Strategy vs. Direct-Response Method: Evidence from a Large Online Experiment on Simple Social Dilemmas

Marcus Roel
marcus.roel@nottingham.edu.cn
School of Economics & CeDEx China
University of Nottingham Ningbo China

Zhuoqiong Chen*
chenzq926@gmail.com
School of Economics and Management
Harbin Institute of Technology, Shenzhen

November 7, 2025

Abstract

This paper examines behavioral differences in sequential games that arise when choices are elicited via the direct-response method, where players observe the choices made by those who acted earlier and respond with a single choice, or strategy method, where they make their choice(s) in response to all possible choices by those who acted before regardless of whether such actions were taken. We conducted a 2×2 between-subject large-scale online experiment with over 8,000 participants on Amazon MTurk, manipulating both the elicitation method and participants' ex-ante beliefs about player 1's choices via an information-provision treatment. In neutrally framed binary-choice games, a sequential Prisoners' Dilemma and a mini-Ultimatum Game, we document that the strategy method does not appear to alter player 2's preferences, i.e., their tendency to reward cooperation or reject unfair offers. However, it reduces the tendency to reward defection and reject fair offers, which we interpret as a reduction in mistakes.

Keywords: Strategy method, Direct-response method, Online experiments, Information-provision experiment, Elicitation, Methodology

JEL Codes: C72, C91, C92, D83, D91

*ORCIDs: 0000-0002-8322-3714 (M. Roel) and 0000-0002-7782-7562 (Z. Chen). We would like to thank participants at the Experimental Economics Workshop at Beijing Normal University, and the CeDEx China Workshop at Nottingham Ningbo for their helpful comments and feedback. Zhuoqiong Chen gratefully acknowledges support from the National Natural Science Foundation of China (No. 71903046 and No. 72373034) and the "Shenzhen Peacock Plan" start-up research grant (No. GA11409002).

1. Introduction

Social sciences, such as economics, management, and psychology, often rely on experiments and surveys to collect data. The validity and reliability of such data depend on whether it captures the participants' genuine preferences, feelings, or beliefs. A key methodological question is therefore how such behavior and beliefs can be best elicited – ideally in a way that is simple to understand for participants, as well as inexpensive to implement and easy to conduct for researchers.

For sequential games, the two most common approaches to elicit choices from players who do not act first are the direct-response method and the strategy method. In the former, players directly observe the choices made by players who acted before them and make a (single) choice in response. In the latter, they make their choice(s) without knowing anything about the choices made by those who acted before them. Instead, they are asked to take an action in response to *all possible choices* by those who acted prior, regardless of whether such actions were actually taken. In other words, the direct-response method elicits a single choice[1] whereas the strategy method elicits a full strategy profile. Nevertheless, the general methodological approach is not limited to sequential games; it extends to all contexts involving conditional elicitations (e.g., state-dependent preferences, etc.), emphasizing the importance of understanding when to use each method.

If players' preferences, understanding of the game, and likelihood of making mistakes are not affected by which of the two elicitation methods is used, they will generate the same data. In this case, the strategy method has the advantage of allowing researchers to capture choices at all points in the game for every player, which greatly increases the effective sample size and lowers the cost of running experiments compared to the direct-response method. Moreover, running experiments becomes easier as they can be conducted as a de facto single-user survey with ex-post matching of players to determine payoffs, which is especially useful for large online experiments that have become much more popular recently, partially due to Covid-19.[2]

The strategy method has, however, been criticized for being artificial and psychologically "cold", as forming a contingent plan of actions may not be a natural decision process for most people, and thus may result in different behaviors compared to the direct-response method (e.g., Roth (1995)).[3] In a detailed review of the literature, Brandts and Charness (2011) note that there are generally more studies that find no behavioral differences across elicitation methods than those that do. However, among the cited studies, 13 out of 29 document different behavior or at least partial differences. Additionally, the authors emphasize that they cannot find cases where a treatment effect present under the strategy method is not also observed with the direct-response method, which is noteworthy for experimentalists who are primarily concerned with testing treatment effects.

We conduct a large online experiment on Amazon MTurk with over 8000 participants to investigate the role of the strategy method and the direct-response method in shaping behaviors in social dilemmas. Participants play two sequential, binary-choice games (neutrally framed and in random order), corresponding to a sequential Prisoners' Dilemma (sPD) and a mini-Ultimatum Game (mUG; with an unfair 85-15 or a fair 50-50 split). The choice of games was kept intentionally simple in order to allow for a systematic analysis of the elicitation method across games where social preferences may give rise to player 2 rewarding a helpful action (i.e., cooperate after cooperation) or punishing an unfair choice (i.e., reject the 85-15 offer) of player

[1] This may happen more than once if the player needs to respond to other's choices at multiple points in time.

[2] Figure A.1 in the Appendix documents the recent popularity of online experiments.

[3] It is well understood that beliefs and choices may (partially) be a result of how they are elicited (Gigerenzer and Hoffrage (1995)), or how the problem is presented or framed (Gigerenzer et al. (1988), Tversky and Kahneman (1981)). It is not always obvious, however, what the "true data" is or how behavior will be affected. Both, observing a particular fair/unfair choice that stirs the player's emotions (direct-response), or an increased fairness concern due to considering all nodes of the game (strategy method), may lead to more non-selfish choices. Indeed, Roth (1995) suggests that it is not clear whether forcing participants to think about all their information sets is an advantage or disadvantage compared to the direct-response method.

1, and which may vary with hot-or-cold emotional states. The choice of simpler games also aligns with findings that games with fewer contingent choices are more likely associated with behavioral differences across elicitation methods (Brandts and Charness (2011)). We adopt a 2×2 between-subject design, where, in addition to varying the elicitation method, we manipulate the ex-ante belief of all participants about player 1's choice by providing them with information about typical player 1 behavior in the same games from past published studies. In particular, such information highlighted that the majority of first movers either took the "selfish" (defect, 85-15 offer) or "non-selfish" action (cooperate, 50-50 offer). We refer to the two respective information treatments as the *selfish* and *non-selfish belief treatment*.

This design is motivated by the conjecture that ex-ante beliefs may have a different effect on player 2's response depending on the elicitation method employed, and that this orthogonal dimension explains why behavior across the two elicitation methods is at times similar and at times different in the experimental literature. While multiple behavioral explanations (preferences and/or biases) could account for this differential behavior, the concept of *incomplete conditional thinking* particularly shaped our thinking. In *sequential games*, player 2 faces uncertainty with regards to player 1's choice when her choice is elicited with the strategy method, but not when the direct-response method is used. We conjecture that behavior elicited by the strategy method can be different, due to incomplete conditional thinking: By not fully conditioning on the particular node player 2 is at, her preferences and choices will be influenced by her expectations of what player 1 is going to do. For example, if she believes that player 1 is likely to behave (non-) selfishly, her preferences at any hypothetical choice node will be pushed towards favoring the (non-) selfish choice. Consequently, we hypothesize that player 2's behavior is affected by her ex-ante expectation of player 1's action in the strategy method. Exogenously varying our participants' expectations about player 1's likely choice allows us to test this hypothesis.[4]

More broadly, this experiment is inspired by the literature on the failure of contingent reasoning (see the survey by Niederle and Vespa (2023)), which shows that assisting subjects to focus on all relevant contingent states may help them make optimal choices.[5] However, in games where social preferences play a role, what is the optimal action is less clear, as some subjects may not consider taking the selfish action to be optimal even if they were assisted to focus on all relevant states, i.e., the actions taken by player 1. Hence, despite

[4] Previous first-order explanations for behavioral differences across elicitation methods are generally independent of variation in beliefs (or other orthogonal dimensions) and thus cannot explain the existing experimental results. For example, initial beliefs play no role if the emotional response to a certain action is stronger in the direct-response method (hot-versus-cold, e.g., Brandts and Charness (2000)). Of course, manipulating ex-ante beliefs may have a *direct* effect on preferences (i.e., norms, kindness-evaluations, etc.). Incomplete conditional thinking, among other ideas, suggests this manipulation has a *differential* effect across the elicitation methods. For further details, please consult the theory section in the Online Appendix. Here, the reader will find our theoretical framework that motivated our design, formalizes a simple model of social preferences with incomplete conditional thinking, and discusses various other types of preferences relevant to our setting.

[5] Our experimental design is, however, distinct from a typical test of the failure of contingent reasoning (hereafter, FCR), which often compares choices in simultaneous games to choices in the same setting after exposing participants to certain measures that help them focus on all relevant states (e.g., playing a sequential version of the games, predicting what the first mover would do, or imagining what their own choice would be given that the first mover has hypothetically made a choice). One study that inspired us significantly is Martínez-Marquina et al. (2019). In their creative design, subjects in one treatment choose a price to purchase a firm that may be of high or low value with equal chance, and those in another treatment choose a uniform price that either purchases a low-value firm or, if it is high enough, two firms, one with high and the other with low value. Parameters are designed such that the theoretical optimal prices are the same across treatments, yet subjects were significantly more likely to choose the optimal price in the latter than the former setting, indicating the "power of certainty" in contingent reasoning. We borrow this idea and conjecture that the fact that there is uncertainty in what player 1 would choose in the strategy method but not in the direct response method may lead to different responses by player 2. In fact, "reason-based choice" (Shafir et al., 1993; Tversky and Shafir, 1992; Croson, 1999), one mechanism behind the FCR provides similar predictions as our conjecture in that observing player 1's actual move in the direct response method provides player 2 a "stronger" reason than observing his hypothetical move in the strategy method to take some actions. We highly appreciate Emanuel Vespa for helping us clarify these points. See also the following literature that has documented people's difficulties with contingent reasoning, for example, Shafir and Tversky (1992); Friedman (1998); Charness and Levin (2009); Rabin and Weizsäcker (2009); Esponda and Vespa (2014, 2018, 2024); Cason and Plott (2014); Louis (2015); Eyster and Weizsacker (2016); Enke (2020); Araujo et al. (2018); Moser (2019) and Ngangoué and Weizsäcker (2021).

the relationship to and inspiration by this literature, our findings will be more relevant to the literature that compares elicitation methods.

Overall, we find that MTurk workers' behavior is broadly consistent with conditional cooperation in the *sPD*, with over 60% of player 2 choosing to cooperate after player 1 cooperates and over 75% opting to defect after player 1 defects in all treatments. Yet despite this strong pro-social behavior in the *sPD*, only few participants are willing to punish unfair offers in the *mUG*: rejection rates of unfair offers are below 15% in all treatments and never higher than 4% for even splits. However, participants who do act in a non-selfish manner mainly do so in a way consistent with conditional reciprocity.

Our belief manipulation strongly affects players' beliefs about player 1's choices and, to a slightly lesser extent, player 1's choices themselves – in line with the information provided. Its direct effect on player 2's propensity to take a non-payoff maximizing choice is either very small or non-existent, however.

Moreover, we find no evidence that the belief treatment exerts a significant differential effect (by elicitation methods) on player 2's response to cooperation or unfair offers. Furthermore, for these choices, the elicitation method appears inconsequential: there is no significant difference in player 2's tendency to reward cooperation by player 1 and only a small, albeit significant, 4 percentage point difference of rejecting unfair offers between the two methods. That does not mean that the choice of elicitation method is immaterial, however. Our data highlights that the strategy method reduces both the frequency of cooperation in response to player 1 defecting (rewarding the unhelpful, payoff-minimizing choice) and rejecting fair offers (punishing the payoff-maximizing choice). This effect is large, at around 10 percentage points and constant across belief treatments in the *sPD*, whereas it mainly affects the selfish-belief treatment in the *mUG*. While it is possible that these behavioral differences reflect genuine differences in preferences across elicitation methods, we believe a better interpretation for our findings is that the strategy method reduces "mistakes". We arise at this conclusion based on four pieces of evidence. First, behavior that is consistent with preferences for efficiency or spitefulness is unlikely to be a driving force behind this result as it can, at most, explain only 1/3 of the mistakes we observe. Second, mistakes in our games are positively correlated with inattention, which is captured by whether participants make a mistake in a control questionnaire, and which is completed right after the experimental instruction. Third, the strategy method is particularly useful in reducing mistakes among those who are more likely to make them, namely our inattentive participants. Fourth, response time data highlights that longer deliberation reduces mistakes but is unrelated to conditional cooperation or rejecting unfair offers.

We contribute to the literature in several ways. Our paper is the largest experiment to this date that analyzes the important question of whether eliciting choices via the direct-response method or the strategy method fundamentally influences behavior. Prior studies, which are discussed in detail in the next section, that adopt both methods often show that subjects are more likely to punish selfish behavior under the direct-response method while other studies found no difference between the two methods. Our results suggest that preferences seem to be unaffected by the elicitation method yet mistakes are significantly lowered by the strategy method.

This insight may be particularly important for online experiments which are increasingly relied upon nowadays.[6] For (at the very least) the typical simple experiments that are run online, the strategy method

[6] As of right now, the social science literature that uses online experiments has become too large to provide an exhaustive list of references. Consequently, we will limit ourselves here to a few key related papers as well as broader reviews. In economics, Horton et al. (2011) provide an early discussion of conducting experiments in online labor markets. Relevant to our work, they (i) find similar levels of cooperation in a prisoner's dilemma on MTurk and a traditional laboratory environment and (ii) show that MTurkers respond to framing (Tversky and Kahneman, 1981). Other studies that evaluated the use of MTurk for laboratory research typically conclude that it provides consistent, reliable, high-quality data (Paolacci et al., 2010; Buhrmester et al., 2011; Amir et al., 2012; Paolacci and Chandler, 2014; Johnson and Ryan, 2020). More broadly, Fréchette et al. (2022) review the past and future of experimental economics, highlighting the growing role of online methodologies. Hunt and Scheetz

appears to be the superior choice as it reduces mistakes, lowers costs, and is, in many cases, easier to run, without distorting participants' preferences. We hope our large-scale experiment provides the empirical support for using the strategy method in these settings.

Furthermore, our data highlights a small, yet noteworthy detail, which suggests that the strategy method appears to facilitate the strategic thinking of participants. In particular, we observe that player 1's behavior is more responsive to the information, which indirectly provides information about player 2's likely response, in the strategy method treatment. This effect may prove particularly useful for online experiments that focus on one-shot games.[7]

Finally, we contribute to the literature on information-provision experiments (Haaland et al. (2023)) by providing a very simple, yet powerful method to shape participants' beliefs about the likely choices of others in games. Similarly to the literature on framing in games (Ellingsen et al. (2012), Dreber et al. (2013), Ockenfels and Werner (2014)), the belief treatment appears to have little direct effect on players' preferences (as documented by player 2 behavior) and thus seems to operate mostly through beliefs when shaping player 1s' choices.

1.1. Literature Review

Bargaining games. Studies on the Direct Response (DR) versus Strategy Method (SM) in bargaining contexts, particularly in ultimatum games, have yielded two main insights: the DR appears to elicit stronger emotional, "hot" reactions, while the SM often induces more deliberative, "cooler" responses. Güth and Kocher (2014) provide a comprehensive review of over three decades of ultimatum experiments, noting that the DR commonly captures immediate emotional responses, leading to higher rejection rates of unfair offers, whereas the SM generally prompts more reasoned, monotonic acceptance thresholds. Oxoby and McLeish (2004) explicitly compare the two elicitation methods. While they find no statistically significant difference in mean offers or acceptance rates, they observe that low offers are rejected more frequently under the DR, which may suggest an amplified emotional response. Aina et al. (2020) provide additional support for the "hot" nature of the DR by testing the frustration-and-anger model of Battigalli et al. (2019). Their data show that the DR triggers more rejections of "greedy" offers, especially when responders initially expect fair treatment. The authors link this to the salience of proposers' self-interested actions under the DR.

Differences between elicitation methods are absent in Güth et al. (2001), who are employing a mini-ultimatum game with binary offers. They find that under the DR, responders are marginally less likely to reject an "unfair" offer when the alternative is a nearly-equal split, compared to when the proposer's choice set includes a lopsided offer and an exactly even split. While this effect is absent under the SM, there is no statistically significant difference in acceptance and rejection rates between elicitation methods.

A few papers on multi-stage bargaining tasks seem to indicate that the distinction between emotional and deliberative decision-making may become more pronounced in multi-stage games. Chuah et al. (2014)

(2019) provide an overview of how online experiments have been utilized in the field of information systems, whereas Fink (2022) elaborates on their specific benefits and applications for advancing research. Aguinis et al. (2020) examine the role of online experiments in management science, particularly in addressing the reproducibility and replicability crisis. In marketing, Goodman and Paolacci (2017) discuss the use of crowdsourcing platforms for consumer research. Holzmeister et al. (2024) explore the replicability of online experiments across different disciplines, while Krefeld-Schwalb et al. (2024) provide a broader, more critical examination of why effect sizes differ across various social sciences.

[7] In an early survey of the bargaining literature, Güth and Tietz (1990) make a similar observation when discussing Güth et al. (1982), who included a treatment with role-reversal and strategy method as a "consistency check" in their experiment, pointing out that subjects' attention to strategic aspects can be shaped by the experimental setting. Surveying the impact of players' role-reversal more generally, Brandts and Charness (2011) conclude that it is difficult to draw any conclusion about its effect. Indeed, when Selten (1967) first proposed the elicitation of choices via strategies, he suggested letting subjects become familiar with the environment first – using repeated play given the direct-response method. Only afterwards were their full preferences elicited via a complete strategy.

examine a four-stage bargaining experiment and show that bargaining failures arise more often in the DR than the SM, possibly because confrontation in real time encourages escalatory moves. An alternative interpretation of their findings offered by the authors is that the complexity of decisions in the SM leads to quantal response. This aligns with the observations of Brandts and Charness (2011), who noted that the SM may elicit errors in games with more decision nodes.

In a separate line of inquiry, Chen and Schonger (2024a) provide a formal theoretical analysis demonstrating that the SM can yield systematically different outcomes than the DR due to structural differences in decision-making. They argue that participants in the SM must pre-commit to choices across all information sets, whereas in the DR, decisions are made sequentially in response to actual events. This commitment effect results in choices at one node being influenced by psychological costs tied to committing to certain actions in alternative nodes. Consequently, off-equilibrium motivations – such as self-image concerns, duty, or aversion to unfairness – can shape behavior differently under the SM compared to the DR. Chen and Schonger (2024b) confirm these predictions: acceptance rates in ultimatum games are higher under the DR than under the SM. The DR likewise prompts stronger reciprocity in trust games and greater punishment in a three-player prisoner's dilemma. Chen and Schonger caution that the SM's sensitivity to framing can distort treatment effects, urging pilot testing to detect method biases. They conclude that the SM and the DR are not equivalent, especially in emotionally or socially complex contexts. Although our own experimental findings partially concur with their theory in showing that the SM encourages rational reflection and better understanding of instructions, we also observe that the SM reduces errors and heightens attentiveness in online experiments.

Prisoners' Dilemma and reward-punishment games. Studies comparing the SM and the DR in Prisoners' Dilemma (PD) games often investigate whether elicitation methods affect cooperation and defection. Brandts and Charness (2000) examine a sequential PD in which second movers see first movers' choices in the DR, and pre-specify contingent actions in the SM, which is equivalent to the sPD in our setting. They find no statistically significant differences in cooperative or defecting behavior. Reuben and Suetens (2012) extend this inquiry to a repeated sequential PD with a probabilistic endpoint. Despite manipulating whether participants knew the final period and whether second movers observed first movers' behavior, they find almost no difference in cooperation across treatments or elicitation methods, suggesting that, in comparatively straightforward PD environments, the SM and the DR can produce similar outcomes. It is interesting to highlight that the differences in elicitation methods in multi-period bargaining mentioned previously (Chuah et al., 2014), do not simply translate to multi-period PDs.

Falk et al. (2005) study a prisoners'-dilemma-like setting in which participants can punish defectors. Although the overall punishment behavior is similar under the DR and the SM, they show that the average penalty inflicted on defectors under the DR is nearly twice as large as under the SM. The authors attribute this difference to the more emotionally charged atmosphere of the DR. Their results align with a broader theme that the DR often elicits stronger "hot" responses for punishment-like behavior.

Beyond canonical prisoners' dilemmas, mixed findings emerge from studies comparing elicitation methods in reward-punishment games. Brandts and Charness (2003), in a one-shot reward-punishment setting, observe more intense punishment under the DR, even though reward rates do not differ. Li et al. (2024) extend that design over 20 fixed-matching rounds and find that the SM participants actually punish and reward more frequently than those under the DR, implying how repeated interaction and strategic deliberation can reverse some one-shot patterns. Zhao et al. (2018) switch to random rematching and also find higher punishment under the SM, suggesting that the reversal remains salient even when participants are not paired with the same partners.

Jordan et al. (2016) note that in dictator games with third-party punishment (i.e., punishment from a non-participant of the game itself), where emotional involvement is typically lower than in second-party punishment (i.e., when a participant of the game punishes the counterparty), the SM and the DR generate similar punishment levels. Their findings reinforce the point that the emotional stakes of a decision are crucial to whether elicitation method matters. In contexts of reduced emotional salience, the SM and the DR seem to converge. Indeed, Zhao et al. (2016) find no behavioral difference in a capacity-allocation game and observe similar neural activation under the SM and the DR, indicating that in certain settings, the "cooling" effect of the SM makes little difference.

Other games. Several additional lines of research compare the SM and the DR in trust, cheap talk, and public goods games. Amdur and Schmick (2013) show that trust and reciprocity remain stable across different elicitation methods, whereas Davis (2018) demonstrates a "cooling" effect of the SM in a gift-exchange game: when agents pre-specify effort levels, they become less resentful toward "controlling" contracts (i.e., contracts that impose stricter monitoring or minimum effort requirements), which ultimately promotes higher effort and more cooperative outcomes. Di Bartolomeo and Papa (2016) find that the SM promotes more consistent reciprocity in a triadic setting combining investment and dictator games. Minozzi and Woon (2020) explore cheap talk, expecting the SM to yield more equilibrium-like communication through counterfactual thinking, but in fact discover that the SM participants overcommunicate to a greater extent than predicted by equilibrium theory. Lin and Palfrey (2022) demonstrate that in centipede games, the DR leads to early "take" decisions, while the SM prolongs play, yielding higher overall payoffs. In public goods games, Fischbacher et al. (2012) note that the SM consistently identifies conditional cooperators and free-riders, matching the DR in classification. In an experiment comparing the effects of leading by example and by suggesting an effort level on team effort contribution, Dong et al. (2018) found no difference between elicitation methods in followers' behavior. Zhao and Zhao (2018) confirm the minimal impact of the SM and the DR in a competing newsvendor setting.[8]

The rest of the paper is structured as follows: section 2 details our experimental design. The results are presented in section 3. Our paper concludes with a discussion in section 4. Other supportive tables and figures (e.g., variable definitions, summary statistics, etc.) can be found in appendix A. The instructions for our experiments are provided in appendix B.

2. Experimental Design

The study was pre-registered and has been approved by the (School of Economics and Management) Research Ethics Committee at the Harbin Institute of Technology, Shenzhen.[9] The experimental instructions can be found in section B of the appendix. Screenshots of the experiment can be found in the Online Appendix.

Our participants played two sequential games, the sequential Prisoner's Dilemma (sPD) and the mini Ultimatum Game (mUG). The payoffs in the sPD were (\$1, \$1), (\$1.5, \$0), (\$0, \$1.5), and (\$0.5, \$0.5). In

[8] In addition to studies that compare the SM to the DR, there are a few studies that compare the SM to simultaneous play in prisoners' dilemma and public goods games, with the primary purpose of investigating how beliefs affect behavior. Note that simultaneous play is different from the DR in sequential games in that subjects move simultaneously. Columbus and Böhm (2021) compare the SM and simultaneous play in a continuous Prisoner's Dilemma with belief elicitation, finding higher cooperation under simultaneous play, driven primarily by low-trust players. Fischbacher et al. (2001) let subjects play a public goods game both conditionally (SM) and unconditionally (simultaneous play) in order to identify conditional cooperators; follow-up work by Martinsson et al. (2013) adopts a similar approach to examine conditional cooperation in developing countries. Finally, meta-analyses by Karakostas et al. (2022) suggest that employing the SM can sometimes inflate destructive behaviors, implying that the SM may bias destruction rates upward compared to simultaneous play.

[9] Chen, Zhuoqiong and Marcus Roel. 2019. "Strategy vs. direct-response Method." AEA RCT Registry. September 20. https://doi.org/10.1257/rct.4737.

the *mUG*, the proposer could either split $2 equally or according to 85-15 ($1.7, $0.3). Rejection resulted in a payoff of 0 for both parties. See Figure 1 for the respective payoff matrices.

1\2	cooperate	defect
cooperate	1, 1	0, 1.5
defect	1.5, 0	0.5, 0.5

1\2	accept	reject
offers 50–50	1, 1	0, 0
offers 85–15	1.7, 0.3	0, 0

Figure 1: Payoffs for the sequential Prisoner's Dilemma (left) and mini Ultimatum Game (right).

The games were presented in a neutral frame, referred to as task 1 or 2, with their order randomized. The only discernible difference between the games was their payoffs, which were presented in a matrix-like format using actual $-values.[10] Player 1 (he), referred to as the first mover in the experiment, chose between actions $\{A, B\}$ and player 2 (she), referred to as the second mover, chose between actions $\{C, D\}$.

We chose the two games for several reasons. Both games are simple and well-suited for large-scale online experiments, where attention spans may be limited and interruptions frequent. They are also commonly used in the literature to study social preferences such as fairness and reciprocity, aligning with our focus on how different elicitation methods affect these preferences. Previous literature has also used these games to compare the strategy and direct-response methods and found mixed results (see Section 1.1 for more detail).

2.1. Treatments

We implemented a 2×2 between-subject design, in which we manipulated both the beliefs of participants, by providing them with behavioral data from past experiments and how the second mover' choices were elicited.

Belief-Manipulation. Participants were randomly assigned to the *selfish* or *non-selfish beliefs* treatment, which determined what information was presented to a player for a given game. In particular, a short sentence that described player 1's behavior in a past study was displayed in addition to their particular payoff matrix and their particular role. For the *sPD* in the non-selfish belief treatment, the sentence read: *"In a well-known study of this task by Watabe, Terai, Hayashi, and Yamagishi, published in the year 1996, 82.6% of the first movers chose **A**."* For the selfish belief treatment, participants were informed that *"In a well-known study of this task by Bolle and Ockenfels, published in the year 1990, 82.7% of the first movers chose **B**."*[11]

In order to keep the belief treatments as similar as possible, and to prevent potential confusion, we emphasized the more likely action in the sentence instead of highlighting the same action. In the *sPD*, action **A** represents cooperation whereas **B** stands for defection. In other words, players in the *non-selfish* belief manipulation were informed that the action of player 1 that improves player 2's (set of) payoffs is the more commonly chosen one, and vice versa in the *selfish* belief treatment.[12] For the non-selfish belief treatment in the *mUG*, we cited Güth et al. (2001), where 70.6% offer an equal split, using the same format as in the *sPD*.

[10] Many of our design choices were aimed towards making games as easy and as quickly to understand as possible in view of the online nature of our experiment. One implication of this design principle was using real currency values in the games over some fake experimental-currency with some specific exchange rate. This also guided our thinking in how to present payoffs. We opted for a payoff-table, where the subject was always the "row player", indicated with *You*, and the other player was the column player, indicated with *Other Participant*. The participant's payoffs were preceded with "*You earn:*", whereas the other player's payoff were preceded with "*Other earns:*", for each element in the table.

[11] Due to an unfortunate Excel mistake when calculating fractions, this number differs marginally from the 85.2% of the actual study, which reported that 9 players cooperated and 52 defected in the role of player 1. We sincerely apologize for this.

[12] Throughout this paper, we use the terms *selfish* and *non-selfish* beliefs treatments mainly for the sake of expositional convenience. Whether player 1's action is selfish or non-selfish from his perspective generally depends on player 2's response.

We selected these papers based on two criteria: (1) a similar payoff structure as our experiment in order to provide accurate information to our participants, and (2) similar frequencies of the two opposing actions.[13] Given these restrictions, we unfortunately did not find a suitable data point for the *mUG* and the selfish belief treatment. The closest study in this regard involved Chimpanzees (Jensen et al. (2007)), who documented a 75% offer rate of unequal splits. Such a study is unsuitable for an online experiment, however, where subjects can look up data provided to them – even if such data were predictive of human behavior. We had used this study in a small classroom pilot that tested how our belief manipulation affects beliefs. Among first-year economics and business majors, who were unfamiliar with the games, 80% opted for unequal splits.[14] Instead of citing any alternative study with more balanced offer rates, we truthfully told participants in the selfish belief treatment in the *mUG* *"In our previous experiment of this task, 80% of the first movers chose **B**."* We note that this belief manipulation satisfies our previous criteria and involves no deception.[15] The only remaining concern is thus whether it is as powerful as citing an existing published study. In the next section, we will see that it is.

Behavior-Elicitation. Participants were randomly assigned to the *direct-response* or *strategy method* treatment at a rate of 3-to-1. We use a larger sample for the direct-response treatment, in which we only observe player 2's response to player 1's chosen action, in order to balance the observations of player 2 at each node.[16] While the choice of elicitation method makes no difference for how player 1's choices are elicited, all participants were informed in the common set of instructions on how choices are elicited for player 1 and player 2. Indeed, our goal was to create common knowledge among players regarding the game played, including the fact that they will receive the same information as the other person they are matched with.

2.2. Experimental Setting and other Procedures

The experiment was conducted online, with participants recruited from Amazon's online job platform mechanical turk (MTurk). In order to be eligible to participate in our experiment, MTurk workers had to have completed at least 100 jobs, possess an approval rate of at least 99%, and be located in the USA or Canada.[17] In the job description on MTurk, we invited subjects to take part in a large online experiment, informed potential participants that they would be paid a participation fee of $1 upon successful completion, with a possible additional payment of up to $3, and highlighted that the task can be finished within 10 minutes.[18]

After accepting the job on MTurk, interested participants were redirected to our experiment on an external website, created with oTree (Chen et al., 2016), that provided more general details about the job's nature, the experiment itself, and elicited consent. At this point in time, workers who were uninterested to

[13] The payoff used in Watabe et al. (1996) were (10, 10), (0, 15), (15, 0) and (5, 5), in Bolle and Ockenfels (1990) (50, 50), (0, 75), (75, 0), and (10, 10). In Güth et al. (2001) 20 units were either equally split or in (17, 3).

[14] 80% offered unequal splits in the selfish-belief treatment and 79% did so in the non-selfish belief treatment.

[15] The observed variation in cooperation rates and fair offers across these studies is likely driven by cultural differences (German vs. Japanese or Chinese subjects), framing effects (neutral vs. competitive market framing), sample size, and presentation format of games (extensive form vs. strategic form). These factors influence participants' expectations, strategic reasoning, and perceived norms and incentives, leading to different behavioral patterns across treatments.

[16] If player 1 chooses each action with probability 0.5, a randomization of 66.6% to 33.3% results in the same number of observations for player 2 at each node. Ex-ante, we did not expect player 1's behavior to be balanced, which led us to use a randomization of 75%/25% in order to increase the power in the less likely node.

[17] Such quality requirements are typical to ensure attentive and high-quality responses. A good resource in this regard is Hauser et al. (2019), who outline common concerns regarding MTurk experiments, provide empirical evidence about them, and offer practical solutions on how to run experiments. Note, that ineligible MTurk workers cannot see or access the job-ad itself.

[18] We set payoffs and participation fees so that most people earn a good hourly-wage, expecting most participants to finish the experiment within 10 minutes or less. If anything, we expected our round payoffs to be too generous, but preferred them over scaling payments down. Payments varied across subjects but were reasonable overall: the hourly wage was $14.79 and $21.91 at the 10th and 25th percentile, and $31.91 for the median earner. 90% of participants finished within 10 minutes. Note that these hourly rates overestimate pay as they only account for the time participants spent on our website.

participate in our experiment could quit the website and return the job on the MTurk platform.[19] Those who chose to continue were provided with a detailed instruction about the nature of the games played, including how choices are elicited. After completing a short test that checked whether they understood the instructions, they proceeded to the play stage. After both games were played, we elicited their beliefs about player 1's behavior while reminding them about the particular games.[20] The experiment concluded with a short survey that asked participants about their gender, age, degree, household income, as well as whether they have participated in similar experiments before.

Our sessions were conducted in two stages, from October to November 2019, and again in October 2021, with up to 500 participants per session.[21] We will refer to data from the first time frame as the *pre-Covid* sample and from the second as the *Covid* sample. In order to ensure an ideal user-experience with no wait-times, we ran player 1 and player 2 separately and matched them afterwards to determine their payoffs.[22] When paying subjects after the experiment, we explicitly informed them of their opponents' choices and whether their incentivised beliefs resulted in additional payments. We also reminded them about their role, choices, and provided them with a link to their particular payoff tables.[23]

3. Results

In this section, we present our main findings. The definitions and descriptions of our variables can be found in Table A.1. Summary statistics are provided in Table A.2. In our regressions, we include the personal information elicited by the post-experiment survey as categorical control variables. We also control for the subjects' mistakes in the instruction test, the task order, and whether they are part of the Covid sample. Finally, we conducted a randomization analysis (Table A.3) that shows that the randomization was a success.[24]

Our main analysis, Table 3, 4, 5 and 6 (and the accompanying graphs), follows the pre-registered analysis

[19] MTurk workers can only accept and work on a single task at any given time. Hence, if they are uninterested in participating in our particular experiment, e.g., because they don't want to interact with other MTurk workers, they can return the un-completed task without any repercussions. Instead of returning a task themselves, they may (choose to) let it time out, in which case the task is returned after a pre-determined time. It is important to know that MTurk only displays the job-ad to potential workers if there are outstanding jobs (the total number is set upon posting the ad). When a job is accepted, the number of outstanding jobs is reduced by one. It is increased by one if an un-completed task is returned.

[20] Belief elicitation was incentivized, with an additional $0.25 paid for beliefs within 5 percentage points of the correct answer. We opted to elicit beliefs after the play stage to keep the play stage simple and comparable to the usual environment where beliefs are typically not elicited before play. The downside to this approach is that the second mover's elicited beliefs in the direct-response method can be (and are) influenced by their particular experience. As the belief measure is not a primary outcome measure of interest, we viewed this to be a sensible tradeoff to make.

[21] In our Pre-Analysis Plan, we had committed to a sample of 4000 in case of no order-effects and 8000 otherwise. The seemingly large sample size was determined following a power analysis. It takes into account that our primary research question focuses (i) on the behavior of the second-mover and (ii) the estimation on the difference-in-difference effect between the elicitation methods and the belief treatments. A preliminary check of the data, however, revealed the existence of simple order-effects, and so we continued the data collection. Indeed, order-effects are still present in our full data. However, our main conclusion is robust. As a result, we report results from both tasks jointly, noting that the task order is a control variable in all regressions. For a detailed discussion of the order effects, please consult the Online Appendix. By November 2019 we unfortunately ran into technical issues due to the fact that Amazon had ceased operations in China and hence no longer officially accepted payments from China. We resumed the experiment in the fall of 2021 (with the delay due to Covid-19), cooperating with a large British University to facilitate payments. Except for the change in the official MTurk account, all details of the procedures and implementation remained the same.

[22] For a given batch of players, we first ran the experiment for player 1. We then used their actual behavior in each treatment arm/task order to determine the action that player 2 was informed of. Workers were never able to retake the experiment. The nature of running the experiment on MTurk typically leads to samples that are not perfectly balanced, e.g., there are a total of 4009 first and 4020 second movers.

[23] In the program we used to match Player 1 and Player 2, we mistakenly applied an 80-20 split (USD 1.60, USD 0.40) for payoffs in the mUG, rather than the 85-15 split (USD 1.70, USD 0.30) that was shown to participants. This error affected the calculation of mTurk payments made a few days after the experiment. We apologize to our participants for this oversight.

[24] In particular, we ran pairwise (across four treatments) Kolmogorov-Smirnov tests of equality of distributions for the end-of-experiment questionnaire responses. We find no significant differences at conventional levels for all but one out of 36 tests.

plan.[25] In order to provide further insight into these results and our interpretation thereof, we explore mistakes in more detail (some of which were suggested by reviewers at various stages of this project), framed the analysis using a basic selfish and social theory, and classified player 2 types. These additional analyses are presented in Figure 3 and Table 7-11. We also include a heterogeneity analysis (section 3.5) upon request from reviewers. None of these analyses were prespecified.

For our main results, we require a significance level of 5% or lower. For other, more supportive analyses, we will also report and discuss those estimates that are significant at the less demanding 10% level. We believe this approach strikes a balance between ensuring the replicability of our main results and gaining additional insights from further exploratory work, which are often based on smaller samples. Note that the respective table notes always specify the significance levels reported.

Before presenting our findings, we outline two classical theories that will help us categorize player 2's behavior. According to the *selfish theory*, players choose those actions that maximize their material payoffs. For our games, the selfish theory predicts that player 2 always defects in the *sPD* and always accepts any offer in the *mUG*. The alternative explanation we consider is a *social theory*, according to which players are motivated by reciprocity, i.e., they want to reward nice and punish nasty behavior. In particular, player 2 cooperates after cooperation but defects after defection in the *sPD*, accepts the fair offer (50-50) but rejects unfair splits of 85-15. In line with these theories, we will refer to players who act according to the selfish theory (social theory) prediction as selfish/selfish-types (social/social-types).[26]

Table 1: Predictions for Player 2 in the Selfish and Social Theories

seq. Prisoners' Dilemma	after P1 cooperates	after P1 defects
Selfish theory	defect	defect
Social theory	cooperate	defect
mini Ultimatum Game	after P1 offers 50-50	after P1 offers 85-15
Selfish theory	accept	accept
Social theory	accept	reject

Table 1 summarizes these predictions and illustrates two key observations that will guide our data analysis in this section and beyond. First, in both games, player 2's behavior after one particular action of player 1 can be used to classify her as either selfish or socially motivated. In the *sPD*, in response to cooperation, cooperation is only consistent with the social theory whereas defection is only consistent with the selfish theory. In the *mUG*, player 2's acceptance of the unequal offer is predicted by the selfish theory while the social theory predicts subsequent rejection. Second, after player 1's other action, there is an action for player 2 that can be explained by neither theory, namely cooperation after defection and rejecting the fair offer. For a selfish player, such behavior is clearly a *mistake* as it results in lower payoffs. It can similarly be viewed as a mistake for a socially minded player, since it would reward player 1 for choosing the action that results in

[25] Chen, Zhuoqiong and Marcus Roel. 2019. "Strategy vs. direct-response Method." AEA RCT Registry. September 20. https://doi.org/10.1257/rct.4737. In summary, the analysis plan stipulated that we test whether the belief manipulation affects participants' beliefs, and whether these beliefs differ from the provided information or actual behavior. It also outlined that we compare the behaviors of player 1 and player 2 across the elicitation methods, belief treatments, and the interactions of the two at each node of the game. Our primary hypothesis, as detailed in the introduction and the theory in the Online Appendix, is that the impact of the belief manipulation on player 2's response to cooperation or unfair offers varies across the elicitation methods. Every pre-specified analysis was carried out, with less relevant ones relegated to the Appendix (Table A.5), the Online Appendix (order-effects), or to a short footnote (e.g., which action maximizes player 1's monetary payoffs).

[26] We use those two theories to classify behavior in the 'as if' sense. In reality, it is often not one or the other, but rather, a question of how much players are motivated by these different notions. Note also that the selfish theory can easily be rejected when non-selfish choices are observed. It is much more difficult, however, to reject social preferences when only selfish choices are taken as social preferences may only induce non-selfish behavior if the social motivation is sufficiently strong and/or the material cost of the non-payoff maximizing action is not too prohibitive, all of which depends on the game played.

a strictly lower set of feasible payoffs in the *sPD* and punish the strictly more generous offer in the *mUG*. In the rest of the paper, we refer to these two actions as *mistakes*. Before we continue, we do want to recognize however that there are theories that predict what we view as mistakes as normal behavior. For example, preferences for efficiency lead player 2 to always cooperate in the *sPD* (Charness and Rabin, 2002) while spiteful preferences (Levine, 1998) may result in the rejection of all offers in the *mUG*. Indeed, we would have a difficult time arguing that all such behaviors represent genuine mistakes in our data. However, it is a useful notion to categorize this type of behavior and we will show how elicitation methods may help players to either avoid or facilitate them. In section 3.4, we will also provide both direct and indirect evidence that this behavior is, at least partially, driven by mistakes.

3.1. Overall Behavior

We begin by describing the overall frequency of cooperation, fair offers, and rejections, which are summarized in Table 2. The majority of first movers cooperate in the *sPD* (57%) and offer 50-50 in the *mUG* (66%), which indicates that a large fraction of player 1 expects player 2 to act in a non-selfish manner. On average, the behavior of player 2 in the *sPD* is more consistent with the reciprocal behavior suggested by the social theory: 65% cooperate in response to cooperation yet only 18% cooperate after defection. For the *mUG*, player 2's behavior is largely consistent with the selfish theory as we observe very few rejections, although most of which occur in response to unequal offers as predicted by the social theory. In light of the high rates of cooperation after cooperation in the *sPD*, we find the low rejection rates of the unequal offer in the *mUG* to be a surprising data point.[27] The distribution of the two player types by games and elicitation methods is provided in Table A.4 of the Appendix.

