore who played as the dictator and who played as the receiver in the first stage dictator game were determined at the second step. The next steps only affected the matching between the two dictators and the two receivers. In the third step, we implemented the *chosen player*'s (A) *choice-that-counts*. Suppose A chose B's dictator offerings in that choice, then A and B were matched with B as the dictator. Fourth, the partner whom was not picked by the *chosen player* A, D in our example, was matched with the remaining player C. D's dictator transfer decision was carried out, and C received D's offerings as the receiver. The fact that D was the dictator but not C was determined in step 2. Therefore, who got dictator roles was completely determined by whom was randomly selected to be the *chosen player* in the matching protocol, choosing a player as the dictator in stage 3 did not raise her chances of being the dictator in the dictator game.

The dictator's decisions made in Stage 1 were then carried out. For example, suppose B chose to give x cents to the random receiver in the first stage, and if A and B are matched with B being the dictator, then A gets x cents and B gets $(200-x)$ cents. Putting the dictator decision ahead of the second-stage games solves the institution shaping people's prosociality confound. The idea is that, at the third stage, the dictators had already made their decisions about how much to transfer in the first stage. Therefore, what happened at the second stage could not have an impact on them. Even if the benign-game player became a nicer person after playing the game, her choice in the first stage remained the same.

Treatments

There are four treatments in the experiment and they only differ in the second stage. What separates them from each other is how many games each subject plays and how much information they are given.

In Treatment 1 (as indicated in Figure 1), each player only plays one game, either the benign or the malign game, and is not aware of the existence of the other game. In the third stage, subjects are asked to choose between receiving the dictator givings of the person they play this one game with, or those from a random participant in the study.

In Treatment 2, as already described, each player again only plays one game. However, in this condition, in addition to the outcomes and the action of the opponent in the game he/she plays, the subject also observes the action of one other player who plays the other game, as well as receiving full information about the game itself. Then, the focal subject chooses whether to receive the dictator givings of the player who he/she actually played with, or the player who he/she only received information about.

Treatment 3 is the same as Treatment 2, except that subjects actually play both games (the benign game and the malign game) with different other subjects in their group. In the information stage, they learn the actions of both their opponents and their outcomes in the two games. The difference between Treatment 3 and Treatment 2, therefore, is that in Treatment 3 all information is gathered through "experience" instead of partly through "observation" as in Treatment 2.

Treatment 4 is the same as Treatment 3 except that subjects are also informed of the behaviors of their opponents in the game they do not play together. For example, if A plays the benign game with B, she also learns how B behaved playing the malign game with D.

Predictions

The four-treatment design helps us investigate the causes of correspondence bias and the potential ways to reduce or even eliminate it.

Prediction 1. There exists a benign premium in Treatment 2, that is, the average WTP towards the benign-game player is larger than that towards the malign-game player.

Treatment 2 is our baseline treatment, and we can test the existence of correspondence bias by looking at the benign premium in this treatment.[5]

[5]We choose Treatment 2 as our baseline for two reasons. First, in daily life, people often draw inferences about others' characteristics based on their personal experience with those people, but with only second-hand knowledge of

Prediction 2. In Treatment 1, when choosing between a benign-game player and a random stranger, subjects are on average willing to pay more for the benign-game player; when choosing between a malign-game player and a stranger, they are on average willingness to pay more for the stranger.

Treatment 1 aims to decompose the benign premium. As no information is provided on the stranger, the chance of her being the Good type is equal to the prior, p_0. Thus, Treatment 1 helps us separate the benign premium $E[\tilde{\pi} \mid \tau = b] - E[\tilde{\pi} \mid \tau = m]$ into two parts: underestimation of the chance of the malign-game player being the Good type $p_0 - E[\tilde{\pi} \mid \tau = m]$ and overestimation of the chance of the benign-game player being the Good type $E[\tilde{\pi} \mid \tau = b] - p_0$. While Bayesian inference predicts that $E[\pi \mid \tau = m] = E[\pi \mid \tau = b] = p_0$, we predict that for correspondence-biased agents $E[\tilde{\pi} \mid \tau = m] < p_0 < E[\tilde{\pi} \mid \tau = b]$.

Prediction 3. The benign premium is smaller in Treatment 3 than in Treatment 2.

Treatment 3 is set to test whether inattention to strategic motives is a cause of correspondence bias. As participants play both games in this treatment, they have a better understanding of the incentives in the two games. In Treatment 2, the subject may only pay attention to behaviors without understanding the incentives behind them. Consequently, she tends to treat cooperation in the two games equally even though it is a much stronger signal of the Good type to cooperate in the malign game. When she plays both games herself in Treatment 3, she is more likely to know that choosing cooperation does not mean the same thing across the two games.

Prediction 4. The benign premium is smaller in Treatment 4 than in Treatment 3.

In Treatment 4, we test whether providing counterfactual information reduces correspondence bias. In treatments 2 and 3, the participant was not able to know how the benign-game player performed in the malign game, and *vise versa*. However, in Treatment 4, such information was available, and subjects could clearly see how other's actions changed according to the incentives. If correspondence bias is caused by failing to fully account for the impact of the incentives on actions, then enabling people to compare opponents' behaviors in the same game with the same incentives should

---

those people's behavior in other environments. Second, Treatment 2 is directly comparable to Treatment 3 and 4, as in all three of these treatments subjects chose, in stage 3, between a benign-game player and a malign-game player. In Treatment 1, in contrast, they chose between a benign or malign player and a stranger.

reduce the bias significantly.

Implementation

The experiment was conducted on Amazon Mechanical Turk between October 12, 2018 and December 7, 2018. As our experiment is rather complicated, we only recruited subjects who had at least a two-year associate degree. We also restricted participation to residents of the United States who had completed at least 100 tasks prior to our study and had an approval rating of at least 95%. We advertised the experiment as a 20-minutes academic decision-making study with an average payment of 2.5 dollars. On average, the experiment lasted 20.1 minutes and subjects earned 2.77 dollars.

Overall, we recruited 817 subjects in our online experiment, with 121 in Treatment 1, 246 in Treatment 2, 223 in Treatment 3, and 227 in Treatment 4.[6] We randomly assign fewer subjects to Treatment 1 based on a power calculation. We need more subjects in the other 3 treatments because we need to test whether the benign premium is significantly different between two treatments, whereas in Treatment 1, we only need to test whether the average WTP is significantly different from 0 or not.

Table 3 shows summary statistics both in aggregate and across treatment conditions. All of the non-outcome behaviors and demographics are balanced. On average subjects shared 67 cents in the dictator game. 95.2% of subjects chose to cooperate in the benign game and 38.9% defected in the malign game. A natural concern is that subjects may behave differently in Treatment 2 and in treatments 3 and 4 as the number of games they play is different. Reassuringly, the cooperation rate in the malign game in Treatment 2 is not significantly different from the average cooperation rate in treatments 3 and 4 (p-value=0.486). We collected subjects' demographic information in a voluntary follow-up survey. 735 out of 817 subjects (90%) completed the survey, and there is no significant difference in the take-up rates across treatments. Survey respondents have an average age of 38, 57% are female, and 80% have jobs (either employed or self-employed).

[6] We received a total of 857 responses, but dropped 40 subjects (4.67%) who exhibited multiple switching points in the multiple price-list questions at the third stage.

4 Results

The objective of this study is to examine whether when people making inferences about others based on their behaviors, they over-attribute behaviors to other's characteristics, but underestimate the impact of incentives on behaviors. To do so, we look at how an individual's randomly assigned game, which is orthogonal to her characteristics, affect other people's perception of her. We first confirm that the game a subject is assigned to play is indeed orthogonal to her prosociality, which is measured by her dictator givings in Stage 1. Figure 2 illustrates that subjects who play the benign game transfer an average of 66.90 cents, which is, as would be expected if randomization was successful, almost identical to the average dictator givings from malign-game players (66.56 cents).

