Julian Detemple

Thoughts about the Dictator and Trust Game

SAFE Working Paper No. 422 | June 2024

Leibniz Institute for Financial Research SAFE
Sustainable Architecture for Finance in Europe
info@safe-frankfurt.de | www.safe-frankfurt.de

Thoughts about the Dictator and Trust Game

Julian Detemple*

June 2024

Abstract

Experiments are an important tool in economic research. However, it is unclear to which extent the control of experiments extends to the perceptions subjects form of such experimental decision situations. This paper is the first to explicitly elicit perceptions of the dictator and trust game and shows that there is substantial heterogeneity in how subjects perceive the same game. Moreover, game perceptions depend not only on the game itself but also on the order of games (i.e., the broader experimental context in which the game is embedded) and the subject herself. This highlights that the control of experiments does not necessarily extend to game perceptions. The paper also demonstrates that perceptions are correlated with game behavior and moderate the relationship between game behavior and field behavior, thereby underscoring the importance and relevance of game perceptions for economic research.

Keywords: dictator game, trust game, game perceptions
JEL Classification: C90, D01, D91

*E-mail: detemple@econ.uni-frankfurt.de. I thank Michael Kosfeld, Simon Heß, Florian Hett, David Yang, Noam Yuchtman, Na Zou, and seminar participants at Goethe University and the Center for Leadership and Behavior in Organizations for helpful comments and discussions. Julian Detemple gratefully acknowledges support from the Joachim Herz foundation, the Karl Schlecht foundation, and the SAFE Experiment Center at the Leibniz Institute for Financial Research. Jannis Bernschein, Julia Kunz, and Marcel Rindt provided excellent research assistance.

1 Introduction

Experiments based on stylized and controlled decision situations have become a crucial tool in economic research, both in the lab (Falk and Heckman, 2009) and in the field (Viceisza, 2016; Gneezy and Imas, 2017). On the one hand, the control afforded by experiments allows researchers to causally test hypotheses on human behavior by introducing exogenous treatment variation in the underlying decision situation (e.g., Falkinger et al., 2000; Ambrus and Greiner, 2012; Duffy and Puzzello, 2014). On the other hand, the controlled and often abstract environments also lend themselves well to measuring risk, social, and time preferences (Charness, Gneezy and Imas, 2013; Cohen et al., 2020; Charness and Fehr, 2023), with heterogeneity in behavior revealing heterogeneity in preferences and, in strategic games, heterogeneity in beliefs (Fischbacher, Gächter and Fehr, 2001).

Research, however, suggests that subjects in experiments might not think in terms of the game-theoretic model of the experiment, thereby potentially allowing for additional heterogeneity outside the control of experiments to influence decision-making. Consider, for example, the literature on framing (and to some extent priming) effects: different ways to frame the same decision situation can evoke different beliefs, norms, and potentially even preferences (e.g., Ellingsen et al., 2012; Dreber et al., 2013; Chang, Chen and Krupka, 2019). This requires that subjects in the experiment have a perception — or, in the language of that literature, a "frame" — in mind that differs from the game-theoretic model of the underlying decision situation. Another strand in the experimental literature looks at the effect of abstract instructions and argues that in abstract decision situations, subjects project their own "frame" on the decision situation (Engel and Rand, 2014; Alekseev, Charness and Gneezy, 2017). This notion is also expressed by Levitt and List (2007) in their critical assessment of laboratory measures of social preferences. Finally, Henrich et al. (2001) interpret their findings in a cross-cultural study on behavioral experiments that "when faced with a novel situation (the experiment), they [the subjects] looked for analogues in their daily experience, asking ‘What familiar situation is this game like?’ ” (p.76, Henrich et al., 2001).

However, despite this suggestive evidence, little is known about subjects’ perceptions of experimental decision situations and to which extent the control of experiments does extend to them.[1] This question is of particular importance, considering that recent research on decision-making processes more generally points to experience effects (Malmendier, 2021) and effects of associative memory (Gennaioli and Shleifer, 2010; Bordalo, Gennaioli and Shleifer, 2020; Bordalo et al., 2021) on decision-making. Moreover, Ockenfels and Schier (2020) show that the order of experiments influences behavior in them — presumably due to the first experiment influencing the perception of the following experiment. Consequently, the perception of an experiment could depend not only on the experiment itself but also on the broader experimental context in which it is embedded and, most importantly, the subject herself. While treatment-dependent perceptions could provide insights into the mechanisms of treatment effects, subject-dependent perceptions would imply that heterogeneity in preferences across subjects, villages, or even cultures does not immediately follow from heterogeneity in behavior and beliefs in experiments. Moreover, perceptions which also depend on the recent (and therefore more volatile) experiences made before an experiment could provide an explanation for replication failures (Camerer et al., 2016) and instability of individual behavior across time in economic experiments (Chuang and Schechter, 2015), in particular when the environment in which subjects are embedded is more volatile. Since this understanding of experiment perceptions transcends framing effects (i.e., studying the influence of exogenously provided labels or contextualized instructions) but is also more specific than mere experiment impressions (e.g., how interesting, exciting, or boring an

[1] Exceptions are, among others, Falk and Kosfeld (2006), Engel and Rand (2014), and Gächter, Kölle and Quercia (2022), who all explicitly elicit perceptions of intentions and selected features of experiments.

experiment is perceived to be), in this paper, I refer to perceptions of experiments as "mental representations" of such experimental decision situations. This terminology is borrowed from research in neuroscience that studies the mental representations of tasks to better understand models of problem-solving more generally (e.g., Ho et al., 2022).