Table 2: Overall Behavior in sPD and mUG

seq. Prisoner's Dilemma	Freq.	*mini Ultimatum Game*	Freq.
Player 1 cooperates	0.57	Player 1 offers 50-50	0.66
Player 2 cooperates after P1 cooperates	0.65	Player 2 rejects 85-15	0.13
Player 2 cooperates after P1 defects	0.18	Player 2 rejects 50-50	0.03

Notes: the total number of participants is 8029.

3.2. Beliefs and Player 1's Choices

We now turn to whether the belief manipulation influenced beliefs and how it affected player 1's behavior. Part 1 of Table 3 tabulates the data from previous studies that were used for our belief manipulations, part 2 summarizes the elicited belief about player 1's behavior from our participants, while player 1's actual behavior is shown in part 3.

The belief manipulation strongly influenced players' beliefs. In the selfish belief treatment, on average, subjects believe that 34% (33%) of player 1 cooperate (offer 50-50), whereas they expect 74% (74%) to cooperate (offer 50-50) in the non-selfish belief treatment. These beliefs are significantly different from the respective data that subjects were provided with (t-tests: $p < 0.01$), and, with the exception of the non-selfish belief treatment in the *mUG*, tend to be less extreme than the provided data itself.[28]

[27] Note also that rejecting the unequal offer only costs \$0.3, while rewarding the first mover for cooperating costs \$0.5. In addition, the unequal offer was also very skewed at 85 to 15.

[28] Belief histograms for both treatments can be found in Figure A.2 & A.3 in the Appendix. A regression analysis of beliefs can be found in Table A.5. Regarding the regression estimates, it is not surprising that beliefs differ between the two elicitation methods for they are elicited at the end of the experiment following different player experiences.

Table 3: Behavior of and Beliefs about Player 1

	Selfish	Non-Selfish
1. Provided data (from previous studies)		
Player 1 cooperates	0.173	0.826
Player 1 offers 50-50	0.20	0.706
2. Elicited Beliefs		
Belief Player 1 cooperates	0.34	0.74^{***}
Belief Player 1 offers 50-50	0.33	0.74^{***}
3. Player 1's Behavior		
Player 1 cooperates	0.47	0.67^{***}
Player 1 offers 50-50	0.59	0.72^{***}

Notes: statistically significant differences between the belief-treatments (based on t-tests) for part (2) and (3) at significance levels of ** $p < 0.05$, *** $p < 0.01$ are indicated by the respective stars in column (2).

The belief manipulation had a similar effect on player 1's actual behavior, leading to significantly more cooperation and equal offers in the non-selfish belief group compared to the selfish belief treatment. This effect could, for example, be due to indirect learning about player 2's behavior, norms, or experimenter demand effects.[29] Relative to beliefs, player 1's behavior is less responsive to the belief manipulation. Finally, we observe significantly more cooperation (fair offers) in the selfish belief treatment than what players expected (t-tests: $p < 0.01$) but significantly less than expected in the non-selfish belief group (*sPD*: $p < 0.01$, *mUG*: $p = 0.06$).

Figure 2 graphs player 1's behavior for all treatments, showing in detail how the elicitation method impacts their behavior. For the selfish belief group, the average rate of cooperation and fair offers tend to be higher when choices are elicited using the direct-response rather than the strategy method; for the non-selfish belief group, the pattern is reversed. Table 4 shows that this pattern is statistically significant in the *sPD*. It reports the estimates from an OLS regression for player 1's behavior, i.e., whether he cooperates or makes the fair offer, on the non-selfish belief treatment dummy, the strategy method treatment dummy, as well as the interaction term of both dummies. As a result, the baseline is the direct-response, selfish belief treatment. In the *sPD*, we observe a relatively larger increase in cooperation in the strategy method as we move from selfish to non-selfish belief, indicated by the positive interaction term. For selfish belief, there is significantly less cooperation when the strategy method is employed. For the *mUG*, these effects are qualitatively similar but not statistically significant. In the appendix, Table A.6, we also repeat these regressions directly using the four treatment dummies (instead of an interaction term) and report all respective treatment differences.

The differential effect of the elicitation method on player 1's behavior is interesting in the sense that player 1's choice is elicited in the same way regardless of the elicitation method. However, player 1 is cognizant of how choices are elicited since experimental instructions were independent of role assignment, which only occurred thereafter. One reason for this could be that the elicitation method itself affects how player 1 approaches the game (fixing player 2's behavior). Qualitatively, player 1's behavior is consistent with a larger degree of strategic thinking in the strategy method. After all, a selfish-belief signal is not only indicative of player 1's behavior, but also suggestive of a more selfish response by player 2, and vice versa for

[29] It was neither the aim of our study to differentiate between different causes for such behavioral change nor to explore player 1's preferences. Instead, the goal was to leverage the belief manipulation to understand how player 2's behavior depends on the elicitation method.

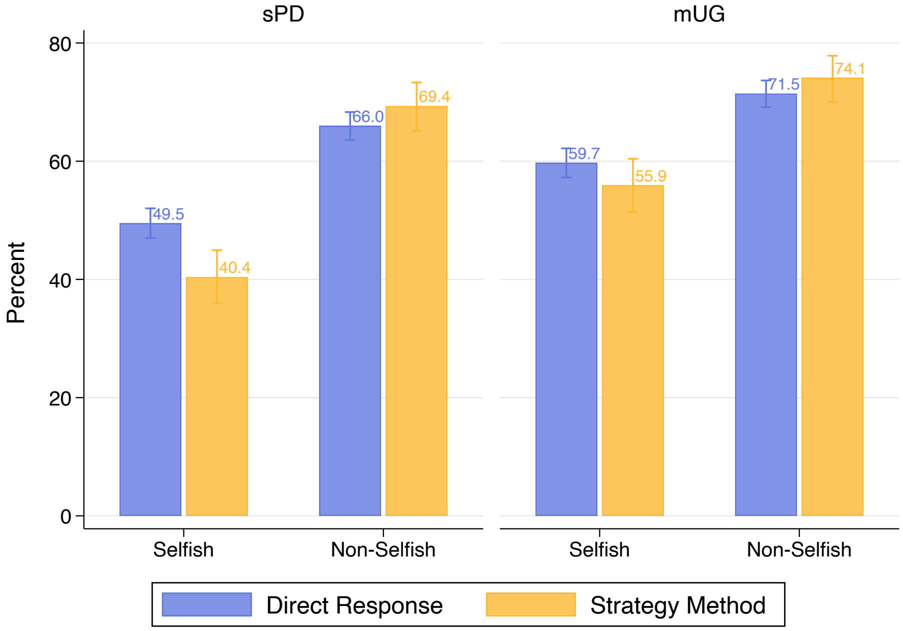


Figure 2: Player 1's Behavior (Frequency of Cooperation or 50/50 Offers)

Notes: the number on each bar represents the mean and the error bar represents the 95% confidence interval.

Table 4: Player 1's Behavior

	sPD		mUG	
Dep. Var: P1 cooperates; offers 50-50	(1)	(2)	(3)	(4)
Non-Selfish Belief	0.165***	0.166***	0.117***	0.115***
	(0.0176)	(0.0175)	(0.0170)	(0.0169)
Strategy Method	-0.0912***	-0.0907***	-0.0380	-0.0317
	(0.0262)	(0.0263)	(0.0263)	(0.0262)
Non-Selfish × Strategy Method	0.125***	0.123***	0.0646	0.0663
	(0.0357)	(0.0357)	(0.0349)	(0.0347)
Controls	No	Yes	No	Yes
Observations	4009	4009	4009	4009

Notes: this table reports estimates from OLS regressions. Control variables for individual characteristics include gender, age, income, highest education, prior participation in experiments, mistakes in instruction test, task order, and the Covid sample dummy. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. ** and *** indicate statistical significance at the 5% and 1% levels.

the non-selfish belief signal. For the strategy method, player 1 appears more responsive to these signals.[30] An alternative explanation would be that player 1 believes that player 2's behavior is directly influenced by the elicitation method, in turn influencing player 1's preferred choice.

Since we did not elicit player 1's belief about player 2's choices, we cannot test this directly with our

[30] In terms of material (expected) payoffs, we find that they are often fairly similar (within $0.1) in the *sPD* based on behavior by-treatment-and-sample (Table E.52). Pre-Covid, cooperation (defection) is payoff maximizing for the strategy method (direct-response) regardless of belief treatment. The same pattern is true for the overall sample. In the Covid sample, defection is optimal except in the selfish belief, strategy method treatment. For the *mUG*, the unfair offer does significantly better given the few rejections. As we do not elicit player 1's belief about player 2, we cannot test whether they act as if they maximize their own payoffs.

data. However, response time data, which measures the time (in seconds) it takes participants to make their choice in each game, offers indirect evidence in favor of the strategic-thinking explanation, as shown in Figure 3 – assuming that longer response time indicates deeper strategic thinking. The pattern is quite striking: it is exactly where we observe much lower cooperation rates in the sPD – for selfish beliefs in the strategy method compared to the direct-response method – that we detect longer response times. The difference is significant at 10% (Table A.7). Moreover, consistent with this observation, there is a strong negative correlation between response time and player 1's likelihood of cooperating in the *sPD* (Table A.8).

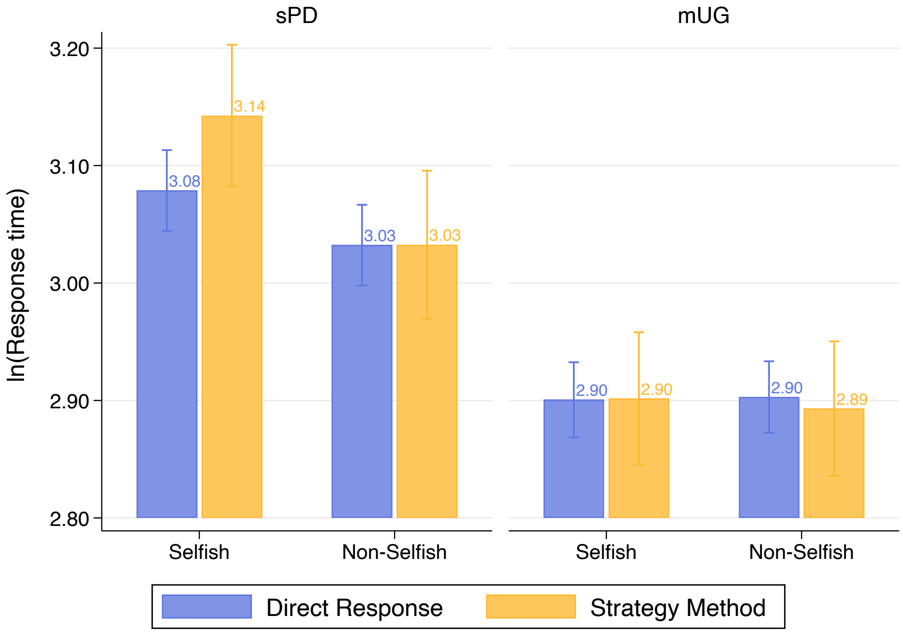


Figure 3: Player 1's Response Time

Notes: the number on each bar represents the mean and the error bar represents the 95% confidence interval.

3.3. Player 2

We now turn to the main focus of this paper: player 2's behavior. Player 2's choices in the *sPD* are depicted in Figure 4. In general, player 2's behavior is more in line with the social than selfish theory: after player 1 cooperates, over 60% of player 2s in all combinations of elicitation methods and belief groups cooperate. In response to defection, more than 75% in all groups defect.

Overall, behavior in response to player 1 cooperating is remarkably similar across our four treatment groups. Indeed, the only significant difference between any of the groups is between the direct-response treatment with selfish and non-selfish belief ($p < 0.05$ when controls are included), where we observe a small uptick in cooperation with non-selfish belief, which can be explained by social norms (among others).[31] In contrast to the economically large response of player 1 to the belief manipulation, the change is slight.[32] Table 5 reports our respective OLS-estimates for this setting. The elicitation method does not appear to have any effect on player 2's response to cooperation.

[31] This positive effect, in turn, also suggests that positive reciprocity due to surprise (Khalmetski et al. (2015)) may not be an important driver of behavior in our sample.

[32] The regression table that reports treatment differences between groups can be found in the Appendix, Table A.9.

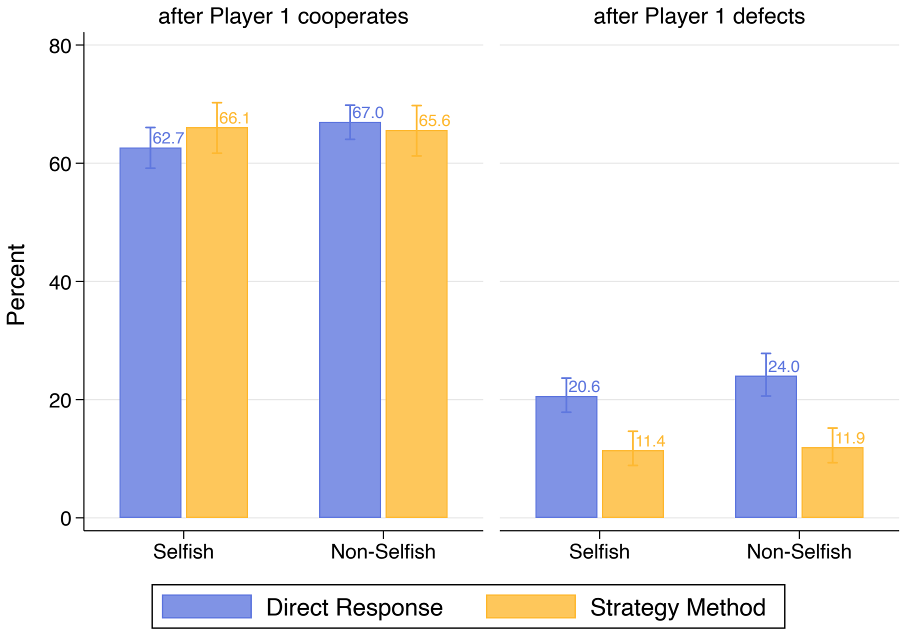


Figure 4: Player 2's Behavior in sPD (Frequency of Cooperation)

Notes: the number on each bar represents the mean and the error bar represents the 95% confidence interval.

Table 5: Player 2's Behavior in sPD

	after P1 cooperates		after P1 defects	
Dep. Var: Player 2 cooperates	(1)	(2)	(3)	(4)
Non-Selfish Belief	0.0433	0.0472**	0.0342	0.0302
	(0.0230)	(0.0226)	(0.0235)	(0.0232)
Strategy Method	0.0344	0.0401	-0.0917***	-0.110***
	(0.0280)	(0.0279)	(0.0208)	(0.0212)
Non-Selfish × Strategy Method	-0.0481	-0.0580	-0.0291	-0.0173
	(0.0384)	(0.0383)	(0.0315)	(0.0315)
Controls	No	Yes	No	Yes
Observations	2722	2722	2247	2247

Notes: this table reports estimates from OLS regressions, with control variables identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. ** and *** indicate statistical significance at the 5% and 1% levels.

In contrast, the elicitation method strongly affects player 2's response to player 1 defecting: the strategy method cuts the rate of cooperation in half; subjects make significantly less mistakes if the strategy method is used. The best explanation for this effect is that the strategy method forces player 2 to pay more attention when they are required to make two choices, reducing (random) mistakes. For example, they are less likely to look at the incorrect node of the game/payoff column or think that player 1 took a different action.[33]

[33] The alternative hypothesis that player 2 simply understands the game better when she is forced to look at both nodes of the game cannot explain this behavior, however. While in isolation, this may explain why there is less cooperation in response to defection as player 2 realizes that player 1 took the worst action for her, she would also have to come to the opposite conclusion after cooperation. This in turn would predict more conditional cooperation in the strategy method, which is not true in the data.

It is important to emphasize that this reduction in mistakes cannot be explained by simple alternative theories that view cooperation in response to defection as true reflections of preferences. For example, while preferences for efficiency predict cooperation after defection, they cannot explain why these preferences vary with the elicitation method. Moreover, given that we do not detect any effect of the elicitation method *after cooperation*, i.e., the part of the game where we would expect and indeed observe "positive" social preferences to operate, it is difficult to conceive why there should be such effect on preferences after defection. Finally, we do not observe any difference in behavior after defection due to different beliefs.

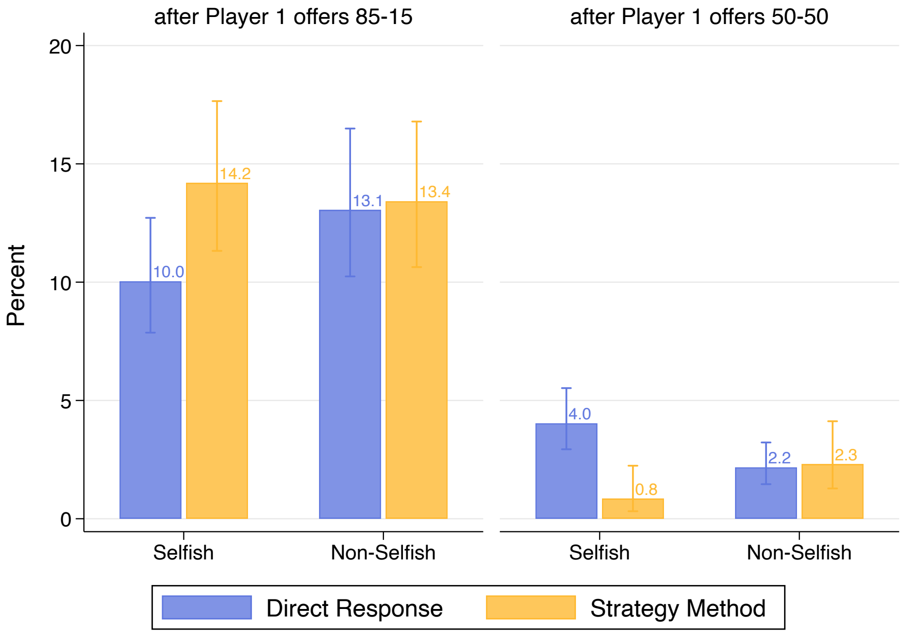


Figure 5: Player 2's Behavior in mUG (Frequency of Rejection)

Notes: the number on each bar represents the mean and the error bar represents the 95% confidence interval.

Figure 5 depicts player 2's behavior in the *mUG*. Overall, rejection rates are low, below 15% in all treatment groups. Consequently, behavior is well predicted by the selfish theory. What the social theory gets correct, is the relative behavior across the two nodes: rejection rates of unequal offers are no less than 10% while they are no higher than 4% for equal splits. Before we look at the treatment effects in detail, it is worth emphasizing that in contrast to the *sPD*, differences between treatment groups are very small. Consequently, the economic significance of the elicitation method is minor when it comes to behavior.

From the left-hand panel of Figure 5, we see that the selfish belief, direct-response group features the least rejections of unequal splits.[34] [35] Consistent with this, our regression estimates in Table 6 indicate that the strategy method (dummy) results in more rejections. Like in the *sPD*, there is no evidence for a differential effect on behavior as the interaction term remains insignificant. Moreover, beliefs do not affect rejections of unfair offers.

With regard to player 2's behavior after equal splits, the picture is more complicated than in the *sPD*. When choices are elicited using the strategy method, player 2 is significantly less likely to reject offers in

[34] The difference is significant at $p < 0.05$ compared to the strategy method with selfish belief. For a comparison between all treatment groups, consult Table A.10 in the Appendix.

[35] This pattern is clearly inconsistent with visceral reactions such as anger (Loewenstein (1996)) that arise (more strongly) when directly observing player 1's selfish action.

Table 6: Player 2's Behavior in mUG

	after P1 offers 85-15		after P1 offers 50-50	
Dep. Var: Player 2 rejects	(1)	(2)	(3)	(4)
Non-Selfish Belief	0.0302	0.0277	-0.0186**	-0.0195**
	(0.0201)	(0.0202)	(0.00785)	(0.00773)
Strategy Method	0.0416**	0.0410**	-0.0319***	-0.0363***
	(0.0202)	(0.0201)	(0.00775)	(0.00810)
Non-Selfish × Strategy Method	-0.0380	-0.0399	0.0332***	0.0362***
	(0.0301)	(0.0300)	(0.0113)	(0.0116)
Controls	No	Yes	No	Yes
Observations	1999	1999	2970	2970

Notes: this table reports estimates from OLS regressions, with control variables identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. ** and *** indicate statistical significance at the 5% and 1% levels.

the selfish-belief manipulation compared to the direct-response method. This observation is consistent with our previous argument that the strategy method results in less mistakes, as rejection of equal offers can neither be explained by selfish preferences nor typical social preferences. Interestingly, we do not observe this pattern for non-selfish belief groups, where rejection rates are essentially identical for the two elicitation methods.[36]

Across both games, player 2's behavior displays an interesting similar pattern: when the belief manipulation points towards the other action, we observe slightly more mistakes in the direct-response method. In the *sPD*, we observe more mistakes after defection in the direct-response method when the belief manipulation points towards cooperation (i.e., the non-selfish belief). In the *mUG*, player 2 rejects more equal offers in the direct-response method when the belief manipulation points towards the unequal offer (i.e., the selfish belief). Note that this effect occurs on top of the general increase in mistakes for the direct-response method relative to the strategy method. One reason for the higher rates of mistake could be that player 2 does not solely respond to the hypothetical choice of player 1 that is shown to her, but instead is influenced by her subjective (ex-ante) belief of what player 1 is going to do, leading to confusion regarding which node she is at. Alternatively, it is not her belief that leads to such confusion but the information that is provided by the belief treatment, i.e., she may confuse the displayed choice of player 1 with the action emphasized by the information treatment. In other words, it could be that our particular experimental setup is partially responsible for inducing more mistakes by highlighting two, possibly different, choices of player 1 in the direct-response method. Table 7 below lends credence to the first interpretation, alleviating such concerns. It shows that the belief treatment itself is not correlated to the rate of mistakes but, if anything, that the belief in the alternative node is associated with more mistakes for players in the direct-response method.

3.4. Mistakes or Preferences?

So far, we observed how the strategy method leads to less cooperation after defection and rejections of fair offers. Guided by the theory of either selfish or reciprocal preferences, we classified such choices as mistakes and concluded that their frequency was reduced by the strategy method. We now provide several pieces of

[36] For those interested, we also report the frequencies with which all game end-nodes are reached by treatments, Table E.50, and by elicitation methods, Table E.54. These combine player 2's behavior with player 1's action, which, as one may recall, was also affected by the elicitation method in the *sPD*. In particular, player 1 appeared to be more cautious in the strategy method under selfish-beliefs, exhibiting less cooperative behavior. Nevertheless, the total frequency of outcomes with mistakes, (D, c), is about 4 percentage points lower in the strategy method.

Table 7: Player 2's Behavior: Mistakes and Beliefs

	sPD: after P1 defects		mUG: after P1 offers 50-50	
Dep. Var: Player 2 makes mistake	(1)	(2)	(3)	(4)
Non-Selfish Belief	0.00367	-0.00180	-0.00389	-0.00598
	(0.0288)	(0.0281)	(0.00755)	(0.00756)
Belief Player 1 cooperates	0.0798*	0.0757*		
	(0.0422)	(0.0413)		
Belief Player 1 offers 85-15			0.0399***	0.0372***
			(0.0131)	(0.0130)
Controls	No	Yes	No	Yes
Observations	1298	1298	2021	2021

Notes: this table reports estimates from OLS regressions, with control variables identical to those in Table 4, for player 2s in the direct-response method. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

evidence in support of this idea. Together, these suggest that what we observe in the data is unlikely to be a result of differences in genuine preferences for cooperation after defection and rejections of fair offers across the two elicitation methods.

Taking seriously the idea that the strategy method reduces the likelihood of actual mistakes, we would expect players who are more prone to mistakes be more strongly affected by the strategy method than those who are not. To test this hypothesis, we can leverage our experimental design, where after the introduction of our experiment and all related rules, subjects were required to answer a series of control questions. We expect that those who make mistakes in the questionnaire are more likely to be inattentive and/or generally care less about (their performance in) the experiment. Consequently, we expect them to make (i) more mistakes when playing the games and (ii) that the strategy method is particularly useful for mitigating their mistakes. We refer to this type of participants as *inattentive* and test our hypotheses by regressing player 2's choice, i.e., whether they make a mistake in the *sPD* or *mUG*, on the strategy method dummy, whether the player is considered inattentive, and the interaction of the two.

The results are reported in Table 8. The positive and significant coefficient of the *inattentive* dummy in column (1) and (3) supports conjecture (i) that inattentive people make more mistakes in general. By itself, this provides direct support for the viewpoint that cooperation after defection and rejecting an equal offer are simple mistakes since it is unlikely that such preferences are correlated with making more mistakes in a control questionnaire. With regards to hypothesis (ii), we observe that the interaction term in column (1) and (3) are negative and significant. In other words, the strategy method reduces the likelihood of making a mistake relatively more when the player is inattentive, confirming our prediction.[37] It is also important to highlight that the estimate for the strategy method dummy itself in column (1) remains strongly negative for the *sPD*, the game where this effect was most pronounced. This suggests that even for the baseline, here, attentive participants who make no mistakes in the questionnaire, the strategy method still helps in limiting mistakes.

In column (2) and (4) we expand our analysis by considering where subjects made a mistake in the questionnaire. Question 1, 2, and 3 test simple procedural rules (player anonymity, player roles across tasks, who they interact with in each task) whereas question 4 evaluates their understanding of the payoff

[37] When presenting this table, we opted to omit the belief-treatment dummy as it greatly improves the presentation of the interactions, and allows us to focus on the main question. The pattern remains true for more involved regressions. To keep the table short, we also do not report estimates without controls. For completeness, estimates of Table 8 without controls can be found in the Online Appendix, Table D.21. The results are unaffected by the inclusion/exclusion of such controls.

Table 8: Player 2's Behavior: Mistakes and Inattention

	sPD: after P1 defects		mUG: after P1 offers 50-50	
Dep. Var: Player 2 makes mistake	(1)	(2)	(3)	(4)
Strategy Method	-0.0772***	-0.0773***	-0.00625	-0.00624
	(0.0179)	(0.0179)	(0.00541)	(0.00540)
Inattentive	0.135***		0.0451***	
	(0.0270)		(0.0101)	
Inattentive × Strategy Method	-0.118***		-0.0319**	
	(0.0351)		(0.0140)	
Inattentive (Q123)		0.110***		0.0343***
		(0.0285)		(0.0102)
Inattentive (Q4)		0.260***		0.0970***
		(0.0647)		(0.0320)
Inattentive (Q123) × Strategy Method		-0.119***		-0.0254
		(0.0387)		(0.0156)
Inattentive (Q4) × Strategy Method		-0.214***		-0.0791**
		(0.0720)		(0.0349)
Controls	Yes	Yes	Yes	Yes
Observations	2247	2247	2970	2970

Notes: this table reports estimates from OLS regressions, with control variables identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

table.[38] Since question 4 is more directly linked to playing the games, we predict that players who answer it incorrectly, identified by the dummy variable *Inattentive (Q4)*, are (i') more likely to make mistakes when playing the games and (ii') that they benefit more from the strategy method in terms of a reduction of their mistakes than those who answer question 1, 2, or 3 incorrectly, indicated by *Inattentive (Q123)*.

The data broadly supports this hypothesis. For both games, the estimates for the dummy variables (i') are larger for participants who answer question 4 incorrectly compared to those who answer questions 1-3 incorrectly. Similarly, the corresponding interaction terms (ii') are more negative for question 4. Moreover, the difference between the Inattentive (Q4) and the Inattentive (Q123) dummy is statistically significant for the *sPD* ($p = 0.028$) and marginally significant for the *mUG* ($p = 0.063$). This, however, is not true for the interaction terms ($p = 0.221$ and $p = 0.152$).

The next piece of evidence comes from the response time data, the time (in seconds) it takes participants to make their choice in each game. Table 9 documents the relationship between player 2's behavior and their response time.[39] The estimates in column (2) and (4) highlight that choices that are taken more slowly result in fewer mistakes while column (1) and (3) emphasize that conditional cooperation (column 1) or rejection of unfair offers (column 3) are unrelated to response time. Since preferences and deliberation time may generally be correlated, we are usually agnostic with regards to their relationship. Nevertheless, the striking negative relationship between response time and mistakes together with a lack of relationship between the two at the other two nodes points to the most natural conclusion that deliberation reduces mistakes.[40]

[38] Questions 1, 2, 3, and 4 were answered incorrectly by 1.28%, 15.62%, 6.17%, and 8% of participants, respectively.

[39] Estimates without controls can be found in the Online Appendix, Table D.22 and D.23.

[40] Recalde et al. (2018)'s findings align with ours. They investigate whether the positive correlation between fast decision-making and high contributions in public-good games reflects genuine intuitive generosity or systematic decision errors. Using a 2×3 experimental design, they manipulate the location of the equilibrium (low vs. high) and decision timing constraints (self-paced, time pressure, and time delay), which exogenously vary the pace of decision-making. The results show that fast decision-makers contribute more than slow ones when the equilibrium is low but contribute less when the equilibrium is high—a pattern inconsistent with stable prosocial preferences but consistent with fast decisions leading to more mistakes. Their analysis further reveals that fast decision-makers are more prone to errors, as they make choices that reduce both individual and group

Indeed, such deeper and longer deliberation would be the most natural mechanism behind the impact of the strategy method on mistakes. In Appendix Tables A.11 and A.12, we document (the admittedly obvious fact given that more choices must be made) that the strategy method indeed gives rise to longer response times.

Table 9: Player 2's Behavior in sPD and mUG: Response Time

	sPD		mUG	
	After Player 1			
	cooperates	defects	offers 85-15	offers 50-50
Dep. Var: P2 cooperates/rejects	(1)	(2)	(3)	(4)
ln(Response time)	-0.00560	-0.0362**	-0.0178	-0.0228***
	(0.0163)	(0.0146)	(0.0188)	(0.00801)
Non-Selfish Belief	0.0465**	0.0325	0.0291	-0.0206***
	(0.0227)	(0.0232)	(0.0202)	(0.00770)
Strategy Method	0.0421	-0.0983***	0.0468**	-0.0266***
	(0.0284)	(0.0216)	(0.0207)	(0.00837)
Non-Selfish × Strategy Method	-0.0577	-0.0222	-0.0433	0.0348***
	(0.0383)	(0.0315)	(0.0300)	(0.0115)
Controls	Yes	Yes	Yes	Yes
Observations	2722	2247	1999	2970

Notes: this table reports estimates from OLS regressions, with control variables identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

For our final piece of evidence, we turn to the usual suspects for preferences that could explain cooperation after defection and rejection of fair offers, namely preferences for efficiency or spitefulness. We classify player 2 as having preferences for efficiency if she always cooperates or always accepts. Her preferences are classified as spiteful if she always defects or always rejects.[41] We categorize players' preferences both by the game as well as overall using the data from the strategy method, which provides us with their choices for all nodes of the game.[42] Table 10 highlights that efficiency preferences in the *sPD* and spitefulness in the *mUG* may explain at most 1/3 of the mistakes in each game (6% vs. 18% and 1% vs. 3%; see Table 2), and even less if we rely on both-games for the classification of such preferences. In other words, it is doubtful that what we consider mistakes are reflections of deep preferences. Moreover, even if behavior appears to be consistent with such preferences, it may still result from an honest mistake.[43]

While none of these four pieces of evidence is conclusive proof by themselves that the strategy method reduces mistakes – or at least of mistakes being the primary reason behind such reduction - together, they greatly increase our confidence that it does. We hope future research explores this observation further.

payoffs. In contrast, slow decision-makers are more responsive to incentives and adjust their contributions accordingly.

[41] More generally, equal splits are at times also rejected in experiments on the "full" ultimatum game, where offers range from $0, 0.1, 0.2, \ldots 1$ (such as Roth et al. (1991)). Such behavior is typically not considered a mistake. However, the equal split is the *best available offer* in our mini ultimatum game and its rejection implies forgoing any positive payoffs, whereas in the general ultimatum game there are many more potential offers that could be preferred by the responder over the equal split. This leaves us with the general notion of spitefulness to explain this behavior in our context.

[42] Note that these preferences are only different from selfish preferences for one of the two games, i.e., efficiency (spiteful) preferences match selfish preferences for the mUG (sPD), or when we consider both games jointly.

[43] A more general response to any preference-based argument is that it is difficult to see why preferences for efficiency and spitefulness would be more pronounced in the direct-response method than the strategy method.

Table 10: Frequency of Player 2 with Preferences for Efficiency or Spite

	sPD only	mUG only	Both Games
Efficiency	0.06	0.85	0.05
Spite	0.28	0.01	0.00

Notes: Player 2 has preferences for efficiency if she always cooperates and always accepts. Her preferences are classified as spiteful if she always defects and always rejects. A player classified as efficiency-seeking or spiteful for *sPD*, *mUG*, or *Both Games* only if their behavior matches these classifications for that particular game or for both respectively. Type frequencies are computed based on data from the strategy method only.

3.5. Heterogeneity Analysis: Attentive vs. Inattentive Players

As part of our analysis of mistakes in the previous subsection, we introduced the notion of attentive and inattentive participants, i.e., those who do and those who do not make a mistake in the control questions. We will now move beyond the narrow idea of mistakes and analyze how, if any, our overall results vary with how attentive ($n = 2754$) or inattentive ($n = 1266$) player 2s are. We will examine this by highlighting the similarities, differences, and general patterns illustrated in our usual graphs, Figure 6 and 7 along with our standard regressions, now combined in the single Table 11.[44] In view of the relatively small numbers of players in the subgroups Inattentive (Q123) and Inattentive (Q4), we will concentrate on the broader category of inattentive players as a whole.[45]

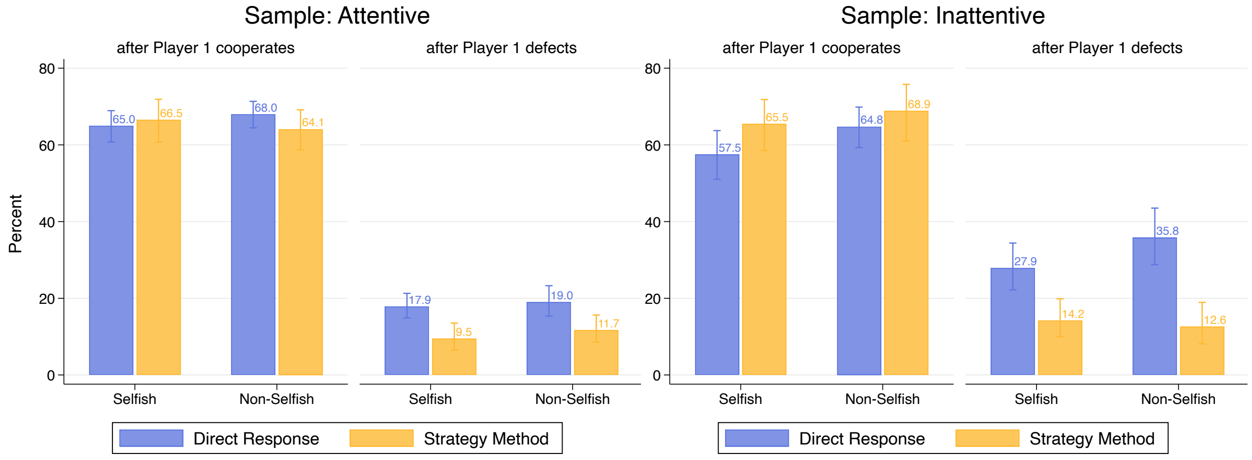


Figure 6: Player 2's Behavior in sPD for Attentive and Inattentive Players

Notes: the number on each bar represents the mean and the error bar represents the 95% confidence interval.

For attentive players, the patterns across elicitation methods and belief-treatments in the *sPD*, Figure 6, closely mirror those observed in the full sample of Figure 4. Our main finding – that the strategy method reduces mistakes – remains strongly noticeable. This is confirmed by the regressions in Table 11. Indeed,

[44] Estimates without controls, which are consistent with the estimates reported here, are relegated to the Online Appendix, Table D.24 and D.25, for conciseness.

[45] In part D.4 of the Online Appendix, we also split the inattentive sample along question 1, 2, 3, and question 4. Due to the relatively small size of these subgroups, the resulting figures D.5 and D.6 have fairly large confidence intervals, which makes it challenging to make more than basic observations with confidence. Overall, the figures and regressions, Table D.48 and D.49 echo the results of Table 8, that mistakes are higher and differences between the direct-response and the strategy method larger for the Inattentive (Q4) sample compared to the Inattentive (Q123) group.

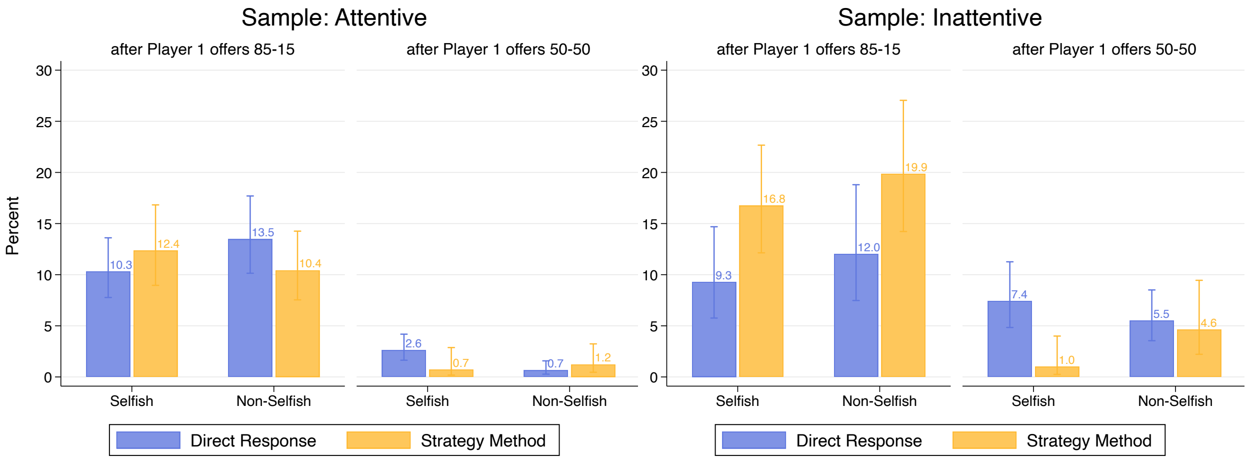


Figure 7: Player 2's Behavior in mUG for Attentive and Inattentive Players

Notes: the number on each bar represents the mean and the error bar represents the 95% confidence interval.

these regressions further strengthen an observation we made for the full sample, namely that "behavior in response to player 1 cooperating is remarkably similar across our four treatment". While we observed a slight increase in cooperation rates for the direct-response method under non-selfish beliefs in the full sample, this result is no longer present.

Table 11: Player 2's Behavior in sPD and mUG – for Attentive and Inattentive Players

	sPD		mUG	
after Player 1	cooperates	defects	85-15	50-50
Dep. Var: Player 2 cooperates / rejects	(1)	(2)	(3)	(4)
Sample: Attentive				
Non-Selfish Belief	0.0315	0.00672	0.0264	-0.0202***
	(0.0269)	(0.0259)	(0.0244)	(0.00698)
Strategy Method	0.0209	-0.0831***	0.0214	-0.0201**
	(0.0350)	(0.0245)	(0.0246)	(0.00830)
Non-Selfish × Strategy Method	-0.0660	0.0108	-0.0500	0.0265**
	(0.0475)	(0.0363)	(0.0360)	(0.0108)
Controls	Yes	Yes	Yes	Yes
Observations	1826	1529	1346	2009
Sample: Inattentive				
Non-Selfish Belief	0.0878**	0.0756	0.0228	-0.0197
	(0.0418)	(0.0495)	(0.0375)	(0.0195)
Strategy Method	0.0806*	-0.170***	0.0858**	-0.0616***
	(0.0475)	(0.0401)	(0.0360)	(0.0181)
Non-Selfish × Strategy Method	-0.0515	-0.0861	-0.00295	0.0486*
	(0.0667)	(0.0624)	(0.0554)	(0.0280)
Controls	Yes	Yes	Yes	Yes
Observations	896	718	653	961

Notes: this table reports estimates from OLS regressions for attentive players who don't make any mistake in the control questions, and inattentive players who do. Control variables identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Given the small difference between the full and attentive sample, it is difficult to draw further conclusions with much confidence. Nevertheless, it is interesting to observe that cooperation rates in response to player 1 cooperating are (marginally) higher in all treatments except for the strategy method with non-selfish beliefs but are always lower after player 1 defects. This may suggest that attentive player's choices more accurately reveal their social preferences.

For the *mUG*, Figure 7 suggests that the behavior of attentive players is fairly similar to that of the full sample in Figure 5. The main observation holds: mistakes (player 2 rejecting 50-50 offers) occur most frequently in the selfish-belief treatment of the direct response method. Interestingly, just as in the *sPD* sample, behavior following unequal offers are now statistically indistinguishable across all treatments, Table 11. In conclusion, our results are robust when only looking at attentive players, for whom the strategy method reduces mistakes.

Next, we turn to the inattentive players in Figure 6 and 7. We observe that mistakes increase significantly, particularly in the direct-response treatments for both games, compared to the full sample or the subset of attentive players. In the *sPD*, this rise in mistakes is accompanied, if anything, by lower cooperation rates in response to player 1 cooperating in the direct-response method. One interpretation of this opposing pattern extends beyond our usual assertion about mistakes, that occur in response to player 1 defecting. Specifically, this pattern may suggest that lower cooperation rates in response to player 1 cooperating could also be due to mistakes, implying that the strategy method may better reveal genuine preferences in our data. In the *mUG*, a similar pattern emerges, where the social action of rejecting unequal offers is more common with the strategy method than the direct-response method, and which, unlike in the full sample, is now statistically significant.