Then as a manipulation check, we look at whether the two games induce different behaviors (Table 4). While almost everyone chooses to cooperate in the benign game (95.2%), the frequency of cooperation is much lower in the malign game (38.9%), so the game structure can indeed affect subjects' choices. The choices in the malign game are also informative for identifying types of subjects. Figure 2 shows that subjects who choose to cooperate in the malign game transfer 76.59 cents in the first stage, while subjects who choose to defect only transfer 60.17 cents, a statistically significant difference (p-value$<$0.01, rank-sum test).

Result 1. *Correspondence bias exists in the baseline treatment when subjects experience the action of one player and observe the action of another player. The existence of the bias leads to a clear welfare loss.*

Turning into the main results of the paper, we first look at the existence of correspondence bias in the baseline treatment, Treatment 2. A rational Bayesian model predicts that subjects will be, in expectation, indifferent between receiving the dictator offerings from either the benign-game player or the malign-game player. However, supporting the first prediction of our model, there is a positive benign premium: subjects are willing to pay, on average to receive the dictator game offerings from the benign-game player rather than those from the malign-game player. Using the multiple price list, we define the willingness-to-pay (WTP) for the benign-game player as the switch point between option 1 and option 2 in Table 2. We further code it as positive if a subject chooses the benign-game player in the first choice, and negative otherwise. Since the multiple price list can only elicit intervals

of WTP, we use the mid-point of the interval as the WTP for the benign-game player.[7] For example, if subject A chooses B's (the benign-game player) transfers over D's (the malign-game player) transfers plus 10 cents bonus, and switches to D's transfers plus 20 cents when choosing between it and B's transfers, then A's WTP for the benign-game player is coded as 15 cents.

As shown in Figure 3, the average WTP for the benign-game player's dictator givings is 11.67 cents higher than that for the malign-game player's givings in Treatment 2, which is significantly larger than 0 at the 1% level. The Bayesian model is rejected. One way to interpret this result is that subjects believe that the benign-game player on average transferred 11.67 cents more in Stage 1 than the malign-game player. To put those numbers into perspective, one can compare them with the maximum plausible benign premium of 100 cents. A completely selfish individual transfers 0 in Stage 1, while an altruistic individual who weights other's utility exactly as much as her own transfers 100 cents in Stage 1. Therefore, although larger values are possible (up to 200 cents), the largest plausible difference between the two potential opponents' transfer is 100 cents.

The benign premium can also be interpreted as a measure of the welfare loss caused by correspondence bias. To see this, consider the case when the expected dictator givings of the malign-game player are higher than that of the benign-game player from a Bayesian's perspective but the difference between the two is smaller than the benign premium. While a risk-neutral Bayesian would choose the malign-game player, a risk-neutral correspondence-biased agent would still choose the benign-game player, leading to an expected welfare loss. The larger the benign premium, the more likely a correspondence-biased agent would forfeit a gain from choosing the malign-game player's givings.

On the aggregate level we confirm that subjects are correspondence-biased, a natural next question is how many subjects are correspondence-biased. This question is hard to answer when the malign-game player chooses to defect. Both the Bayesian model and our model predict that in this situation subjects should choose the benign-game player, and the only difference is that our model predicts a larger WTP towards the benign-game player. However, the case when the malign-game player chooses to cooperate is clear-cut. While a Bayesian subject should choose the malign-game player,

[7] The results are robust if we use the lower or upper bound of the interval as the WTP for the benign-game player (Appendix Figure A1).

regardless of her prior, our model predicts that a fully correspondence-biased subject is indifferent between the two players and may choose the benign-game player. Our data show that 52% of subjects choose the benign-game player over the malign-game player when the latter choose to cooperate in Treatment 2 (Panel A of Table 6).

Result 2. *Evidence suggests that correspondence bias is caused by both an overestimation of the prosociality of the benign-game player and an underestimation of the prosociality of the malign-game player.*

In Treatment 1, subjects only play one game, and are asked to choose between receiving the dictator givings of the person they play this one game with, and those from a random participant. As predicted by the model, a Bayesian subject should be indifferent between her partner and a stranger in expectation regardless of which game she is assigned to play. However, the game an individual plays does have an impact on her WTP towards her partner.

Treatment 1 is more comparable to previous studies in psychology on correspondence bias. We randomly assigned subjects to interact with someone in a benign environment (corresponding to the "against an opinion" condition in the psychology literature) or a malign environment (corresponding to the "in favor of an opinion" condition), and we test whether this randomly assigned environment had an impact on a subject's evaluation of their partner or not (corresponding to asking subjects to rate the attitudes of the speaker towards that opinion). Our results show that the orthogonal environment has a strong effect on a subject's WTP towards her partner. When the game played together is the benign game, the average WTP for partners over strangers is 12.62 cents; when it is the malign game, the average WTP for the partner is -7.24 cents, meaning subjects are willing to pay to receive the dictator givings from random strangers, rather than from their partners. The two WTPs are significantly different from each other (p-value$<$0.01, Wilcoxon rank-sum test), which serves as another piece of evidence of correspondence bias.

Treatment 1 also serves as a test of the formulation of correspondence bias. If the bias is caused by people's failure to fully account for the degrees to which incentives affect actions, then we would predict a preference for the benign-game player to the stranger and a preference for the stranger rather than the malign-game player. The results are consistent with this prediction. As shown above, the average WTP for the benign-game player is positive and is significantly different from 0, with a p-value of 0.025. Meanwhile, the average WTP for the malign-game player is negative (p-value=0.155).

The negative WTP for the malign-game player is unlikely to be a mistake as subjects do respond to the malign-game player's actions. When the malign-game player chooses to cooperate, the average WTP towards her is 11.67 cents; when the malign-game player chooses to defect, the average WTP is -20.59 cents.

Result 3. *Direct experience with both games reduces correspondence bias, but, alone, is not sufficient to eliminate the bias.*

So far these results show that there exists a correspondence bias: subjects tend to believe that someone who is randomly assigned to play a benign game is more prosocial than someone who is randomly assigned to play a malign game. The next question is: can we alleviate this bias? By comparing Treatment 2 with Treatment 3, we can see the effect of letting subjects experience both regimes so as to better understand the strategic motives. The only difference between the two treatments is that subjects only play one game but observe the other one in Treatment 2, while in Treatment 3 they play both. The average benign premium decreases from 11.67 cents in Treatment 2 to 7.78 cents in Treatment 3, with a p-value of 0.263. However, experience alone is not sufficient to eliminate correspondence bias. The benign premium in Treatment 3 is still significantly larger than 0 (p-value=0.003, t-test).

The reduction in the benign premiums from Treatment 2 to Treatment 3 is mainly driven by the reduction in the WTP for the benign-game player of subjects whose malign-game player chooses to defect. As shown in Figure 4, when the malign-game player chooses D, the average WTP for the benign-game player decreases from 20.68 cents to 15.16 cents (p-value=0.159). Meanwhile, the average WTP for the benign-game player only decreases from -0.05 cents to -2.16 cents when the malign-game player chooses to cooperate. The results suggest that experience is better at reducing the overestimation of the niceness of the benign-game player. It has little effect on reducing the underestimation of the niceness of the malign-game player.