3.6. Robustness

In the Online Appendix, we further investigate the robustness of our results as well as compare the pre-Covid (October to November 2019) and Covid sample (October 2021) in terms of participants' characteristics and behavior.

In particular, we first check for the robustness of our OLS estimates by re-estimating all tables from this section using Probit and Logit regressions (D.1). The results are robust to such a change in estimator. We also revisit the topic of order effects in our data (D.2). We show that, in general, order effects exist, yet that our analysis restricted to task 1 yields qualitatively similar results despite the reduction in sample size by 50%. Since our experiment was conducted online, we also carefully analyze attrition (D.3). We detect a small yet statistically significant difference in attrition between the elicitation methods, originating from question 4 in our control questionnaire, which participants in the strategy method were more likely to answer incorrectly. Although this question had the same correct answer in both elicitation methods, it asked subjects to identify payoffs either for a given action profile in the direct response method or a strategy profile in the strategy method. This approach was consistent with how the task and payoffs were explained to the respective participants in the experimental instructions. Our additional analyses show that our results are robust when this difference in attrition is taken into account.

The sample comparison (E) highlights that participants in the Covid sample tend to be slightly more educated, have higher incomes, and are slightly older. The most notable change in terms of behavior is the higher rejection rates of unfair offers in the later sample.

4. Conclusion

We conducted a large-scale online experiment, where, in addition to varying the elicitation method, we also manipulated the ex-ante beliefs of participants about player 1's likely choices via an information-provision treatment. In neutrally-framed sequential games, a sequential Prisoners' Dilemma and a mini-Ultimatum Game, we found that the elicitation method does not alter player 2's preferences (their tendency to reward cooperation or reject unfair offers) yet significantly reduces mistakes (rewarding defection or rejecting fair offers). The takeaway from our study is that unless researchers have a particular reason for using the direct-response method, they should opt for the more economical strategy method as their go-to method for eliciting behavior in sequential games.

This conclusion may be particularly relevant for many experiments conducted online nowadays. Here, the strategy method appears to be the superior choice as it reduces mistakes, lowers costs, and, in many cases, is easier to run. While we expect these insights to apply also to laboratory experiments (the strategy method is always more economical) – especially when researchers are interested in one-shot behavior – we also recognize the inherent difference between this setting and unsupervised online experiments. Running multiple practice rounds with opportunities for Q&As is simply less feasible online. Maintaining participants' attention and limiting dropouts over an extended amount of time is more challenging in an online context, although recent research suggests it might be possible (Arechar et al. (2018)). Nonetheless, more research in this area is needed.

One limitation of our study is its focus on simple one-shot binary-choice games, which may restrict the external validity of our findings. This limitation was intentional, however, as we considered this setting to be most relevant for testing differences in elicitation methods in general and ideally suited for examining our initial conjecture that ex-ante beliefs may differentially affect behavior across elicitation methods. As cautiously noted by Brandts and Charness (2011, p. 391) in their literature review, differences between elicitation methods may be more likely with fewer contingent choices and less likely for games with multiple periods.

What surprised us was how participants showed a strong tendency for positive reciprocity in the *sPD* yet were very unwilling to punish low offers in the *mUG*, despite the low cost of doing so ($0.3). The findings by Amir et al. (2012) suggests that MTurk workers are generally not averse to rejecting low offers, with the average minimum-acceptable offer falling in the mid-30s for both a no stakes and a $1 treatment.[46] Our findings are reminiscent of those by Charness and Rabin (2002), who test social preferences in a series of simple games, which were neutrally framed and elicited with the strategy method. Similar to our results, they documented very weak negative reciprocity. One might conjecture that the typical framing of an ultimatum bargaining game, with "offers" and the opportunity to "reject", may trigger the participants' willingness to forgo monetary payoffs to punish others, compared to its fully neutrally framed counterpart.

Another limitation relates to our choice of using a between-subject design, in which choices are elicited for some participants with the strategy method and for other participants with the direct-response method. We opted for this approach as it provides a clean test for our research question of how ex-ante beliefs influence player 2's response depending on the elicitation method employed and whether this orthogonal dimension could explain why behavior across the two elicitation methods is at times similar and at times different. More generally, it allows us to directly relate our findings to the majority of the literature, which typically employs a between-subject design when investigating elicitation methods.

This design falls short in terms of what we can learn about the types of preferences or biases player 2

[46] Their Ultimatum Game was the standard discrete version, with offers made from $\{0, 10, \ldots, 100\}$. Player 2 indicated whether they accept/reject each possible offer, i.e., their preferences were elicited by the strategy method.

might exhibit based on population behavior. Fundamentally, behavioral differences between the strategy method and direct-response method can be understood as *dynamically inconsistent* preferences (Machina (1989)).[47] A useful way to conceptualize dynamic inconsistency is by a decision maker who devises a plan (based on her current preferences) but deviates from it in the future. In the strategy method, the decision maker makes a choice for all potential nodes of the game without knowing player 1's choice; in other words, she creates a plan. Since the strategy method does not allow her to adjust her plan in response to player 1's actual choice, she is fully committed to it. Comparing the planned actions with her actual choice, as elicited through the direct-response method, allows us to test if her preferences are dynamically inconsistent.

The ideal test of any theory that results in dynamic inconsistency hence requires observing both the player's plan (in the strategy method) and their actual choices (in the direct-response method), i.e., relies on within-subject variation in the elicitation methods. Between-subject designs become problematic when different types of players behave inconsistently in exactly opposite ways. In such cases, mistakes cancel out at the population level, meaning that behavior that appears fully consistent could be anything but.[48]

In the introduction, we highlighted a behavioral bias, namely incomplete conditional thinking as a motivation for our 2×2 design. In fact, other forms of preferences and/or behavioral biases can also produce similar predictions at the population level, which is one reason we chose not to investigate any specific theory in a more complex design.[49] While incomplete conditional thinking could be one potential source for the interaction of beliefs with the elicitation method, our previous discussion highlights that its absence at the population level does not necessarily imply that it is unimportant in games of social dilemmas. The resulting behavior may simply be canceled by other forms of inconsistencies. In this regard, further research is needed. We believe that a more comprehensive understanding of player motivations, reasoning and potential errors in the context of different elicitation methods will best be achieved through creative designs that observe behavior under both methods (for at least some participants). We emphasize *creative* designs, as players' behavior may be influenced by learning or other effects over time, which could operate differently across the two elicitation methods.

What strikes us as a more intriguing avenue for future research, is to move away from more complex preferences, which depend on multiple dimensions, and focus on a simpler two-type model: one that includes the classic hot/cold and a cold/hot players. A hot/cold player displays social preferences in the direct-response method but acts selfishly in the strategy method, which is often motivated with more emotional reactivity when observing actual choices in the direct-response method. In contrast, the cold/hot player behaves selfishly in the direct-response method but adheres to social preferences in the strategy method. For example, this player may wish to behave socially but is tempted by monetary incentives "in the moment" under the direct-response method. Here, the strategy method allows her to pre-commit to actions that align with what she considers to be her better self.[50] Since social preferences are mirrored across these two player types, experiments that use a between-subject design will not detect difference in behavior across elicitation methods when both player types occur in equal frequencies, and do if the population frequencies differ. Using within-subject variation could therefore provide new valuable insights.

[47] We appreciate the Editor, Emanuel Vespa, for pointing this out.

[48] See C.1 in the Online Appendix for a more detailed discussion on this point.

[49] Our theoretical framework in the Online Appendix develops these preferences in details and discusses the implication of other types of preferences on our design. It also covers some of the ideas raised in this conclusion in more detail.

[50] This may also arise if preference for fairness or kindness are more prominent when players (are forced to) consider the whole game (-tree), which is required in the strategy method.

A. Supporting Tables and Figures

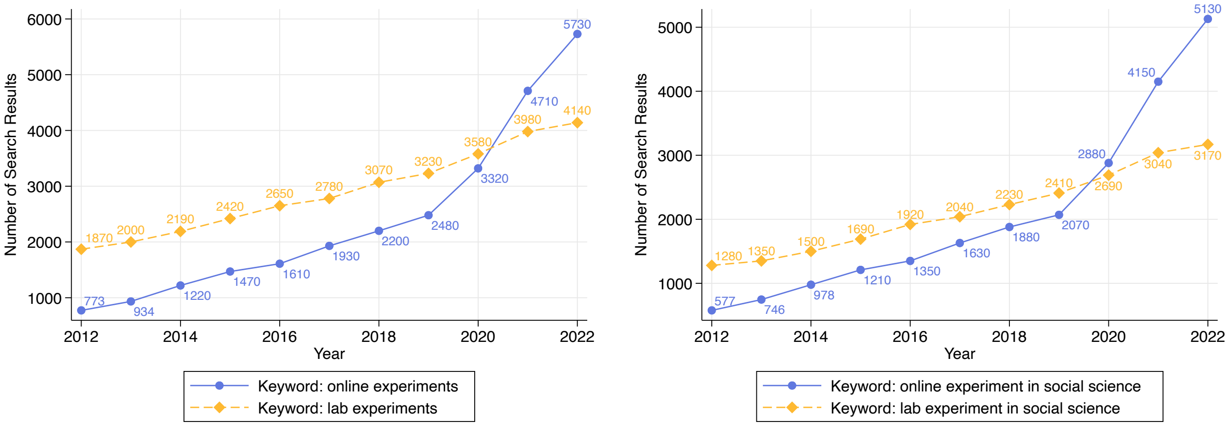


Figure A.1: Number of Search Results for Lab/Online Experiments on Google Scholar

Notes: the figure displays the numbers of search results from a Google Scholar Search using the keywords "online experiment" or "lab experiment", with or without being restricted to the social sciences, which relied on the additional keywords: economics, psychology, management, business, or politics (*date accessed: 3. April 2023*).

Table A.1: Variable Definitions

Variable Name	Definitions
Behavior and Beliefs	
Player 1 cooperates	= 1 if player 1 chooses to cooperate in the *sPD*; 0 otherwise
Player 2 cooperates after P1 cooperates	= 1 if player 2 chooses to cooperate in response to player 1 cooperating in the *sPD*; 0 otherwise
Player 2 cooperates after P1 defects	= 1 if player 2 chooses to cooperate in response to player 1 defecting in the *sPD*; 0 otherwise
Belief Player 1 cooperates	A player's belief about the % of player 1 that cooperate in the *sPD*
Player 1 offers 50-50	= 1 if player 1 offers 50-50 in the *mUG*; 0 otherwise
Player 2 rejects 50-50	= 1 if player 2 rejects the 50-50 offer in the *mUG*; 0 otherwise. [Often written as: Player 2 rejects after P1 offers 50-50]
Player 2 rejects 85-15	= 1 if player 2 rejects the 85-15 offer in the *mUG*; 0 otherwise. [Often written as: Player 2 rejects after P1 offers 85-15]
Belief Player 1 offers 50-50	A player's belief about the % of player 1 that offer 50-50 in the *mUG*
Player 2 makes mistake	= 1 if player 2 cooperates after defection in the *sPD* (rejects the 50-50 offer in the *mUG*); 0 otherwise
Treatment Variables and Other Indicator Variables	
Strategy Method (SM)	= 1 if in strategy method treatment, i.e., behavior of player/opponent is elicited using the strategy method; 0 otherwise
Direct Response (DR)	= 1 if in direct response treatment, i.e., behavior of player/opponent is elicited using the direct response method; 0 otherwise
Selfish Belief	= 1 if in selfish belief treatment; 0 otherwise
Non-Selfish Belief	= 1 if in non-selfish belief treatment; 0 otherwise
Direct Response, Selfish	= 1 if in direct-response *and* selfish belief treatment; 0 otherwise
Direct Response, Non-Selfish	= 1 if in direct-response *and* non-selfish belief treatment; 0 otherwise
Strategy Method, Selfish	= 1 if in strategy method *and* selfish belief treatment; 0 otherwise
Strategy Method, Non-Selfish	= 1 if in strategy method *and* non-selfish belief treatment; 0 otherwise
Player 2	= 1 if participants plays in the role of the second mover; 0 otherwise
P1 (P2)	Short version for Player 1 (2)
Task 2	= 1 if a given game (*sPD* or *mUG*) was played second; 0 otherwise
pre-Covid	= 1 if observation is from the pre-Covid sample; 0 otherwise
Covid	= 1 if observation is from the Covid sample; 0 otherwise
Inattentive	= 1 if participants makes a mistake in the instruction test; 0 otherwise
Inattentive(Q123)	= 1 if participants makes a mistake in question 1, 2, or 3 of the instruction test; 0 otherwise
Inattentive(Q4)	= 1 if participants makes a mistake in question 4 of the instruction test; 0 otherwise
Response Time	Time (in seconds) it takes participants to answer a particular game (and click "Next" to proceed to the next page)
x × y	Interaction term: = 1 if both x and y is true (i.e., = 1); 0 otherwise
x × y × z	Interaction term: = 1 if x, y, and z are true (i.e., = 1); 0 otherwise

Table A.1: Variable Definitions (continued)

Variable Name	Definitions
Personal Information from Survey	
Gender	Categorical variable that indicates whether participant is female, male, other/prefer not to say
Age	Categorical variable that indicates whether participant is $<$ 12, 12-17, 18-24, 25-34, 35-44, 45-54, 55-64, 65-74, $\geq$ 75 years old; or prefers not to say
Income	Categorical variable that indicates whether the participant's annual household income (in USD) is $<$ 20 000, 20 000 - 34 999, 35 000 - 49 999, 50 000 - 74 999, 75 000 - 99 999, 100 000 - 140 999, or $\geq$ 150 000; or prefers not to say
Education	Categorical variable that indicates whether the highest degree the participant is holding or currently pursuing is no degree, High School, Bachelor, Master, Doctorate, Other post-graduate degree; or prefers not to say
Participated in experiments before	=1 if participant has participated in a similar experiment before; 0 otherwise

Table A.2: Summary Statistics

	Median	Mean	Std. Dev.	Min.	Max.	Obs.
Sequential Prisoner's Dilemma						
Player 1 cooperates	1.00	0.572	0.49	0.0	1.0	4009
Player 2 cooperates after P1 cooperates	1.00	0.654	0.48	0.0	1.0	2722
Player 2 cooperates after P1 defects	0.00	0.177	0.38	0.0	1.0	2247
Belief Player 1 cooperates	0.64	0.541	0.32	0.0	1.0	8029
Response time	24.90	31.521	50.10	1.3	3387.9	8029
mini Ultimatum Game						
Player 1 offers 50-50	1.00	0.655	0.48	0.0	1.0	4009
Player 2 rejects 50-50	0.00	0.026	0.16	0.0	1.0	2970
Player 2 rejects 85-15	0.00	0.125	0.33	0.0	1.0	1999
Belief Player 1 offers 50-50	0.62	0.536	0.32	0.0	1.0	8029
Response time	21.10	26.438	31.65	1.6	1585.3	8029
Other Game Outcomes						
Total earnings from games (in USD)	2.00	1.752	0.60	0.0	3.6	8029
Total time (in sec.)	304.30	367.037	280.80	31.7	4846.3	8029
Number of mistakes in understanding test	0.00	0.311	0.46	0.0	1.0	8029
Treatments						
Direct Response, Selfish	0.00	0.379	0.49	0.0	1.0	8029
Direct Response, Non-Selfish	0.00	0.385	0.49	0.0	1.0	8029
Strategy Method, Selfish	0.00	0.116	0.32	0.0	1.0	8029
Strategy Method, Non-Selfish	0.00	0.120	0.32	0.0	1.0	8029
Players, Game Order, Sample						
Player 2	1.00	0.501	0.50	0.0	1.0	8029
Player 2 × Strategy Method	0.00	0.118	0.32	0.0	1.0	8029
Played sPD as 2nd task	1.00	0.501	0.50	0.0	1.0	8029
Covid	0.00	0.421	0.49	0.0	1.0	8029
Inattentive	0.00	0.311	0.46	0.0	1.0	8029
Inattentive (Q123)	0.00	0.231	0.42	0.0	1.0	8029
Inattentive (Q4)	0.00	0.080	0.27	0.0	1.0	8029

Table A.2: Summary Statistics (continued)

	Median	Mean	Std. Dev.	Min.	Max.	Obs.
Personal Information from Survey						
Participated in experiments before	1.00	0.742	0.44	0.0	1.0	8029
Gender:						
Female	1.00	0.517	0.50	0.0	1.0	8029
Male	0.00	0.475	0.50	0.0	1.0	8029
Other / Prefer not to say	0.00	0.008	0.09	0.0	1.0	8029
Age:						
$<$ 12 years	0.00	0.000	0.00	0.0	0.0	8029
12-17 years old	0.00	0.000	0.01	0.0	1.0	8029
18-24 years old	0.00	0.073	0.26	0.0	1.0	8029
25-34 years old	0.00	0.370	0.48	0.0	1.0	8029
35-44 years old	0.00	0.277	0.45	0.0	1.0	8029
45-54 years old	0.00	0.153	0.36	0.0	1.0	8029
55-64 years old	0.00	0.091	0.29	0.0	1.0	8029
65-74 years old	0.00	0.031	0.17	0.0	1.0	8029
$\geq$ 75 years	0.00	0.003	0.05	0.0	1.0	8029
Prefer not to say	0.00	0.002	0.05	0.0	1.0	8029
Income:						
Less than 20 000	0.00	0.000	0.00	0.0	0.0	8029
20 000 to 34 999	0.00	0.152	0.36	0.0	1.0	8029
35 000 to 49 999	0.00	0.178	0.38	0.0	1.0	8029
50 000 to 74 999	0.00	0.241	0.43	0.0	1.0	8029
75 000 to 99 999	0.00	0.155	0.36	0.0	1.0	8029
100 000 to 140 999	0.00	0.108	0.31	0.0	1.0	8029
over 150 000	0.00	0.048	0.21	0.0	1.0	8029
Prefer not to say	0.00	0.022	0.15	0.0	1.0	8029
Education:						
No Degree	0.00	0.009	0.09	0.0	1.0	8029
High School Degree	0.00	0.262	0.44	0.0	1.0	8029
Bachelor Degree	1.00	0.503	0.50	0.0	1.0	8029
Master Degree	0.00	0.158	0.37	0.0	1.0	8029
Other Post-Grad Degree	0.00	0.031	0.17	0.0	1.0	8029
Doctorate Degree	0.00	0.025	0.16	0.0	1.0	8029
Prefer not to say	0.00	0.012	0.11	0.0	1.0	8029

Table A.3: Randomization Check

	Direct Response, Selfish (1)			Direct Response, Non-Selfish (2)			Strategy Method, Selfish (3)			Strategy Method, Non-Selfish (4)		
	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.
Female	3,040	0.52	0.50	3,094	0.52	0.50	935	0.51	0.50	960	0.53	0.50
Age	3,036	39.41	12.08	3,087	39.14	12.20	935	39.35	12.12	954	39.29	11.80
Participated in experiments before	3,040	0.74	0.44	3,094	0.74	0.44	935	0.74	0.44	960	0.74	0.44
Household income	2,206	65,607.43	38,187.38	2,273	65,849.10	38,399.20	675	64,662.96	36,258.21	680	63,382.35	36,091.66
No Degree	3,040	0.01	0.10	3,094	0.01	0.10	935	0.01	0.10	960	0.01	0.07
High School Degree	3,040	0.27	0.44	3,094	0.27	0.44	935	0.24	0.42	960	0.26	0.44
Bachelor Degree	3,040	0.50	0.50	3,094	$0.49^{(3)**}$	0.50	935	0.54	0.50	960	0.52	0.50
Master or above	3,040	0.21	0.41	3,094	0.22	0.42	935	0.20	0.40	960	0.21	0.41

Notes: this table reports the Kolmogorov-Smirnov pairwise randomization test results between treatments. To keep the table succinct, we aggregate the age and household incom categories. Moreover, small category values such as "Other" or "Prefer not to say" for gender and highest education are omitted, noting that none of these subcategories display significar differences. The superscript next to the mean of each treatment shows the column number to which treatment (column) is compared, and the asterisks mark the significance level the difference following the conventional manner. If, for a given variable, two treatments are not significantly different at conventional levels, no superscript is added. This compariso is only conducted to the "right" to avoid double counting, i.e., Direct Response, Selfish (1) is compared to Direct Response, Non-Selfish (2), Strategy Method, Selfish (3), and Strateg Method, Non-Selfish (4). Next, Direct Response, Non-Selfish (2) is compared to Strategy Method, Selfish (3) and Strategy Method, Non-Selfish (4), etc. *, **, and *** indicate statistic significance at the 10%, 5%, and 1% levels, respectively.

Table A.4: Frequency of Player 2 Types

Strategy Method	sPD only	mUG only	Both Games
Social	0.60	0.13	0.10
Selfish	0.28	0.85	0.27
Other	0.12	0.02	0.63
Direct Response	sPD only	mUG only	Both Games
Social	0.71	0.68	0.48
Selfish	0.53	0.94	0.51
Other	0.09	0.02	0.22

Notes: a player classified as social/selfish for *sPD* or *mUG* only if their behavior is consistent with the respective theoretical prediction for that particular game. If it is consistent with neither, they are classified as *Other*. Similarly, if their behavior is consistent with the predictions of a theory for *both games*, then they are classified into the respective category. Further note that for the strategy method, the three categories are distinct and exhaustive. However, given that only one action of player 2 is observed in the direct response method, a player may be categorized as more than one type, and as such, frequencies do not need to sum to 1.

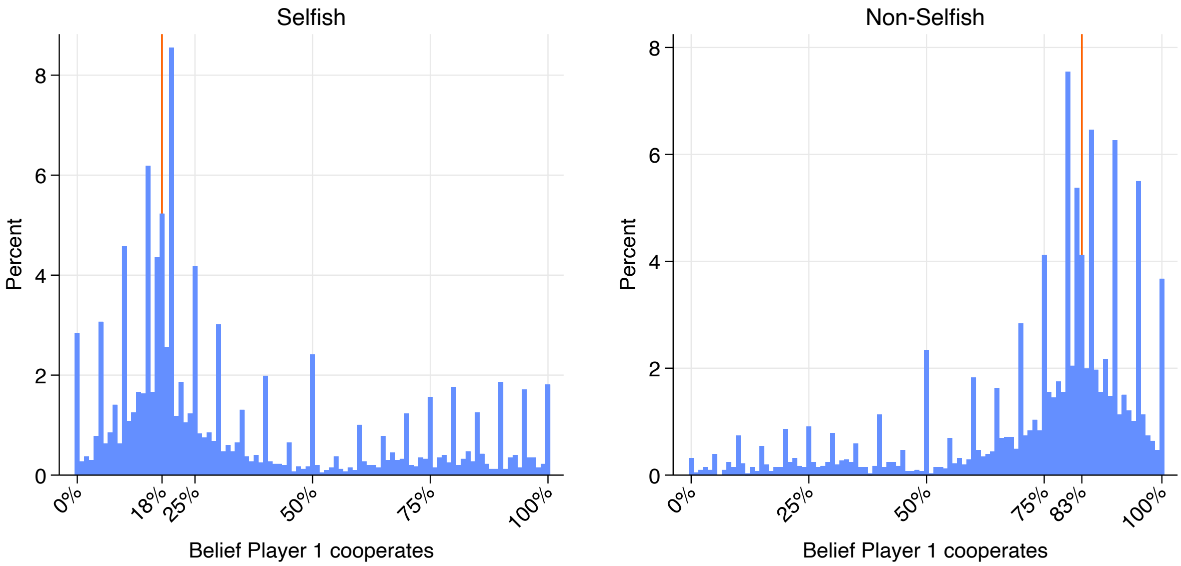


Figure A.2: Distribution of the Belief that Player 1 Cooperates in the seq. Prisoner's Dilemma

Notes: The thin orange vertical lines at 18% and 83% indicates the probabilities provided in the belief treatments.

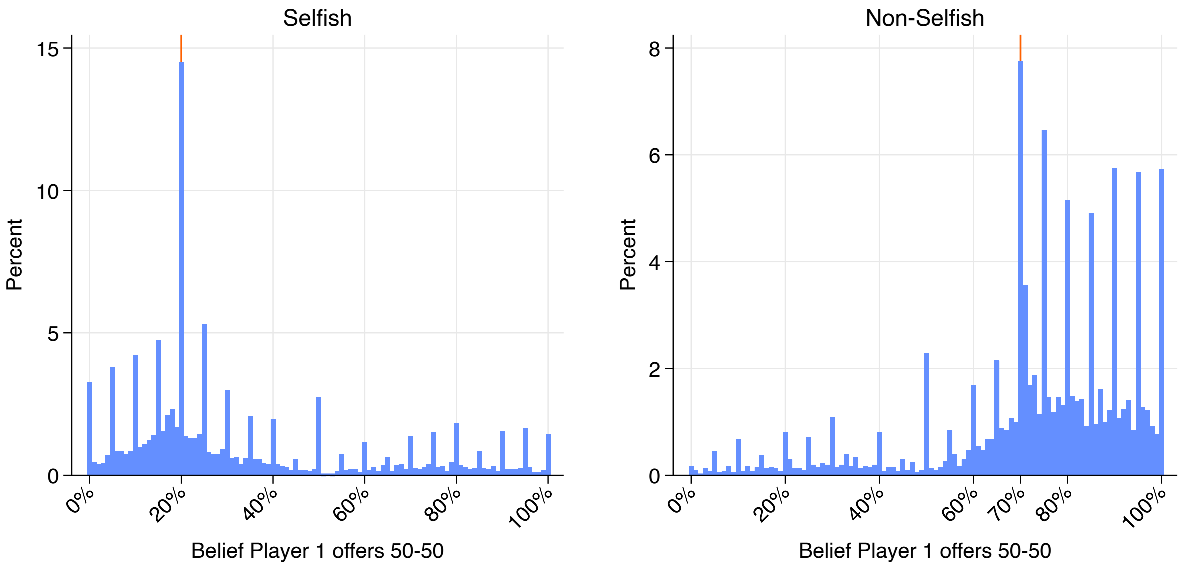


Figure A.3: Distribution of the Belief that Player 1 Offers 50/50 in the mini Ultimatum Game

Notes: The thin orange vertical lines at 20% and 70% indicates the probabilities provided in the belief treatments.

Table A.5: Beliefs about Player 1's Behavior

	sPD				mUG			
Dep. Var: Belief P1 takes Non-Selfish Action	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Non-Selfish Belief	0.399***	0.399***	0.399***	0.434***	0.402***	0.402***	0.401***	0.407***
	(0.00649)	(0.00649)	(0.00641)	(0.00860)	(0.00623)	(0.00623)	(0.00619)	(0.00843)
Strategy Method	-0.0187*	-0.0196*	-0.0201**	-0.0142	-0.0205**	-0.0209**	-0.0213**	-0.00224
	(0.0101)	(0.0101)	(0.00989)	(0.0122)	(0.00976)	(0.00977)	(0.00976)	(0.0128)
Non-Selfish × Strategy Method	0.0149	0.0156	0.0166	0.0191	0.0140	0.0144	0.0150	0.0118
	(0.0130)	(0.0130)	(0.0128)	(0.0170)	(0.0125)	(0.0125)	(0.0125)	(0.0169)
Player 2			0.0785***	0.118***			0.0509***	0.0650***
			(0.00555)	(0.0100)			(0.00537)	(0.00974)
Player 2 × Strategy Method				-0.0121				-0.0378*
				(0.0197)				(0.0195)
Player 2 × Non-Selfish				-0.0707***				-0.0117
				(0.0128)				(0.0124)
Player 2 × Strategy Method × Non-Selfish				-0.00515				0.00610
				(0.0256)				(0.0249)
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Observations	8029	8029	8029	8029	8029	8029	8029	8029

Notes: this table reports estimates from OLS regressions with the dependent variable being a player's belief regarding the % of player 1 that either cooperate in the *sPD* or offer 50-50 in the mUG. The control variables are identical to those in Table 4. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively. For column (4), the linear combinations indicate that player 2 is more optimistic for the Strategy Method and Selfish Beliefs (diff = 0.1056, p-value < 0.01) and Strategy Method and Non-Selfish Beliefs (diff = 0.0298, p-value = 0.036) than player 1 in the respective treatments. In column (8), while positive, neither of the two linear combinations is significant at 10% (with p-values of 0.107 and 0.109).

Table A.6: Player 1's Behavior – Treatment Differences

	sPD		mUG	
Dep. Var: P1 cooperates; offers 50-50	(1)	(2)	(3)	(4)
Direct Response, Non-Selfish	0.165***	0.166***	0.117***	0.115***
	(0.0176)	(0.0175)	(0.0170)	(0.0169)
Strategy Method, Selfish	-0.0912***	-0.0907***	-0.0380	-0.0317
	(0.0262)	(0.0263)	(0.0263)	(0.0262)
Strategy Method, Non-Selfish	0.198***	0.198***	0.144***	0.150***
	(0.0246)	(0.0245)	(0.0236)	(0.0232)
Estimated Differences				
SM, Selfish − DR, Non-Selfish	-0.256***	-0.256***	-0.155***	-0.147***
	(0.0258)	(0.0260)	(0.0258)	(0.0258)
SM, Non-Selfish − DR, Non-Selfish	0.0336	0.0328	0.0266	0.0345
	(0.0242)	(0.0242)	(0.0230)	(0.0228)
SM, Non-Selfish − SM/Selfish	0.290***	0.289***	0.182***	0.181***
	(0.0310)	(0.0311)	(0.0305)	(0.0303)
Controls	No	Yes	No	Yes
Observations	4009	4009	4009	4009

Notes: this table reports estimates from OLS regressions in the top panel and estimated difference between the treatments, with control variables identical to those in Table 4. Estimates for control variables are not reported. The omitted category is the Direct Response, Selfish Beliefs treatment, meaning that the regression estimates in the top panel represent the difference to this treatment group. Robust standard errors are reported in parentheses. ** and *** indicate statistical significance at the 5% and 1% levels.

Table A.7: Player 1's Response Time in sPD and mUG

	sPD		mUG	
Dep. Var: ln(Response time)	(1)	(2)	(3)	(4)
Direct Response, Non-Selfish	-0.0466*	-0.0445*	0.00231	0.00753
	(0.0248)	(0.0232)	(0.0225)	(0.0212)
Strategy Method, Selfish	0.0637*	0.0873***	0.00107	0.0203
	(0.0354)	(0.0330)	(0.0330)	(0.0308)
Strategy Method, Non-Selfish	-0.0462	-0.0196	-0.00741	0.0165
	(0.0366)	(0.0351)	(0.0334)	(0.0309)
Estimated Differences				
SM, Selfish − DR, Non-Selfish	0.110***	0.132***	-0.00124	0.0128
	(0.0354)	(0.0329)	(0.0327)	(0.0306)
SM, Non-Selfish − DR, Non-Selfish	0.000366	0.0249	-0.00972	0.00897
	(0.0366)	(0.0350)	(0.0330)	(0.0307)
SM, Non-Selfish − SM/Selfish	-0.110**	-0.107**	-0.00848	-0.00382
	(0.0444)	(0.0420)	(0.0409)	(0.0379)
Controls	No	Yes	No	Yes
Observations	4009	4009	4009	4009

Notes: this table reports estimates from OLS regressions in the top panel and estimated difference between the treatments, with control variables identical to those in Table 4. Estimates for control variables are not reported. The omitted category is the Direct Response, Selfish Beliefs treatment, meaning that the regression estimates in the top panel represent the difference to this treatment group. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table A.8: Player 1's Behavior in sPD and mUG: Response Time

	sPD		mUG	
Dep. Var: P1 cooperates; offers 50-50	(1)	(2)	(3)	(4)
ln(Response time)	-0.0817***	-0.108***	0.0273**	-0.00371
	(0.0113)	(0.0118)	(0.0125)	(0.0132)
Non-Selfish Belief	0.161***	0.161***	0.117***	0.115***
	(0.0175)	(0.0174)	(0.0170)	(0.0169)
Strategy Method	-0.0860***	-0.0812***	-0.0380	-0.0317
	(0.0262)	(0.0263)	(0.0262)	(0.0262)
Non-Selfish × Strategy Method	0.120***	0.117***	0.0649*	0.0662*
	(0.0355)	(0.0354)	(0.0349)	(0.0347)
Controls	No	Yes	No	Yes
Observations	4009	4009	4009	4009

Notes: this table reports estimates from OLS regressions, with control variables identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table A.9: Player 2's Behavior in the sPD – Treatment Differences

	after P1 cooperates		after P1 defects	
Dep. Var: Player 2 cooperates	(1)	(2)	(3)	(4)
Direct Response, Non-Selfish	0.0433	0.0472**	0.0342	0.0302
	(0.0230)	(0.0226)	(0.0235)	(0.0232)
Strategy Method, Selfish	0.0344	0.0401	-0.0917***	-0.110***
	(0.0280)	(0.0279)	(0.0208)	(0.0212)
Strategy Method, Non-Selfish	0.0295	0.0292	-0.0866***	-0.0968***
	(0.0280)	(0.0279)	(0.0209)	(0.0210)
Estimated Differences				
SM, Selfish − DR, Non-Selfish	-0.00893	-0.00707	-0.126***	-0.140***
	(0.0263)	(0.0262)	(0.0235)	(0.0235)
SM, Non-Selfish − DR, Non-Selfish	-0.0138	-0.0179	-0.121***	-0.127***
	(0.0263)	(0.0263)	(0.0236)	(0.0234)
SM, Non-Selfish − SM/Selfish	-0.00483	-0.0108	0.00509	0.0129
	(0.0308)	(0.0309)	(0.0209)	(0.0212)
Controls	No	Yes	No	Yes
Observations	2722	2722	2247	2247

Notes: this table reports estimates from OLS regressions in the top panel and estimated difference between the treatments, with control variables identical to those in Table 4. Estimates for control variables are not reported. The omitted category is the Direct Response, Selfish Beliefs treatment, meaning that the regression estimates in the top panel represent the difference to this treatment group. Robust standard errors are reported in parentheses. ** and *** indicate statistical significance at the 5% and 1% levels.

Table A.10: Player 2's Behavior in the mUG – Treatment Differences

	after P1 offers 85-15		after P1 offers 50-50	
Dep. Var: Player 2 rejects	(1)	(2)	(3)	(4)
Direct Response, Non-Selfish	0.0302	0.0277	-0.0186**	-0.0195**
	(0.0201)	(0.0202)	(0.00785)	(0.00773)
Strategy Method, Selfish	0.0416**	0.0410**	-0.0319***	-0.0363***
	(0.0202)	(0.0201)	(0.00775)	(0.00810)
Strategy Method, Non-Selfish	0.0338	0.0289	-0.0173	-0.0195**
	(0.0199)	(0.0198)	(0.00946)	(0.00943)
Estimated Differences				
SM, Selfish − DR, Non-Selfish	0.0114	0.0133	-0.0133**	-0.0168***
	(0.0226)	(0.0225)	(0.00609)	(0.00637)
SM, Non-Selfish − DR, Non-Selfish	0.00364	0.00118	0.00132	-0.0000752
	(0.0223)	(0.0222)	(0.00816)	(0.00824)
SM, Non-Selfish − SM/Selfish	-0.00778	-0.0122	0.0146	0.0167**
	(0.0224)	(0.0223)	(0.00807)	(0.00844)
Controls	No	Yes	No	Yes
Observations	1999	1999	2970	2970

Notes: this table reports estimates from OLS regressions in the top panel and estimated difference between the treatments, with control variables identical to those in Table 4. Estimates for control variables are not reported. The omitted category is the Direct Response, Selfish Beliefs treatment, meaning that the regression estimates in the top panel represent the difference to this treatment group. Robust standard errors are reported in parentheses. ** and *** indicate statistical significance at the 5% and 1% levels.

Table A.11: Player 2's Response Time in sPD

	after P1 cooperates				after P1 defects			
Dep. Var: ln(Response time)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Strategy Method	0.386***	0.394***	0.346***	0.363***	0.257***	0.250***	0.313***	0.313***
	(0.0241)	(0.0224)	(0.0362)	(0.0340)	(0.0265)	(0.0247)	(0.0356)	(0.0334)
Non-Selfish Belief			-0.119***	-0.112***			0.0677*	0.0639*
			(0.0314)	(0.0291)			(0.0385)	(0.0356)
Non-Selfish × Strategy Method			0.0632	0.0450			-0.123**	-0.134***
			(0.0487)	(0.0455)			(0.0536)	(0.0499)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	2722	2722	2722	2722	2247	2247	2247	2247

Notes: this table reports estimates from OLS regressions, with control variables identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table A.12: Player 2's Response Time in mUG

	after P1 offers 85-15				after P1 offers 50-50			
Dep. Var: ln(Response time)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Strategy Method	0.245***	0.231***	0.326***	0.323***	0.388***	0.397***	0.425***	0.426***
	(0.0269)	(0.0248)	(0.0375)	(0.0349)	(0.0230)	(0.0213)	(0.0351)	(0.0327)
Non-Selfish Belief			0.0496	0.0793**			-0.0441	-0.0487*
			(0.0401)	(0.0368)			(0.0288)	(0.0260)
Non-Selfish × Strategy Method			-0.170***	-0.193***			-0.0763*	-0.0615
			(0.0540)	(0.0499)			(0.0462)	(0.0429)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	1999	1999	1999	1999	2970	2970	2970	2970

Notes: this table reports estimates from OLS regressions, with control variables identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

B. Instructions of Experiment

Below, you will find the exact instructions of our experiment.[51]

Introduction

Thank you for accepting this HIT. If you choose to continue with this job, you will participate in an experiment on decision making. This experiment has three parts: two decision tasks and a short survey with 7 questions. The entire experiment will take 10 minutes to complete. You will earn $1 for completing the HIT and, depending on your choices and the choices of other MTurk workers, an additional amount of up to $3. If you wish to continue with this HIT, please ensure you have sufficient time to complete the whole study. In each decision task, you will be randomly matched with another participant. The interaction is completely anonymous. Neither you nor the other worker will know the other person's worker ID. This experiment follows a no-deception policy. All tasks are implemented exactly as outlined in the instructions. The instructions are the same for all participants that you may interact with. All participants are real MTurk workers. Finally, your earnings and decision in each part of the experiment do not depend on earnings and decisions in other parts. If you wish to continue with this HIT, please ensure you have sufficient time to complete the whole study. **Please do not close this page during the experiment.** If you leave the website during the experiment, you will **not** receive any earnings. Moreover, you will only be able to participate in this experiment once. By clicking the next button, you consent to taking part in this experiment and promise to do your best to complete the whole experiment. [Next Page Button]

Instructions

We now explain how the decision tasks work. Please read these instructions carefully as we will ask you some simple questions about it on the next page.

Who you interact with

In each of the two decision tasks, you will be randomly matched with another MTurk worker. The interaction is completely anonymous. Neither you nor the other worker will know the other person's worker ID. Moreover, you will not face the same worker twice, i.e. you will interact with one participant in task 1 and another in task 2.

The amount of money that you earn in these tasks will depend on your choice and the other participant's choice. For each task, you will be given a table, similar to the one below, that summarizes your potential earnings. The numbers in the table represent real dollars.

An example of your tasks (slightly different from the actual tasks)

We will now walk you through an example to illustrate the finer details. Note that you will not be paid for this particular example and that the earnings associated with the actual tasks will be quite different.

		Other Participant	
		C	**D**
You	**A**	You earn: $2.00 Other earns: $3.00	You earn: $1.00 Other earns: $2.00
	B	You earn: $0.50 Other earns: $0.50	You earn: $6.00 Other earns: $5.00

[51] In order to save space, we have omitted many line-breaks as well as spacing between lines that were used to display the text unless it facilitates readability and/or improves understanding. All other formatting is exactly the same as in the experiment itself. Screenshots of our experiments can be found in the Online Appendix.

In this example, you can choose between option **A** and **B** (the rows) while the other participant decides between **C** and **D** (the columns). If, for example, you choose **B** and the other participant chooses **D**, you will earn $6 while the other participant will earn $5.

Who acts when

Either you or the other participant will move first. You will be randomly assigned to be the **first mover** or the **second mover**. Your role will be the same for both tasks, that is, you will be either a first mover for both tasks or a second mover for both tasks. The difference between these two roles is as follows:

`[Direct response instructions]` The first mover makes his or her decision first. Afterwards, the second mover will be informed about the first mover's choice and decides how to respond.

`[end of direct response instructions]`

`[Strategy method instructions]` The **first mover** makes his or her decision first. The **second mover** needs to make two choices, one in response to each of the first mover's possible decisions. For example, if you are the **second mover**, you will make the following choices:

If the first mover chooses **C**, I respond with [select **A** or **B**]

If the first mover chooses **D**, I respond with [select **A** or **B**]

The actual outcome will be determined by the first mover's choice and how the second mover responds to that particular choice. For instance, suppose the first mover chose **C** and you, as the second mover, chose **A** in response to **C** and **B** in response to **D**. In this case you earn $2 and the other participant earns $3.

`[end of strategy method instructions]`

Note: All information that you see as the first or second mover will also be available to the other participant

Your earnings Your total earnings from participating in this HIT will be sum of your earnings from the two decision tasks, money earned in the survey, and the participation fee. [Next Page Button]

Control Questions

Before we start with task 1, we want to ensure that you have understood the instructions. **In order to continue with this study, you will need to get at least 3 out of 4 questions correct.** If you aren't quite sure about your answers, have a look at the instructions at the bottom of this page again. Please answer the following questions:

1. Do you know the identity, i.e. their MTurk ID or any other personal information, of the participant you are matched with? [Yes/No (multiple choice list)]

2. Imagine you assume the role of the second mover in task 1. Will your role change in task 2? [Yes/No (multiple choice list)]

3. In the two decision tasks, will you interact with the same MTurk worker? [Yes/No (multiple choice list)]

4. Suppose you are the first mover and earnings are determined by the following table:

`[Same earnings table as in the introduction is displayed here]`

`[Direct response treatment]`

If you choose **A** and the second mover responds with **C**, how much do you and the other participant earn in this task?

`[end of direct response treatment]`

`[Strategy method treatment]`

Suppose you choose **A** and the second mover takes the following conditional choices:

In response to **A**, the second mover chooses **C**
In response to **B**, the second mover chooses **D**

How much do you and the other participant earn in this task? [`end of strategy method treatment`]
You earn [select among 1,2,3,4,5,6 (dropdown menu without default value)]
The other participant earns [select among 1,2,3,4,5,6 (dropdown menu without default value)]

[Next Page Button]

[`previous instructions are again displayed in a box at this position`]

Control Questions – Review
[`this page is displayed only if the participant answered at least one question incorrectly. The following are the statements for the respective questions if they were answered incorrectly. Subjects who answered at least two questions incorrectly are not able to continue.`][52]

Dear participant, you answered at least one control question incorrectly. Before you continue, please have a quick look at the correct answer(s) given below:
Question 1: Do you know the identity of the participant you are matched with? You answered this question with *Yes*. This is *incorrect*. You will never learn any information about the identity of the MTurk workers that you will interact with. Neither will they learn any information about your identity.
Question 2: Imagine you assume the role of the second mover in task 1. Will your role change in task 2? You answered this question with *Yes*. This is *incorrect*. Your role will never change. If you are the first mover in task 1, you will also be the first mover in task 2. Similarly, if you are the second mover in task 1, you will also be the second mover in task 2.
Question 3: In the two decision tasks, will you interact with the same MTurk worker? You answered this question with *Yes*. This is *incorrect*. In task 2, you will be randomly matched with a different worker.
Question 4: Suppose you are the first mover and earnings are determined by the following table:

[`Earnings table from control question 4 is displayed here`]

[`Direct response treatment`] If you choose **A** and the second mover responds with **C**, how much do you and the other participant earn in this task? You answered this question with: You earn [`their answer`], the other participant earns [`their answer`]. This is incorrect. As the first mover, which is you in this example, chooses **A** and the second mover responds with **C**, you will earn $2 and the other participant will earn $3. [`end of direct response treatment`]

[`Strategy method treatment`] Suppose you choose **A** and the second mover takes the following conditional choices:

- In response to **A**, the second mover chooses **C**
- In response to **B**, the second mover chooses **D**

You answered this question with: You earn [`their answer`], the other participant earns [`their answer`]. This is incorrect. As the first mover, which is you in this example, chooses **A** and the second mover responds

[52] Those who failed the control questions were redirected to a simple feedback page that informed them that the experiment has ended and offered them an opportunity to write down any feedback or complaints in a text field.

to **A** with **C**, you will earn $2 and the other participant will earn $3.[end of strategy method treatment]

Decision Task 1 out of 2
When you are ready, please press the next button [Next Page Button]

[The task order is random. For the purpose of presentation, we use the sPD (mUG) for task 1(2). We present task 1(2) from the perspective of player 1(2).

Decision Task 1

		Other Participant			
		C		**D**	
You	**A**	You earn:	$1.00	You earn:	$0.00
		Other earns:	$1.00	Other earns:	$1.50
	B	You earn:	$1.50	You earn:	$0.50
		Other earns:	$0.00	Other earns:	$0.50

Your role: you are the **first mover.**

[Non-Selfish belief treatment] **Background Information:** In a well-known study of this task by Watabe, Terai, Hayashi, and Yamagishi, published in the year 1996, 82.6% of the first movers chose **A**. [end of non-selfish belief treatment]

[Selfish Belief treatment] **Background Information:** In a well-known study of this task by Bolle and Ockenfels, published in the year 1990, 82.7% of the first movers chose **B**. [end of selfish belief treatment]

As the first mover, I choose: [A/B (multiple choice list)]
[Next Page Button]

Decision Task 2 out of 2
When you are ready, please press the next button [Next Page Button]

Decision Task 2

		Other Participant			
		A		**B**	
You	**C**	You earn:	$1.00	You earn:	$0.30
		Other earns:	$1.00	Other earns:	$1.70
	D	You earn:	$0.00	You earn:	$0.00
		Other earns:	$0.00	Other earns:	$0.00

Your role: you are the **second mover.**

[Non-Selfish Belief treatment] **Background Information:** In a well-known study of this task by Güth, Huck, and Müller, published in the year 2001, 70.6% of the first movers chose **A**.

[end of non-selfish belief treatment]

[Selfish belief treatment] **Background Information:** In our previous experiment of this task, 80% of the first movers chose **B**. [end of selfish belief treatment]

[Direct response treatment]
The other participant chose: **A**

As the second mover, I respond with [C/D (multiple choice list)] [end of direct response treatment]

[Strategy method treatment] As the second mover,
if the first mover chooses **A**, I respond with: [C/D (multiple choice list)]
if the first mover chooses **B**, I respond with: [C/D (multiple choice list)] [end of strategy method treatment]

[Next Page Button]

Survey - page 1/3

The first decision task you completed today was the following interaction:
[The payoff matrix, role assignment and background information from task 1 is displayed inside a gray box, appearing exactly how they say it before]

Among the MTurk workers who participated in this experiment with you today, what percentage of first movers do you think will choose [**A** or **B** is shown depending on which action was highlighted in the belief-treatment]?
[slider from 0 to 100 is shown, with a default at 50]
Note: If you are within 5% of the correct answer you will receive an additional $0.25.
[Next Page Button]

Survey - page 2/3
[same as page 1 but for the second task]

Survey - page 2/3
Before finishing the experiment, we would like to know more about you. All answers will be processed anonymously and will not be connected to your MTurk worker ID.
What is your gender? [Male, Female, Other, Prefer not to say (dropdown menu)]
What is your age (in years)? [Under 12 years old, 12-17 years old, 18-24 years old, 25-34 years old, 34-44 years old, 45-54 years old, 55-64 years old, 65-74 years old, 75 years or older, Prefer not to say (dropdown menu)]
What is the highest degree you are holding or currently pursuing? [High School, Bachelor, Master, Doctorate, Other post-graduate degree, None, Prefer not to say (dropdown menu)]
What is the annual household income (in USD) you have at your disposal? [Less than $20,000, $20,000 to $34,999, $35,000 to $49,999, $50,000 to $74,999, $75,000 to $99,999, $100,000 to $140,999, Over $150,000 (dropdown menu)]

Have you participated in another similar experiment as this before? [Yes, No (dropdown menu)]
[Next Page Button]

End of Experiment[53]
Thank you for completing this HIT. Before you continue, please copy-paste the following survey completion-code into MTurk.
Completion Code: [`participant's completion code is shown`]
[check-box] I have copy-pasted the completion code.
Have a good day [Finish HIT button]

Feedback[54]
Thanks again for participating. If you have copy pasted the survey code to MTurk, you are done. We will calculate your earnings shortly and will provide you with a detailed summary of your choices, as well as the choices of the participants you were matched with, in the message that is sent alongside the bonus payment. If you encountered any technical or other difficulties today, it would be great if you would let us know that we can fix them. You can type in here: [large textfield]
Thank you and have a great day! [Exit button]

[53] At this point, the experiment had formally ended. No further button or check-box click was required.

[54] This page was fully optional. No button-click was required.

References

Herman Aguinis, Wayne F Cascio, and Ravi S Ramani. Science's reproducibility and replicability crisis: International business is not immune. Research Methods in International Business, pages 45–66, 2020.

Chiara Aina, Pierpaolo Battigalli, and Astrid Gamba. Frustration and anger in the ultimatum game: An experiment. Games and Economic Behavior, 122:150–167, 2020.

Maurice Allais. Le comportement de l'homme rationnel devant le risque, critique des postulats et axiomes de l'ecole americaine. Econometrica, 21:503–546, 1953.

David Amdur and Ethan Schmick. Does the direct-response method induce guilt aversion in a trust game? Economics Bulletin, 33(1):687–693, 2013.

Ofra Amir, David G. Rand, and Ya'akov Kobi Gal. Economic games on the internet: The effect of $1 stakes. PLOS ONE, 7(2):1–4, 2012.

Felipe A Araujo, Stephanie W Wang, and Alistair J Wilson. The times they are a-changing: Dynamic adverse selection in the laboratory. Working Paper, 2018.

Antonio A Arechar, Simon Gächter, and Lucas Molleman. Conducting interactive experiments online. Experimental Economics, 21:99–131, 2018.

Pierpaolo Battigalli, Martin Dufwenberg, and Alec Smith. Frustration, aggression, and anger in leader-follower games. Games and Economic Behavior, 117:15–39, 2019.

Friedel Bolle and Peter Ockenfels. Prisoners' dilemma as a game with incomplete information. Journal of Economic Psychology, 11(1):69–84, 1990.

Pablo Brañas-Garza, Marisa Bucheli, María Paz Espinosa, and Teresa García-Muñoz. Moral cleansing and moral licenses: Experimental evidence. Economics & Philosophy, 29(2):199–212, 2013.

Jordi Brandts and Gary Charness. Hot vs. cold: Sequential responses and preference stability in experimental games. Experimental Economics, 2:227–238, 2000.

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

Jordi Brandts and Gary Charness. The strategy versus the direct-response method: A first survey of experimental comparisons. Experimental Economics, 14:375–398, 2011.

Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science, 6(1):3–5, 2011.

Timothy N Cason and Charles R Plott. Misconceptions and game form recognition: Challenges to theories of revealed preference and framing. Journal of Political Economy, 122(6):1235–1270, 2014.

Gary Charness and Dan Levin. The origin of the winner's curse: A laboratory study. American Economic Journal: Microeconomics, 1(1):207–236, 2009.

Gary Charness and Matthew Rabin. Understanding social preferences with simple tests. Quarterly Journal of Economics, 117(3):817–869, 2002.

Daniel L. Chen and Martin Schonger. Invariance of equilibrium to the strategy method i: Theory. Journal of the Economic Science Association, 10(1):15–30, 2024a.

Daniel L. Chen and Martin Schonger. Invariance of equilibrium to the strategy method ii: Experimental evidence. Journal of the Economic Science Association, 10(1):31–43, 2024b.

Daniel L. Chen, Martin Schonger, and Chris Wickens. oTree – an open-source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance, 9:88–97, 2016.

Swee-Hoon Chuah, Robert Hoffmann, and Jeremy Larner. Elicitation effects in a multi-stage bargaining experiment. Experimental Economics, 17:335–345, 2014.

Simon Columbus and Robert Böhm. Norm shifts under the strategy method. Judgment and Decision Making, 16(5):1267–1289, 2021.

Rachel TA Croson. The disjunction effect and reason-based choice in games. Organizational Behavior and Human Decision Processes, 80(2):118–133, 1999.

Jing L Davis. Cooler heads prevail: An experimental study on the "cooling" effect of the strategy method on agent resentment. Working Paper, 2018.

Giovanni Di Bartolomeo and Stefano Papa. Trust and reciprocity: extensions and robustness of triadic design. Experimental Economics, 19:100–115, 2016.

Lu Dong, Maria Montero, and Alex Possajennikov. Communication, leadership and coordination failure. Theory and Decision, 84:557–584, 2018.

Anna Dreber, Tore Ellingsen, Magnus Johannesson, and David G. Rand. Do people care about social context? Framing effects in dictator games. Experimental Economics, 16(3):349–371, 2013.

Tore Ellingsen, Magnus Johannesson, Johanna Mollerstrom, and Sara Munkhammar. Social framing effects: Preferences or beliefs? Games and Economic Behavior, 76(1):117–130, 2012.

Benjamin Enke. What you see is all there is. Quarterly Journal of Economics, 135(3):1363–1398, 2020.

Ignacio Esponda and Emanuel Vespa. Hypothetical thinking and information extraction in the laboratory. American Economic Journal: Microeconomics, 6(4):180–202, 2014.

Ignacio Esponda and Emanuel Vespa. Endogenous sample selection: A laboratory study. Quantitative Economics, 9(1):183–216, 2018.

Ignacio Esponda and Emanuel Vespa. Contingent thinking and the sure-thing principle: Revisiting classic anomalies in the laboratory. Review of Economic Studies, 91(5):2806–2831, 2024.

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

Erik Eyster and Georg Weizsacker. Correlation neglect in portfolio choice: Lab evidence. Working Paper Available at SSRN 2914526, 2016.

Armin Falk, Ernst Fehr, and Urs Fischbacher. Driving forces behind informal sanctions. Econometrica, 73 (6):2017–2030, 2005.

Lior Fink. Why and how online experiments can benefit information systems research. Journal of the Association for Information Systems, 23(6):1333–1346, 2022.

Urs Fischbacher, Simon Gächter, and Ernst Fehr. Are people conditionally cooperative? Evidence from a public goods experiment. Economics Letters, 71(3):397–404, 2001.

Urs Fischbacher, Simon Gächter, and Simone Quercia. The behavioral validity of the strategy method in public good experiments. Journal of Economic Psychology, 33(4):897–913, 2012.

Guillaume R Fréchette, Kim Sarnoff, and Leeat Yariv. Experimental Economics: Past and future. Annual Review of Economics, 14(1):777–794, 2022.

Daniel Friedman. Monty hall's three doors: Construction and deconstruction of a choice anomaly. American Economic Review, 88(4):933–946, 1998.

Gerd Gigerenzer and Ulrich Hoffrage. How to improve bayesian reasoning without instruction: frequency formats. Psychological Review, 102(4):684, 1995.

Gerd Gigerenzer, Wolfgang Hell, and Hartmut Blank. Presentation and content: The use of base rates as a continuous variable. Journal of Experimental Psychology: Human Perception and Performance, 14(3): 513–525, 1988.

Uri Gneezy, Alex Imas, and Kristóf Madarász. Conscience accounting: Emotion dynamics and social behavior. Management Science, 60(11):2645–2658, 2014.

Joseph K. Goodman and Gabriele Paolacci. Crowdsourcing consumer research. Journal of Consumer Research, 44(1):196–210, 2017.

Werner Güth and Martin G Kocher. More than thirty years of ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature. Journal of Economic Behavior & Organization, 108: 396–409, 2014.

Werner Güth and Reinhard Tietz. Ultimatum bargaining behavior: A survey and comparison of experimental results. Journal of Economic Psychology, 11(3):417–449, 1990.

Werner Güth, Rolf Schmittberger, and Bernd Schwarze. An experimental analysis of ultimatum bargaining. Journal of Economic Behavior & Organization, 3(4):367–388, 1982.

Werner Güth, Steffen Huck, and Wieland Müller. The relevance of equal splits in ultimatum games. Games and Economic Behavior, 37(1):161–169, 2001.

Ingar Haaland, Christopher Roth, and Johannes Wohlfart. Designing information provision experiments. Journal of Economic Literature, 61(1):3–40, 2023.

David J. Hauser, Gabriele Paolacci, and Jesse Chandler. Common concerns with mturk as a participant pool: Evidence and solutions. In F. R. Kardes, P. M. Herr, and N. Schwarz, editors, Handbook of research methods in consumer psychology, pages 319–337. Routledge/Taylor & Francis Group, 2019.

Felix Holzmeister, Magnus Johannesson, Colin F. Camerer, Yan Chen, Teck-Hua Ho, Susan Hoogeveen, and Anna Dreber. Examining the replicability of online experiments selected by a decision market. Nature Human Behaviour, pages 1–15, 2024.

John J. Horton, David G. Rand, and Richard J. Zeckhauser. The online laboratory: conducting experiments in a real labor market. Experimental Economics, 14(3):399–425, 2011.

Nicholas C. Hunt and Andrea M. Scheetz. Using mturk to distribute a survey or experiment: Methodological considerations. Journal of Information Systems, 33(1):43–65, 2019.

Keith Jensen, Josep Call, and Michael Tomasello. Chimpanzees are rational maximizers in an ultimatum game. Science, 318(5847):107–109, 2007.

David Johnson and John Barry Ryan. Amazon mechanical turk workers can provide consistent and economically meaningful data. Southern Economic Journal, 87(1):369–385, 2020.

Jillian Jordan, Kathleen McAuliffe, and David Rand. The effects of endowment size and strategy method on third party punishment. Experimental Economics, 19:741–763, 2016.

Alexandros Karakostas, Nhu Tran, and Daniel John Zizzo. Experimental insights on anti-social behavior: Two meta-analyses. Working Paper, 2022.

Kiryl Khalmetski, Axel Ockenfels, and Peter Werner. Surprising gifts: Theory and laboratory evidence. Journal of Economic Theory, 159:163–208, 2015.

Antonia Krefeld-Schwalb, Eli Rosen Sugerman, and Eric J Johnson. Exposing omitted moderators: Explaining why effect sizes differ in the social sciences. Proceedings of the National Academy of Sciences, 121(12): e2306281121, 2024.

David K Levine. Modeling altruism and spitefulness in experiments. Review of Economic Dynamics, 1(3): 593–622, 1998.

Luyao Li, Xiaobo Zhao, Dong Xie, and Xue Xiao. On difference between direct-response method and strategy method in decision-making: Behavioural and neural evidence in a reward-punishment game. Journal of the Operational Research Society, 75(9):1681–1698, 2024.

Po-Hsuan Lin and Thomas R Palfrey. Cognitive hierarchies in extensive form games. Caltech Social Science Working Paper, 2022.

George Loewenstein. Out of control: Visceral influences on behavior. Organizational Behavior and Human Decision Processes, 65(3):272–292, 1996.

Philippos Louis. The barrel of apples game: Contingent thinking, inferences from observed actions, and strategic heterogeneity. Working Paper, 2015.

Mark J. Machina. Dynamic consistency and non-expected utility models of choice under uncertainty. Journal of Economic Literature, 27(4):1622–1668, 1989.

Alejandro Martínez-Marquina, Muriel Niederle, and Emanuel Vespa. Failures in contingent reasoning: The role of uncertainty. American Economic Review, 109(10):3437–3474, 2019.

Peter Martinsson, Nam Pham-Khanh, and Clara Villegas-Palacio. Conditional cooperation and disclosure in developing countries. Journal of Economic Psychology, 34:148–155, 2013.

William Minozzi and Jonathan Woon. Direct response and the strategy method in an experimental cheap talk game. Journal of Behavioral and Experimental Economics, 85:101498, 2020.

Johannes Moser. Hypothetical thinking and the winner's curse: An experimental investigation. Theory and Decision, 87(1):17–56, 2019.

M. Kathleen Ngangoué and Georg Weizsäcker. Learning from unrealized versus realized prices. American Economic Journal: Microeconomics, 13(2):174–201, 2021.

Muriel Niederle and Emanuel Vespa. Cognitive limitations: Failures of contingent thinking. Annual Review of Economics, 15(1):307–328, 2023.

Axel Ockenfels and Peter Werner. Scale manipulation in dictator games. Journal of Economic Behavior & Organization, 97:138–142, 2014.

Robert J. Oxoby and Kendra N. McLeish. Sequential decision and strategy vector methods in ultimatum bargaining: Evidence on the strength of other-regarding behavior. Economics Letters, 84(3):399–405, 2004.

Gabriele Paolacci and Jesse Chandler. Inside the turk: Understanding mechanical turk as a participant pool. Current Directions in Psychological Science, 23(3):184–188, 2014.

Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. Running experiments on amazon mechanical turk. Judgment and Decision Making, 5(5):411–419, 2010.

Matthew Rabin and Georg Weizsäcker. Narrow bracketing and dominated choices. American Economic Review, 99(4):1508–1543, 2009.

María P Recalde, Arno Riedl, and Lise Vesterlund. Error-prone inference from response time: The case of intuitive generosity in public-good games. Journal of Public Economics, 160:132–147, 2018.

Ernesto Reuben and Sigrid Suetens. Revisiting strategic versus non-strategic cooperation. Experimental Economics, 15:24–43, 2012.

Marcus Roel and Zhuoqiong Chen. Strategy vs. direct-response method: Evidence from a large online experiment. Working Paper, 2024. Available at SSRN: https://ssrn.com/abstract=4735811 or http://dx.doi.org/10.2139/ssrn.4735811.

Alvin E. Roth. Bargaining experiments. In John H. Kagel and Alvin E. Roth, editors, Handbook of Experimental Economics, volume 1. Princeton University Press, Princeton, 1995.

Alvin E Roth, Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. American Economic Review, pages 1068–1095, 1991.

Reinhard Selten. Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In H. Sauermann, editor, Beiträge zur experimentellen Wirtschaftsforschung, pages 136–168. J. C. B. Mohr, Tübingen, 1967.

Eldar Shafir and Amos Tversky. Thinking through uncertainty: Nonconsequential reasoning and choice. Cognitive Psychology, 24(4):449–474, 1992.

Eldar Shafir, Itamar Simonson, and Amos Tversky. Reason-based choice. Cognition, 49(1-2):11–36, 1993.

Stefanie Stantcheva. How to run surveys: A guide to creating your own identifying variation and revealing the invisible. Annual Review of Economics, 15:205–234, 2023.

Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. Science, 211 (4481):453–458, 1981.

Amos Tversky and Eldar Shafir. The disjunction effect in choice under uncertainty. Psychological Science, 3(5):305–310, 1992.

Motoki Watabe, Shigeru Terai, Nahoko Hayashi, and Toshio Yamagishi. Cooperation in the one-shot prisoner's dilemma based on expectations of reciprocity. The Japanese Journal of Experimental Social Psychology, 36(2):183–196, 1996.

Yukun Zhao and Xiaobo Zhao. On elicitation-method effect in game experiments: A competing newsvendor perspective. Journal of the Operational Research Society, 69(4):541–555, 2018.

Yukun Zhao, Xiaobo Zhao, Lihong Wang, Yefen Chen, and Xinmeng Zhang. Does elicitation method matter? behavioral and neuroimaging evidence from capacity allocation game. Production and Operations Management, 25(5):919–934, 2016.

Yukun Zhao, Xiaobo Zhao, and Zuo-Jun Max Shen. The hot-versus-cold effect in a punishment game: A multi-round experimental study. Annals of Operations Research, 268:333–355, 2018.

Strategy vs. Direct-Response Method: Evidence from a Large Online Experiment on Simple Social Dilemmas

- Online Appendix -

Marcus Roel
marcus.roel@nottingham.edu.cn
University of Nottingham Ningbo China

Zhuoqiong Chen
chenzq926@gmail.com
Harbin Institute of Technology, Shenzhen

List of Tables

List of Figures

C. Theory

In this section, we present a simple theoretical framework that motivated our 2×2 between-subject experimental design, in which we vary the elicitation method and manipulate participants' ex-ante belief about player 1's choices by providing them with information about typical player 1 behavior in the same games from previous studies. Our premise is that initial beliefs may have a different effect on player 2's response depending on the elicitation method employed, and that this orthogonal dimension explains why behavior across the two elicitation methods is at times similar and at times different.

Our framework features only one type of player, a player with social preferences who suffers from a lack of conditional thinking. We will show that what we refer to as incomplete continent reasoning generates an interaction effect between the two dimensions we vary in our setup. The goal of this framework is to formalize a potential explanation for the interaction effect using the fewest player types necessary while presenting a mechanism we viewed as most plausible. We acknowledge that other potential behavioral explanations exist and may also account for this effect.[55] Our primary concern when developing this model framework is to make predictions for our two games, the sequential Prisoner's Dilemma (sPD) and the mini Ultimatum Game (mUG) in a (somewhat) concise and simple way. Nevertheless, the fundamental ideas should generalize with appropriate adjustments.[56]

After presenting our model, we will briefly discuss potential extensions in the form of alternative types of preferences such as norm-based, hot/cold, and cold/hot social preferences. As we will see, these preferences will result in a constant shift in behavior either along the dimension of the belief treatment or the elicitation method, but without any interaction between the two. We will also touch upon more general ideas such as dynamic inconsistency that could be at the heart of behavioral difference between elicitation methods and what we can infer from our data about the existence of certain player types.

The Model. Player 1 (he) and player 2 (she) interact in a simple sequential game, where player 1 moves first and player 2 observers his action and responds. Assume that each player only has two choices and denote their respective actions and choice sets by $a_1 \in A_1$ and $a_2 \in A_2$. Player 2's choice set at each decision node of hers, $h = a_1$, is assumed to be identical to the other. For the sPD, $A_1 = \{C, D\}$ and $A_2 = \{c, d\}$.[57] For the mUG, player 1's action denotes his offer, $A_1 = \{0.5, 0.15\}$. Player 2 either accepts or rejects her offer, $A_2 = \{a, r\}$.

Our focus will be on player 2. Her ex-ante belief about player 1's behavior (at the population level) is denoted by σ_1, with $\sigma_1(a_1)$ being the probability with which she thinks a_1 will be played. The likelihood with which player 2 takes action a_2 at node h is expressed by $\sigma_2(a_2|h, \sigma_1)$, which, as we will see, may depend on her ex-ante belief about player 1's choices. Finally, let $e \in \{dr, sm\}$, indicate the elicitation method with which choices are elicited, where dr stands for direct response (method) and sm for strategy method.

A Social Player with Reciprocity Preferences. At the core, player 2 has preferences for reciprocity; she wants to be nice (nasty) to those players who treat her nicely (nastily). As such, she may be willing to lower her own material payoff to increase (reduce) player 1's material payoff. In our two games, this leads

[55] Indeed, in the Pre-Analysis Plan, we highlighted that a signaling model can also generate a similar effect, e.g., if player 2 prefers to signal their social type – either to herself or to other players – at the decision node she views as unlikely. More recently, Chen and Schonger (2024a) made a similar observation in a more general context.

[56] For those who are interested in a more detailed theory, please see an earlier version of this paper (Roel and Chen (2024), which models social-preferences with and without incomplete conditional thinking (and also nests hot/cold preferences) at the utility level in terms of outcome based preferences. The model could also be written in terms of player 1's types, where player 2 may want to give up material payoffs to rewards (punish) the type of players she views as nice (nasty) (Levine, 1998). In such a model, a failure to properly conditionally update can be written similarly to Eyster and Rabin (2005).

[57] The use of small and large letters in the sPD in this part differs slightly from our main text, where no such distinction is made. As we feel it improves this section's readability, we accept this inconsistency.

to preferences for conditional cooperation in the *sPD* and only rejecting unequal offers in the *mUG*. We will implicitly assume utilities $u(\cdot)$ that generate such preferences and denote the respective *fundamentally preferred choice* in response to action a_1 by

$$a_2^*(a_1) = \underset{a_2 \in A_2}{\arg\max}\ u_2(a_2, a_1). \tag{1}$$

Incomplete Conditional Thinking. The key idea is that player 2 suffers from incomplete conditional thinking. In the context of sequential games, incomplete conditional thinking implies that at a particular node $h = a_1$, player 2 may not fully recognize that a_1 has been taken. As a result, her social preferences might be influenced by player 1's alternative choice(s). One way to capture this in a simple way for our two games is as follows: rather than taking her preferred action $a_2^*(a_1)$ in response to a_1, with probability λ_e, player 2 takes the action that is ex-ante optimal based on her ex-ante belief about player 1's choices instead:

$$a_2^*(\sigma_1) = \underset{a_2 \in A_2}{\arg\max} \sum_{a_1 \in A_1} \sigma_1(a_1) \cdot u_2(a_2, a_1) \tag{2}$$

The ex-ante optimal choice, $a_2^*(\sigma_1)$, combines player 2's decision problem at the correct, actual node with that at the alternative node, with the importance of each determined by her ex-ante beliefs. Intuitively, it captures that if player 2 expects player 1 to take the nice (nasty) action, she will have an inherent tendency to react more nicely (nastily) irrespective of the actual choice taken when she fails to fully condition her action on this very choice. This lack of conditioning effectively turns the sequential game into a simultaneous one, as highlighted by equation 2.[58]

Crucially, the frequency of taking the ex-ante optimal action will depend on the elicitation method. When choices are elicited using the strategy method, player 2 must consider multiple hypothetical choices simultaneously. The hypothetical natures of this situation makes it difficult for her to fully recognize that a particular action a_1 must have been taken when she considers a response at node $h = a_1$. In the direct response method, player 1's chosen action is fully observed by player 2. This significantly simplifies her choice as she no longer needs to consider multiple hypothetical nodes. As the actual node becomes more salient, it results in $\lambda_{dr} < \lambda_{sm}$. For convenience, we will assume $\lambda_{dr} = 0$ going forward and omit the subscript for the likelihood of taking the ex-ante optimal action in the strategy method's, i.e., $\lambda \equiv \lambda_{sm}$. Our predictions generalize to a non-zero λ_{dr}.

The question now becomes: when does incomplete conditional thinking result in mistakes at the level of behavior, or, in other words, when does the ex-ante action $a_2^*(\sigma_1)$, that is sometimes chosen instead of the fundamentally preferred conditional choice $a_2^*(a_1)$, lead to an action that is different from this very conditional choice? Given player 2's fundamental preferences for reciprocity, player 2 prefers different choices at each of her two nodes (in both the *sPD* and the *mUG*). As there will generically be only a single choice that is ex-ante optimal, incomplete conditional thinking only affects the player's choices at one of the two nodes, specifically the node where $a_2^*(a_1) \neq a_2^*(\sigma_1)$.

In our experimental design, we exogenously vary player 2's ex-ante belief about player 1's likely choice. In particular, players in the *Selfish Belief treatment* come to hold a stronger belief that player 1 defects

[58] It is not entirely clear whether it is sensible to describe incomplete continent reasoning via equation 2 when an action $a_2 \in A_2$ does *not* operate in a similar way to every $a_1 \in A_1$ (in terms of affecting the player's own as well as the other player's payoff). Admittedly, for the *sPD* and *mUG*, this is but a hypothetical digression as player 2's actions work in the same way regardless of what player 1 does. The current equation may generalize well if incomplete continent reasoning leads to more "muddled thinking" that effectively combines multiple decision nodes. If, on the other hand, player 2 may give up material payoffs to reward (punish) *a type of players* she views as nice (nasty), with incomplete conditional thinking meaning insufficient updating about these types, it is doubtful that it does.

(makes an unequal offer) compared to those in the *Non-Selfish Belief treatment*.[59] Let σ_1^S and σ_1^{NS} denote the induced selfish and non-selfish beliefs. For this belief manipulation to have an effect in our particular model and experiment, it must be that the induced ex-ante optimal actions differ between the two beliefs: $a_2^*(\sigma_1^S) \neq a_2^*(\sigma_1^{NS})$. This assumption is fairly weak. Fundamentally, what is required for this assumption to hold is that (i) it is either true for some subset of player 2, or (ii) the likelihood that a player's choice is influenced by the alternative node differs between the two induced beliefs, i.e., that mistakes are function of the player's ex-ante belief. In the context of our two games, this leads to the following assumption: In the *sPD*, $a_2^*(\sigma_1^S) = d$ and $a_2^*(\sigma_1^{NS}) = c$, whereas in the mUG, $a_2^*(\sigma_1^S) = r$ and $a_2^*(\sigma_1^{NS}) = a$.

Behavioral Predictions. After this long exposition, we are finally ready to analyze how this player behaves across treatments and games. In addition to a difference-in-difference style Table C.13, which best illustrates the theory's implication for our experimental design, we will also list her choices in a slightly more easy-to-read Table C.14. Note that differences across columns or rows in Table C.13 are recorded in the respective Δ column or row, with the first column or row being subtracted from the second column or row. We will begin with deriving behavior for all possible cases before looking at the overall picture.

Table C.13: Predictions for Player 2's Behavior by Treatments - DiD Style

seq. Prisoners' Dilemma: $\sigma_2(c|h)$

	after P1 cooperates			after P1 defects		
Elicitation \ Belief Treatment	Selfish	Non-Selfish	Δ	Selfish	Non-Selfish	Δ
Direct Response	1	1	0	0	0	0
Strategy Method	$1-\lambda$	1	λ	0	λ	λ
Δ	$-\lambda$	0	λ	0	λ	λ

mini Ultimatum Game: $\sigma_2(r|h)$

	after P1 offers 85-15			after P1 offers 50-50		
Elicitation \ Belief Treatmeant	Selfish	Non-Selfish	Δ	Selfish	Non-Selfish	Δ
Direct Response	1	1	0	0	0	0
Strategy Method	1	$1-\lambda$	$-\lambda$	λ	0	$-\lambda$
Δ	0	$-\lambda$	$-\lambda$	λ	0	$-\lambda$

Notes: this table tabulates the frequencies with which the social player 2 is predicted to cooperate in the *sPD* and reject an offer in the mUG. $\lambda > 0$ describes the degree the social player suffers from incomplete conditional thinking. Δ computes the difference across columns or rows, with the first column or row being subtracted from the second column or row.

Since player 2 does not experience problems with incomplete conditional thinking when her choices are elicited using the direct response method, her actions fully reflect her fundamental preferences for conditional cooperation and for rejecting only unfair offers.

This is not the case under the strategy method, however. Here, we have to distinguish between two cases: (1) for any node in the game $h = a_1$ where the ex-ante action $a_2^*(\sigma_1)$ matches the conditionally optimal action $a_2^*(a_1)$, player 2's chosen action will fully reflect her true preferences. Loosely speaking, this is true whenever the induced beliefs point to the very node at which the decision maker is actually making her choice. This occurs at $h = C$ or $h = 0.5$ for the non-selfish belief treatment or at $h = D$ or $h = 0.15$ for the selfish belief treatment, i.e., $\sigma_2(c|C, \sigma_1^{NS}) = 1, \sigma_2(c|D, \sigma_1^S,) = 0$ and $\sigma_2(r|0.15, \sigma_1^S) = 1, \sigma_2(r|0.5, \sigma_1^{NS}) = 0$.

[59] Indeed, our manipulation is so effective that player 2 views defection (an unequal offer) to be more likely than cooperation (an equal offer) in the selfish belief treatment while she sees cooperation (an equal offer) to be more likely than defection (an unequal offer) in the non-selfish belief treatment.

(2) For the nodes where the ex-ante action and conditionally optimal action do not match, player 2's behavior becomes a mixture of these two choices. Consequently, the model predict less cooperation after C in the Selfish Belief treatment and fewer rejections of unfair offers in the Non-Selfish Belief treatment under the strategy method. The model also predicts some cooperation after D in the Non-Selfish belief treatment and that some fair offers are mistakenly rejected in the Selfish Belief treatment. In summary, $\sigma_2(c|C, \sigma_1^S) = 1 - \lambda, \sigma_2(c|D, \sigma_1^S) = \lambda$ and $\sigma_2(r|0.15, \sigma_1^{NS}) = 1 - \lambda, \sigma_2(r|0.5, \sigma_1^S) = \lambda$.

We now turn our eye to the overall picture. Examining the nodes at which we are typically measure non-selfish behavior – after player 1 cooperates or offers an uneven split – our model predicts that behavior across the two elicitation methods may be the same or different depending on player 2's (induced) belief about player 1's behavior as illustrated in the bottom Δ-row of the two games in Table C.13. This pattern arises from the interaction between the orthogonal belief dimension and the elicitation method as indicated by our measure of incomplete conditional thinking λ.

It is also important to note that the predicted differential effect in the mUG goes in the opposite direction compared to the sPD. This opposing effect was part of the reasons that motivated us to select these two games. Furthermore, we also observe that the strategy method exhibits less non-selfish behavior when behavior between to elicitation methods differ. This matches the frequently cited assertion that this method elicits a less emotional (or "cold") response.

Looking at the two other nodes, after player 1 defects (offers an even split), behavior coincides across the elicitation methods and belief treatments, with the exception of the strategy-method under non-selfish beliefs (selfish beliefs). In this case, there is some degree of cooperation (rejections) that arises from confusion stemming from incomplete conditional thinking. Admittedly, we did not expect this prediction be important when designing the experiment.[60] Interestingly, we ended up observing the opposite effect in our experiment, namely less mistakes in the strategy method at these two nodes (and which is constant across belief treatments in the sPD).

Table C.14: Predictions for Player 2's Behavior by Treatments

	sPD: $\sigma_2(c\|h, \sigma_1)$		*mUG:* $\sigma_2(r\|h, \sigma_1)$	
After Player 1:	cooperates	defects	offers 85-15	offers 50-50
	— Social Player —			
Direct Response, Selfish Beliefs	1	0	1	0
Direct Response, Non-Selfish Beliefs	1	0	1	0
Strategy Method, Selfish Beliefs	$1 - \lambda$	0	1	λ
Strategy Method, Non-Selfish Beliefs	1	λ	$1 - \lambda$	0

Notes: this table tabulates the frequencies with which the social player 2 is predicted to cooperate in the sPD and reject an offer in the mUG. $\lambda > 0$ describes the degree the social player suffers from incomplete conditional thinking.

C.1. Extensions and Discussion

In our theoretical framework, we chose to focus on the minimal number of player types necessary to motivate our experimental design. In the context of an actual experiment with real participants, we acknowledge that

[60] With our simple model of mistakes, the frequencies of these mistakes after player 1 defects (offers 50-50), coincides with the differential effect following player 1 cooperating (offering 50-50). We don't expect this prediction to generalize. For example, if a player's preferences for d over c at $h = D$ are relatively stronger than for c over d at $h = C$, something we view as reasonable, it will be easier for "a mistake" to turn $c|C$ into $d|C$ than to turn $d|D$ into $c|D$.

player 2's motivations and preferences may extend beyond this single type. We will now discuss a variety of these.

We begin our discussion with a player who is motivated by (social) *norms*. This player type highlights the direct impact that our belief treatments may have on behavior. Given the importance of such a type to our design, we will cover her in slightly more detail than other types. Imagine a fundamentally selfish player but who is willing to take non-selfish actions to comply with social expectations or norms. Importantly, this player derives an understanding of social norms not from the single observable action from her opponent, but from data about population behavior. In other words, this player takes cues about social norms from the information provided by our belief treatments. While the belief treatments inform player 2 about the average behavior of player 1, they also enable her to infer the typical reactions of the player 2s.

Looking at the *sPD* to illustrate this further, suppose the norm player considers which of the two predominant strategies – conditional cooperation or full defection – is the socially acceptable strategy to adopt. In the Non-Selfish Belief treatment, she concludes that player 2 typically cooperates in response to C, as this behavior motivates the majority of player 1s to cooperate. Thus, she identifies conditional cooperation as the social norm and adopts this strategy. Conversely, in the Selfish Belief treatment, she infers that player 2 typically defects in response to C, which explains why the majority of player 1s choose to defect. As a result, she adopts a strategy of full defection. The norm mechanics work similarly in the *mUG*.[61] In both games, the Non-Selfish Belief treatment induces the norm player to adopt a "non-selfish" strategy, independently of the elicitation method.

While the norm player's behavior is affected by the belief treatment, it is important to recognize that simpler, more common social preferences, such as a preference for efficiency or spiteful preferences, will affect (the frequency of non-selfish) behavior in a way that is independent of the elicitation method and the belief treatment. Fundamentally, these preferences would function similarly as a social player in Table C.13 who has no problems with conditional thinking, i.e., $\lambda = 0$, except that their non-selfish behavior would be expressed at different nodes of the game. Adding a purely selfish player 2, who always defects and accepts any offer, would also not affect our predictions, for it would simply scale down the frequency of non-selfish behavior at the population level.

A common suggestion in the literature to explain potential differences in behavior across elicitation methods is that preferences vary with the method used. For instance, if players are more emotionally reactive when they directly observe their opponent's choice in the direct response method, they would act upon their other-regarding motivations more strongly.[62] Such "hot/cold" players would lead to behavioral differences between the direct response and strategy method that are independent of the belief treatment in Table C.13. For example, this type may act as a conditional cooperator (rejects only uneven offers) in the direct response method but always defects (accepts all offers) in the strategy method. It is clear that this player type alone is insufficient to explain why some experiments find no differences in behavior across elicitation methods, while others do observe such differences.

Admittedly, one could propose a hot/cold player whose social preferences are also influenced by some orthogonal dimension. For instance, a hot/cold norm player, whose behavior will be affected by their ex-ante

[61] Detail: suppose the norm player considers which of the two predominant strategies – accepting any offer or only accepting the equal split – is the socially acceptable strategy. In the Non-Selfish Belief treatment, the player learns that player 2 typically rejects unequal offers, which in turn motivates player 1 to offer an equal split. Consequently, she adopts a strategy of only accepting equal splits. In contrast, in the Selfish Belief treatment, she concludes that player 2 generally accepts unequal offers, which explains why most player 1s make such unequal offers. As a result, she accepts all offers.

[62] These forms of preferences can easily be described by a utility function that combines player 2's standard material payoff $\pi_2(a_2, a_1)$ with an other-regarding component $g(\cdot)$, that may motivate her to lower her own material payoff for alternative concerns, e.g., being nice to people that have treated her nicely, etc. For choices to depend on the elicitation method, all that is required is for the weight attached to $g(\cdot)$ to depend on the elicitation method.

beliefs, would also result in a differential effect and thus to similar predictions as Table C.13.[63] It follows that there exists more than one player type that can generate similar predictions. We do not view this to be a flaw of our theory, however. If anything it reinforces the idea behind our experimental design, namely that player 2's ex-ante beliefs about player 1's behavior is one of the likely dimensions varies across experiments, and that this belief could influence behavior in ways that differ across the elicitation methods employed.