One potential concern is that the difference between treatments 2 and 3 can also be driven by inattention to the partner's choices instead of inattention to the strategic motives. Subjects might only pay attention to the game they played and might ignore the other game. To deal with this issue, we further look at how subjects who only played the benign-game respond to the malign-game players' choices. If they are not paying any attention to the malign-game player's choices, then the average WTP for the benign-game player should be the same regardless of the malign-game player's choices.

Table 6 shows that for subjects who only played the benign-game, when the malign-game player chooses to cooperate, their average WTP for the benign-game player is 13.89 cents; while when the malign-game player chooses to defect, the average WTP for the benign-game player increases to 27.88 cents. The two amounts are significantly different from each other (p-value=0.042).

Result 4. *Providing counterfactual information in addition to letting subjects experience both games can eliminate correspondence bias. The result is mainly driven by a reduction in overestimation of the niceness of the benign-game player.*

By comparing treatments 3 and 4, we can study the effect of informing the subjects of "counterfactuals." When we not only let subjects learn the behaviors of two partners by playing games with them, as in Treatment 3, but also inform them of the behaviors of their partners in the game they did not play together in Treatment 4, the benign premium further decreases to 2.14 cents, which is not significantly different from zero (p=0.407). The difference between Treatment 3 and Treatment 4 in the benign premium is significant at the 10% level (p-value=0.095), suggesting that providing counterfactual can alleviate correspondence bias. The difference between Treatment 2 and Treatment 4 is significant at the 1% level (p-value=0.007), which indicates that experience plus counterfactual information can jointly eliminate the bias.

The reduction in the benign premiums from Treatment 3 to Treatment 4 is mainly driven by the reduction in the benign premiums when the malign-game player chose to defect (Figure 4 and Table 6). In this situation, the average WTP for the benign-game player reduces from 15.16 cents in Treatment 3 to 6.37 cents in Treatment 4 (p-value=0.103). The average WTP in Treatment 4 (6.37 cents) is very close to the Bayesian level with the correct prior, 6.94 cents.[8] It suggests that the overestimation of the niceness of the benign-game player is almost gone in Treatment 4. At the same time, when the malign-game player chose to cooperate, the benign premium declines from -2.16 cents in Treatment 3 to -3.65 cents in Treatment 4. Again, it is also closer in Treatment 4 than in Treatment

[8] When subjects are Bayesian with correct priors, the WTP for the benign-game player should equal to the conditional expected differences in the two players' dictator givings. As 2 illustrates, the difference in the dictator givings from the benign-game player who chooses to cooperate (67.11 cents) and the malign-game player who chooses to defect (60.17 cents) is 6.94 cents.

3 to the Bayesian amount with the correct prior, -9.48 cents. [9]

The finding that providing "counterfactuals" reduces the correspondence bias helps to explain its robustness in daily life: it is usually impossible to observe the "counterfactual" behavior of the people we interact with. For example, in a society with low mobility, the rich are born rich and the poor typically remain poor. It is hard to see how the rich would behave if they were poor, and it is hard to observe how the poor would behave if they were rich. Even if some people experienced both cases, it is hard for others to witness how they behaved in the two different situations.

Interestingly, even though the benign premium becomes smaller in Treatment 3 and especially in Treatment 4, the proportion of subjects who are biased remains quite stable. Around 52.81 % and 47.31% of subjects in treatments 3 and 4 respectively still choose the benign-game player in the first choice when the malign-game player chooses cooperation, which is inconsistent with the predictions of the Bayesian model but consistent with our model of correspondence bias. One plausible interpretation is being correspondence-biased is a relatively stable trait, and experience and counterfactual information can only reduce the extent of the bias.

Robustness Checks

One potential concern is our results are driven by the complexity of the design or by subjects' inattention. We use education level to proxy mathematical/computational skills, and test whether people who have fewer years of education show a stronger sign of correspondence bias. For inattention, We use how long subjects stay in each stage as a proxy. People who pay more attention to the study may stay longer in each stage before they making their decisions. We present the results in Table 7. In this analysis, we only include observations in treatments 2, 3, and 4, as the definition of the benign premium is slightly different in Treatment 1. When looking at the effect of education on the level of bias, we continue the analysis with a subsample of subjects who finished the voluntary follow-up survey. As shown in Table 7 column 3, the level of education has no significant impact on the WTP for the benign-game player. The same applies to all the stay duration variables, which suggests that the results are not driven by people with a relatively low level of education or people who did not pay

[9] -9.48 is the difference in the dictator givings between the benign-game player who chooses to cooperate (67.11 cents) and the malign-game player who also chooses to cooperate (76.59 cents).

enough attention to the study.

5 Conclusion

This paper investigates people's tendency to underestimate the influence of immediate incentives when making sense of others' behavior. The key intuition is that failing to fully appreciate the impact of incentives on actions leads individuals to over-attribute others' behaviors to their enduring characteristics.

We test the predictions of the model in an experiment with 817 subjects. We first ask subjects to decide how much to transfer as the dictator in a dictator game. Then, we let them play the benign game and the malign game, and inform them of the actions of a benign-game player and a malign-game player. Lastly, we ask them to choose, as a receiver in the dictator game, between the benign-game player and the malign-game player's first-stage transfers. We find strong evidence of correspondence bias. Subjects are willing to pay 11.67 out of 100 more for the benign-game player's givings in the baseline treatment, which is significantly larger than what the Bayesian model predicts, 0. Allowing subjects to experience both games instead of playing one and observing the other one reduces correspondence bias, but the benign premium is still significantly above 0. However, if we inform subjects of how their benign-game partner behaves in the malign-game and *vice versa*, correspondence bias disappears.

Both treatments 3 and 4 suggest that correspondence bias is caused by ignorance of the effect of incentives on actions. In Treatment 1, we directly test the predictions of our model. We find that when choosing between a benign-game player and a random stranger, subjects are on average willing to pay more for the benign-game player; when choosing between a malign-game player and a stranger, they are on average willingness to pay more for the stranger.

Our findings shed light on why correspondence bias is widely observed in real life as well as on potential ways to reduce or eliminate it. First, in reality, we often only experience one environment and observe other environments, which makes it hard for us to understand how alternative environments affect other people's behaviors. Accordingly, one potential route to speed up social cohesion is to encourage social interactions between different groups and let them experience other people's lives (Rao, 2019; Lowe, 2020). Second, counterfactual information about how the people we encounter behave in other environments is rarely available; the broader the range of situations in which we observe

another person, the current research suggests, the more we are likely to appreciate how contingent the individual's behavior is on the situation they are in.

6 REFERENCES

Alstadsæter, Annette, Niels Johannesen, and Gabriel Zucman, "Tax evasion and inequality," *American Economic Review*, 2019, *109* (6), 2073–2103.

Ambuehl, Sandro and Shengwu Li, "Belief updating and the demand for information," *Games and Economic Behavior*, 2018, *109*, 21–39.

Andreoni, James, Nikos Nikiforakis, and Jan Stoop, "Are the rich more selfish than the poor, or do they just have more money? A natural field experiment," *NBER Working Paper*, 2017.

Barreda-Tarrazona, Iván, Ainhoa Jaramillo-Gutiérrez, Marina Pavan, and Gerardo Sabater-Grande, "Individual characteristics vs. experience: an experimental study on cooperation in prisoner's dilemma," *Frontiers in Psychology*, 2017, *8*, 596.

Benjamin, Daniel J, "Errors in probabilistic reasoning and judgment biases," in "Handbook of Behavioral Economics: Applications and Foundations 1," Vol. 2, Elsevier, 2019, pp. 69–186.

Bó, Ernesto Dal, Pedro Dal Bó, and Erik Eyster, "The demand for bad policy when voters underappreciate equilibrium effects," *The Review of Economic Studies*, 2018, *85* (2), 964–998.