Looking at our experimental results, this conjecture is not supported by the data. Despite strongly shifting player's beliefs, our belief treatment does not differentially affect player 2's behavior across elicitation methods. In fact, in the largest study on elicitation methods to date, our data suggests that the elicitation method appears to have minimal impact on player 2's preferences.

Our findings are based on a between-subject design, in which choices are elicited for some participants with the strategy method and for other participants with the direct response method. As such, our work speaks directly to the majority of the literature, which typically employs a between-subject design when investigating elicitation methods. Additionally, it also addresses the primary concern for experimentalists, who need to decide between adopting either the direct response method or the strategy method for their study. Ideally, this choice should not impact the observed treatment effect, which Brandts and Charness (2011) suggest that it does not, and, if possible, should also not affect the level of behavior itself, which is addressed in our paper.

What our design cannot fully address, however, is whether the lack of differential effect implies that the social types, which are central to our framework, do not suffer from incomplete conditional thinking.[64] [65] Fundamentally, a between-subject design is simply not an ideal design choice for detecting inconsistent behavior as different people may act inconsistently in opposing ways, which can give rise to an appearance of consistency at the population level.[66]

More generally, behavioral differences between the strategy method and direct response method can be understood as *dynamically inconsistent* preferences. A useful way to conceptualize dynamic inconsistency is by a decision maker who devises a plan (based on her current preferences) but deviates from it in the future. In the strategy method, the decision maker makes a choices for all potential nodes of the game without knowing player 1's choice; in other words, she creates a plan. Since experiments utilizing the strategy method do not allow her to adjust her plan following player 1's actual choice, she is fully committed to it. In contrast, when her choices are elicited through the direct response method, it is as if she can revise her plan. Although such a plan was never actually elicited in the direct response method, this concept serves as a useful way to highlight how behavioral differences between the two methods related to dynamic inconsistency.

Returning to our discussion of social types with incomplete conditional thinking, dynamic inconsistency would arise from subjects not fully conditioning on player 1's potential actions in the strategy method,

[63] The differential effect would be limited to the left-hand side of the table, i.e., after P1 cooperates or offers 85-15. We are not aware of this idea being discussed in the literature previously.

[64] As highlighted in Table A.4, the majority of behavior is consistent with the social type (i.e, a conditional cooperator) in *sPD*. In the *mUG*, the majority of players act like a selfish player.

[65] We thank one of our reviewers, Emanuel Vespa, for highlighting this point as well as the relationship to dynamic inconsistency.

[66] One of the classic investigations of inconsistent preferences is the Allais' Paradox (Allais (1953)), where subjects are first asked to choose between a lottery A and B, and then between two different lotteries C and D. The beauty behind this setup is that for an expected utility maximizer, the two choices are identical so that choosing A must imply choosing C, and conversely, B must imply D. We use this example to illustrate why it is incorrect to only look at the overall choice frequencies (which is equivalent to using a between-subject design) instead of calculating whether an individual subject switches from A (B) to D (C). To see this, suppose each choice is taken by exactly 50% of participants in both questions. At first glance, it would appear that behavior is fully consistent with expected utility due to the lack of differences between the two questions. However, this data could be generated by half participants switching from A to D and the other switching from B to C, resulting in a total of 100% of players whose preferences are inconsistent with expected utility. Admittedly, overall frequencies of the respective choices in this experiment typically indicate strong inconsistencies in such an experiment as the majority of subjects switch from one to the other choice.

leading to incorrect plans, but who respond to an observed action by player 1 under the direct response method in line with their true preferences. For our between-subject design not to detect this effect, there must thus exist another player type whose dynamic inconsistency operates in the exact opposite direction of what we predicted, i.e., this player errs in the opposite way in response to our belief-treatment. Moreover, the population frequency of such a player and their error rates would need to jointly match those of our social player. Without a more comprehensive understanding of player motivations and their potential errors, it is difficult to assess whether this is a likely scenario or whether incomplete conditional thinking surrounding other people's actions is not very important for social preferences in our simple games.

Nevertheless, we agree that these questions are best explored with creative designs that observe behavior under the strategy method and the direct response method for every or at least some participants. The designs should be *creative* in the sense that they should avoid potential learning or order effects, especially if such effects work differently across the two elicitation methods. For instance, if the strategy method is played first, participants have an opportunity to think about each contingent state before responding to Player 1's action under the direct-response method. According to the literature on the failure of contingent reasoning that we discussed in footnote 4, simply imagining each contingent state can alter how Player 2 approaches the game and, hence, may alter their behavior in the direct-response treatment. Alternatively, if participants first experience the direct-response method, they may observe their counterpart's actions and update their beliefs about a "norm" of cooperation in the sequential Prisoners' Dilemma. This updated information will carry into any subsequent strategy-method decisions, making it impossible to tell whether a change in behavior reflects dynamic inconsistency or simply updated beliefs. There are other potential reasons why a within-subject design, in which participants choose under the direct-response method before the strategy method, is problematic. First, conscience accounting (Gneezy et al. (2014)) predicts that choosing to defect in response to cooperation first may increase the likelihood that the subject subsequently chooses to cooperate in both nodes of the game under the strategy method. This may happen because acting selfishly may create a threat to conscience, and acting pro-socially in subsequent choices may be considered compensation for the conscience. Second, moral licensing (e.g., Brañas-Garza et al. (2013)) indicates that choosing to cooperate in response to cooperation first may lead the participant to consider themselves "licensed" to be pro-social and decide that it would not hurt to act selfishly just once.

D. Robustness Checks

D.1. Probit and Logit

In this section, we replicate the main tables from our paper, estimating Probit and Logit regressions. We note that the numbers of observations for these results vary (and hence vary from our OLS-regressions) when controls for individual characteristics are included. This is due to the fact that some of the categories for some control variables feature very few observations. As a result, behavior may be perfectly predicted by their respective category dummy. In Table D.15, for example, a single participant said to be between 12-17 years old, and in Table D.17, 4 participants, who prefer to not reveal their age, all cooperated after cooperation. For player 2's behavior in the *mUG*, the effect is slightly more pronounced given that most players behave in the same way. Here, the omitted observation again come from categories with very few observations, such those who used "other" for their gender or preferred not to disclose it, those who are "75 years or older" or prefer not to reveal their age, etc. Unlike in OLS-regressions, Probit and Logit regressions (necessary) omit such observations, while their variation is fully explained by the respective dummies in the OLS regression, which essentially "drops" these observations as well. Given that our results are consistent between estimates that control or not control for characteristics, it is clear that the results are unaffected by this (which is unsurprising as these should be orthogonal to the randomization by design).[67]

More generally, this subsection shows that our main results, Table 4 - 6 (D.15 - D.18) are fully robust to a change in estimator. The same is true for tables 7, 9, and 11 (D.19, D.21-D.25). Signs and significance levels also match fully for Table 8 (D.20) in the sPD. For the *mUG*, they also match for the Strategy Method and the Inattentive dummies, and are qualitatively similar for their respective interactions. This is also true for Table A.8, player 1's behavior and response time (D.16)

[67] Needless to say, we prefer to keep the characteristic dummies as they are (provided by the participants) for transparency reason over arbitrarily reclassifying them into some bigger category.

Table D.15: Robustness. Player 1's Behavior – Probit / Logit

	sPD				mUG			
Dep. Var: P1 cooperates; offers 50-50	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Non-Selfish Belief	0.425***	0.432***	0.683***	0.697***	0.320***	0.322***	0.523***	0.529***
	(0.0460)	(0.0463)	(0.0743)	(0.0751)	(0.0469)	(0.0475)	(0.0769)	(0.0783)
Strategy Method	-0.231***	-0.233***	-0.370***	-0.374***	-0.0971	-0.0826	-0.156	-0.131
	(0.0671)	(0.0681)	(0.108)	(0.110)	(0.0669)	(0.0680)	(0.107)	(0.110)
Non-Selfish × Strategy Method	0.325***	0.326***	0.523***	0.527***	0.177*	0.192*	0.290*	0.307*
	(0.0957)	(0.0966)	(0.156)	(0.158)	(0.0970)	(0.0981)	(0.160)	(0.162)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	4009	4008	4009	4008	4009	4008	4009	4008
Estimator	Probit	Probit	Logit	Logit	Probit	Probit	Logit	Logit

Notes: this table reports estimates from Probit and Logit regressions, with control variables identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table D.16: Robustness. Player 1's Behavior: Response Time – Probit / Logit

	sPD				mUG			
Dep. Var: P1 cooperates; offers 50-50	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ln(Response time)	-0.217***	-0.293***	-0.355***	-0.480***	0.0729**	-0.0157	0.123**	-0.0216
	(0.0306)	(0.0330)	(0.0505)	(0.0549)	(0.0338)	(0.0360)	(0.0565)	(0.0598)
Non-Selfish Belief	0.419***	0.427***	0.676***	0.692***	0.320***	0.322***	0.523***	0.530***
	(0.0462)	(0.0466)	(0.0748)	(0.0759)	(0.0470)	(0.0475)	(0.0770)	(0.0783)
Strategy Method	-0.218***	-0.209***	-0.352***	-0.339***	-0.0972	-0.0823	-0.156	-0.131
	(0.0677)	(0.0688)	(0.109)	(0.112)	(0.0669)	(0.0681)	(0.107)	(0.110)
Non-Selfish × Strategy Method	0.315***	0.312***	0.507***	0.507***	0.178*	0.192*	0.292*	0.307*
	(0.0961)	(0.0972)	(0.157)	(0.159)	(0.0971)	(0.0981)	(0.160)	(0.162)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	4009	4008	4009	4008	4009	4008	4009	4008
Estimator	Probit	Probit	Logit	Logit	Probit	Probit	Logit	Logit

Notes: this table reports estimates from Probit and Logit regressions, with control variables identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table D.17: Robustness. Player 2's Behavior in sPD – Probit / Logit

	after P1 cooperates				after P1 defects			
Dep. Var: Player 2 cooperates	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Non-Selfish Belief	0.117*	0.131**	0.190*	0.215**	0.115	0.101	0.198	0.170
	(0.0618)	(0.0626)	(0.100)	(0.102)	(0.0784)	(0.0797)	(0.135)	(0.138)
Strategy Method	0.0922	0.108	0.150	0.178	-0.383***	-0.471***	-0.698***	-0.876***
	(0.0755)	(0.0770)	(0.123)	(0.126)	(0.0917)	(0.0971)	(0.170)	(0.180)
Non-Selfish × Strategy Method	-0.130	-0.162	-0.212	-0.263	-0.0888	-0.0460	-0.148	-0.0543
	(0.104)	(0.106)	(0.170)	(0.175)	(0.132)	(0.136)	(0.243)	(0.250)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	2722	2718	2722	2718	2247	2242	2247	2242
Estimator	Probit	Probit	Logit	Logit	Probit	Probit	Logit	Logit

Notes: this table reports estimates from Probit and Logit regressions, with control variables identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table D.18: Robustness. Player 2's Behavior in mUG – Probit / Logit

	after P1 offers 85-15				after P1 offers 50-50			
Dep. Var: Player 2 rejects	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Non-Selfish Belief	0.156	0.155	0.297	0.292	-0.272**	-0.342***	-0.638**	-0.707***
	(0.102)	(0.105)	(0.195)	(0.198)	(0.113)	(0.120)	(0.266)	(0.270)
Strategy Method	0.208**	0.215**	0.394**	0.421**	-0.641***	-0.746***	-1.593***	-1.816***
	(0.1000)	(0.102)	(0.190)	(0.193)	(0.198)	(0.234)	(0.530)	(0.556)
Non-Selfish × Strategy Method	-0.191	-0.219	-0.363	-0.420	0.666***	0.761***	1.654**	1.813***
	(0.144)	(0.147)	(0.271)	(0.275)	(0.249)	(0.286)	(0.645)	(0.680)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	1999	1957	1999	1957	2970	2625	2970	2625
Estimator	Probit	Probit	Logit	Logit	Probit	Probit	Logit	Logit

Notes: this table reports estimates from Probit and Logit regressions, with control variables identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table D.19: Robustness. Player 2's Behavior: Mistakes and Beliefs – Probit / Logit

	sPD: after P1 defects				mUG: after P1 offers 50-50			
Dep. Var: Player 2 makes mistake	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Non-Selfish Belief	0.0107	-0.00829	0.0213	-0.0230	-0.0412	-0.0943	-0.130	-0.172
	(0.0955)	(0.0988)	(0.163)	(0.173)	(0.124)	(0.140)	(0.289)	(0.311)
Belief Player 1 cooperates	0.275*	0.268*	0.468*	0.495*				
	(0.144)	(0.150)	(0.245)	(0.262)				
Belief Player 1 offers 85-15					0.575***	0.620***	1.282***	1.248***
					(0.183)	(0.203)	(0.408)	(0.459)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	1298	1293	1298	1293	2021	1766	2021	1766
Estimator	Probit	Probit	Logit	Logit	Probit	Probit	Logit	Logit

Notes: this table reports estimates from Probit and Logit regressions, with control variables identical to those in Table 4, for player 2s in the Direct Response method. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table D.20: Robustness. Player 2's Behavior: Mistakes and Inattention – Probit / Logit

	sPD: after P1 defects				mUG: after P1 offers 50-50			
Dep. Var: Player 2 makes mistake	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Strategy Method	-0.358***	-0.358***	-0.660***	-0.659***	-0.165	-0.167	-0.488	-0.492
	(0.0851)	(0.0850)	(0.160)	(0.160)	(0.185)	(0.184)	(0.467)	(0.467)
Inattentive	0.443***		0.766***		0.622***		1.379***	
	(0.0846)		(0.145)		(0.124)		(0.281)	
Inattentive × Strategy Method	-0.337**		-0.559**		-0.306		-0.482	
	(0.139)		(0.254)		(0.248)		(0.606)	
Inattentive (Q123)		0.374***		0.654***		0.527***		1.215***
		(0.0911)		(0.156)		(0.134)		(0.304)
Inattentive (Q4)		0.755***		1.259***		0.902***		1.835***
		(0.171)		(0.279)		(0.192)		(0.406)
Inattentive (Q123) × Strategy Method		-0.410**		-0.712**		-0.276		-0.483
		(0.171)		(0.319)		(0.298)		(0.725)
Inattentive (Q4) × Strategy Method		-0.521**		-0.819**		-0.525		-0.790
		(0.218)		(0.375)		(0.325)		(0.754)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2242	2242	2242	2242	2625	2625	2625	2625
Estimator	Probit	Probit	Logit	Logit	Probit	Probit	Logit	Logit

Notes: this table reports estimates from Probit and Logit regressions, with control variables are identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Finally, we also estimate Table 8 (as well as Table D.20 in one go) without controls. These estimates were skipped despite all other tables reporting results with and without controls in order to limit the table size in the main part of the paper, and, as we had seen and will again see, because such controls have little effect on the estimates themselves.

Table D.21: Robustness. Player 2's Behavior: Mistakes and Inattention – No Controls

	sPD: after P1 defects						mUG: after P1 offers 50-50					
Dep. Var: P2 makes mistake	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Strategy Method	-0.0767***	-0.0767***	-0.348***	-0.348***	-0.640***	-0.640***	-0.00564	-0.00564	-0.160	-0.160	-0.457	-0.457
	(0.0179)	(0.0179)	(0.0840)	(0.0840)	(0.158)	(0.158)	(0.00524)	(0.00524)	(0.181)	(0.181)	(0.466)	(0.466)
Inattentive	0.130***		0.429***		0.736***		0.0480***		0.629***		1.464***	
	(0.0273)		(0.0835)		(0.141)		(0.0104)		(0.118)		(0.273)	
Inattentive × SM	-0.102***		-0.295**		-0.483*		-0.0321**		-0.277		-0.529	
	(0.0352)		(0.138)		(0.250)		(0.0140)		(0.244)		(0.599)	
Inattentive (Q123)		0.102***		0.352***		0.609***		0.0356***		0.518***		1.243***
		(0.0287)		(0.0899)		(0.153)		(0.0103)		(0.128)		(0.296)
Inattentive (Q4)		0.276***		0.778***		1.296***		0.108***		0.991***		2.146***
		(0.0651)		(0.169)		(0.276)		(0.0323)		(0.182)		(0.374)
Inattentive (Q123) × SM		-0.103***		-0.377**		-0.654**		-0.0234		-0.254		-0.499
		(0.0388)		(0.169)		(0.316)		(0.0156)		(0.292)		(0.717)
Inattentive (Q4) × SM		-0.215***		-0.497**		-0.776**		-0.0882**		-0.554*		-1.028
		(0.0724)		(0.217)		(0.371)		(0.0352)		(0.313)		(0.721)
Controls	No	No	No	No	No	No	No	No	No	No	No	No
Observations	2247	2247	2247	2247	2247	2247	2970	2970	2970	2970	2970	2970
Estimator	OLS	OLS	Probit	Probit	Logit	Logit	OLS	OLS	Probit	Probit	Logit	Logit

Notes: this table reports estimates from OLS, Probit and Logit regressions for Table 8 and D.20 without controls. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

In Table D.22 and D.23 we report the OLS estimates of Table 9 without controls as well as the respective Probit and Logit estimates. Estimates across OLS, Probit and Logit are consistent. In the *mUG*, response time estimates are not affected by control variables. The only noteworthy difference in this regard is player 2's choice in the *sPD* after player 1 has cooperated, where we observe a positive relationship between response time and conditional cooperation without but not with controls (see Table 9). In other words, the part of the preferences towards conditional cooperation that are captured by our participants' characteristics are positively correlated with the time it takes these participants to make their choice.

Table D.22: Robustness. Player 2's Behavior in sPD: Response Time – OLS w/o Controls, Probit / Logit

	after P1 cooperates					after P1 defects				
Dep. Var: Player 2 cooperates	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
ln(Response time)	0.0361**	0.0973**	-0.0141	0.159**	-0.0235	-0.0246*	-0.0924*	-0.141**	-0.166*	-0.244**
	(0.0149)	(0.0404)	(0.0445)	(0.0661)	(0.0734)	(0.0138)	(0.0509)	(0.0550)	(0.0930)	(0.0987)
Non-Selfish Belief	0.0476**	0.128**	0.129**	0.209**	0.212**	0.0359	0.122	0.111	0.209	0.187
	(0.0230)	(0.0620)	(0.0628)	(0.101)	(0.103)	(0.0235)	(0.0785)	(0.0799)	(0.135)	(0.139)
Strategy Method	0.0219	0.0581	0.113	0.0945	0.187	-0.0839***	-0.356***	-0.428***	-0.646***	-0.798***
	(0.0284)	(0.0768)	(0.0785)	(0.125)	(0.129)	(0.0212)	(0.0933)	(0.0988)	(0.174)	(0.183)
Non-Selfish × Strategy Method	-0.0504	-0.136	-0.161	-0.222	-0.261	-0.0322	-0.0991	-0.0627	-0.169	-0.0878
	(0.0384)	(0.104)	(0.106)	(0.170)	(0.175)	(0.0315)	(0.132)	(0.137)	(0.244)	(0.252)
Controls	No	No	Yes	No	Yes	No	No	Yes	No	Yes
Observations	2722	2722	2718	2722	2718	2247	2247	2242	2247	2242
Estimator	OLS	Probit	Probit	Logit	Logit	OLS	Probit	Probit	Logit	Logit

Notes: this table reports estimates from OLS regressions identical to Table 9 for the *sPD* but without controls, as well as the respective Probit and Logit regressions. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table D.23: Robustness. Player 2's Behavior in mUG: Response Time – OLS w/o Controls, Probit / Logit

	after P1 offers 85-15					after P1 offers 50-50				
Dep. Var: Player 2 rejects	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
ln(Response time)	-0.0240	-0.101	-0.0585	-0.222	-0.142	-0.0290***	-0.411***	-0.240**	-1.111***	-0.611**
	(0.0168)	(0.0716)	(0.0771)	(0.155)	(0.160)	(0.00801)	(0.111)	(0.104)	(0.272)	(0.240)
Non-Selfish Belief	0.0314	0.160	0.159	0.308	0.302	-0.0199**	-0.312***	-0.355***	-0.644**	-0.710***
	(0.0201)	(0.103)	(0.105)	(0.195)	(0.199)	(0.00779)	(0.114)	(0.119)	(0.268)	(0.269)
Strategy Method	0.0494**	0.244**	0.235**	0.469**	0.468**	-0.0195**	-0.441**	-0.627***	-1.071**	-1.543***
	(0.0208)	(0.103)	(0.104)	(0.195)	(0.198)	(0.00794)	(0.205)	(0.231)	(0.532)	(0.548)
Non-Selfish × Strategy Method	-0.0420	-0.212	-0.232	-0.399	-0.448	0.0310***	0.616**	0.720**	1.547**	1.724**
	(0.0300)	(0.144)	(0.147)	(0.271)	(0.275)	(0.0111)	(0.257)	(0.288)	(0.640)	(0.675)
Controls	No	No	Yes	No	Yes	No	No	Yes	No	Yes
Observations	1999	1999	1957	1999	1957	2970	2970	2625	2970	2625
Estimator	OLS	Probit	Probit	Logit	Logit	OLS	Probit	Probit	Logit	Logit

Notes: this table reports estimates from OLS regressions identical to Table 9 for the *mUG* but without controls, as well as the respective Probit and Logit regressions. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

In Table D.24 and D.25 we report the OLS estimates of Table 11 without controls as well as the respective Probit and Logit estimates. Estimates across OLS, Probit and Logit are consistent for both games.

Table D.24: Robustness. Player 2's Behavior in sPD for Attentive and Inattentive Players – OLS w/o Controls, Probit / Logit

	after P1 cooperates					after P1 defects				
Dep. Var: Player 2 cooperates	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Sample: Attentive										
Non-Selfish Belief	0.0305	0.0837	0.0884	0.137	0.148	0.0115	0.0431	0.0278	0.0763	0.0332
	(0.0273)	(0.0748)	(0.0761)	(0.122)	(0.125)	(0.0260)	(0.0972)	(0.0986)	(0.172)	(0.176)
Strategy Method	0.0159	0.0434	0.0581	0.0708	0.0943	-0.0840***	-0.392***	-0.387***	-0.733***	-0.754***
	(0.0353)	(0.0963)	(0.0983)	(0.157)	(0.162)	(0.0241)	(0.122)	(0.126)	(0.234)	(0.243)
Non-Selfish × Strategy Method	-0.0548	-0.150	-0.188	-0.244	-0.306	0.0106	0.0779	0.0564	0.158	0.166
	(0.0476)	(0.129)	(0.133)	(0.211)	(0.219)	(0.0361)	(0.169)	(0.173)	(0.319)	(0.328)
Controls	No	No	Yes	No	Yes	No	No	Yes	No	Yes
Observations	1826	1826	1816	1826	1816	1529	1529	1509	1529	1509
Estimator	OLS	Probit	Probit	Logit	Logit	OLS	Probit	Probit	Logit	Logit
Sample: Inattentive										
Non-Selfish Belief	0.0725*	0.190*	0.239**	0.306*	0.390**	0.0792	0.223	0.212	0.366	0.372
	(0.0422)	(0.110)	(0.112)	(0.178)	(0.183)	(0.0490)	(0.137)	(0.145)	(0.225)	(0.246)
Strategy Method	0.0797*	0.209*	0.212*	0.338*	0.352*	-0.137***	-0.485***	-0.661***	-0.848***	-1.126***
	(0.0470)	(0.124)	(0.128)	(0.200)	(0.210)	(0.0399)	(0.144)	(0.151)	(0.256)	(0.271)
Non-Selfish × Strategy Method	-0.0386	-0.0956	-0.135	-0.152	-0.217	-0.0955	-0.298	-0.236	-0.507	-0.456
	(0.0660)	(0.179)	(0.184)	(0.291)	(0.304)	(0.0613)	(0.219)	(0.230)	(0.391)	(0.415)
Controls	No	No	Yes	No	Yes	No	No	Yes	No	Yes
Observations	896	896	895	896	895	718	718	708	718	708
Estimator	OLS	Probit	Probit	Logit	Logit	OLS	Probit	Probit	Logit	Logit

Notes: this table reports estimates from OLS regressions identical to Table 11 for the *sPD* but without controls, as well as the respective Probit and Logit regressions. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table D.25: Robustness. Player 2's Behavior in mUG for Attentive and Inattentive Players – OLS w/o Controls, Probit / Logit

	after P1 offers 85-15					after P1 offers 50-50				
Dep. Var: Player 2 cooperates	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Sample: Attentive										
Non-Selfish Belief	0.0315	0.159	0.148	0.302	0.261	-0.0197***	-0.540***	-0.593***	-1.403***	-1.489***
	(0.0242)	(0.120)	(0.123)	(0.229)	(0.231)	(0.00694)	(0.190)	(0.192)	(0.512)	(0.518)
Strategy Method	0.0203	0.106	0.129	0.203	0.231	-0.0190**	-0.504*	-0.541*	-1.302*	-1.419*
	(0.0248)	(0.127)	(0.129)	(0.243)	(0.244)	(0.00811)	(0.274)	(0.284)	(0.751)	(0.795)
Non-Selfish × Strategy Method	-0.0509	-0.259	-0.275	-0.494	-0.517	0.0247**	0.735**	0.808**	1.931*	2.119**
	(0.0356)	(0.181)	(0.185)	(0.345)	(0.351)	(0.0106)	(0.371)	(0.376)	(1.009)	(1.027)
Controls	No	No	Yes	No	Yes	No	No	Yes	No	Yes
Observations	1346	1346	1317	1346	1317	2009	2009	1620	2009	1620
Estimator	OLS	Probit	Probit	Logit	Logit	OLS	Probit	Probit	Logit	Logit
Sample: Inattentive										
Non-Selfish Belief	0.0273	0.149	0.135	0.288	0.252	-0.0191	-0.152	-0.235	-0.318	-0.347
	(0.0360)	(0.194)	(0.204)	(0.374)	(0.409)	(0.0202)	(0.159)	(0.167)	(0.331)	(0.346)
Strategy Method	0.0745**	0.358**	0.416**	0.674**	0.835**	-0.0642***	-0.877***	-1.055***	-2.058***	-2.108***
	(0.0347)	(0.170)	(0.179)	(0.325)	(0.351)	(0.0176)	(0.288)	(0.314)	(0.748)	(0.756)
Non-Selfish × Strategy Method	0.00389	-0.0312	-0.0724	-0.0789	-0.165	0.0553**	0.791**	0.963**	1.874**	1.849**
	(0.0554)	(0.250)	(0.258)	(0.467)	(0.495)	(0.0275)	(0.355)	(0.390)	(0.875)	(0.895)
Controls	No	No	Yes	No	Yes	No	No	Yes	No	Yes
Observations	653	653	640	653	640	961	961	760	961	760
Estimator	OLS	Probit	Probit	Logit	Logit	OLS	Probit	Probit	Logit	Logit

Notes: this table reports estimates from OLS regressions identical to Table 11 for the *mUG* but without controls, as well as the respective Probit and Logit regressions. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

D.2. Order Effects

In this section, we discuss the presence of order effects in our data (which had, for example, been highlighted previously in footnote 21). The general takeaway is that order effects exist and that our analysis restricted to task 1 yields qualitatively similar results despite the reduction in sample size by 50%.

Table D.26 provides the average beliefs and choice frequencies for all four treatment groups by task-order, with significant differences between task 1 and task 2 for a given belief/elicitation group highlighted by the respective star-indicators in the task 1 columns. The null-hypothesis that behavior (and beliefs) is constant across tasks is rejected. Moreover, behavior is neither constant for the direct-response method nor the strategy method for either player.[68] Consequently, we repeat our analysis using behavior of task 1 only in Tables D.27 to D.31 below.

For player 1 (Table D.27), the treatment effect estimates are very similar.[69] If anything, the impact of the strategy method in the *mUG* follows more closely that of the *sPD* when the data is restricted to task 1 – which we previously interpreted as a higher degree of strategic sophistication.

Regarding player 2's behavior in the *sPD* (Table D.28), we see that treatment effects remain similar, with all having the same sign. Crucially, the effect of the strategy method in reducing mistakes remains strongly significant. Somewhat surprisingly, the impact of the non-selfish belief treatment is no longer significant after cooperation (we interpreted this small yet significant effect previously as a 'norms') but is significant after defection. This effect may be interpreted in line with our previous argument regarding mistakes (see our analysis and respective discussion of Table 7).

For player 2's behavior in the *mUG* (Table D.31), the strategy method dummy remains significant after unequal offers (and, as before, with other estimates being insignificant). After equal splits, all estimates are qualitatively similar to before. In light of their small coefficients, and the 50% reduction in sample size, all but the strategy method dummy with controls are no longer significant, however.

We also repeat the analysis of mistakes (Table D.30 and D.31), which result in a similar takeaway as before.

Table D.26: Beliefs and Behavior by Task Order

	Task 1				Task 2			
	Selfish		Non-Selfish		Selfish		Non-Selfish	
	DR	SM	DR	SM	DR	SM	DR	SM
Beliefs about Player 1								
Belief Player 1 cooperates	0.34	0.33	0.75**	0.74	0.34	0.32	0.73	0.74
Belief Player 1 offers 50-50	0.32***	0.30	0.75***	0.76***	0.35	0.33	0.73	0.71
Behavior: Player 1								
Player 1 cooperates	0.53**	0.45**	0.67	0.69	0.46	0.35	0.65	0.70
Player 1 offers 50-50	0.62**	0.56	0.77***	0.79**	0.57	0.56	0.66	0.69
Behavior: Player 2								
Player 2 cooperates after C	0.68***	0.71**	0.70**	0.69*	0.57	0.62	0.64	0.62
Player 2 cooperates after D	0.22	0.12	0.30***	0.13	0.19	0.11	0.19	0.11
Player 2 rejects 50-50	0.03*	0.01	0.02	0.01	0.05	0.00	0.03	0.03
Player 2 rejects 85-15	0.10	0.17	0.09**	0.16	0.10	0.12	0.16	0.11

Notes: *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively, testing for differences between the respective columns of among the pre-Covid and Covid samples.

[68] We also note that the direct-respond appears to feature more occurrences of significant difference between tasks for both player 1 (despite the identical set of information for player 1 in both elicitation methods) and player 2. The latter leads us to explore the effect of prior experience in the direct response method at the end of this section (Table D.32). We find no indication for this.

[69] We will usual refer to (two) estimates as being similar when they are (i) both statistically significant (with similar signs), or (ii) the directional effect is at least similar (subject to neither being too close to zero).

Table D.27: Robustness. Player 1's Behavior – Task 1

	sPD		UG	
Dep. Var: P1 cooperates; offers 50-50	(1)	(2)	(3)	(4)
Non-Selfish Belief	0.140***	0.144***	0.141***	0.138***
	(0.0248)	(0.0248)	(0.0233)	(0.0234)
Strategy Method	-0.0744**	-0.0798**	-0.0616*	-0.0584
	(0.0374)	(0.0382)	(0.0371)	(0.0371)
Non-Selfish × Strategy Method	0.0942*	0.0937*	0.0898*	0.0906*
	(0.0506)	(0.0510)	(0.0480)	(0.0479)
Controls	No	Yes	No	Yes
Observations	2011	2011	1998	1998

Notes: this table reports estimates from OLS regressions based only on task 1 data. Control variables for individual characteristics include gender, age, income, highest education, prior participation in experiments, mistakes in instruction test, and the Covid sample dummy. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table D.28: Robustness. Player 2's Behavior in sPD – Task 1

	after P1 cooperates		after P1 defects	
Dep. Var: Player 2 cooperates	(1)	(2)	(3)	(4)
Non-Selfish Belief	0.0211	0.0260	0.0728**	0.0663*
	(0.0308)	(0.0307)	(0.0363)	(0.0359)
Strategy Method	0.0296	0.0397	-0.103***	-0.129***
	(0.0380)	(0.0384)	(0.0312)	(0.0319)
Non-Selfish × Strategy Method	-0.0356	-0.0559	-0.0664	-0.0489
	(0.0523)	(0.0529)	(0.0474)	(0.0472)
Controls	No	Yes	No	Yes
Observations	1391	1391	1074	1074

Notes: this table reports estimates from OLS regressions based only on task 1 data, with control variables identical to those in Table D.27. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table D.29: Robustness. Player 2's Behavior in UG – Task 1

	after P1 offers 85-15		after P1 offers 50-50	
Dep. Var: Player 2 rejects	(1)	(2)	(3)	(4)
Non-Selfish Belief	-0.0121	-0.0129	-0.0140	-0.0139
	(0.0277)	(0.0285)	(0.00918)	(0.00927)
Strategy Method	0.0618**	0.0724**	-0.0168	-0.0229**
	(0.0299)	(0.0299)	(0.0105)	(0.0113)
Non-Selfish × Strategy Method	0.00421	-0.0104	0.0144	0.0208
	(0.0437)	(0.0438)	(0.0137)	(0.0148)
Controls	No	Yes	No	Yes
Observations	953	953	1551	1551

Notes: this table reports estimates from OLS regressions based only on task 1 data, with control variables identical to those in Table D.27. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table D.30: Robustness. Player 2's Behavior: Mistakes and Beliefs – Task 1

	sPD: after P1 defects		mUG: after P1 offers 50-50	
Dep. Var: Player 2 makes mistake	(1)	(2)	(3)	(4)
Non-Selfish Belief	0.0450	0.0376	0.00123	0.000473
	(0.0452)	(0.0443)	(0.0120)	(0.0124)
Belief Player 1 cooperates	0.0697	0.0750		
	(0.0662)	(0.0652)		
Belief Player 1 offers 85-15			0.0703***	0.0687***
			(0.0241)	(0.0238)
Controls	No	Yes	No	Yes
Observations	602	602	947	947

Notes: this table reports estimates from OLS regressions based only on task 1 data, with control variables identical to those in Table D.27, for player 2s in the direct-response method. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table D.31: Robustness. Player 2's Behavior: Mistakes and Inattention – Task 1

	sPD: after P1 defects		mUG: after P1 offers 50-50	
Dep. Var: Player 2 makes mistake	(1)	(2)	(3)	(4)
Strategy Method	-0.0974***	-0.0974***	-0.00162	-0.00161
	(0.0272)	(0.0273)	(0.00729)	(0.00729)
Inattentive	0.163***		0.0356***	
	(0.0425)		(0.0124)	
Inattentive × Strategy Method	-0.157***		-0.0308*	
	(0.0532)		(0.0175)	
Inattentive (Q123)		0.136***		0.0230**
		(0.0452)		(0.0117)
Inattentive (Q4)		0.295***		0.102**
		(0.0989)		(0.0460)
Inattentive (Q123) × Strategy Method		-0.155***		-0.0252
		(0.0593)		(0.0166)
Inattentive (Q4) × Strategy Method		-0.263**		-0.0887*
		(0.108)		(0.0503)
Controls	Yes	Yes	Yes	Yes
Observations	1074	1074	1551	1551

Notes: this table reports estimates from OLS regressions, with control variables identical to those in Table D.27. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Finally, in Table D.32, we explore the role of past experience on behavior for task 2. One of the key difference between task 1 and task 2 in the direct-response treatment is that player 2 gets to observe player 1's action in the prior round. If this experience influences her choices in the second round (despite the fact that the person behind player 1 is different for task 2), then order effects will shape player 2's behavior in the direct response method in a different way than in the strategy method for task 2. We find no indication for this when we relate player 2's behavior (cooperate/reject) in task 2 at any node of the game to a dummy that indicates whether player 1 takes the Non-Selfish action (cooperates in the *sPD* or offers 50-50 in the mUG) for task 1.[70]

Table D.32: Regression Table: Player 2 Behavior based on their Experience in the first Task

	Player 1's choice in task 2:							
	cooperates		defects		offers 85-15		offers 50-50	
Dep. Var: Player 2 cooperates / rejects	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
P1 takes Non-Selfish Action in t=1	0.0364	0.0351	-0.0210	-0.0108	0.0306	0.0258	0.00769	0.00220
	(0.0386)	(0.0381)	(0.0318)	(0.0317)	(0.0278)	(0.0279)	(0.0135)	(0.0133)
Non-Selfish Belief	0.0660*	0.0754**	-0.000544	-0.0153	0.0577**	0.0542*	-0.0244*	-0.0237*
	(0.0345)	(0.0346)	(0.0305)	(0.0308)	(0.0285)	(0.0289)	(0.0136)	(0.0135)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	854	854	696	696	574	574	947	947

Notes: this table reports estimates from OLS regressions for player 2 behavior in the second task in the direct response treatment(s) for all possible choices of player 1. Control variables are identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

[70] We also checked if player 1's behavior "matched" the belief-treatment, but found no evidence that this to affect player 2's behavior for task 2 (Results available upon request).

D.3. Attrition

Like many online experiments, our study required participants to complete a short understanding test after the experimental instructions. Participants were informed that they must correctly answer at least 3 out of 4 questions in order to proceed. This standard procedure served two important purposes. First, it gives participants a strong incentive to carefully read and understand the instructions (or re-read the instructions during the test as they were also displayed at the bottom of the test screen). Second, it eliminates those participants from the experiment who either show little care or for whom the instructions are too difficult, e.g., due to poor language skills. This procedure also excludes non-human participants, i.e., bots, which weren't sophisticated enough yet at the time we conducted this study. Overall, this procedure aims to ensure high-quality data. It is especially important for online experiments, where the researcher has less control over participants compared to traditional laboratory settings.

By design, this enforced dropout will induce attrition. Following good practices for online (survey) experiments (e.g., Stantcheva (2023)), we will carefully analyze this issue in this section. After explaining our understanding test in more detail, we will document that it results in *some differential attrition across treatments*. We then provide evidence that this effect has only minor consequences for our main results, highlighting that our results are robust to attrition.

There are four questions in the understanding test. The first three were identical for everyone. They asked about the participant's anonymity (question 1), their assigned player role (question 2), and whether they will be rematched with the same person in task 2 (question 3).[71] In question 4, we asked participants to identify their own as well as the other player's payoff for a given action profile or strategy from a payoff table. For this question, everyone took on the role of player 1.[72] Payoffs were selected separately for each player using a dropdown menu. As the experimental instructions differed with the elicitation method (we needed to explain how their choices would be elicited after all), the most natural way to ask this question was to rely on the action profile in the direct response and the strategy profile in the strategy method. In the direct response method, we asked participants to identify the payoffs if "you choose **A** and the second mover responds with **C**". In the strategy method, we relied on the following phrase:

"Suppose you choose **A** and the second mover takes the following conditional choices:
- In response to **A**, the second mover chooses **C**
- In response to **B**, the second mover chooses **D**"

We considered the two versions to be of similar difficulty – especially for careful participants who diligently take the understanding test, and who are able to re-read the instructions at the bottom of the test-page if they were unsure.

Table D.33 provides an overview of the frequency with which any participant who takes and submits the understanding test gets a particular question correct and how many mistakes they make in total.[73] Participants perform decently across the four questions. Surprisingly, participants perform worst on question number 2, which asked whether their player role changes across the two tasks. In general, participants demonstrated a good understanding of the payoff question (question 4). Unlike the binary choice questions

[71] The exact wording is (Q1) "Do you know the identity, i.e., their MTurk ID or any other personal information, of the participant you are matched with?" (correct answer: no), (Q2) "Imagine you assume the role of the second mover in task 1. Will your role change in task 2?" (correct answer: no), (Q3) "In the two decision tasks, will you interact with the same Mechanical Turk worker?" (correct answer: no).

[72] Participants would only learn their actual role once the real game started.

[73] In this section, we omit observations from participants that completed the experiment but did not qualify for payments. This typically occurred if they participated more than once or if they did not submit the completion code shown at the end of the experiment in oTree on mTurk (meaning they could not be payed for they did not formally complete their task mTurk). 30 observations are dropped as a result.

in question 1 to 3, this question was almost impossible to get correct by accident since participants had to select a payoff for each player from a dropdown menu with options $\{1, 2, \ldots, 6\}$. Overall, 83.4% passed our hurdle of making at most 1 mistake.

Table D.33: Frequency of Correct Questions and Number of Mistakes for Everyone who Submits Understanding Test

	Mean	Obs.
Understanding test: Q1 correct	0.893	9738
Understanding test: Q2 correct	0.742	9738
Understanding test: Q3 correct	0.832	9738
Understanding test: Q4 correct	0.815	9738
Understanding test fully correct	0.571	9738
Understanding test: 1 mistake	0.263	9738
Understanding test: 2 mistakes	0.081	9738
Understanding test: 3 mistakes	0.048	9738
Understanding test: 4 mistakes	0.037	9738

Notes: the statistics in this table are based on data from every person who submits their answers in the understanding test, regardless of whether or not they passed it and were allowed (not allowed) to proceed to the main task

Table D.34 tabulates the frequency of correct questions and number of mistakes by elicitation methods.[74] While question 1, 2, and 3 – the questions that are identical for all subjects – are answered correctly with similar frequencies, this is not true for question 4 – the question that differs marginally between treatments. Fewer participants get question 4 correct in the strategy method.[75]

It is worth emphasizing that it is not immediately obvious from the data whether question 4 itself is more difficult under the strategy method or whether (the instructions about) the strategy method is more difficult to understand.[76]

To understand the implications of having a relatively more difficult question in the strategy method, we now turn to the conditional dropout rates of anyone who enters a particular stage of the experiment, Table D.35. In addition to the dropout rates, the table also reports the total number of participants that enter each respective part of the experiment in the columns titled "Obs." (e.g., those that arrive at the very first page of the experiment, the introduction page, those that continue to the instructions, etc). We start with a total number of 10797 participants (= 8091+2706). Around 5.6% choose not to participate in the experiment after they read the general introduction. Another 2% dropout when the details of the experiments are outlined in the instruction page. Up to this stage, and despite the instruction page being slightly different across elicitation methods, dropout is essentially identical across elicitation methods. At the understanding check, we observe a significant difference. The dropout rate in the direct response method is around 6 percentage

[74] We do not report mistakes by participants eventually assigned role (player 1 or 2) or their belief treatment (selfish or non-selfish). These dimensions were unknown to the participant at this point in the experiment and mistakes are not significantly different across them.