Cason, Timothy N, Sau-Him Paul Lau, and Vai-Lam Mui, "Prior interaction, identity, and cooperation in the Inter-group Prisoner's Dilemma," *Journal of Economic Behavior and Organization*, 2019.

Charness, Gary, Luca Rigotti, and Aldo Rustichini, "Social surplus determines cooperation rates in the one-shot Prisoner's Dilemma," *Games and Economic Behavior*, 2016, *100*, 113–124.

Christian, Charles W, "Voluntary compliance with the individual income tax: results from the 1988 TCMP study," *The IRS Research Bulletin*, 1994, *1500* (9-94), 35–42.

Cooper, Russell, Douglas V DeJong, Robert Forsythe, and Thomas W Ross, "Cooperation without reputation: Experimental evidence from prisoner's dilemma games," *Games and Economic Behavior*, 1996, *12* (2), 187–218.

Edwards, Ward, "Conservatism in human information processing," in Kleinmuntz B, ed., *Formal Reresentation of Human Judgement*, New York: Wiley, 1968, p. 17–52.

Eyster, Erik and Matthew Rabin, "Cursed equilibrium," *Econometrica*, 2005, *73* (5), 1623–1672.

Gilbert, Daniel T and Patrick S Malone, "The correspondence bias.," *Psychological Bulletin*, 1995, *117* (1), 21.

Haggag, Kareem, Devin G Pope, Kinsey B Bryant-Lees, and Maarten W Bos, "Attribution bias in consumer choice," *The Review of Economic Studies*, 2019, *86* (5), 2136–2183.

Jones, Edward E and Victor A Harris, "The attribution of attitudes," *Journal of Experimental Social Psychology*, 1967, *3* (1), 1–24.

Kahneman, Daniel and Amos Tversky, "Subjective probability: A judgment of representativeness," *Cognitive Psychology*, 1972, *3* (3), 430–454.

_ and _ , "On the psychology of prediction.," *Psychological Review*, 1973, *80* (4), 237.

Lowe, Matt, "Types of contact: A field experiment on collaborative and adversarial caste integration," *CESifo Working Paper*, 2020.

Mengel, Friederike, "Risk and Temptation: A Meta-study on Prisoner's Dilemma Games," *The Economic Journal*, 2018, *128* (616), 3182–3209.

Möbius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat, "Managing self-confidence," *NBER Working Paper*, 2014.

Peysakhovich, Alexander and David G Rand, "Habits of virtue: Creating norms of cooperation and defection in the laboratory," *Management Science*, 2015, *62* (3), 631–647.

Phillips, Lawrence D and Ward Edwards, "Conservatism in a simple probability inference task.," *Journal of Experimental Psychology*, 1966, *72* (3), 346.

Piff, Paul K, Daniel M Stancato, Stéphane Côté, Rodolfo Mendoza-Denton, and Dacher Keltner, "Higher social class predicts increased unethical behavior," *Proceedings of the National Academy of Sciences*, 2012, *109* (11), 4086–4091.

Rao, Gautam, "Familiarity does not breed contempt: Generosity, discrimination, and diversity in Delhi schools," *American Economic Review*, 2019, *109* (3), 774–809.

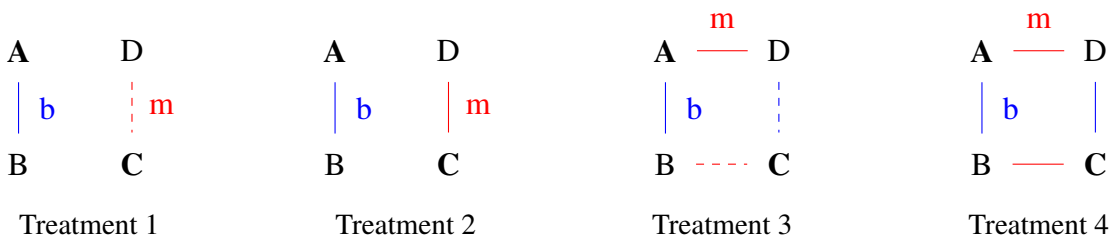
Ross, Lee, "The intuitive psychologist and his shortcomings: Distortions in the attribution process," *Advances in Experimental Social Psychology*, 1977, *10*, 173–220.

Saez, Emmanuel and Gabriel Zucman, *The triumph of injustice: How the rich dodge taxes and how to make them pay*, WW Norton & Company, 2019.

Simonsohn, Uri, Niklas Karlsson, George Loewenstein, and Dan Ariely, "The tree of experience in the forest of information: Overweighing experienced relative to observed information," *Games and Economic Behavior*, 2008, *62* (1), 263–286.

Figures

Figure 1: Overview of Four Treatments

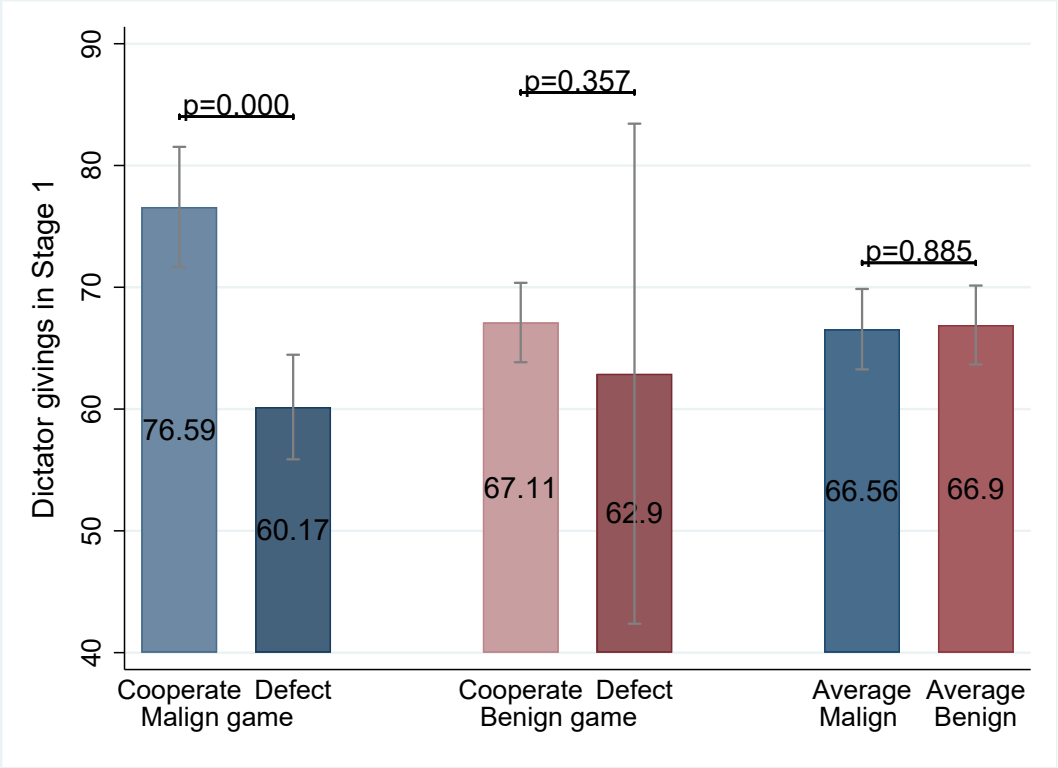


b: benign game m: malign game

: observed by A :not observed by A

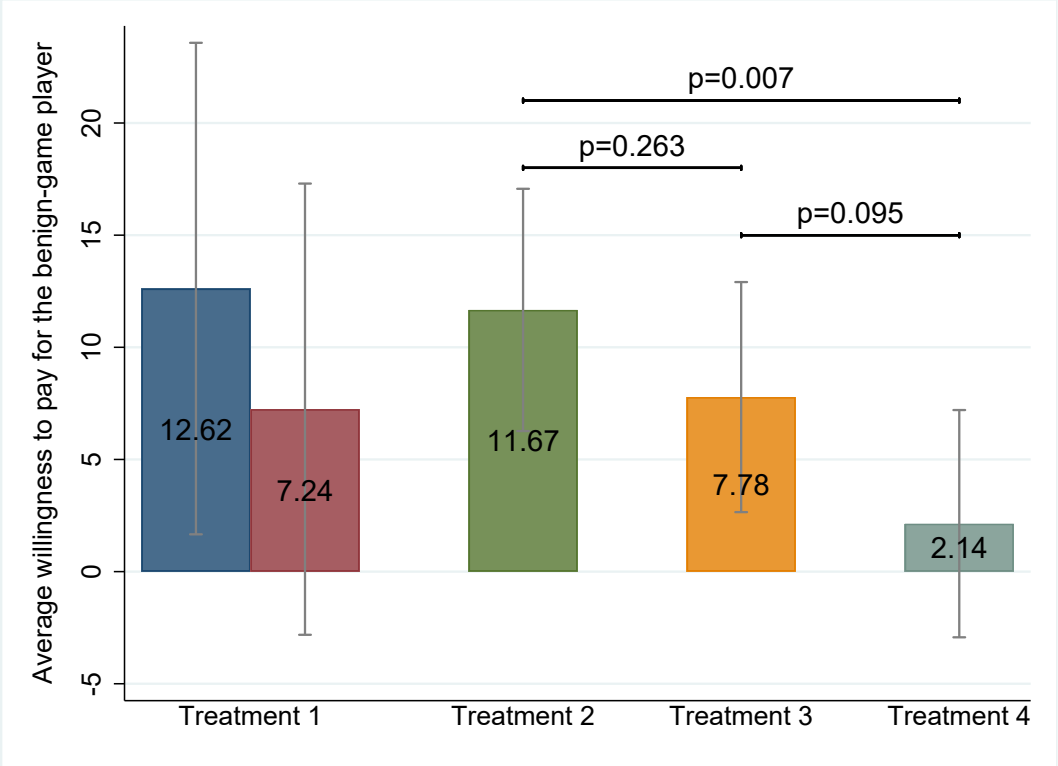
Notes: The figure displays the four treatments from subject A's perspective. The solid line denotes that A is able to observe (the outcome of) a game, and the dashed line denotes that A is not able to observe a game. But of course, A is not the only active player in the game. The games faced by B, C, and D are symmetric in treatments 2, 3, and 4. For example, player D plays the benign game with C and the malign game with A in Treatment 3. She cannot observe the game played between A and B or the game played between C and B in Treatment 3. The game is not symmetric in Treatment 1. In that treatment, A and B only play the benign game, and C and D only play the malign game.

Figure 2: Dictator Givings as a Function of Action in the Malign Game



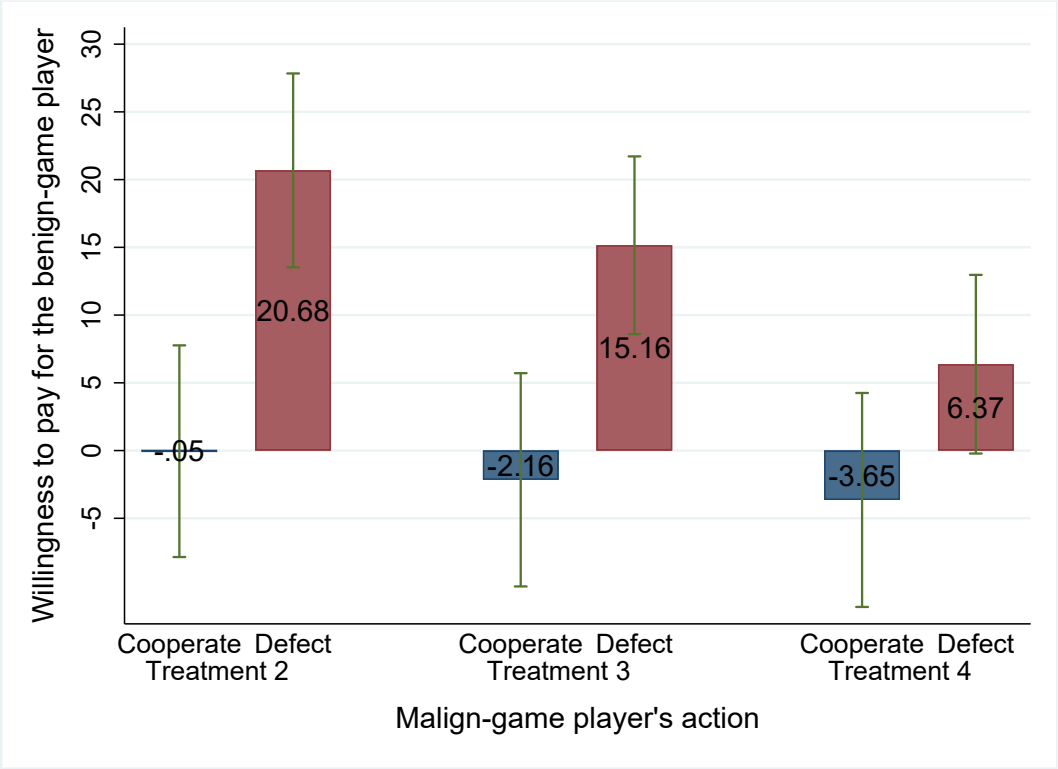
Notes: The figure plots the average dictator givings in the dictator game in Stage 1 depending on their own actions in the malign game. Recall that subjects were asked to divide 200 cents between themselves and a random receiver in the first stage. The bars show means of dictator givings, and the vertical lines report 95% confidence intervals.

Figure 3: Benign Premiums across Treatments



Notes: The figure plots the benign premiums across treatments. The bars show means of WTP for a benign-game player in different treatments. The vertical lines report 95% confidence intervals. The left bar in Treatment 1 represents the average WTP for a benign-game player when choosing between her and a stranger, and the right bar in Treatment 1 represents the average WTP for a stranger when choosing between him and a malign-game player.

Figure 4: Average WTP for Benign-game Player As a Function of the Malign-game Player's Action



Notes: The figure plots the average WTP for a benign-game player in treatments 2 to 4, depending on their malign-game player's action. The bars show means of WTP and the vertical lines report 95% confidence intervals.

Tables

Table 1: The Benign and Malign Games

Harmony Game

	C	D
C	40,40	10,30
D	30,10	0,0

Prisoner's Dilemma

	C	D
C	40,40	20,120
D	120,20	30,30

Notes: The harmony game is the benign game, and the prisoner's dilemma is the malign game.

Table 2: The Multiple Price List

	Option 1	Option 2
Choice 1	Amount transferred to me by B	Amount transferred to me by D+¢10
Choice 2	Amount transferred to me by B	Amount transferred to me by D+¢20
Choice 3	Amount transferred to me by B	Amount transferred to me by D+¢30
Choice 4	Amount transferred to me by B	Amount transferred to me by D+¢40
Choice 5	Amount transferred to me by B	Amount transferred to me by D+¢50
Choice 6	Amount transferred to me by B	Amount transferred to me by D+¢60
Choice 7	Amount transferred to me by B	Amount transferred to me by D+¢70
Choice 8	Amount transferred to me by B	Amount transferred to me by D+¢80
Choice 9	Amount transferred to me by B	Amount transferred to me by D+¢90
Choice 10	Amount transferred to me by B	Amount transferred to me by D+¢100

Notes: The table shows the multiple price list shown to subject A if she chose B over D in the first choice.