[75] One potential error subjects in the strategy method could make is to select the payoffs arising from the two actions highlighted in the other conditional strategy from player 2: (B,D). Indeed, this is the second most common combination of payoffs selected among participants in the strategy method; 9.85% do so. However, even if we would consider such an answer to also be "correct" for those in the strategy method, the difference between this adjusted rate and that in the direct response method is still significantly different at $p < 0.01$.

[76] Unfortunately, it is also not immediately obvious what the ideal mechanism to test understanding in a setting such as ours is. Not asking (or not scoring their answer) about how choices, as they appear in the game, translate into payoffs invites participants into the study who do not understand the game. Maybe a good compromise is to give participants more than one attempt for this type of question in the future.

Table D.34: Frequency of Correct Questions and Number of Mistakes for Everyone who Submits Understanding Test by Elicitation Method

	Direct Response		Strategy Method		
	Mean	Obs	Mean	Obs	Δ
Understanding test: Q1 correct	0.8913	7325	0.8985	2413	-0.0071
Understanding test: Q2 correct	0.7405	7325	0.7484	2413	-0.0080
Understanding test: Q3 correct	0.8330	7325	0.8293	2413	0.0038
Understanding test: Q4 correct	0.8560	7325	0.6913	2413	0.1647^{***}
Understanding test fully correct	0.5943	7325	0.5010	2413	0.0932^{***}
Understanding test: 1 mistake	0.2524	7325	0.2934	2413	-0.0410^{***}
Understanding test: 2 mistakes	0.0700	7325	0.1144	2413	-0.0443^{***}
Understanding test: 3 mistakes	0.0464	7325	0.0543	2413	-0.0079
Understanding test: 4 mistakes	0.0369	7325	0.0369	2413	-0.0000

Notes: the statistics in this table are based on data from every person who submits their answers in the understanding test, regardless of whether or not they passed it and were allowed (not allowed) to proceed to the main task for each elicitation method. Differences between the elicitation methods are shown in the column labelled Δ, with *, **, and *** indicating statistical significant differences (based on t-tests) at the 10%, 5%, and 1% levels, respectively.

points lower than in the strategy method.[77] Finally, for participants that started the games, dropouts for the remaining parts of the experiments are neglectable and statistically indistinguishable across elicitation methods. The total number of participants that complete the survey, which represents our actual "final" sample is 8029 (= 6134 + 1895).

Table D.35: Conditional Dropout Frequency of Participants

	Direct Response		Strategy Method		
	Freq.	Obs.	Freq.	Obs.	Δ
Drop out: Introduction Page	0.0565	8091	0.0562	2706	0.0003
Drop out: Instruction Page	0.0221	7634	0.0223	2554	-0.0002
Drop out: Understanding Test	0.1725	7465	0.2351	2497	-0.0625^{***}
Drop out: Games	0.0044	6177	0.0047	1910	-0.0003
Drop out: Survey	0.0026	6150	0.0032	1901	-0.0006
Final Sample		6134		1895	

Notes: the conditional frequency of dropout refers to the frequency with which participants that entered a given stage do not continue to the next part. The column "Obs." tabulates to the number of participants that enter a specific page/part of the experiment. Differences between the elicitation methods are shown in the column labelled Δ, with *, **, and *** indicating statistical significant differences (based on t-tests) at the 10%, 5%, and 1% levels, respectively.

Table D.36 provides these dropout frequencies in a different format, namely in the form of the percentage of participants remaining at each relevant stage of the experiment relative to those who started the experiment, i.e., arrived at the introduction page. Overall, we retain 75.8% of participants in the direct response method and 70% in the strategy method, which is within the usual dropout rates observed in other studies (Stantcheva (2023)).

[77] The dropout rates in Table D.35 at the Understanding Test are not inconsistent with those implied by the number of mistakes greater than 1 of Table D.34. This is because dropouts can also occur from participants who do not submit the test (140 for DR vs. 84 for SM) or stop on the page that provides them feedback on their mistakes despite being allowed to continue (25 for DR vs. 7 for SM).

Table D.36: Retention of Participants over Stages of Experiment

Dep. Var.: % of Participants Remaining	Direct Response	Strategy Method	Δ
Started Experiment	1.0000	1.0000	0.0000
Started Instruction Page	0.9435	0.9438	-0.0003
Started Understanding Test	0.9226	0.9228	-0.0001
Started Games	0.7634	0.7058	0.0576^{***}
Started Survey	0.7601	0.7025	0.0576^{***}
Completed Experiment	0.7581	0.7003	0.0578^{***}

Notes: differences between the elicitation methods are shown in the column labelled Δ, with *, **, and *** indicating statistical significant differences (based on t-tests) at the 10%, 5%, and 1% levels, respectively.

In general, these dropouts result in a sample in which only 23.6% of participants are in the strategy method – instead of the intended 25%. In terms of the subgroups, 11.6% (12%) are in the selfish (non-selfish) belief, strategy method treatment compared to the intended 12.5% (see Summary Statistics, Table A.2).

Next, we will show that our main results are robust to differential dropouts. We will tackle this in three ways. First, we restrict our original sample. Second, we add hypothetical players to compensate for the dropouts. Third, we return to ideas and observations that had already been raised in the paper's main section and which are inconsistent with attrition being the cause of our results. In what follows, we will primarily focus on player 2, as their results are central to our paper. Nevertheless, we will cover the results of player 1 alongside.

1. Sample restriction. In the main part of the paper, we introduced the notion of *attentiveness*. We observed that attentive participants, who make no mistake in the understanding test, are less likely to make a mistake in the *sPD* or *mUG* than inattentive participants, who make a mistake in the understanding test. We now generalize this idea but use a slightly different term in order to avoid confusion with the attentiveness notion. Suppose any participant that takes our experiment can be described by a degree of *carefulness*. The likelihood with which a participant makes a mistake, in both the games as well as the instruction test, is assumed to be decreasing in this carefulness measure. The observed differential dropouts would pose a danger to our study's validity if (1) it leads to a lower carefulness among subjects in the direct response treatment compared to the strategy method treatment, and (2) if this lower level of carefulness is the main cause behind the observed fewer mistakes under the strategy method.

By adjusting the relevant sample of participants in the direct response and/or strategy method, we can address and refute this concern. Returning to Table D.34, we can make two observations. First, as relatively more participants in the direct response method get the understanding test fully correct, it must be that the test in the direct response method (particularly, question 4), is relatively easier. As a result, the (minimum) level of carefulness must be relatively lower in the direct response method as fewer careless people are forced to dropout. For the second observation, we note that the % of participants in the direct response method who do not make a mistake in the instruction test (59.4%) is smaller than those who make *at most* 1 mistake in the strategy method (79.4% = 50.1% + 29.3%), and so their (minimum) level of carefulness must be relatively higher. It follows that if we restrict our analysis to participants in the direct response method that make no mistakes in the understanding test, i.e., attentive participants, and everyone in the strategy method (recall, subject with more than 1 mistakes are kicked out), we ensure that the degree of carefulness is relatively higher in the direct response group. Consequently, if we still observe fewer mistakes under the strategy method in this subsample, we can be confident that this effect is not driven by any differential degree of carefulness.[78]

Table D.37: Player 2's Behavior in sPD and mUG for Attentive Players in DR and all Players in SM

	sPD		mUG	
after Player 1	cooperates	defects	85-15	50-50
Dep. Var: Player 2 cooperates / rejects	(1)	(2)	(3)	(4)
Non-Selfish Belief	0.0309	0.00846	0.0267	-0.0201***
	(0.0269)	(0.0259)	(0.0245)	(0.00696)
Strategy Method	0.00587	-0.0778***	0.0129	-0.0261***
	(0.0326)	(0.0234)	(0.0233)	(0.00865)
Non-Selfish × Strategy Method	-0.0378	-0.000682	-0.0350	0.0372***
	(0.0411)	(0.0335)	(0.0330)	(0.0111)
Controls	Yes	Yes	Yes	Yes
Observations	2174	1877	1694	2357

Notes: this table reports estimates from OLS regressions for a sample that comprises all attentive player 2s in the direct response method and all player 2s in the strategy method. Control variables are identical to those in Table 4. Estimates for control variables are not reported. For estimates without control variables, refer to Table D.47. Robust standard errors are reported in parentheses. **, and *** indicate statistical significance at the 5% and 1% levels.

Table D.37 presents the results of this subsample analysis. Estimates from the adjusted sample strongly suggests that our result are robust: the strategy method still reduces mistakes. As this sample introduces the fewest number of careless subjects into the direct response group but the most number of careless subjects into the strategy method, it effectively represents the most conservative test to analyze the potential impact of attrition on our findings.[79] In terms of other behavior, we observe that difference in behavior after cooperation or unequal offers mostly vanish, something which we had already noted when analyzing attentive players in subsection 3.5. Repeating this analysis for player 1, Table D.38, reaffirms our previous finding that the strategy method facilitates strategic thinking of participants, making them more responsive to information. [80]

2. Adding hypothetical players to balance attrition. Instead of restricting the sample by limiting the type of players, we now tackle the problem of dropouts from the other direction, namely by adding hypothetical observations to the strategy method. In particular, we add 74 hypothetical player 2s to the strategy method, divided equally across the belief treatments, to recover the originally targeted 75% to 25% randomization, undoing differential dropouts at the level of player 2.[81]

We will explore three different variations in terms of the behavioral frequencies that are assigned to additional hypothetical players. For a given belief treatment, the hypothetical player 2 (added to the strategy method) is assumed to behave like (a) any generic player in the direct response method, (b) any inattentive player in the direct response method, or (c) any inattentive (Q4) player in the direct response method. Specifically, we set the behavior of our additional hypothetical player 2s in the strategy method by

[78] This sample features a total of 3102 player 2s. 69.41% ($n = 2153$) of those are in the direct response treatment and 30.59% ($n = 949$) are in the strategy method treatment.

[79] Note that any alternative sample that excludes players in the strategy method who make a particular form of mistake in the understanding test (e.g., what kind of answer they give in question 1, 2, 3, or 4) essentially results in an analysis that is closer to the attentive sample analysis in Table 11 and has no bearing on our conclusion.

[80] This sample features a total of 3127 player 1s. 69.75% ($n = 2181$) of those are in the direct response and 30.25% ($n = 946$) are in the strategy method treatment.

[81] Note, our original sample featured 3071 (3063) player 2s (player 1s) in the direct response method and 949 (946) in the strategy method.

Table D.38: Player 1's Behavior in sPD and mUG for Attentive Players in DR and all Players in SM

	sPD		mUG	
Dep. Var: P1 cooperates; offers 50-50	(1)	(2)	(3)	(4)
Non-Selfish Belief	0.155***	0.155***	0.117***	0.113***
	(0.0209)	(0.0209)	(0.0200)	(0.0198)
Strategy Method	-0.0892***	-0.101***	-0.0531	-0.0510
	(0.0274)	(0.0298)	(0.0274)	(0.0294)
Non-Selfish × Strategy Method	0.134***	0.132***	0.0644	0.0660
	(0.0374)	(0.0375)	(0.0365)	(0.0362)
Controls	No	Yes	No	Yes
Observations	3127	3127	3127	3127

Notes: this table reports estimates from OLS regressions for a sample that comprises all attentive player 1s in the direct response method and all player 1s in the strategy method. Control variables are identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. **, and *** indicate statistical significance at the 5% and 1% levels.

matching the respective behavioral frequencies of player 2s in the direct response method across all nodes, games, and belief treatments.[82] Simply put, as we move from version (a) to (c), we increase the frequency of mistakes that the additional player 2s make. In the regressions that follow, we not include controls, allowing us to avoid randomness (and thus the ability to influence the results) that occurs when sampling data for control variables from our existing dataset.

The results in Table D.39 suggest that our results are robust to attrition and any likely shape of behavior that those who dropped out could have engaged in. In the *sPD*, the strategy method strongly reduces the likelihood of making mistakes. The strategy method dummy remains significant at 1%. Interestingly, estimates for player 2 responding to player 1 cooperating hardly change across the three variations. This can be explained by the fact that the imposed behavior (in response to cooperation) doesn't deviate much from baseline behavior in the strategy method in any of the scenarios. The same conclusions also hold true for the *mUG* in (a) to (c), with all estimates being very close to those in the normal, full sample. Table D.40 repeats the same analysis also for player 1. Again, our results are robust.

We like to conclude by commenting on our approach to determining the hypothetical players' behavior in the strategy method. Since players who drop out in the strategy method may be more careless than those who do not, it is not appropriate to sample behavior from those in the strategy method, though differences may be small. Yet, we also know that mistake rates vary significantly between elicitation methods, so estimates for (b) and (c) – and possibly even (a) – might already impose rates of mistakes that are greater than those that would have been observed. As such, these estimates serve as a conservative robustness check, supporting the conclusion that differential dropouts are unlikely behind our main results.

3. Other evidence. Finally, we revisit an idea that was discussed in the main section of the paper, and which is more consistent with the conclusion that the strategy method directly causes a reduction in errors. Results in Table 7 suggested that one cause of higher mistakes in the direct response method are player 2's ex-ante beliefs. The stronger player 2 believes that player 1s take the action that leads to the other node of the game, not the one she is currently at, the more likely she makes a mistake. It appears

[82] The respective frequencies can be found in Figures 4 to 7 as well as D.5 and D.6 in the Online Appendix. As we only add very few observations overall, it is impossible to match frequencies perfectly (1 observation represents $1/37 \approx 2.7\%$ in a particular belief-treatment). As this exercise tries to establish the robustness of our results, we always add one more mistake rather than one less, or rather, one more social choice instead of a selfish choice.

Table D.39: Player 2's Behavior in sPD and mUG with additional Hypothetical Players in SM

	sPD		mUG	
after Player 1	cooperates	defects	85-15	50-50
Dep. Var: Player 2 cooperates / rejects	(1)	(2)	(3)	(4)
(a) Additional player 2s act like typical players in the DR				
Non-Selfish Belief	0.0433	0.0342	0.0302	-0.0186**
	(0.0230)	(0.0235)	(0.0201)	(0.00785)
Strategy Method	0.0335	-0.0843***	0.0392**	-0.0286***
	(0.0274)	(0.0207)	(0.0197)	(0.00807)
Non-Selfish × Strategy Method	-0.0458	-0.0276	-0.0354	0.0302***
	(0.0375)	(0.0314)	(0.0294)	(0.0114)
Controls	No	No	No	No
Observations	2796	2321	2073	3044
(b) Additional player 2s act like inattentive players in the DR				
Non-Selfish Belief	0.0433	0.0342	0.0302	-0.0186**
	(0.0230)	(0.0235)	(0.0201)	(0.00785)
Strategy Method	0.0295	-0.0784***	0.0392**	-0.0266***
	(0.0274)	(0.0209)	(0.0197)	(0.00830)
Non-Selfish × Strategy Method	-0.0438	-0.0238	-0.0354	0.0321***
	(0.0376)	(0.0317)	(0.0294)	(0.0118)
Controls	No	No	No	No
Observations	2796	2321	2073	3044
(c) Additional player 2s act like inattentive (Q4) players in the DR				
Non-Selfish Belief	0.0433	0.0342	0.0302	-0.0186**
	(0.0230)	(0.0235)	(0.0201)	(0.00785)
Strategy Method	0.0354	-0.0686***	0.0490**	-0.0246***
	(0.0274)	(0.0212)	(0.0200)	(0.00853)
Non-Selfish × Strategy Method	-0.0439	-0.0239	-0.0433	0.0360***
	(0.0375)	(0.0321)	(0.0297)	(0.0124)
Controls	No	No	No	No
Observations	2796	2321	2073	3044

Notes: this table reports estimates from OLS regressions without control variables for three scenarios in which 74 hypothetical player 2s are added to the strategy method, split evenly across belief treatments. For more detail, please refer to the accompanying text. Robust standard errors are reported in parentheses. **, and *** indicate statistical significance at the 5% and 1% levels

as if player 2 is confused about the node of the game for which she is to make a choice. Since the strategy method asks player 2 to make a choice for all nodes in the game "at the same time", we conjecture that it is harder for player 2 to be confused by her initial belief about the node she ought to make a choice for. To test this hypothesis, we repeat the analysis of Table 7 for the strategy method. Regression estimates, reported in Table D.41, provide support for our conjecture. Initial beliefs appear unrelated to mistakes in the strategy method. Since this mechanism for mistakes is unrelated to attrition, it strengthens our claim that the strategy method directly reduces mistakes.

D.3.1. Robustness of Auxiliary Analyses

Our preceding analyses highlighted that our main results are robust. We now proceed to discuss the robustness of our auxiliary analyses. Table 8 represented the first of three pieces of evidence, that, taken together, suggested that the strategy method reduces mistakes instead of changing preferences. It showed

Table D.40: Player 1's Behavior in sPD and mUG with additional Hypothetical Players in SM

	sPD	mUG
Dep. Var: P1 cooperates; offers 50-50	(1)	(2)
(a) Additional player 1s act like typical players in the DR		
Non-Selfish Belief	0.165***	0.117***
	(0.0176)	(0.0170)
Strategy Method	-0.0846***	-0.0340
	(0.0255)	(0.0256)
Non-Selfish × Strategy Method	0.117***	0.0594
	(0.0347)	(0.0340)
Controls	No	No
Observations	4083	4083
(b) Additional player 1s act like inattentive players in the DR		
Non-Selfish Belief	0.165***	0.117***
	(0.0176)	(0.0170)
Strategy Method	-0.0846***	-0.0360
	(0.0255)	(0.0256)
Non-Selfish × Strategy Method	0.119***	0.0595
	(0.0347)	(0.0340)
Controls	No	No
Observations	4083	4083
(c) Additional player 1s act like inattentive (Q4) players in the DR		
Non-Selfish Belief	0.165***	0.117***
	(0.0176)	(0.0170)
Strategy Method	-0.0746***	-0.0400
	(0.0256)	(0.0256)
Non-Selfish × Strategy Method	0.109***	0.0597
	(0.0347)	(0.0341)
Controls	No	No
Observations	4083	4083

Notes: this table reports estimates from OLS regressions without control variables for three scenarios in which 74 hypothetical player 1s are added to the strategy method, split evenly across belief treatments. For more detail, please refer to the accompanying text to Table D.39. Robust standard errors are reported in parentheses. **, and *** indicate statistical significance at the 5% and 1% levels.

that inattentive players (i) make more mistakes and (ii) that the strategy method is particularly useful in mitigating their mistakes. It further showed that those who answer the understanding test's question about player payoffs incorrectly, indicated by the *Inattentive (Q4)* dummy variable, were (i') more likely to make mistakes when playing the games and (ii') that they benefit more from the strategy method in terms of a reduction of their mistakes than those who answer question 1, 2, or 3 incorrectly, indicated by the *Inattentive (Q123)* dummy variable. We now repeat this analysis using our method of adding hypothetical players.[83] Since the purpose of adding hypothetical players is to address the differential dropout arising from question 4, these additional players will be classified as *Inattentive (Q4)*, and by extension, as overall *Inattentive* types. The results of this analysis are provided in Table D.42 below.

Based on the results in the top panel, version (a), we can make an important yet obvious observation: the

[83] As the *Inattentive* variable is central to the analysis, we cannot take the approach of Table D.37, where players in the direct response method are limited to attentive participants only.

Table D.41: Player 2's Behavior: Mistakes and Beliefs in the Strategy Method

	sPD: after P1 defects		mUG: after P1 offers 50-50	
Dep. Var: Player 2 makes mistake	(1)	(2)	(3)	(4)
Non-Selfish Belief	0.00784	0.00729	0.0170*	0.0191*
	(0.0250)	(0.0262)	(0.00989)	(0.0105)
Belief Player 1 cooperates	-0.00736	0.00320		
	(0.0398)	(0.0400)		
Belief Player 1 offers 50-50			-0.00578	-0.00494
			(0.0132)	(0.0134)
Controls	No	Yes	No	Yes
Observations	949	949	949	949

Notes: this table reports estimates from OLS regressions, with control variables identical to those in Table 4, for player 2s in the strategy method. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

estimates of the coefficients related to the direct response treatment, i.e., Inattentive, Inattentive (Q123), Inattentive (Q4) are identical to those in our full sample.[84] This observation is important because it provides strong evidence supporting the idea that cooperation after defection and rejecting equal offers are likely simple mistakes. It is unlikely that such preferences are correlated with making more errors in a control questionnaire, either overall (conjecture i) or in relation to the specific types of mistakes made (conjecture i'). This observation is obvious since the additional hypothetical players are only added to the strategy method, so the estimates for inattentive players in the direct response method are unaffected. Consequently, we include these estimates only in part (a) to keep the table concise.

In the *sPD*, the coefficients for the interaction terms between the strategy method and the various inattention measures are negative and statistically significant for but one coefficient, providing support for conjecture (ii). Similarly, the coefficient for the interaction terms of *Inattentive (Q4)* is more negative than that of *Inattentive (Q123)*, yet, just like in our full sample, the two are never significantly different from each other. For the *mUG*, we observe similar qualitative patterns but with fewer statically significant results. As the coefficients were smaller in absolute value in the full sample and featured higher p-values, this is hardly surprising (indeed, the interaction with *Inattentive (Q123)* was not statistically significant in the full sample). Overall, we conclude that our results are robust.

For our next piece of evidence, we documented in Table 9 that choices that are taken more slowly result in less cooperation after defection and rejections of fair offers while the behavior at the other two nodes, conditional cooperation or rejection of unfair offers, are unrelated to response time. We interpreted this behavioral pattern as further evidence that cooperation after defection and rejecting of fair offers are likely mistakes. Moreover, the analysis suggests a potential mechanism by which strategy method reduces mistakes: by inducing longer deliberation times. Generally, it is difficult to see how this very specific relationship between response time and behavior could be caused by attrition. A more plausible explanation is that it is driven by a specific subset of participants. In fact, inattentive players, unsurprisingly, exhibit shorter response times (see Table D.46).

Despite these general reservations, we estimate the relationship between response time and behavior for our restricted sample in Table D.43. For the *sPD*, the negative relationship between response time and mistakes is robust. For the *mUG*, the relationship disappears. Since the restricted sample excludes all

[84] As Table 8 always included control variables for each specifications, please refer to Table D.20 for the respective estimates without control variables.

Table D.42: Player 2's Behavior: Mistakes and Inattention with additional Hypothetical Players in SM

	sPD: after P1 defects		mUG: after P1 offers 50-50	
Dep. Var: Player 2 makes a mistake	(1)	(2)	(3)	(4)
(a) Additional player 2s act like typical players in the DR				
Strategy Method	-0.0767***	-0.0767***	-0.00564	-0.00564
	(0.0179)	(0.0179)	(0.00524)	(0.00524)
Inattentive	0.130***		0.0480***	
	(0.0273)		(0.0104)	
Inattentive × Strategy Method	-0.0852**		-0.0295**	
	(0.0348)		(0.0138)	
Inattentive (Q123)		0.102***		0.0356***
		(0.0287)		(0.0103)
Inattentive (Q4)		0.276***		0.108***
		(0.0651)		(0.0323)
Inattentive (Q123) × Strategy Method		-0.103***		-0.0234
		(0.0388)		(0.0156)
Inattentive (Q4) × Strategy Method		-0.196***		-0.0850**
		(0.0709)		(0.0346)
Controls	No	No	No	No
Observations	2321	2321	3044	3044
(b) Additional player 2s act like inattentive players in the DR				
Strategy Method	-0.0767***	-0.0767***	-0.00564	-0.00564
	(0.0179)	(0.0179)	(0.00524)	(0.00524)
Inattentive × Strategy Method	-0.0662*		-0.0224	
	(0.0352)		(0.0144)	
Inattentive (Q123) × Strategy Method		-0.103***		-0.0234
		(0.0388)		(0.0156)
Inattentive (Q4) × Strategy Method		-0.162**		-0.0725**
		(0.0715)		(0.0353)
Controls	No	No	No	No
Observations	2321	2321	3044	3044
(c) Additional player 2s act like inattentive (Q4) players in the DR				
Strategy Method	-0.0767***	-0.0767***	-0.00564	-0.00564
	(0.0179)	(0.0179)	(0.00524)	(0.00524)
Inattentive × Strategy Method	-0.0425		-0.0130	
	(0.0357)		(0.0151)	
Inattentive (Q123) × Strategy Method		-0.103***		-0.0234
		(0.0388)		(0.0156)
Inattentive (Q4) × Strategy Method		-0.121*		-0.0559
		(0.0721)		(0.0361)
Controls	No	No	No	No
Observations	2321	2321	3044	3044

Notes: this table reports estimates from OLS regressions without control variables for three scenarios in which 74 hypothetical player 2s are added to the strategy method, split evenly across belief treatments. Hypothetical players are inattentive, with Inattentive (Q123) = 0 and Inattentive (Q4) = 1. Estimates for Inattentive, Inattentive (Q123), and Inattentive (Q4) are not reported in panel (b) and (c) as they are identical to those in part (a). Robust standard errors are reported in parentheses. **, and *** indicate statistical significance at the 5% and 1% levels.

Table D.43: Player 2's Behavior in sPD and mUG: Response Time for Attentive Players in DR and all Players in SM

	sPD		mUG	
after Player 1	cooperates	defects	85-15	50-50
Dep. Var: Player 2 cooperates / rejects	(1)	(2)	(3)	(4)
ln(Response time)	-0.00436	-0.0272*	-0.0133	-0.0116
	(0.0187)	(0.0154)	(0.0210)	(0.00838)
Non-Selfish Belief	0.0305	0.0106	0.0275	-0.0209***
	(0.0270)	(0.0259)	(0.0246)	(0.00699)
Strategy Method	0.00739	-0.0698***	0.0163	-0.0216***
	(0.0332)	(0.0240)	(0.0236)	(0.00835)
Non-Selfish × Strategy Method	-0.0376	-0.00465	-0.0373	0.0367***
	(0.0411)	(0.0336)	(0.0330)	(0.0110)
Controls	Yes	Yes	Yes	Yes
Observations	2174	1877	1694	2357

Notes: this table reports estimates from OLS regressions for a sample that comprises all attentive player 2s in the direct response method and all player 2s in the strategy method. Control variables identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. **, and *** indicate statistical significance at the 5% and 1% levels.

inattentive players in the direct response treatment – comprising 29.89% of players in that treatment – it is perhaps less surprising that the effect vanishes in the game where (1) the link between mistakes and response time was the weakest, and (2) the impact of the strategy method in reducing mistakes was also less pronounced.[85]

Next, we look at the analysis of player 1 and response time. Figure 3 offered indirect evidence supporting the interpretation that the strategy method encourages more strategic thinking among first movers. It showed that in the treatment with notably lower cooperation rates, i.e., for selfish beliefs in the strategy method, response times were longer. In Figure D.4a, we graph response times by treatments for the sample with attentive players in the direct response method and all players in the strategy method. Consistent with our earlier finding that attentive players tend to have longer response times, the response times for attentive players in the direct response method now exceed those of the strategy method across all belief in both games.[86] However, aligning with our original result, the smallest difference in response times between the elicitation methods is observed in the selfish belief treatment for the *sPD*. This same pattern is also evident, for example, when the analysis is restricted to attentive players, as shown in Figure D.4b.[87] Since differential dropouts occur at the level of the elicitation method in our experiment and not at treatment level, i.e., the combination of belief manipulation × elicitation method, it is not surprising that the relationship between response time and the treatments is generally robust. Nonetheless, due to the smaller sample size, this relationship is no longer statistically significant, as indicated by the confidence intervals displayed in the graphs.

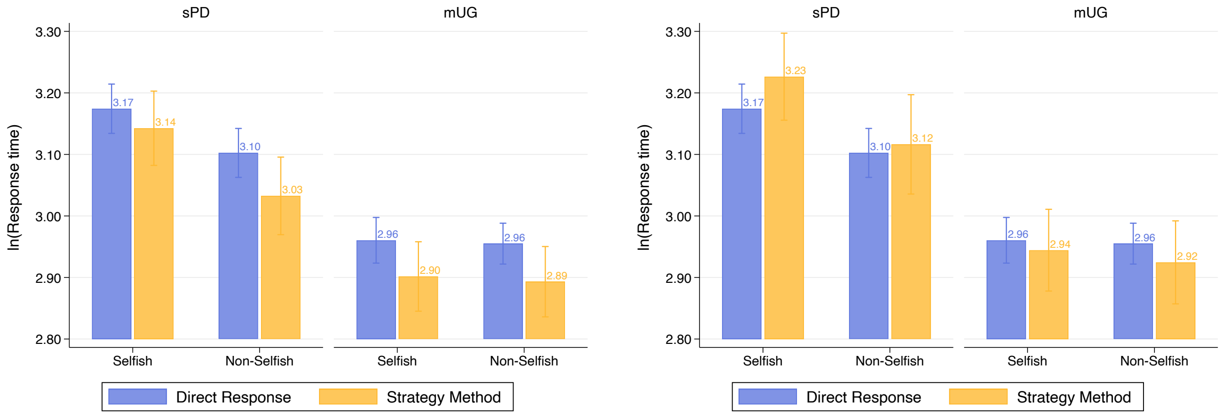
Since our third piece of evidence, Table 10, which computed the frequency of player 2 with preferences for efficiency or spite, only relied on data from the strategy method (and makes no attempt at comparisons between elicitation methods), there is no need to address differential dropouts for this analysis.

This concludes our discussion of the robustness of our auxiliary analyses. Overall, we remain confident

[85] As there is no sensible approach to infer response times for those players in the strategy method who dropped out of the sample, we cannot conduct an analysis with additional hypothetical players.

[86] Recall, this sample eliminates a much larger fraction of inattentive people than necessary to undo the differential attrition, with 69.75% being in the direct response method relatively to the ideal 75%.

[87] Note that this sample does not fully undo attrition as it features 78.43% of people in the direct response treatment.



(a) Sample: Attentive Players in DR, Everyone in SM

(b) Sample: Attentive Players

Figure D.4: Player 1's Response Time for different Sub-Samples

Notes: the number on each bar represents the mean and the error bar represents the 95% confidence interval.

in our interpretation of our main result: that the strategy method reduces mistakes rather than altering underlying preferences.

Finally, we revisit the heterogeneity analysis, which examines player 2's behavior separately for attentive and inattentive players. Since mistakes on question 4 of the understanding check were slightly more common in the strategy method, the set of attentive players ($N = 2754$) includes relatively fewer player 2s ($n = 601$) in the strategy method, accounting for 21.82% of players compared to the randomization target of 25%. In contrast, the set of inattentive player 2s ($N = 1266$) comprises relatively more players in the strategy method ($n = 348$), making up 27.49%.

To address the effects from differential attrition arising from question 4 in the heterogeneity analysis, we thus need to balance both the attentive and inattentive samples. To balance the attentive sample, we ideally classify participants in the strategy method who would not have made a mistake on question 4 – assuming identical difficulty across elicitation methods – as attentive. After this reclassification, we then fill the remaining inattentive sample in the strategy method with hypothetical observations, as we did previously.

Since it is impossible to precisely identify which participants would not have made a mistake, we adopt the simplest and most conservative approach: reclassifying all players in the strategy method who answer question 4 incorrectly ($n = 167$) as attentive. This results in a new sample of attentive players, in which 26.29% are in the strategy method. The inattentive sample for the strategy method is then constructed by adding 126 hypothetical observations to the remaining inattentive players, specifically the *Inattentive (Q123)* types. Note that this approach adds significantly more hypothetical observations to the inattentive sample, compared to the 74 players in our previous analyses of player 2. This is because we shifted more players to the attentive sample than necessary to reach the target of 25%. As a result, the inattentive strategy method sample will exhibit a relatively high number of errors, particular for version (b) and (c), which should be taken into account when interpreting those results.

The estimates for the adjusted attentive subsample are provided in Table D.44. Overall, the new estimates are very similar to those in Table 11.[88] The same is also broadly true for the adjusted inattentive subsample,

[88] Indeed, we had already performed a similar analysis, where we restricted the sample to all attentive player 2s in direct response method and all player 2s in strategy method. This resulted in a sample with even more players in the strategy method at 30.59% than in the current analysis. Nevertheless, the resulting estimates are also close to those of the attentive sample in

as shown in Table D.45. Since we added a larger number of hypothetical observations than necessary, the effect of the strategy method in reducing mistakes appears less pronounced in the inattentive sample compared to the results in Table 11.

Table D.44: Player 2's Behavior in sPD and mUG for Attentive Players in DR and Attentive and Inattentive (Q4) Players in SM

	sPD		mUG	
after Player 1	cooperates	defects	85-15	50-50
Dep. Var: Player 2 cooperates / rejects	(1)	(2)	(3)	(4)
Non-Selfish Belief	0.0303	0.00824	0.0265	-0.0202***
	(0.0269)	(0.0259)	(0.0244)	(0.00697)
Strategy Method	0.0170	-0.0820***	0.0211	-0.0262***
	(0.0337)	(0.0241)	(0.0238)	(0.00888)
Non-Selfish × Strategy Method	-0.0582	0.00821	-0.0498	0.0377***
	(0.0441)	(0.0353)	(0.0341)	(0.0120)
Controls	Yes	Yes	Yes	Yes
Observations	1993	1696	1513	2176

Notes: this table reports estimates from OLS regressions for a sample that comprises all attentive player 2s in the direct response method and player 2s in the strategy method who either get all control questions correct or get Q4 incorrect. Control variables are identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 11.

Table D.45: Player 2's Behavior in sPD and mUG for Inattentive Players in DR and Inattentive (Q123) Players in SM with additional Hypothetical Players in SM

	sPD		mUG	
after Player 1	cooperates	defects	85-15	50-50
Dep. Var: Player 2 cooperates / rejects	(1)	(2)	(3)	(4)
(a) Additional player 2s act like typical players in the DR				
Non-Selfish Belief	0.0725*	0.0792	0.0273	-0.0191
	(0.0422)	(0.0490)	(0.0360)	(0.0202)
Strategy Method	0.0872*	-0.129***	0.0563	-0.0484**
	(0.0501)	(0.0425)	(0.0364)	(0.0205)
Non-Selfish × Strategy Method	-0.0159	-0.0651	0.0260	0.0258
	(0.0676)	(0.0643)	(0.0565)	(0.0280)
Controls	No	No	No	No
Observations	855	677	612	920
(b) Additional player 2s act like inattentive players in the DR				
Non-Selfish Belief	0.0725*	0.0792	0.0273	-0.0191
	(0.0422)	(0.0490)	(0.0360)	(0.0202)
Strategy Method	0.0677	-0.0970**	0.0498	-0.0354
	(0.0505)	(0.0441)	(0.0360)	(0.0224)
Non-Selfish × Strategy Method	-0.00949	-0.0518	0.0259	0.0259
	(0.0681)	(0.0668)	(0.0560)	(0.0307)
Controls	No	No	No	No
Observations	855	677	612	920
(c) Additional player 2s act like inattentive (Q4) players in the DR				
Non-Selfish Belief	0.0725*	0.0792	0.0273	-0.0191
	(0.0422)	(0.0490)	(0.0360)	(0.0202)
Strategy Method	0.100**	-0.0321	0.108***	-0.0224
	(0.0498)	(0.0468)	(0.0393)	(0.0240)
Non-Selfish × Strategy Method	-0.00927	-0.0580	-0.0260	0.0456
	(0.0669)	(0.0700)	(0.0584)	(0.0347)
Controls	No	No	No	No
Observations	855	677	612	920

Notes: this table reports estimates from OLS regressions without control variables for a sample that comprises all inattentive player 2s in the direct response method and all player 2s in the strategy method who get control question 1, 2, or 3 incorrect for three scenarios in which 126 hypothetical player 2s are added to the strategy method, split evenly across belief treatments. For more detail, please refer to the accompanying text. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

D.3.2. Supporting Tables

Table D.46: Response Time and Inattention in sPD and mUG

	sPD		mUG	
Dep. Var: ln(Response time)	(1)	(2)	(3)	(4)
Inattentive	-0.227***	-0.220***	-0.169***	-0.160***
	(0.0168)	(0.0158)	(0.0162)	(0.0153)
Non-Selfish Belief	-0.0497***	-0.0495***	-0.00985	-0.00543
	(0.0172)	(0.0163)	(0.0162)	(0.0152)
Strategy Method	0.222***	0.220***	0.214***	0.209***
	(0.0248)	(0.0234)	(0.0251)	(0.0237)
Non-Selfish × Strategy Method	-0.0473	-0.0468	-0.0667*	-0.0607*
	(0.0354)	(0.0336)	(0.0342)	(0.0321)
Controls	No	Yes	No	Yes
Observations	8029	8029	8029	8029

Notes: this table reports estimates from OLS regressions, with control variables identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table D.47: Player 2's Behavior in sPD and mUG for Attentive Players in DR and all Players in SM; without Control Variables

	sPD		mUG	
after Player 1	cooperates	defects	85-15	50-50
Dep. Var: Player 2 cooperates / rejects	(1)	(2)	(3)	(4)
Non-Selfish Belief	0.0305	0.0115	0.0315	-0.0197***
	(0.0273)	(0.0260)	(0.0242)	(0.00694)
Strategy Method	0.0115	-0.0641***	0.0387	-0.0178**
	(0.0302)	(0.0220)	(0.0218)	(0.00757)
Non-Selfish × Strategy Method	-0.0353	-0.00638	-0.0393	0.0342***
	(0.0412)	(0.0333)	(0.0330)	(0.0106)
Controls	No	No	No	No
Observations	2174	1877	1694	2357

Notes: this table reports estimates from OLS regressions for a sample that comprises all attentive player 2s in the direct response method and all player 2s in the strategy method. Robust standard errors are reported in parentheses. **, and *** indicate statistical significance at the 5% and 1% levels.

D.4. Heterogeneity Analysis: Attentive vs. Inattentive Players

For completeness, we also split the inattentive sample along question 1, 2, 3, and question 4. In Figure D.5 and D.6, we included the attentive players as a natural comparison, splitting the full sample into three distinct subsamples. Given the relatively small number of observations for the two inattentive groups (especially so for Inattentive(Q4)), it is not surprising that the displayed confidence intervals are relatively large.

In view of this, we keep any discussion of the data to a minimum as we cannot confidently make many observations here, nor think one should. What is noticeable in the Figures and the Regression Tables, however, is that mistakes are higher and differences between the direct-response and the strategy method larger for the Inattentive(Q4) sample compared to the Inattentive (Q123) group – which was highlighted in the main paper in Table 8.

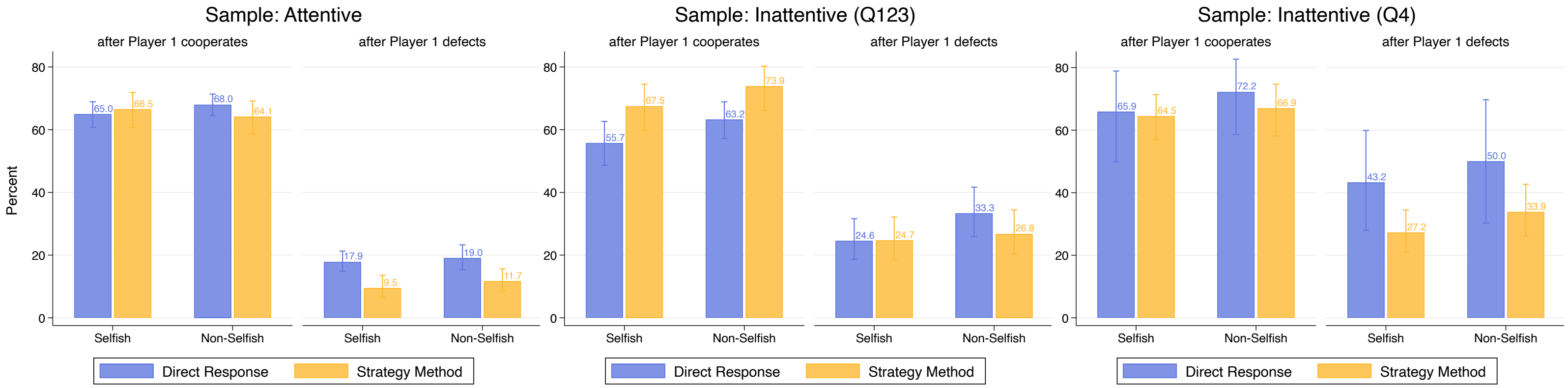


Figure D.5: Player 2's Behavior in sPD for Attentive and Inattentive Players (Q123 and Q4)

Notes: the number on each bar represents the mean and the error bar represents the 95% confidence interval.