Table 3: Summary Statistics

Variable	All Sample	Treatment One	Two	Three	Four
Dictator giving	67.42	71.12	69.13	65.72	65.26
	(41.60)	(39.73)	(40.77)	(42.47)	(42.64)
Cooperation rate in the benign game	0.952	0.921	0.969	0.942	0.960
	(0.215)	(0.272)	(0.175)	(0.235)	(0.196)
Cooperation rate in the malign game	0.389	0.397	0.420	0.359	0.401
	(0.488)	(0.493)	(0.496)	(0.481)	(0.491)
Survey completion rate	0.903	0.901	0.902	0.897	0.912
	(0.296)	(0.300)	(0.297)	(0.305)	(0.284)
Observations	817	121	246	223	227
Follow-up survey					
Income	3.861	3.815	3.914	3.864	3.826
	(1.599)	(1.486)	(1.648)	(1.549)	(1.657)
Female	0.574	0.556	0.584	0.623	0.527
	(0.495)	(0.499)	(0.494)	(0.486)	(0.501)
Age	38.04	37.44	38.43	37.13	38.81
	(10.94)	(10.09)	(11.41)	(9.882)	(11.79)
Employment	0.819	0.824	0.819	0.829	0.807
	(0.385)	(0.383)	(0.386)	(0.377)	(0.396)
Observations	735	108	221	199	207

Notes: The table reports the mean for each variable in the whole sample and across treatments, with standard deviations in parentheses. We collect subjects' demographic information in a voluntary follow-up survey. 735 out of 817 subjects completed the survey. Income is a categorical variable, with categories 1="Less than $25,000", 2="$25,000 to $34,999", 3="$35,000 to $49,999", 4="$50,000 to $74,999", 5="$75,000 to $99,999", 6="$100,000 or more." Employment is defined as the percentage of people who are currently self-employed or employed.

Table 4: Cooperation and Dictator Givings by Games

	Benign game	Malign game
Cooperation rate	0.952	0.389
	(0.215)	(0.488)
Dictator givings	66.90	66.56
	(41.76)	(42.11)
Obs	640	627

Notes: The table reports the average dictator givings (in cents) and cooperation rates in the two games, with standard deviations in parentheses.

Table 5: Benign Premiums across Treatments

Treatment		Obs	Benign Premium	P-value $H_0 : BP = 0$	P-value $H_0 : BP_{Tx} = BP_{T2}$
Treatment 1	benignP VS stranger	63	12.62	0.025	
	stranger VS malignP	58	7.24	0.155	
Treatment 2		246	11.67	0.000	
Treatment 3		223	7.78	0.003	0.263
Treatment 4		227	2.14	0.407	0.007

Notes: The first row in Treatment 1 represents the average WTP for a benign-game player when choosing between the benign-game player and a random stranger, and the second row in Treatment 1 represents the average WTP for a stranger when choosing between the stranger and a malign-game player. BP stands for benign premium. Column (3) reports the p-value of t-tests, and column (4) reports the p-value of rank-sum tests.

Table 6: Benign Premiums and Fractions across Treatments and Malign-game Player's Actions

			Malign-game player				Rank-sum test	
			Cooperate		Defect		p-value	
	Fraction	BP	Fraction	BP	Fraction	BP		
	(1)	(2)	(3)	(4)	(5)	(6)	(3)vs(5)	(4)vs(6)
Panel A								
Treatment 2	0.62	11.66	0.52	-0.05	0.69	20.68	0.008	0.000
	(0.49)	(43.01)	(0.50)	(40.76)	(0.46)	(42.65)		
Obs	246	246	107	107	139	139		
BenignG only	0.74	21.93	0.74	13.89	0.74	27.88	0.990	0.042
	(0.44)	(40.78)	(0.44)	(35.91)	(0.44)	(43.32)		
Obs	127	127	54	54	73	73		
MalignG only	0.49	0.71	0.30	-14.24	0.64	12.73	0.000	0.000
	(0.50)	(42.76)	(0.46)	(40.80)	(0.48)	(40.75)		
Obs	119	119	53	53	66	66		
Panel B								
Treatment 3	0.61	7.78	0.52	-2.16	0.67	15.16	0.019	0.002
	(0.49)	(38.86)	(0.50)	(38.64)	(0.47)	(37.50)		
Obs	223	223	95	95	128	128		
Panel C								
Treatment 4	0.56	2.14	0.47	-3.65	0.62	6.37	0.025	0.013
	(0.50)	(38.73)	(0.50)	(38.98)	(0.49)	(38.14)		
Obs	227	227	96	96	131	131		

Notes: The table shows the fractions of subjects who chose the benign-game player over the malign-game player (*Benign Fraction*) in choice 1 with no bonuses and the benign premiums in treatments 2, 3, and 4. BP stands for the benign premium and Fraction stands for the benign fraction. Standard deviations are in parentheses. Columns 1 and 2 report the benign fractions and benign premiums in the three treatments respectively. Columns 3,4 and Columns 5,6 report the same statistics when the malign-game player chose to cooperate and defect respectively. Column 7 presents the p-value of a rank-sum test that the mean levels are the same for columns 3 and 5; column 8 presents the same test for columns 4 and 6. In Panel A, *BenignG only* denotes subjects who played the benign game and observed the malign game; *malignG only* denotes subjects who played the malign game and observed the benign game.

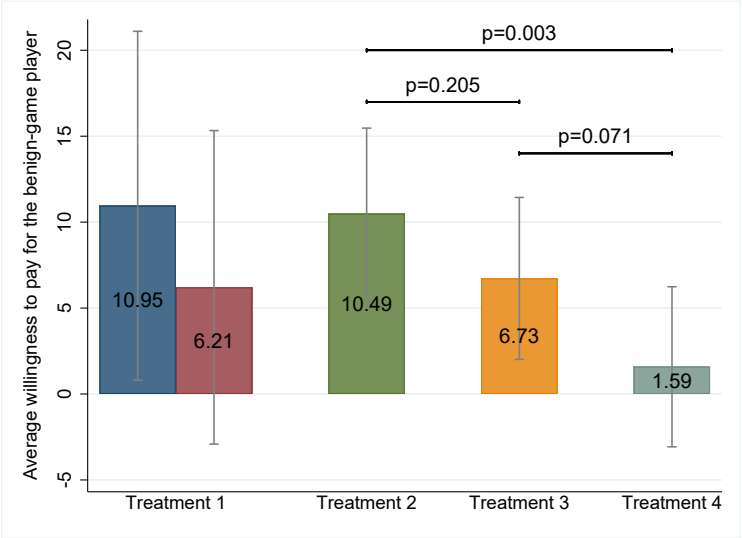
Table 7: Heterogeneous Analysis

Dependent Variable	Benign Premiums		
	All	Survey	Survey
	(1)	(2)	(3)
Treatment 3	-2.959	-4.730	-4.132
	(3.764)	(4.016)	(3.925)
Treatment 4	-8.430**	-8.004**	-7.888**
	(3.812)	(4.034)	(3.952)
Education			1.069
			(1.911)
Stage 1 stay time			0.032
			(0.045)
Stage 2 stay time			0.010
			(0.035)
Stage 2 results stay time			0.008
			(0.012)
Stage 3 stay time			-0.006
			(0.007)
Observations	696	627	627

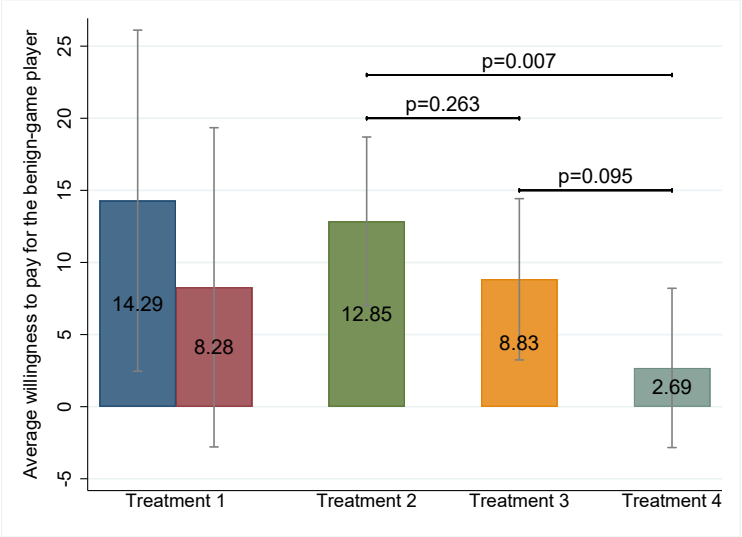
Notes: The table reports results from an interval regression to address the concern that multiple price list only elicits intervals of WTP. Observations are subjects in treatments 2, 3, and 4. The omitted group is Treatment 2. Column 1 includes all subjects in treatments 2, 3, and 4. Columns 2 and 3 include subjects who completed the follow-up survey. All regressions include the date of participation fixed effects. In column 3, we also include subjects' gender, income, risk preference, malign-game player's action. *** $p<0.01$, ** $p<0.05$, * $p<0.1$. Standard errors in parentheses.

Appendices

Figure A1: Robustness Check - Benign Premium across Treatments



Panel (a)



Panel (b)

Notes: The figure plots the average WTPs towards the benign-game player over the four treatments using alternative coding methods. In our main analysis, we code the WTP as the median value of the interval at which subjects switched from one option to the other. In Panel (a), WTP is defined instead as the lower bound of that interval. In Panel (b), WTP is defined as the upper bound of that interval. In both panels, the left bar in Treatment 1 represents the average WTP for a benign-game player when choosing between her and a stranger, and the right bar in Treatment 1 represents the average WTP for a stranger when choosing between him and a malign-game player.

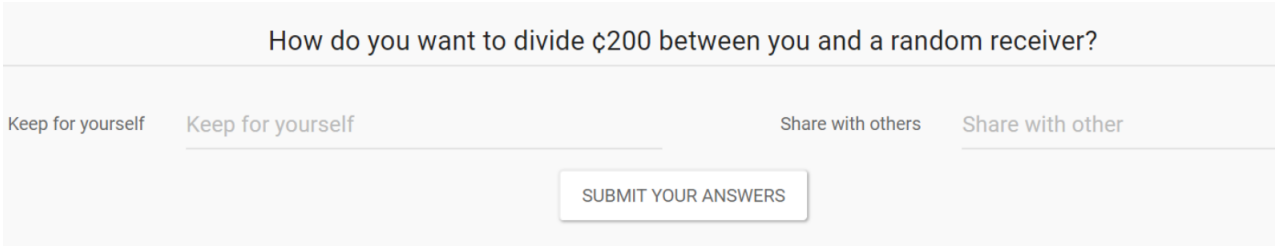
A1 Experimental Instructions

We present experimental instructions for subject D in Treatment 2. Experimental instructions for other subjects in the same treatment and for other treatments are similar and available upon request.

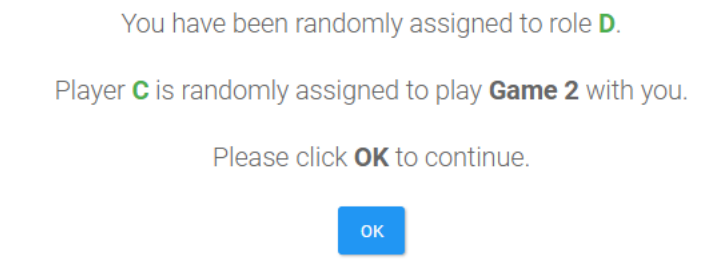
Stage One

Welcome to this experiment! In this stage, each of you will be paired with a different person. Your decisions will be strictly anonymous and cannot be linked to you in any way. Everyone will receive a show-up fee of 50 cents for participating in this study. In addition, either you or the person you are paired with will have access to another 200 cents. Now you have to decide how to divide the 200 cents between you and the person you are paired with. He or she with will also make such a decision. Your decision or his/her decision will be implemented later with 50% chance.

Your choice of how much to give and how much to keep can be any amount between 0 and 200 cents, in 1 cent increments. Your earnings from this stage will be 50 cents show up fee plus the amount you choose to keep out of the 200 cents **if your choice is randomly chosen to be implemented**. If the choice of the person you are paired with is chosen to be implemented, which happens 50 percent of the times, you get 50 cents plus the amount he/she transfers to you. Your decision will not affect his/her decision.



Stage Two

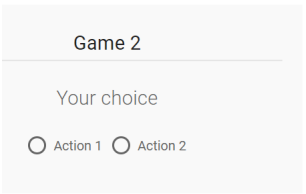


Note: Example screenshot from subject D's perspective.

Welcome to Decision 2, you will receive your earnings from the choices made in this session in addition to your earnings in the first session. How much extra money you earn in Decision 2 depends

on the decision you make and the decision made by the person with whom you are paired.

You are randomly matched with 3 other persons in your own group. You have a role of D, the other three have the role of A, B and C You are grouped anonymously, which means that you will never learn the identity of the others. Here are the games: You (player D) play Game 2 with player C.



Game 2

		You	
		Action 1	Action 2
C	Action 1	40, 40	20, 120
	Action 2	120, 20	30, 30

Both of you have two choices: Action 1 and Action 2 and you two make choices simultaneously. The 1st number in each cell refers to the payoff (in cents) for the C player, while the 2nd number in each cell refers to the payoff (in cents) for you. Thus, if C choose Action 1 and you choose Action 1, C would receive 40 and you would receive 40; If C choose Action 1 and you Action 2, C would receive 20 and you would receive 120; If C choose Action 2 and you Action 1, C would receive 120 and you would receive 20; If both C and you choose Action 2, C would receive 30 and you would receive 30. You only need to make one decision in Game 2. You don't need to play Game 1. You will then learn

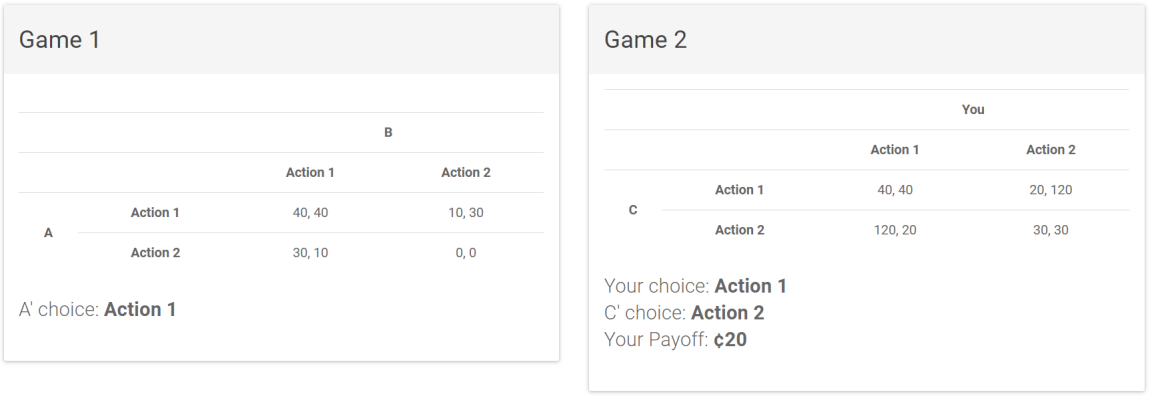
Game 1

		B	
		Action 1	Action 2
A	Action 1	40, 40	10, 30
	Action 2	30, 10	0, 0

your payoff for the period.