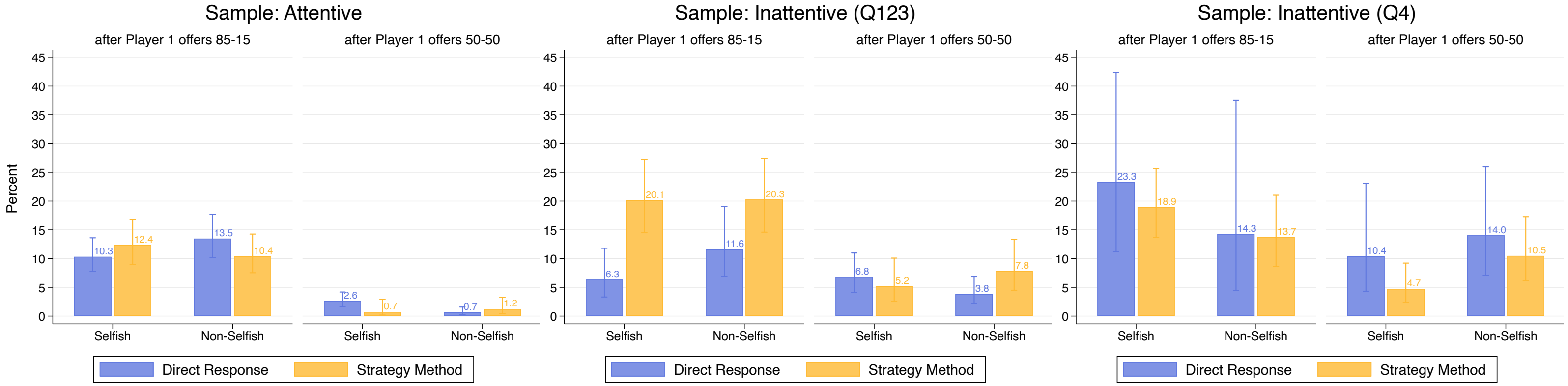


Figure D.6: Player 2's Behavior in mUG for Attentive and Inattentive Players (Q123 and Q4)

Notes: the number on each bar represents the mean and the error bar represents the 95% confidence interval.

Table D.48: Player 2's Behavior in sPD – for Attentive and Inattentive(Q123) and Inattentive(Q123) players

	after P1 cooperates		after P1 defects	
Dep. Var: Player 2 cooperates	(1)	(2)	(3)	(4)
Sample: Attentive				
Non-Selfish Belief	0.0305	0.0315	0.0115	0.00672
	(0.0273)	(0.0269)	(0.0260)	(0.0259)
Strategy Method	0.0159	0.0209	-0.0840***	-0.0831***
	(0.0353)	(0.0350)	(0.0241)	(0.0245)
Non-Selfish × Strategy Method	-0.0548	-0.0660	0.0106	0.0108
	(0.0476)	(0.0475)	(0.0361)	(0.0363)
Controls	No	Yes	No	Yes
Observations	1826	1826	1529	1529
Sample: Inattentive (Q123)				
Non-Selfish Belief	0.0749	0.0813*	0.0877*	0.0807
	(0.0468)	(0.0472)	(0.0521)	(0.0532)
Strategy Method	0.118**	0.118*	0.00114	-0.173***
	(0.0522)	(0.0629)	(0.0480)	(0.0478)
Non-Selfish × Strategy Method	-0.0117	-0.0138	-0.0665	-0.0914
	(0.0699)	(0.0848)	(0.0722)	(0.0724)
Controls	No	Yes	No	Yes
Observations	760	634	616	490
Sample: Inattentive (Q4)				
Non-Selfish Belief	0.0637	0.124	0.0676	0.0155
	(0.0964)	(0.0954)	(0.131)	(0.140)
Strategy Method	-0.0136	-0.00836	-0.160*	-0.296***
	(0.0831)	(0.0893)	(0.0889)	(0.0907)
Non-Selfish × Strategy Method	-0.0393	-0.125	-0.00105	-0.00434
	(0.112)	(0.127)	(0.142)	(0.154)
Controls	No	Yes	No	Yes
Observations	388	262	354	228

Notes: this table reports estimates from OLS regressions for attentive players who don't make any mistake in the control questions, and the two types of inattentive players. Control variables identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Table D.49: Player 2's Behavior in mUG – for Attentive and Inattentive(Q123) and Inattentive(Q123) players

	after P1 offers 85-15		after P1 offers 50-50	
Dep. Var: Player 2 rejects	(1)	(2)	(3)	(4)
Sample: Attentive				
Non-Selfish Belief	0.0315	0.0264	-0.0197***	-0.0202***
	(0.0242)	(0.0244)	(0.00694)	(0.00698)
Strategy Method	0.0203	0.0214	-0.0190**	-0.0201**
	(0.0248)	(0.0246)	(0.00811)	(0.00830)
Non-Selfish × Strategy Method	-0.0509	-0.0500	0.0247**	0.0265**
	(0.0356)	(0.0360)	(0.0106)	(0.0108)
Controls	No	Yes	No	Yes
Observations	1346	1346	2009	2009
Sample: Inattentive (Q123)				
Non-Selfish Belief	0.0527	0.0574	-0.0295	-0.0301
	(0.0367)	(0.0384)	(0.0204)	(0.0195)
Strategy Method	0.138***	0.138***	-0.0159	-0.0549***
	(0.0384)	(0.0466)	(0.0247)	(0.0202)
Non-Selfish × Strategy Method	-0.0514	-0.0112	0.0560	0.0324
	(0.0588)	(0.0714)	(0.0348)	(0.0305)
Controls	No	Yes	No	Yes
Observations	561	435	815	689
Sample: Inattentive (Q4)				
Non-Selfish Belief	-0.0905	-0.179	0.0362	0.0502
	(0.109)	(0.119)	(0.0640)	(0.0700)
Strategy Method	-0.0440	-0.0864	-0.0568	-0.0496
	(0.0834)	(0.100)	(0.0473)	(0.0562)
Non-Selfish × Strategy Method	0.0382	0.134	0.0213	0.00115
	(0.118)	(0.129)	(0.0717)	(0.0850)
Controls	No	Yes	No	Yes
Observations	344	218	398	272

Notes: this table reports estimates from OLS regressions for attentive players who don't make any mistake in the control questions, and the two types of inattentive players. Control variables identical to those in Table 4. Estimates for control variables are not reported. Robust standard errors are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

E. Sample Comparison: Pre-Covid and Covid Sample

In this section, we compare the pre-Covid (October to November 2019) and Covid sample (October 2021) with regards to the participants' behavior and characteristics. For easy comparison to the whole sample, we provide beliefs and behavior by treatments of the whole sample here first. As some may find it interesting, this table also includes the implied frequency with which each end-node is reached.[89]

Table E.50: Beliefs and Behavior by Treatments

	Selfish		Non-Selfish	
	DR	SM	DR	SM
Beliefs about Player 1				
Belief Player 1 cooperates	0.34	0.32	0.74	0.74
Belief Player 1 offers 50-50	0.34	0.32	0.74	0.73
Behavior: Player 1				
Player 1 cooperates	0.50	0.40	0.66	0.69
Player 1 offers 50-50	0.60	0.56	0.71	0.74
Behavior: Player 2				
Player 2 cooperates after C	0.63	0.66	0.67	0.66
Player 2 cooperates after D	0.21	0.11	0.24	0.12
Player 2 rejects 50-50	0.04	0.01	0.02	0.02
Player 2 rejects 85-15	0.10	0.14	0.13	0.13
Frequency of End-nodes				
sPD:				
(cooperate, cooperate)	0.315	0.264	0.442	0.455
(cooperate, defect)	0.185	0.136	0.218	0.235
(defect, cooperate)	0.105	0.066	0.082	0.037
(defect, defect)	0.395	0.534	0.258	0.273
mUG:				
(50-50, accept)	0.576	0.554	0.696	0.725
(50-50, reject)	0.024	0.006	0.014	0.015
(85-15, accept)	0.360	0.378	0.252	0.226
(85-15, reject)	0.040	0.062	0.038	0.034

Notes: for the frequency of end-nodes, the first action refers to player 1's choice, and the second player 2's reply, i.e., (a_1, a_2). These frequencies are calculated based on the respective behavior of player 1 and player 2 at the population level.

Differences in Behavior/Beliefs in the pre-Covid and Covid Sample. Tables E.51 provides a general overview of the frequency of cooperation, fair offers, and rejections, both for our overall sample, as well as for the pre-Covid and Covid samples. When differences between the two subsamples are statistically significant, we indicate this using the typical significance-stars in the Covid column.

Differences in player 1 behavior between the two samples are small, albeit significant in the mUG. Indeed, given our large sample, we will often find differences to be statically significant at conventional levels despite the fact that they are economically small.

In the sPD, player 2's behavior across the samples is fairly stable, albeit with significantly more mistakes being made in the Covid sample. For the mUG, there are significant differences between the two samples. Rejection rate of unequal offers almost double from the pre-Covid days. However, even at 17%, these rejection rates are fairly small, especially in view of the large social behavior in the Prisoner's dilemma. As in the sPD, we again observe significantly more mistakes in the Covid sample.

[89] For those interested in end-nodes frequencies, but who prefer a simpler format, we also list them in Table E.54 by elicitation methods. The table can be found at the end of this section.

Table E.51: Overall Behavior in the pre-Covid and Covid Sample

	Full Sample	pre-Covid	Covid
seq. Prisoner's Dilemma			
Player 1 cooperates	0.57	0.56	0.58
Player 2 cooperates after C	0.65	0.65	0.66
Player 2 cooperates after D	0.18	0.16	0.20^{***}
mini Ultimatum Game			
Player 1 offers 50-50	0.66	0.67	0.64^{**}
Player 2 rejects 85-15	0.13	0.09	0.17^{***}
Player 2 rejects 50-50	0.03	0.01	0.04^{***}
Observations	8029	4647	3382

Notes: statistically significant differences between pre- and Covid behavior (based on t-tests) at significance levels of * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, are indicated by the respective stars in the Covid column.

A more detailed picture is given by Table E.52, which provides beliefs and behavior by treatment for both samples. The table shows that beliefs are quite similar across samples. While beliefs are significantly different for the non-selfish/direct response treatment at 5% and 10%, the total difference is small at 2 percentage points.

For player 1, there is a significant increase in cooperation in the selfish belief/direct-response treatment over the samples, yet a very significant drop in cooperation for the selfish belief/strategy method group. We also see fewer equal offers for the selfish belief/strategy method treatment in the Covid sample. It is interesting that all these differences arise for the selfish belief treatments, suggesting that behavior may be more susceptible to change in a selfish-framing.

In the *sPD*, player 2's tendency to reward cooperation by cooperating themselves is very stable, with no significant differences between the two samples. Cooperation after defection tends to occur more frequently, yet is only significantly higher for the non-selfish/strategy method treatment. The *mUG* is where we see the largest differences between the two samples. Rejecting offers is significantly higher in the Covid sample (for all treatments but the strategy methods groups in response to 50-50 offers).

Table E.52: Beliefs and Behavior in the pre-Covid and Covid Sample by Treatments

	Pre-Covid				Covid			
	Selfish		Non-Selfish		Selfish		Non-Selfish	
	DR	SM	DR	SM	DR	SM	DR	SM
Beliefs about Player 1								
Belief Player 1 cooperates	0.34	0.33	0.73**	0.73	0.34	0.31	0.75	0.75
Belief Player 1 offers 50-50	0.34	0.32	0.73*	0.73	0.33	0.31	0.75	0.74
Behavior: Player 1								
Player 1 cooperates	0.47**	0.46***	0.65	0.70	0.53	0.33	0.68	0.69
Player 1 offers 50-50	0.61	0.60**	0.72	0.76	0.58	0.51	0.71	0.71
Behavior: Player 2								
Player 2 cooperates after C	0.63	0.63	0.66	0.67	0.63	0.70	0.68	0.64
Player 2 cooperates after D	0.19	0.10	0.23	0.09**	0.23	0.14	0.26	0.16
Player 2 rejects 50-50	0.02***	0.01	0.01***	0.02	0.07	0.01	0.04	0.03
Player 2 rejects 85-15	0.08**	0.11***	0.10*	0.10***	0.13	0.19	0.17	0.19

Notes: *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively, testing for differences between the respective columns of the Pre- and Post-Covid Samples.

Differences among Participants. Table E.53 summarizes the participants' (mean) characteristics and highlights their potential differences between our samples. Overall, the Covid sample tends to have slightly more educated participants, with higher incomes. They are also slightly older, which is consistent with more people working from home, using MTurk either as a second source of income or out of interest. More participants have participated in similar experiments before. Despite that, they tend to make more mistakes on the initial understanding test even though they essentially take the same time to complete the experiment.

Table E.53: Characteristics of Participants by Sample

	Pre-Covid	Covid
Participated in experiments before	0.733**	0.756
Gender:		
Female	0.515	0.520
Male	0.478	0.470
Other / Prefer not to say	0.006*	0.010
Age:		
$<$ 12 years	0.000	0.000
12-17 years old	0.000	0.000
18-24 years old	0.081***	0.063
25-34 years old	0.369	0.370
35-44 years old	0.275	0.280
45-54 years old	0.147*	0.162
55-64 years old	0.093	0.088
65-74 years old	0.032	0.029
$\geq$ 75 years	0.002**	0.004
Prefer not to say	0.002	0.003
Income		
Less than 20 000	0.000	0.000
20 000 to 34 999	0.162***	0.137
35 000 to 49 999	0.176	0.182
50 000 to 74 999	0.239	0.244
75 000 to 99 999	0.147**	0.166
100 000 to 140 999	0.113	0.102
over 150 000	0.044*	0.053
Prefer not to say	0.022	0.022
Education:		
No Degree	0.010	0.008
High School Degree	0.298***	0.211
Bachelor Degree	0.487***	0.525
Master Degree	0.133***	0.193
Other Post-Grad Degree	0.033	0.028
Doctorate Degree	0.026	0.024
Prefer not to say	0.012	0.011
Other Game Outcomes		
Total earnings from games (in USD)	1.766**	1.733
Total time (in sec.)	368.737	364.701
Inattentive	0.300**	0.326

Notes: this table reports the mean of various variables for the Pre-Covid and Covid Sample. Statistically significant differences between pre- and Covid data (based on t-tests) at significance levels of * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, are indicated by the respective stars in the Pre-Covid column.

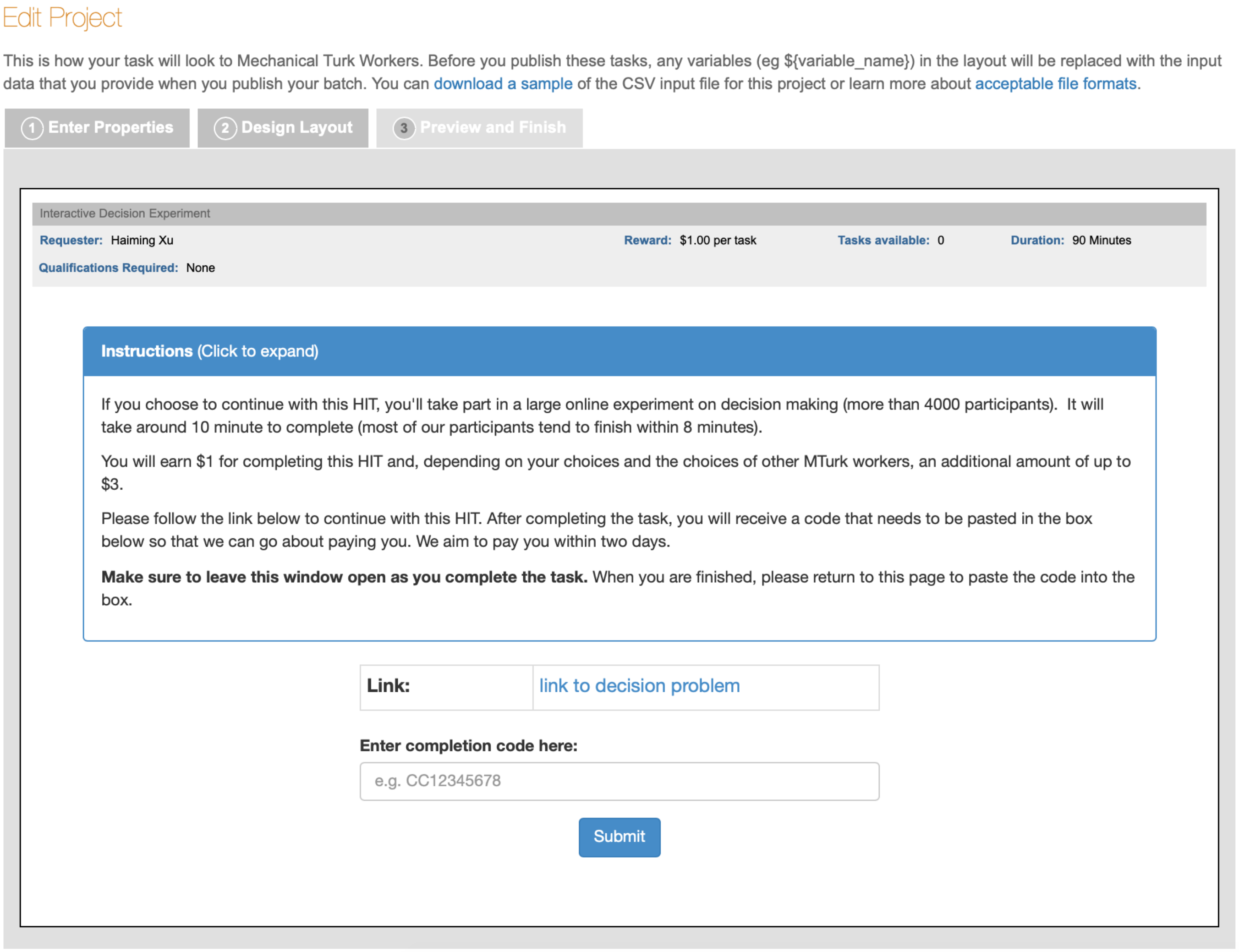
Table E.54: Frequency of End-nodes by Elicitation Methods

Frequency of End-nodes	DR	SM
sPD:		
(cooperate, cooperate)	0.379	0.360
(cooperate, defect)	0.201	0.185
(defect, cooperate)	0.093	0.052
(defect, defect)	0.327	0.403
mUG:		
(50-50, accept)	0.636	0.640
(50-50, reject)	0.019	0.010
(85-15, accept)	0.306	0.030
(85-15, reject)	0.039	0.048

Notes: for the frequency of end-nodes, the first action refers to player 1's choice, and the second player 2's reply, i.e., (a_1, a_2). These frequencies are calculated based on the respective behavior of player 1 and player 2 at the population level.

F. Screenshots of Experiment

In this section, we present all screenshots from the experiment, starting with the screen that mturk workers see on the platform, followed by our landing page (introduction), and so forth. Some screens are shown from multiple perspectives, i.e., player 1, 2 using different elicitation methods and belief treatments.



Edit Project

This is how your task will look to Mechanical Turk Workers. Before you publish these tasks, any variables (eg ${variable_name}) in the layout will be replaced with the input data that you provide when you publish your batch. You can download a sample of the CSV input file for this project or learn more about acceptable file formats.

Interactive Decision Experiment

Requester: Haiming Xu
Reward: $1.00 per task
Tasks available: 0
Duration: 90 Minutes
Qualifications Required: None

Instructions (Click to expand)

If you choose to continue with this HIT, you'll take part in a large online experiment on decision making (more than 4000 participants). It will take around 10 minute to complete (most of our participants tend to finish within 8 minutes).

You will earn $1 for completing this HIT and, depending on your choices and the choices of other MTurk workers, an additional amount of up to $3.

Please follow the link below to continue with this HIT. After completing the task, you will receive a code that needs to be pasted in the box below so that we can go about paying you. We aim to pay you within two days.

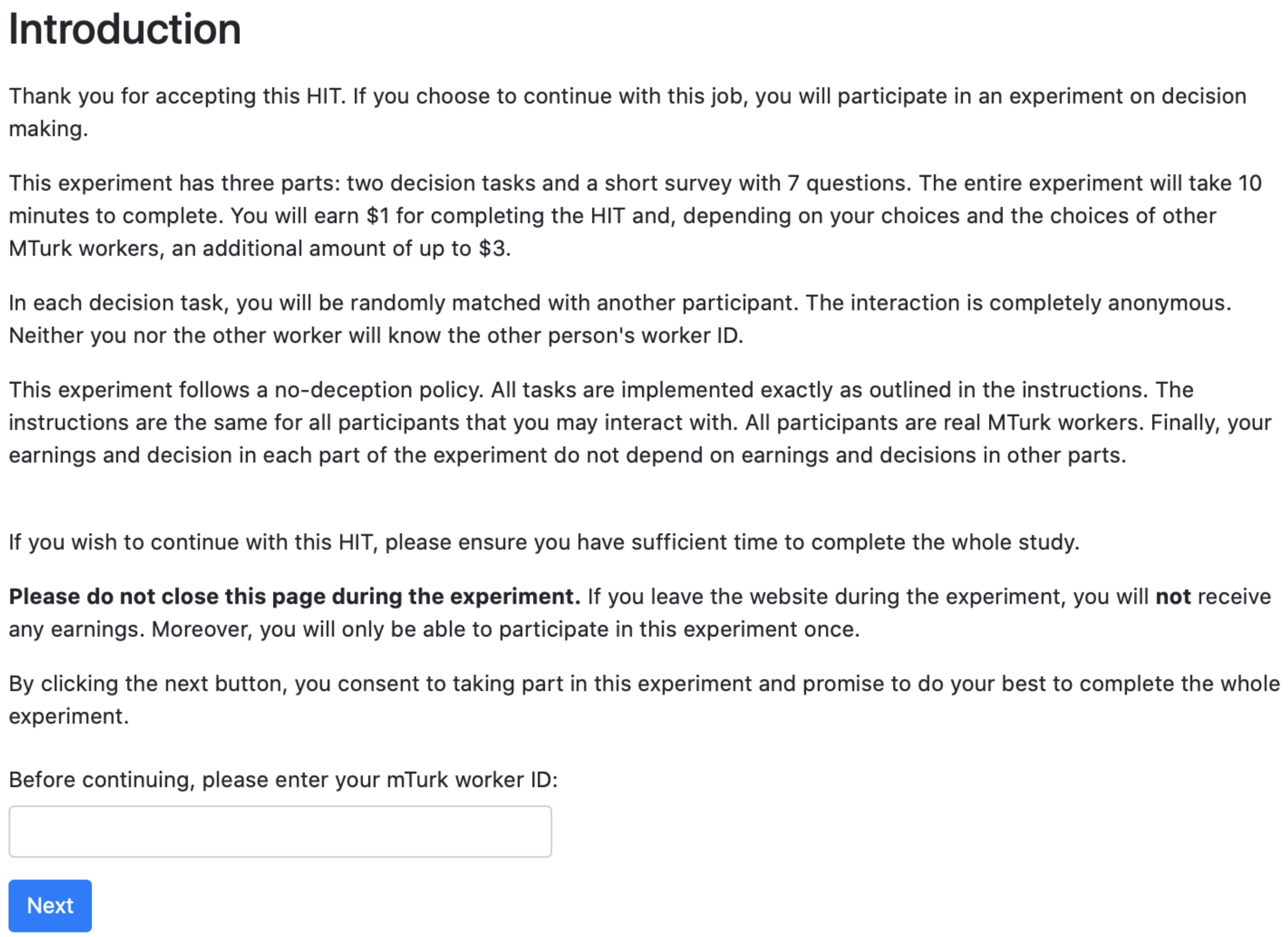
Make sure to leave this window open as you complete the task. When you are finished, please return to this page to paste the code into the box.

Link:	link to decision problem

Enter completion code here:

e.g. CC12345678

Figure F.7: Job-Ad on MTurk



Introduction

Thank you for accepting this HIT. If you choose to continue with this job, you will participate in an experiment on decision making.

This experiment has three parts: two decision tasks and a short survey with 7 questions. The entire experiment will take 10 minutes to complete. You will earn $1 for completing the HIT and, depending on your choices and the choices of other MTurk workers, an additional amount of up to $3.

In each decision task, you will be randomly matched with another participant. The interaction is completely anonymous. Neither you nor the other worker will know the other person's worker ID.

This experiment follows a no-deception policy. All tasks are implemented exactly as outlined in the instructions. The instructions are the same for all participants that you may interact with. All participants are real MTurk workers. Finally, your earnings and decision in each part of the experiment do not depend on earnings and decisions in other parts.

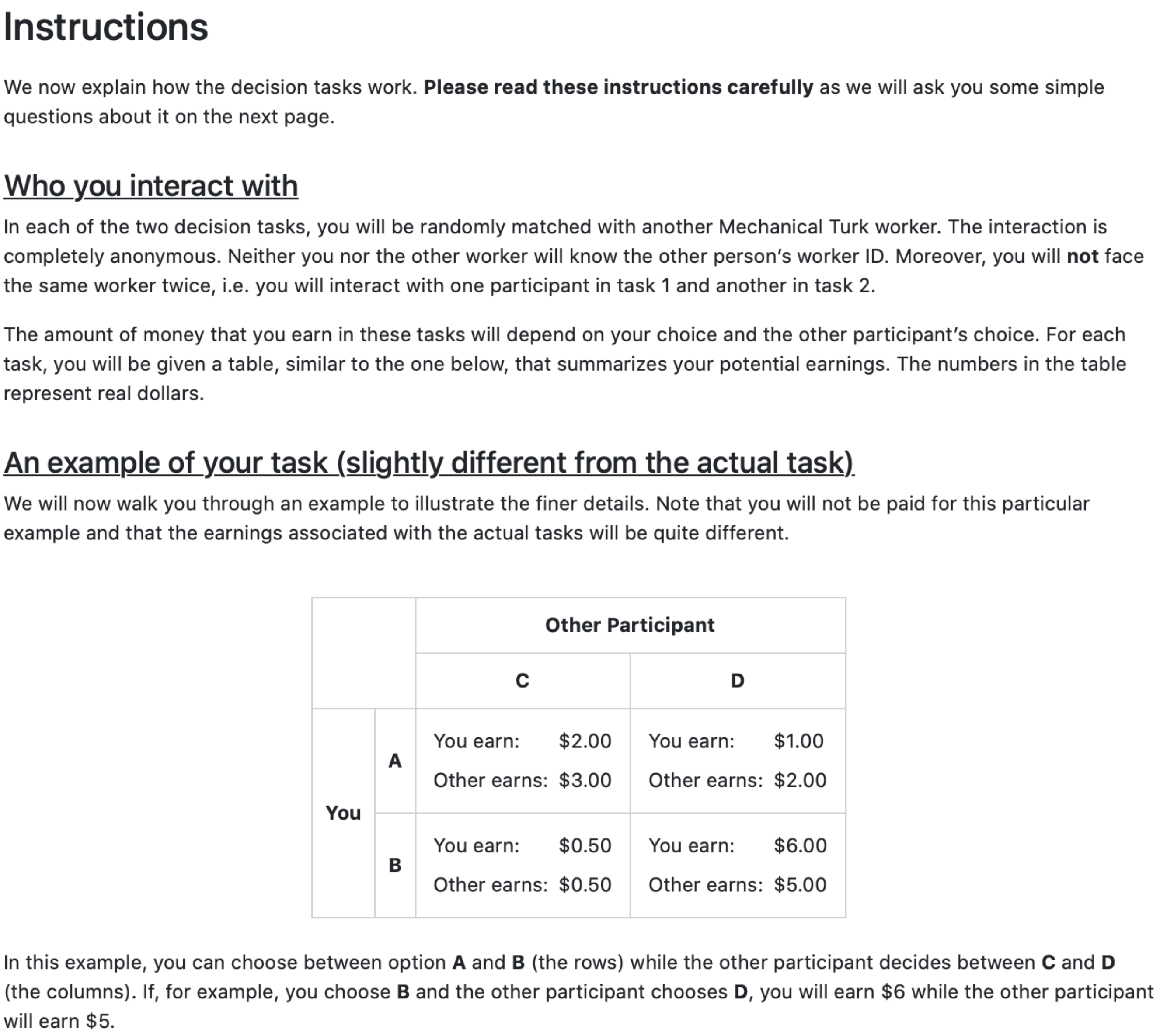
If you wish to continue with this HIT, please ensure you have sufficient time to complete the whole study.

Please do not close this page during the experiment. If you leave the website during the experiment, you will **not** receive any earnings. Moreover, you will only be able to participate in this experiment once.

By clicking the next button, you consent to taking part in this experiment and promise to do your best to complete the whole experiment.

Before continuing, please enter your mTurk worker ID:

Figure F.8: Landing page / introduction



Instructions

We now explain how the decision tasks work. **Please read these instructions carefully** as we will ask you some simple questions about it on the next page.

Who you interact with

In each of the two decision tasks, you will be randomly matched with another Mechanical Turk worker. The interaction is completely anonymous. Neither you nor the other worker will know the other person's worker ID. Moreover, you will **not** face the same worker twice, i.e. you will interact with one participant in task 1 and another in task 2.

The amount of money that you earn in these tasks will depend on your choice and the other participant's choice. For each task, you will be given a table, similar to the one below, that summarizes your potential earnings. The numbers in the table represent real dollars.

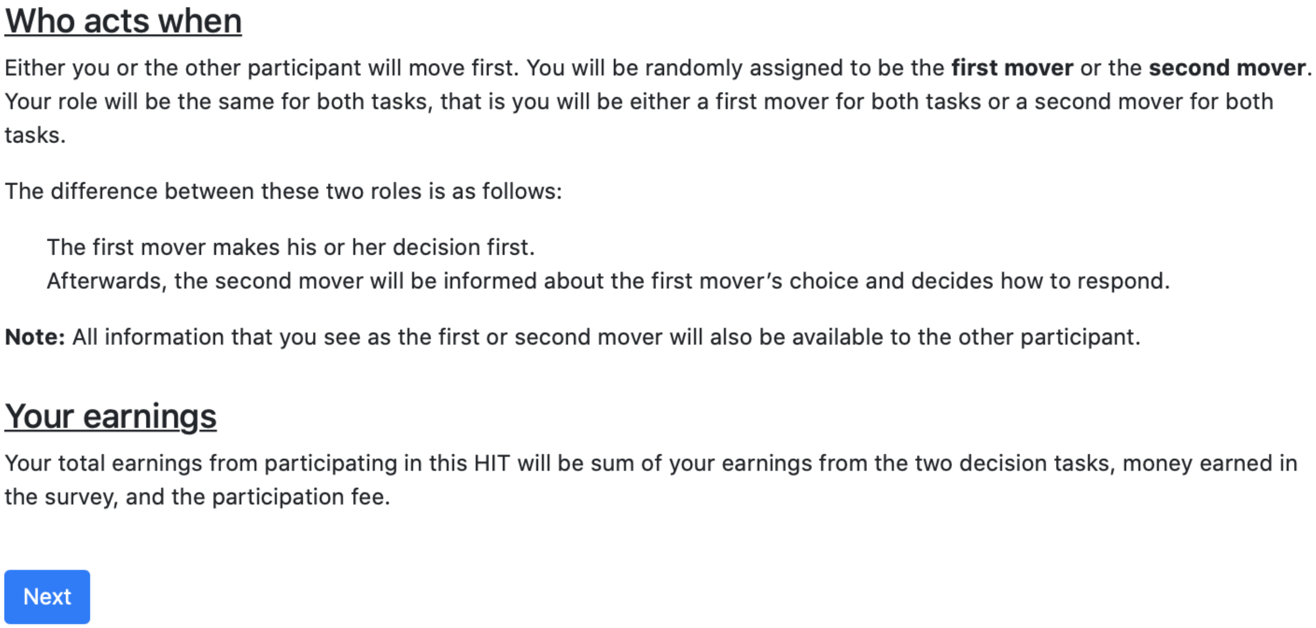
An example of your task (slightly different from the actual task)

We will now walk you through an example to illustrate the finer details. Note that you will not be paid for this particular example and that the earnings associated with the actual tasks will be quite different.

		Other Participant	
		C	**D**
You	**A**	You earn: $2.00 Other earns: $3.00	You earn: $1.00 Other earns: $2.00
	B	You earn: $0.50 Other earns: $0.50	You earn: $6.00 Other earns: $5.00

In this example, you can choose between option **A** and **B** (the rows) while the other participant decides between **C** and **D** (the columns). If, for example, you choose **B** and the other participant chooses **D**, you will earn $6 while the other participant will earn $5.

Figure F.9: Instructions, part 1



Who acts when

Either you or the other participant will move first. You will be randomly assigned to be the **first mover** or the **second mover**. Your role will be the same for both tasks, that is you will be either a first mover for both tasks or a second mover for both tasks.

The difference between these two roles is as follows:

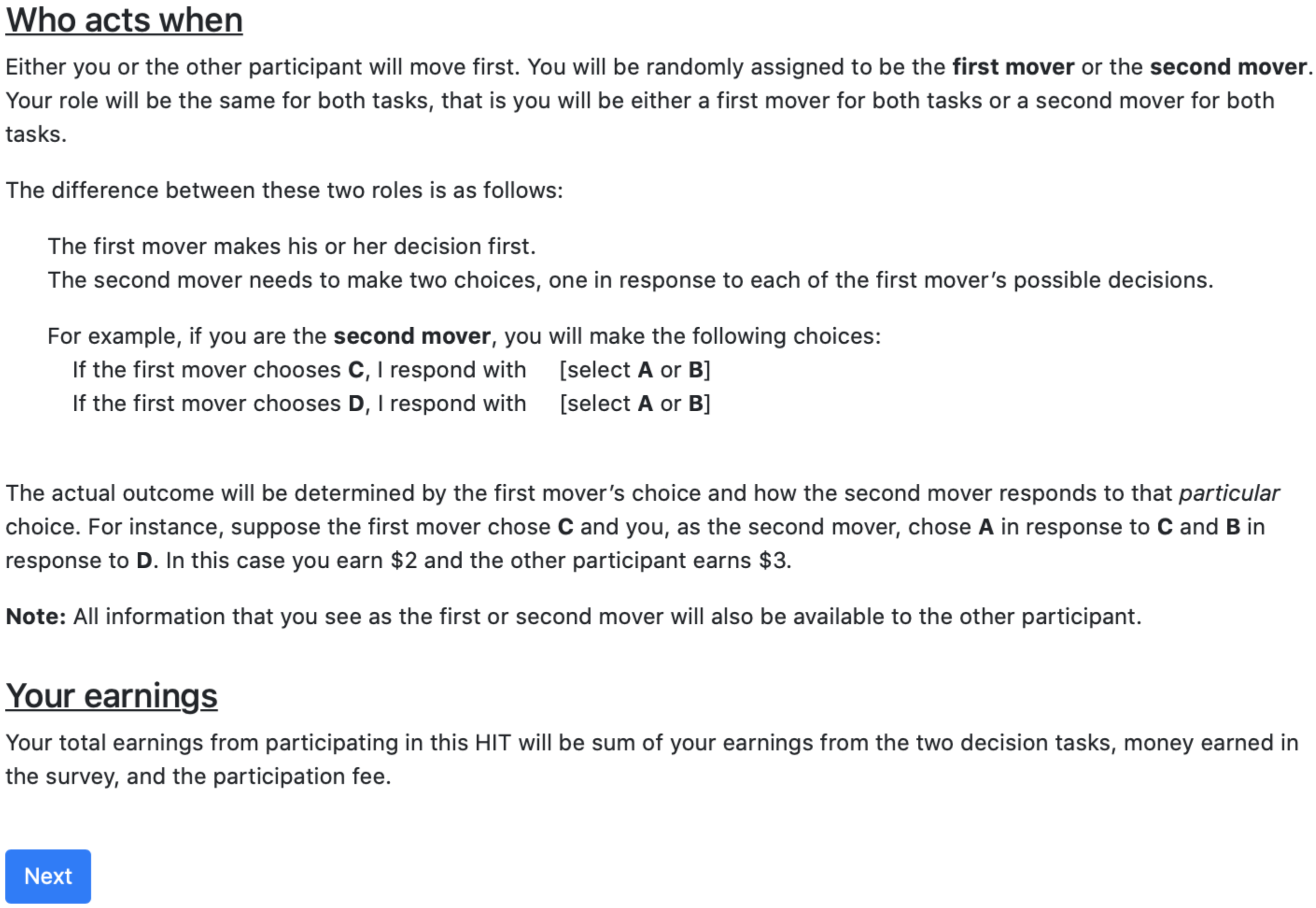
> The first mover makes his or her decision first.
> Afterwards, the second mover will be informed about the first mover’s choice and decides how to respond.

Note: All information that you see as the first or second mover will also be available to the other participant.

Your earnings

Your total earnings from participating in this HIT will be sum of your earnings from the two decision tasks, money earned in the survey, and the participation fee.

Figure F.10: Instructions, part 2 - direct response treatment



Who acts when

Either you or the other participant will move first. You will be randomly assigned to be the **first mover** or the **second mover**. Your role will be the same for both tasks, that is you will be either a first mover for both tasks or a second mover for both tasks.

The difference between these two roles is as follows:

> The first mover makes his or her decision first.
> The second mover needs to make two choices, one in response to each of the first mover’s possible decisions.
>
> For example, if you are the **second mover**, you will make the following choices:
> If the first mover chooses **C**, I respond with [select **A** or **B**]
> If the first mover chooses **D**, I respond with [select **A** or **B**]

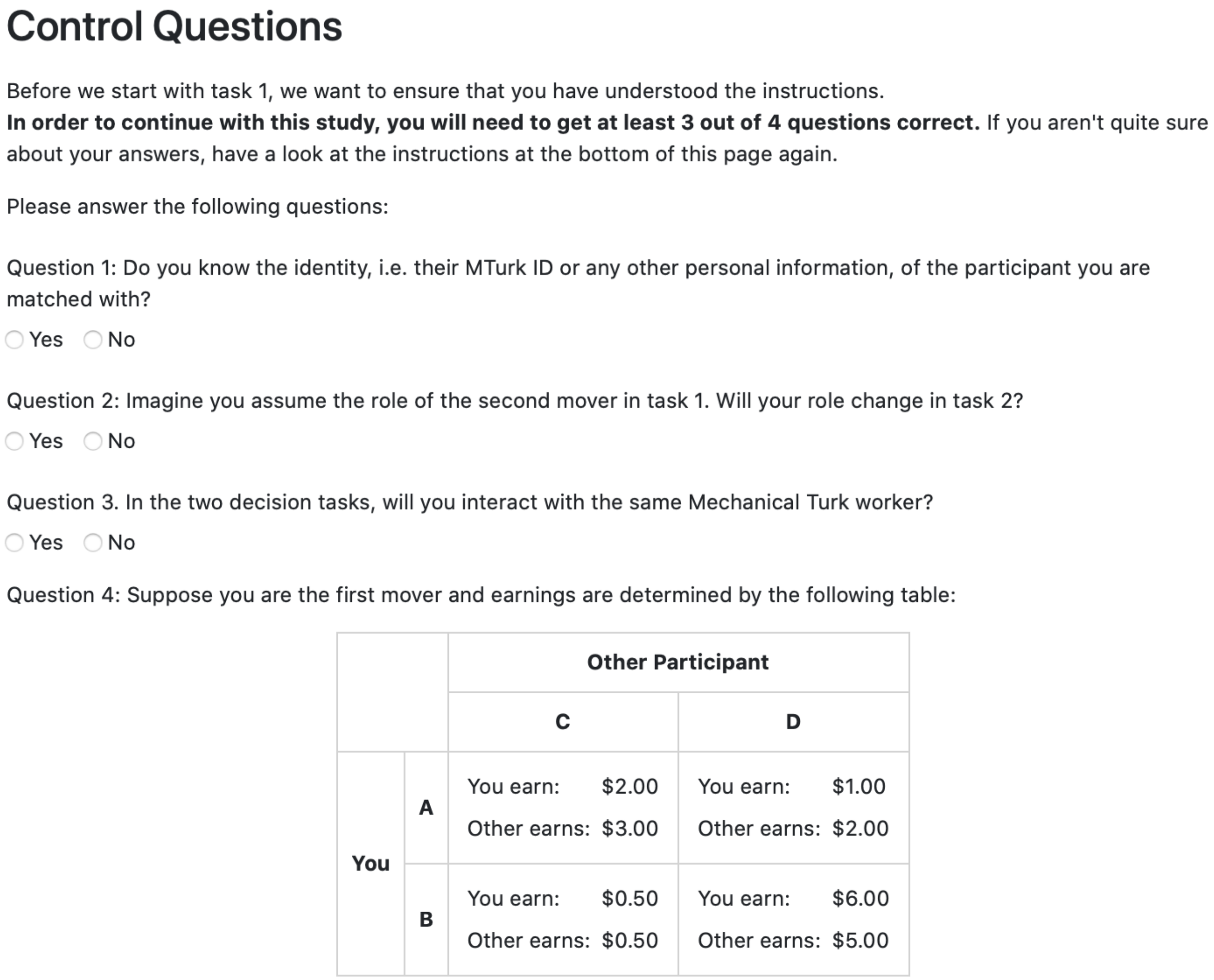
The actual outcome will be determined by the first mover’s choice and how the second mover responds to that *particular* choice. For instance, suppose the first mover chose **C** and you, as the second mover, chose **A** in response to **C** and **B** in response to **D**. In this case you earn $2 and the other participant earns $3.

Note: All information that you see as the first or second mover will also be available to the other participant.

Your earnings

Your total earnings from participating in this HIT will be sum of your earnings from the two decision tasks, money earned in the survey, and the participation fee.

Figure F.11: Instructions, part 2 - strategy method treatment



Control Questions

Before we start with task 1, we want to ensure that you have understood the instructions.
In order to continue with this study, you will need to get at least 3 out of 4 questions correct. If you aren't quite sure about your answers, have a look at the instructions at the bottom of this page again.