Stage Two Results

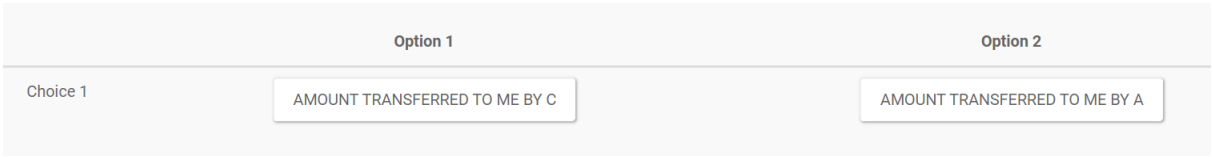
Figure A2: Stage Two Results Seen by Subject D in Treatment 2



Stage Three

In this part you will get to make a series of choices about whether you want to be matched with person A or C from your group.

Remember in Decision 1, everyone was asked to split 200 cents between oneself and a receiver. Both A and C had made that decision. Now you just need to guess how much A decided to transfer and how much C decided to transfer in Decision 1. Imagine you are the receiver, you need to choose your sender. If A transferred X cents to the receiver and C transferred Y cents to the receiver, then if you choose A, you will get X cents and if you choose C, you will get Y cents. You chose C. Will



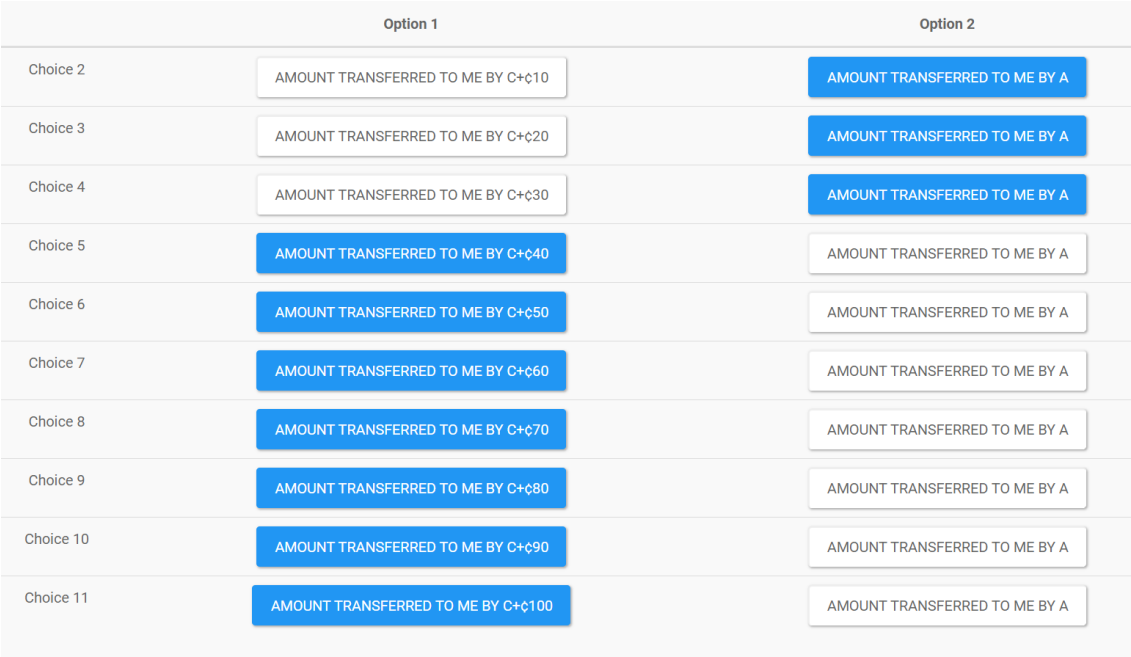
you change your mind if we give you 10 extra cents for choosing the other person A? How about A's transfers in Decision 1 plus 30 cents vs C's transfers? How about 100 extra cents for choosing A, will you still prefer C's transfers? Now, you not only need to guess who transferred more in Decision 1, A or C, but also need to guess how big the gap between the two transfer amounts is. The table below shows the 10 choices between C and A plus some bonuses you will consider. Please make a decision in each of the 10 choices.

1. There is a 25% chance you will be selected to have one of your choices implemented. If you are selected,then we will randomly select one of your 11 choices to implement (the choice you just made plus the 10 choices in the table below). Every choice has the same probability to be implemented.

2. For example, suppose Choice 8 is randomly selected to be implemented. If you chose Option 1 (amount transferred by C) in Choice 8, then you will get Y cents , where Y cents is the amount transferred by C in Decision 1. If you chose Option 2 (amount transferred by A+70 cents) in it, then you will get (70+X) cents, where X cents is the amount transferred by A in Decision 1.

3. At the end of the instruction, you will need to answer several comprehension questions. Answer all of them correctly and you will be able to submit your choice. If one or more of your answers are incorrect, read the instruction again and re-do the test.

4. Between A and C, you are just choosing who to be your sender. If one of your choices is chosen to be implemented, then both A and C will be senders. You cannot reward/punish someone by choosing/not choosing him/her. It's in your best interest to select the option you believe can give you a higher payoff in each choice.



Comprehensive questions.

Question 1: Are the following statements correct?

Statement 1: In Decision 3, I need to choose between A and C as my sender in Decision 1. Both of them had already made their decisions about how much to transfer in Decision 1.

Statement 2: In Decision 3, I need to choose between A and C to play a new game. If the one I choose does not like me due to my choices in Decision 2, he/she may try to hurt me in this new game.

A. Both statements are incorrect

B. Both statements are correct

C. Statement 1 is correct, Statement 2 is incorrect

D. Statement 1 is incorrect, Statement 2 is correct

Question 2: Imagine you need to choose between two hypothetical partners: person 1 and person 2. Person 1 chose to transfer 500 out of 1000 cents in Decision 1 and person 2 chose to transfer 50 out of 1000 cents in Decision 1. There are two options. Option 1: the amount transferred by person 1. Option 2: the amount transferred by person 2 plus 300 cents. How much do you get if you choose Option 1? How about Option 2? (This question is to test your understanding of Decision 3; the numbers and the persons to choose from are intended to be different from that in the study)

A. Option 1: 800 cents; Option 2: 50 cents

B. Option 1: 500 cents; Option 2: 350 cents

C. Option 1: 500 cents; Option 2: 50 cents

D. Option 1: 800 cents; Option 2: 350 cents

Question 3: Suppose in a choice between an apple plus 10 cents and an orange, someone chose an apple plus 10 cents. Consider the following two choices. Choice 1: an apple plus 20 cents or an orange. Choice 2: an apple plus 30 cents or an orange. Which one will she choose in those two choices?

A. Choice 1: an apple plus 20 cents; Choice 2: an apple plus 30 cents

B. Choice 1: an apple plus 20 cents; Choice 2: an orange

C. Choice 1: an orange; Choice 2: an apple plus 30 cents

D. Choice 1: an orange; Choice 2: an orange

E. Not sure

Experimental Results