Please answer the following questions:

Question 1: Do you know the identity, i.e. their MTurk ID or any other personal information, of the participant you are matched with?

○ Yes ○ No

Question 2: Imagine you assume the role of the second mover in task 1. Will your role change in task 2?

○ Yes ○ No

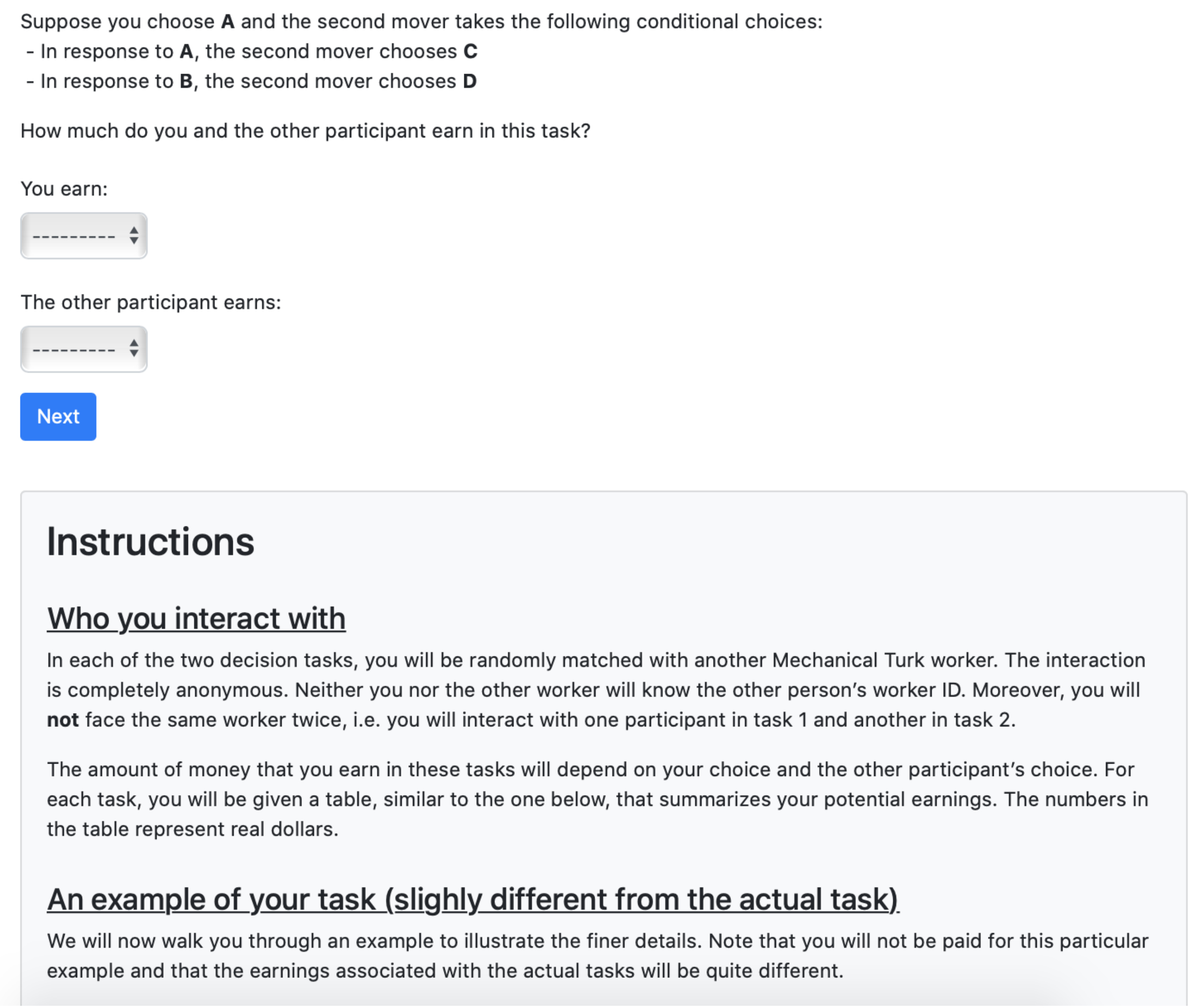
Question 3. In the two decision tasks, will you interact with the same Mechanical Turk worker?

○ Yes ○ No

Question 4: Suppose you are the first mover and earnings are determined by the following table:

		Other Participant	
		C	**D**
You	**A**	You earn: $2.00 Other earns: $3.00	You earn: $1.00 Other earns: $2.00
	B	You earn: $0.50 Other earns: $0.50	You earn: $6.00 Other earns: $5.00

Figure F.12: Control question



Suppose you choose **A** and the second mover takes the following conditional choices:
- In response to **A**, the second mover chooses **C**
- In response to **B**, the second mover chooses **D**

How much do you and the other participant earn in this task?

You earn:

The other participant earns:

Instructions

Who you interact with

In each of the two decision tasks, you will be randomly matched with another Mechanical Turk worker. The interaction is completely anonymous. Neither you nor the other worker will know the other person's worker ID. Moreover, you will **not** face the same worker twice, i.e. you will interact with one participant in task 1 and another in task 2.

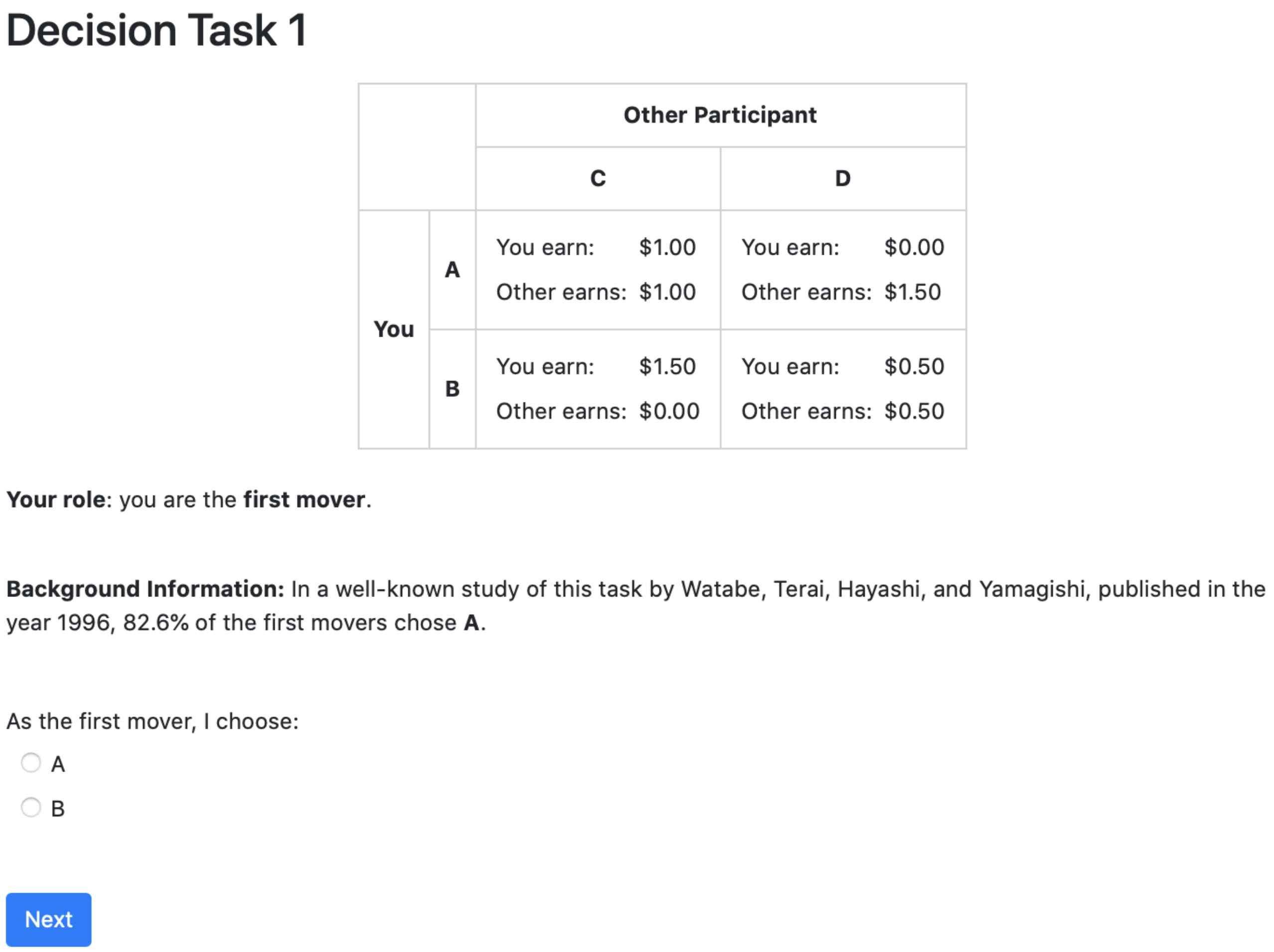
The amount of money that you earn in these tasks will depend on your choice and the other participant's choice. For each task, you will be given a table, similar to the one below, that summarizes your potential earnings. The numbers in the table represent real dollars.

An example of your task (slighly different from the actual task)

We will now walk you through an example to illustrate the finer details. Note that you will not be paid for this particular example and that the earnings associated with the actual tasks will be quite different.

Figure F.13: Control question, continued

Note: in the grey box at the bottom of the page, the full set of instructions (from the prior page) are repeated for the participant, but are cropped for reasons of space in this screenshot.



Decision Task 1

		Other Participant	
		C	D
You	A	You earn: $1.00 Other earns: $1.00	You earn: $0.00 Other earns: $1.50
	B	You earn: $1.50 Other earns: $0.00	You earn: $0.50 Other earns: $0.50

Your role: you are the **first mover**.

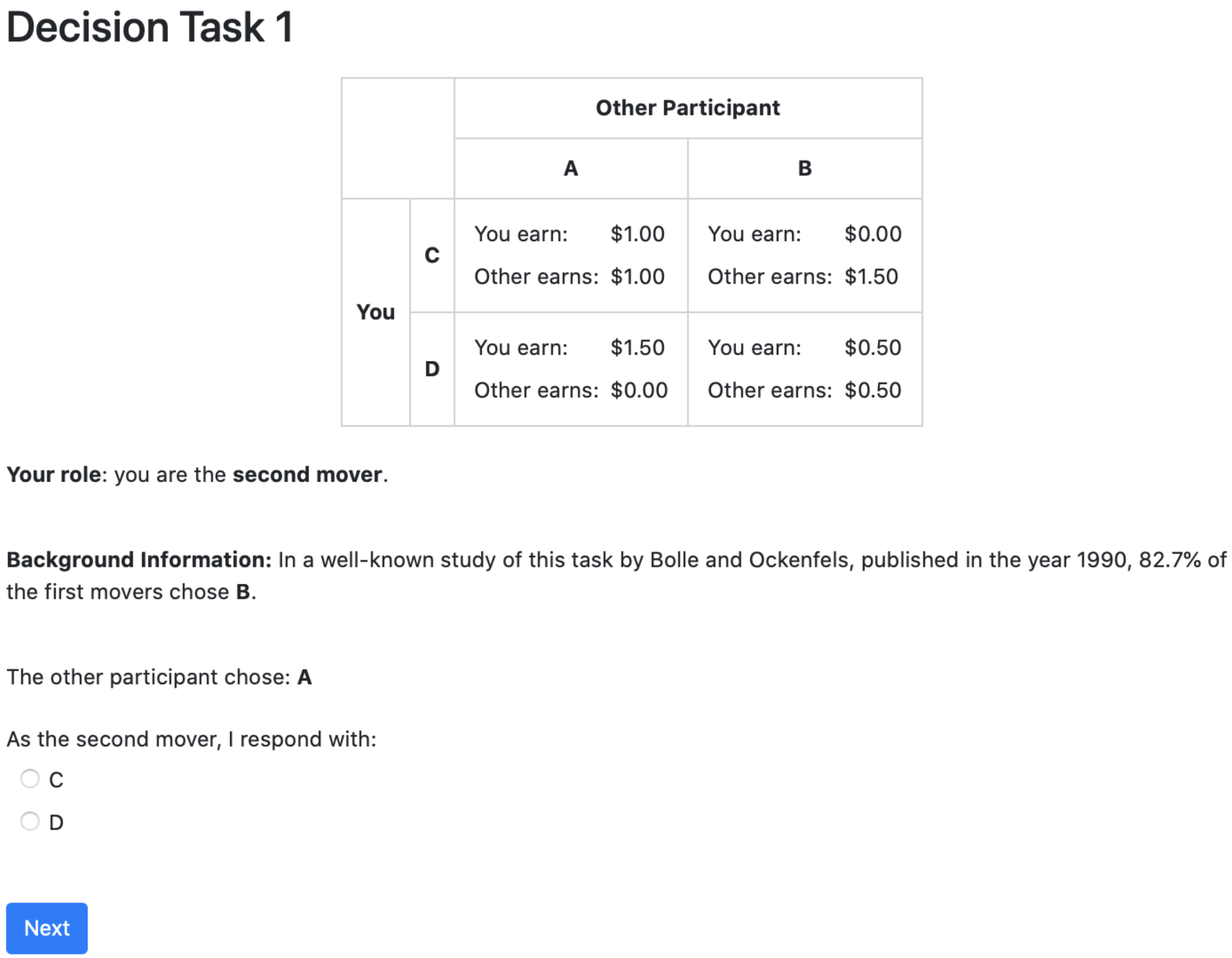
Background Information: In a well-known study of this task by Watabe, Terai, Hayashi, and Yamagishi, published in the year 1996, 82.6% of the first movers chose **A**.

As the first mover, I choose:

○ A

○ B

Figure F.14: Task 1, player 1, non-selfish belief treatment (sPD)



Decision Task 1

		Other Participant	
		A	**B**
You	**C**	You earn: $1.00 Other earns: $1.00	You earn: $0.00 Other earns: $1.50
	D	You earn: $1.50 Other earns: $0.00	You earn: $0.50 Other earns: $0.50

Your role: you are the **second mover**.

Background Information: In a well-known study of this task by Bolle and Ockenfels, published in the year 1990, 82.7% of the first movers chose **B**.

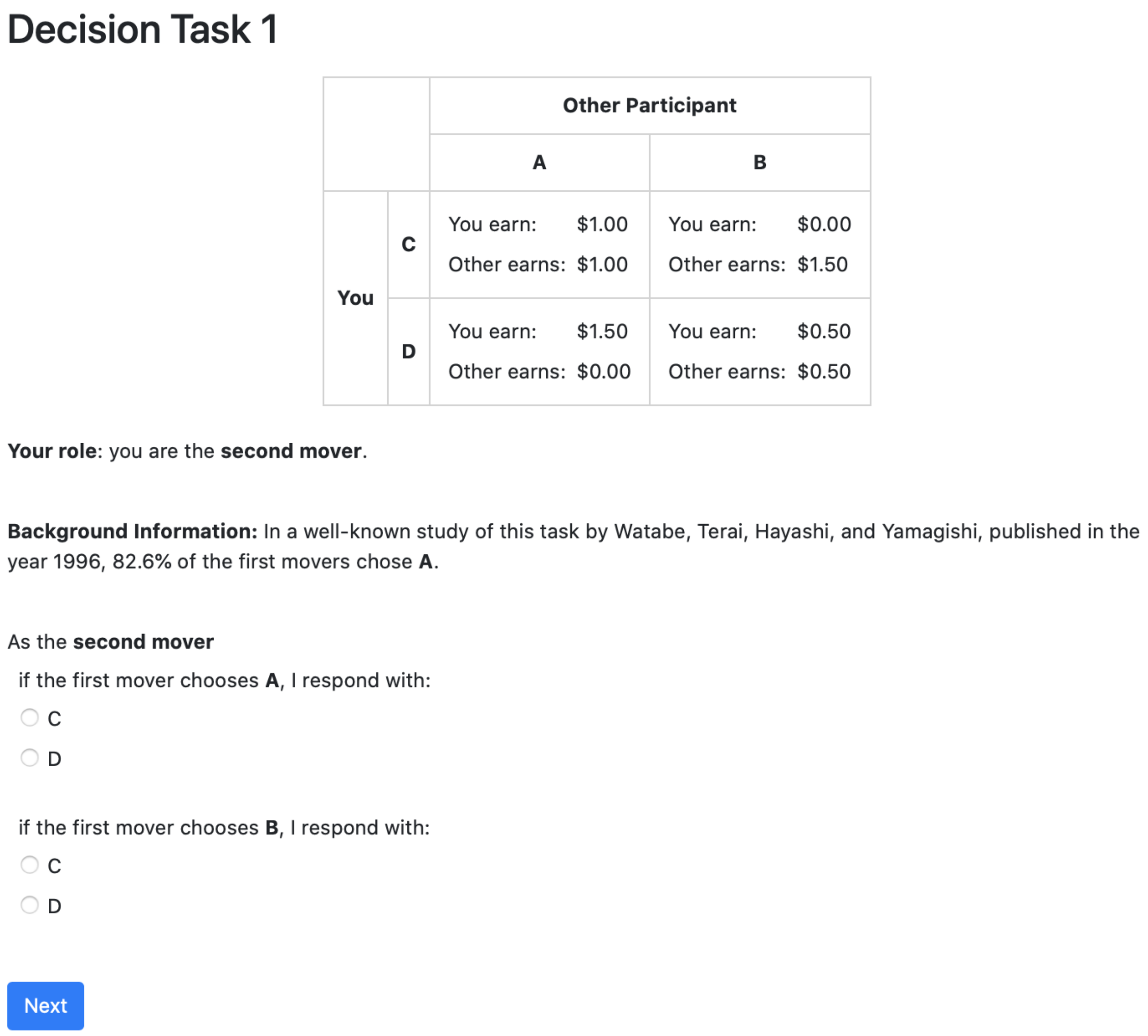
The other participant chose: **A**

As the second mover, I respond with:

○ C

○ D

Figure F.15: Task 1, player 2, direct-response, selfish belief treatment (sPD)



Decision Task 1

		Other Participant	
		A	B
You	C	You earn: $1.00 Other earns: $1.00	You earn: $0.00 Other earns: $1.50
	D	You earn: $1.50 Other earns: $0.00	You earn: $0.50 Other earns: $0.50

Your role: you are the **second mover**.

Background Information: In a well-known study of this task by Watabe, Terai, Hayashi, and Yamagishi, published in the year 1996, 82.6% of the first movers chose **A**.

As the **second mover**

if the first mover chooses **A**, I respond with:

○ C

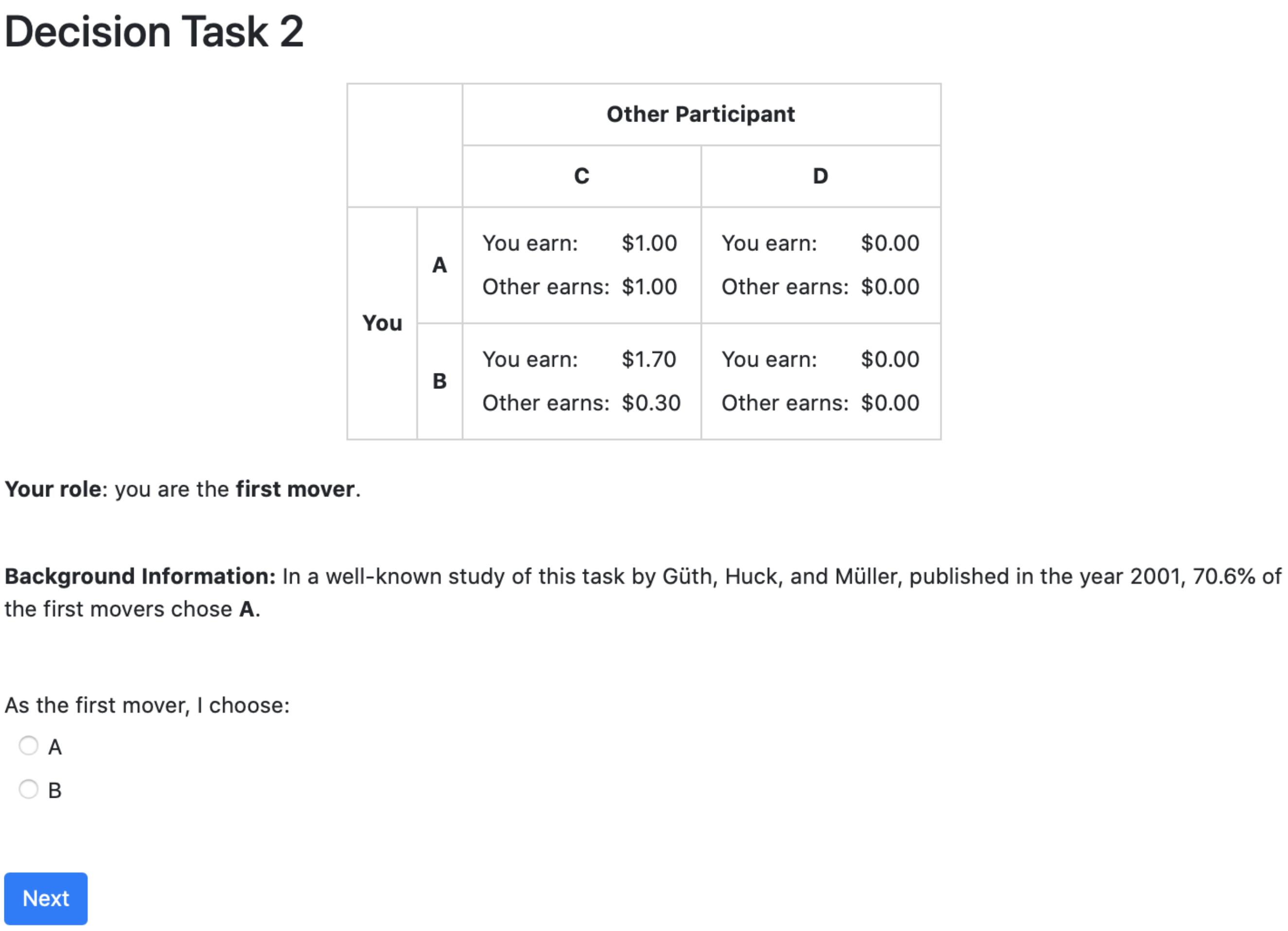
○ D

if the first mover chooses **B**, I respond with:

○ C

○ D

Figure F.16: Task 1, player 2, strategy method, non-selfish belief treatment (*sPD*)



Decision Task 2

		Other Participant	
		C	D
You	A	You earn: $1.00 Other earns: $1.00	You earn: $0.00 Other earns: $0.00
	B	You earn: $1.70 Other earns: $0.30	You earn: $0.00 Other earns: $0.00

Your role: you are the **first mover**.

Background Information: In a well-known study of this task by Güth, Huck, and Müller, published in the year 2001, 70.6% of the first movers chose **A**.

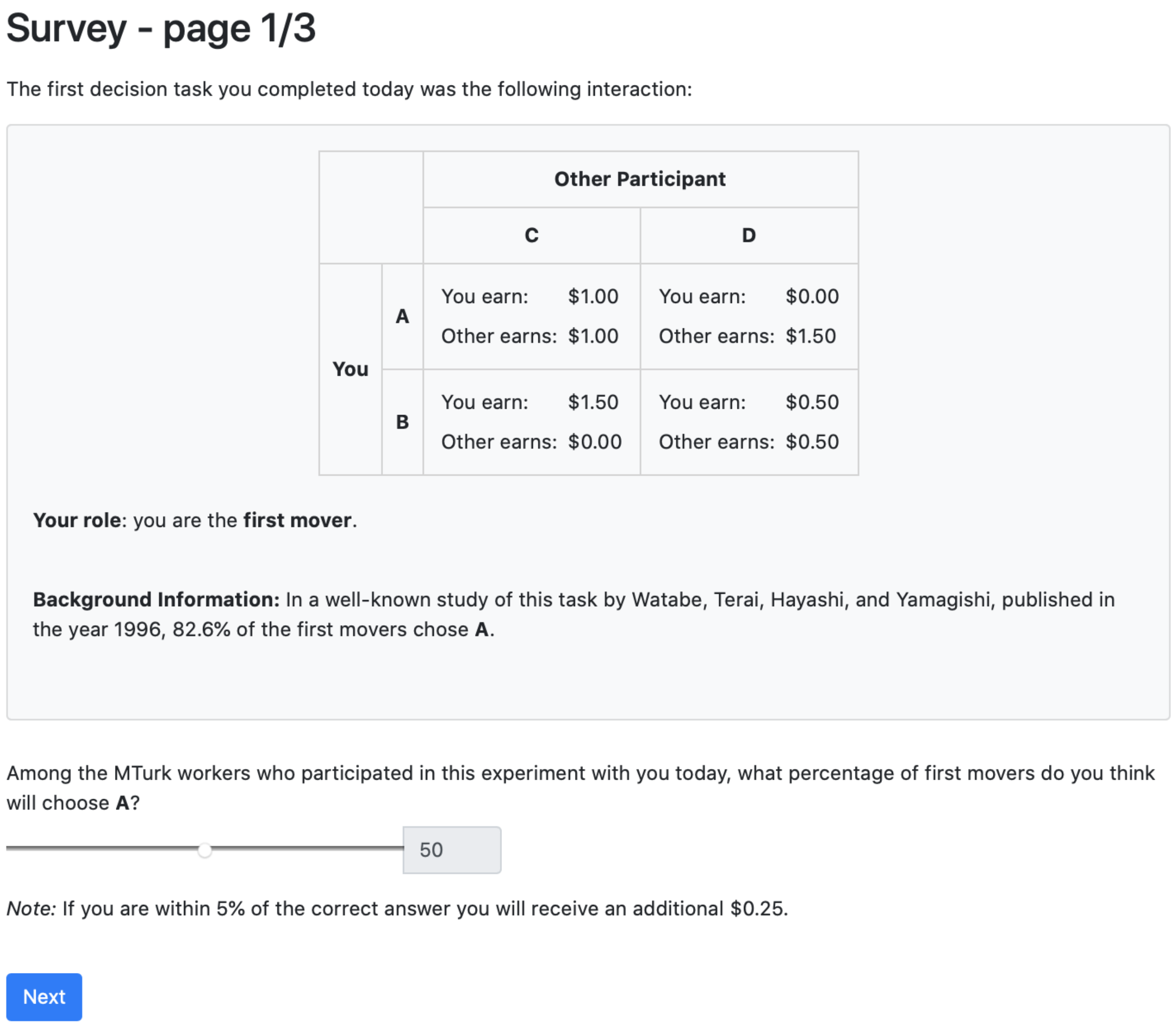
As the first mover, I choose:

○ A

○ B

Next

Figure F.17: Task 1, player 2, strategy method, non-selfish belief treatment (mUG)



Survey - page 1/3

The first decision task you completed today was the following interaction:

		Other Participant	
		C	**D**
You	**A**	You earn: $1.00 Other earns: $1.00	You earn: $0.00 Other earns: $1.50
	B	You earn: $1.50 Other earns: $0.00	You earn: $0.50 Other earns: $0.50

Your role: you are the **first mover**.

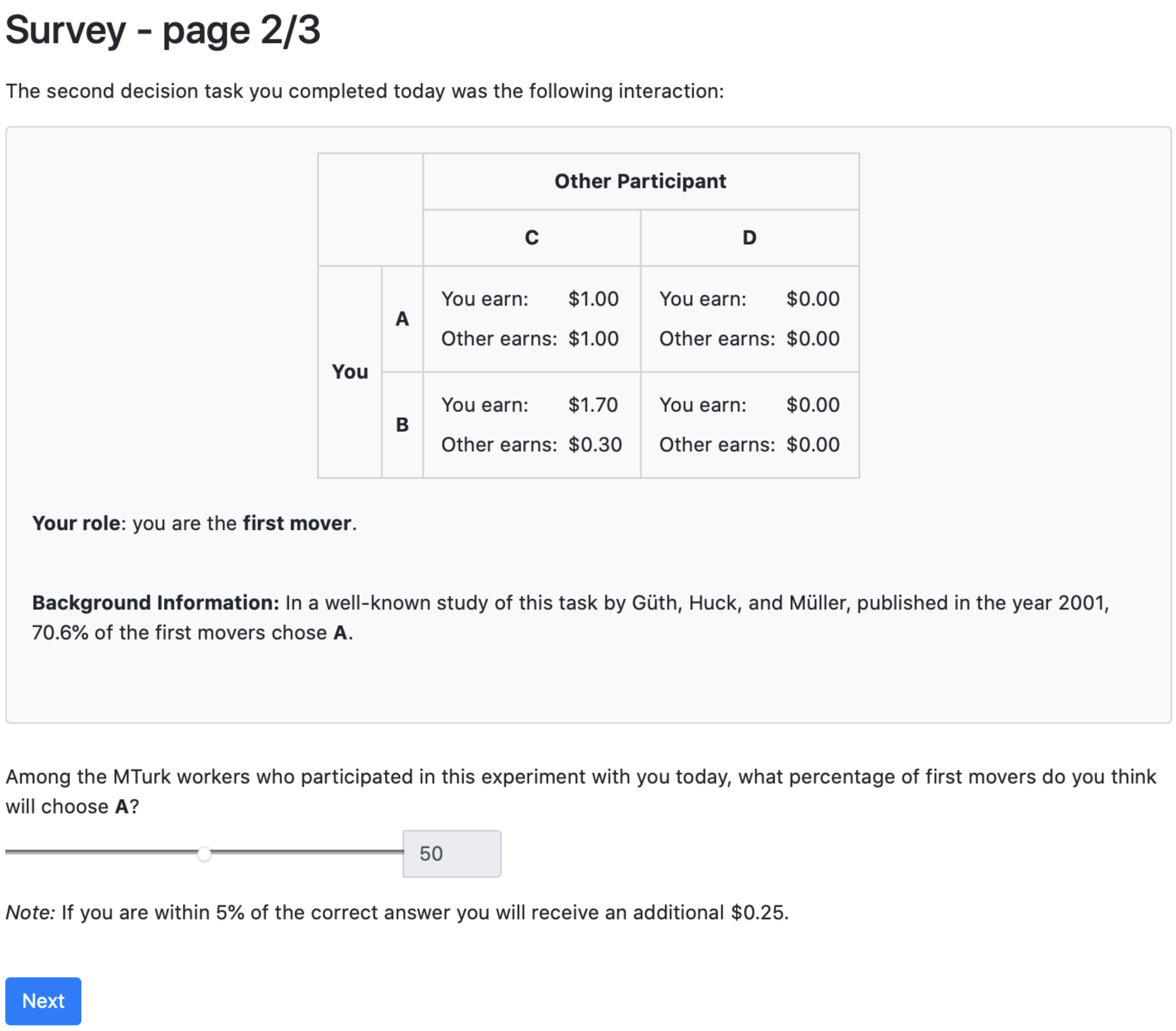
Background Information: In a well-known study of this task by Watabe, Terai, Hayashi, and Yamagishi, published in the year 1996, 82.6% of the first movers chose **A**.

Among the MTurk workers who participated in this experiment with you today, what percentage of first movers do you think will choose **A**?

50

Note: If you are within 5% of the correct answer you will receive an additional $0.25.

Figure F.18: Survey page 1, Belief Elicitation for task 1 of a player 1 in the non-selfish belief treatment (sPD)



Survey - page 2/3

The second decision task you completed today was the following interaction:

		Other Participant	
		C	D
You	A	You earn: $1.00 Other earns: $1.00	You earn: $0.00 Other earns: $0.00
	B	You earn: $1.70 Other earns: $0.30	You earn: $0.00 Other earns: $0.00

Your role: you are the **first mover**.

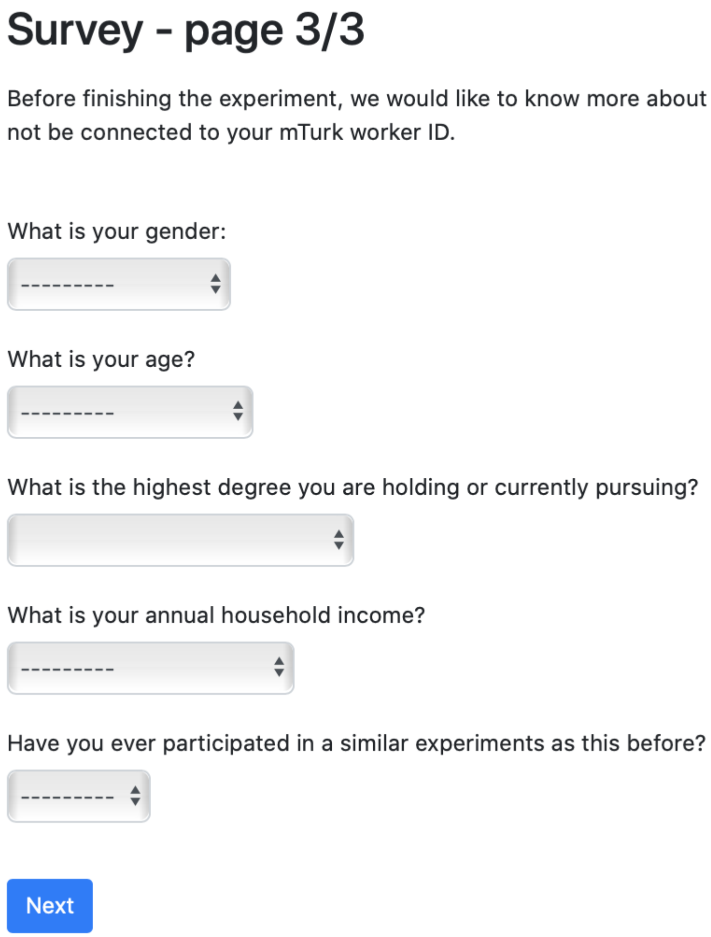
Background Information: In a well-known study of this task by Güth, Huck, and Müller, published in the year 2001, 70.6% of the first movers chose **A**.

Among the MTurk workers who participated in this experiment with you today, what percentage of first movers do you think will choose **A**?

50

Note: If you are within 5% of the correct answer you will receive an additional $0.25.

Figure F.19: Survey page 2, Belief Elicitation for task 2 of a player 1 in the non-selfish belief treatment (mUG)



Survey - page 3/3

Before finishing the experiment, we would like to know more about you. All answers will be processed anonymously and will not be connected to your mTurk worker ID.

What is your gender:

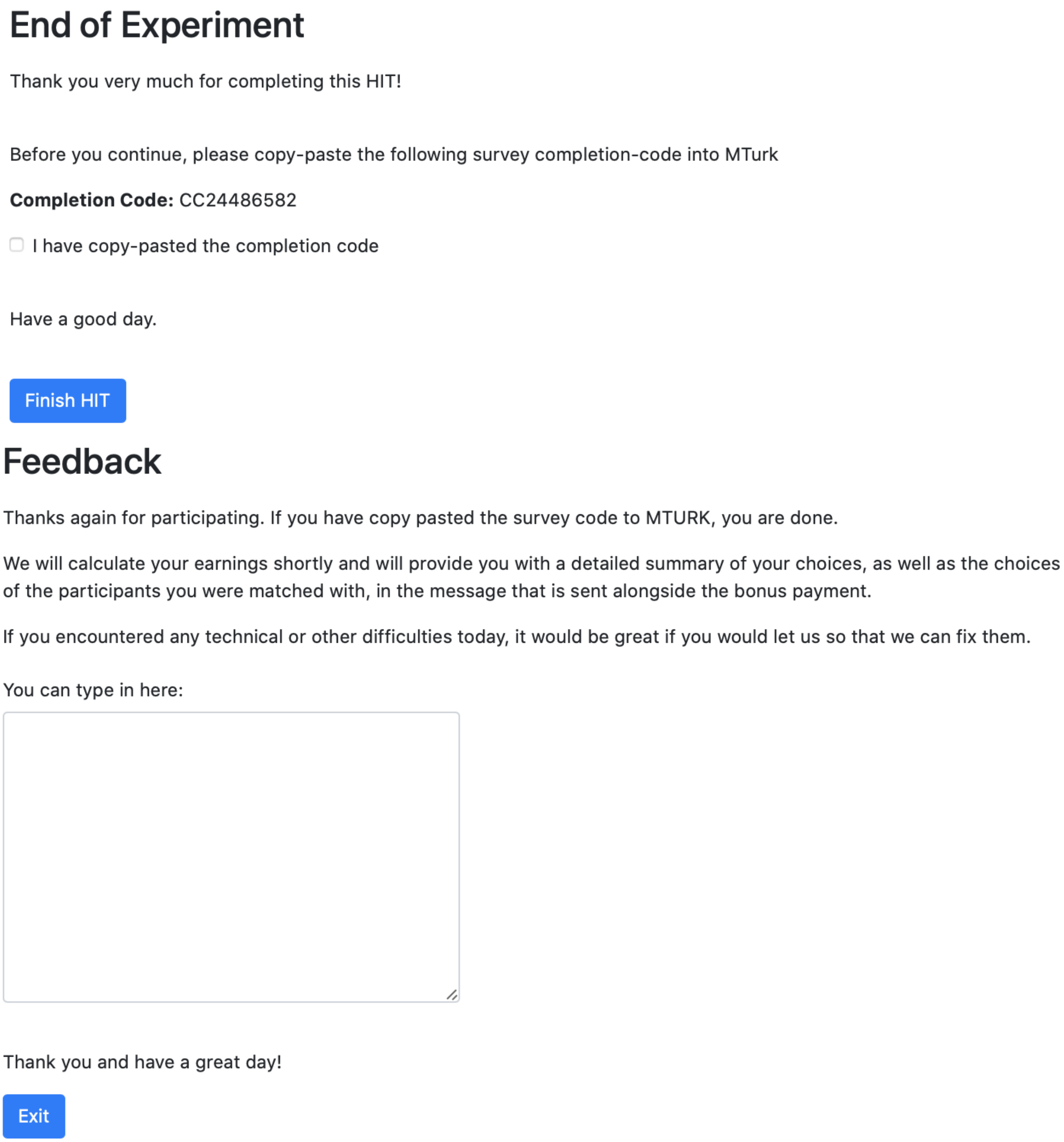
What is your age?

What is the highest degree you are holding or currently pursuing?

What is your annual household income?

Have you ever participated in a similar experiments as this before?

Figure F.20: Survey page 3



End of Experiment

Thank you very much for completing this HIT!

Before you continue, please copy-paste the following survey completion-code into MTurk

Completion Code: CC24486582

□ I have copy-pasted the completion code

Have a good day.

Feedback

Thanks again for participating. If you have copy pasted the survey code to MTURK, you are done.

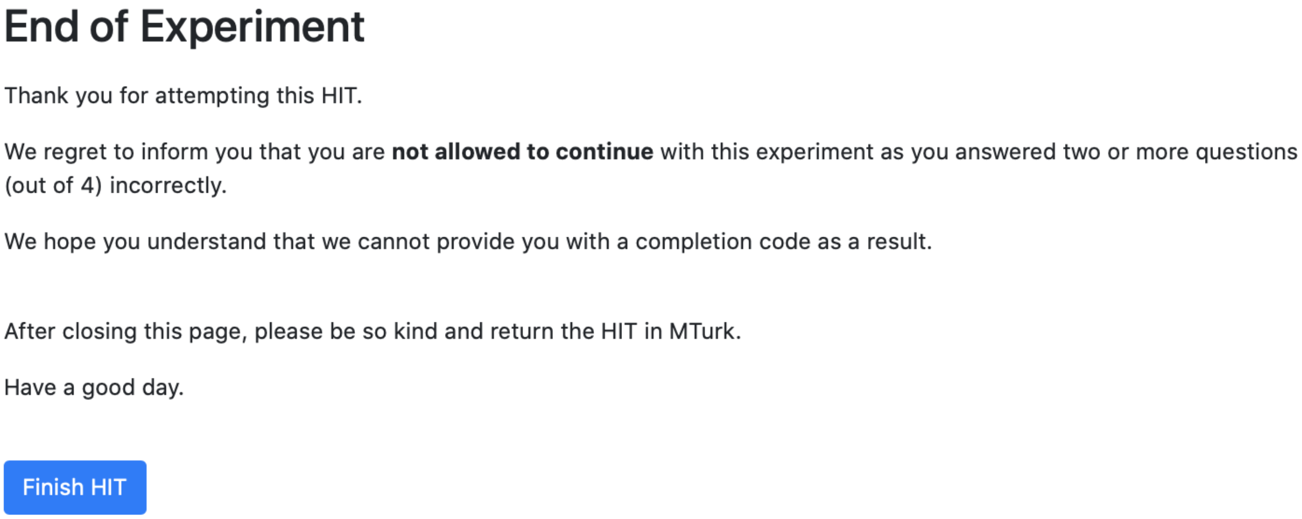
We will calculate your earnings shortly and will provide you with a detailed summary of your choices, as well as the choices of the participants you were matched with, in the message that is sent alongside the bonus payment.

If you encountered any technical or other difficulties today, it would be great if you would let us so that we can fix them.

You can type in here:

Thank you and have a great day!

Figure F.21: Completion page and optional feedback Page



End of Experiment

Thank you for attempting this HIT.

We regret to inform you that you are **not allowed to continue** with this experiment as you answered two or more questions (out of 4) incorrectly.

We hope you understand that we cannot provide you with a completion code as a result.

After closing this page, please be so kind and return the HIT in MTurk.

Have a good day.

Figure F.22: Early termination screen for those who did pass the control questions