This paper studies mental representations of a subclass of economic experiments, so-called economic games that involve interaction between at least two individuals, to shed light on three questions: first, is there variation in mental representations of economic games across individuals (i.e., does the control afforded by these economic games extend to subjects' mental representations)? Second, what is driving the heterogeneity in mental representations? Third, are mental representations of economic games relevant for economic research in that they drive behavior in economic games and moderate the relationship between game behavior and field behavior (i.e., treating behavior as preference estimates)? I conduct an online experiment in a US sample that is representative in terms of age, ethnicity, and sex and directly elicit subjects' mental representations in two economic games, the dictator and the trust game. Inspired by the literature on associative memory, I measure subjects' mental representations based on an open-ended question about associations from everyday life. Three research assistants then assign the answers to different categories. Similar results are obtained when a large language model (GPT-4) is used to classify answers and when I employ a second, closed-ended measure based on which sentences of the instruction text subjects select as influential in the way they think about the game.

I present three results. First, I document substantial heterogeneity in mental representations of the dictator and trust game, both within and across games and spanning social (i.e., representations involving another human, e.g., charitable donations) and non-social domains (i.e., representations not involving another human, e.g., investments into the stock market). The probability that two subjects have a different mental representation, i.e., stemming from two different broader domains, is 84% for the dictator game and can be as high as 88% for the trust game. Importantly, this heterogeneity is not driven by a lack of understanding of or engagement with the economic games.

Second, I show that the mental representation of an economic game depends on the game itself but also on the characteristics of the subject and the order of the games (i.e., what happens outside the game itself). Framing the games as a community decision situation, however, does not lead to systematically different mental representations within a game.

Third, I provide evidence that mental representations are relevant for research with economic games. For this, I show that mental representations are correlated with behavior in the game. This indicates that accounting for mental representations can contribute to a more comprehensive understanding of human behavior in economic games — in particular when considering that exogenous treatment variation can potentially change representations. Moreover, I demonstrate that mental representations moderate the correlation between game and field behavior, i.e., when using economic games to measure preferences.

These findings have important implications for economic research. First and foremost, mental representations of economic games are heterogeneous. This indicates that, at least in the dictator and trust game, the control of experiments does not extend to subjects' mental representation of them. Furthermore, the results in this paper demonstrate that several components contribute to which mental representation is formed: the game itself, the broader (experimental) environment in which the game is embedded, and the subject herself. Second, Result 3 suggests that heterogeneity in mental representations across different studies for the same game (implemented in slightly different ways) could explain the mixed evidence on the ability of game behavior to predict field behavior (e.g., Galizzi and Navarro-Martinez, 2019; Naar, 2020). This motivates the hypothesis that unstable mental representations could explain other experimental "puzzles", such as replication failures (Camerer et al., 2016) and instability of game behavior across time in economic experiments (Chuang and Schechter, 2015). Moreover, heterogeneity in mental representations could account for the documented variation of preferences across different samples and the correlation of game behavior with socio-demographic information (Chapman et al., 2023). Finally, while this paper focuses on economic games as opposed to other types of economic experiments, there is no immediate reason why mental representations should not be heterogeneous in other non-interactive experiments, too. For example, there is evidence pointing to context-dependent risk preferences (Dohmen et al., 2011), while Charness et al. (2020) suggest that different experimental methods (e.g., frames of a task) trigger different mental processes in the risk domain. Moreover, whether subjects perceive choices in money-earlier-or-later tasks as (monetary) income or consumption should matter for experimental behavior and its interpretation (Cohen et al., 2020).

This paper contributes to several strands in the literature. First, it is related to the experimental and behavioral literature that uses economic games to test hypotheses on human behavior by introducing exogenous treatment variation. Closest to this paper is work on framing effects on behavior (Ellingsen et al., 2012; Dreber et al., 2013; Chang, Chen and Krupka, 2019) and perceived kindness of others' game actions (Gächter, Kölle and Quercia, 2022), game order effects (Ockenfels and Schier, 2020), the effect of context (Castillo et al., 2011) as well as contextualized instructions (Engel and Rand, 2014; Alekseev, Charness and Gneezy, 2017) on game behavior, and how misperceptions of a game's incentives can drive behavior (e.g., Cason and Plott, 2014). These studies all — more or less explicitly — build on the idea that subjects play a game that might differ from its game-theoretic model. Mental representations are also related to experimental work that elicits thoughts and perceptions of subjects about outcomes and behavior of others using closed-ended survey questions (e.g., Falk and Kosfeld, 2006). I contribute to this literature by being the first to explicitly elicit mental representations of economic games, documenting that there does exist heterogeneity in them, and producing evidence on which components might contribute to different mental representations. Second, my findings also directly contribute to the literature on measuring social preferences with economic games and predicting field behavior.[2] Closest to this study is recent work by Gächter, Kölle and Quercia (2022), who show that misperceptions of incentives interfere with the identification of cooperation preferences based on behavior. I demonstrate that the correlation between game and field behavior depends on the mental representation of subjects, with the correlation being stronger among subjects who have a mental representation closer to the economic interpretation of a game (e.g., altruism in a dictator game).

Third, this paper is related to research on cognitive processes and how they shape decision-making. In behavioral economics, this includes, among others, prospect theory (Kahneman and Tversky, 1979), the role of associative memory (Gennaioli and Shleifer, 2010; Bordalo, Gennaioli and Shleifer, 2020; Bordalo et al., 2021), and experience effects in finance (Malmendier, 2021). Related to this is also work from neuroeconomics on the biological foundations of strategic thinking, highlighting, for example, that different parts of the brain are active when subjects play against a computer, compared to a human. See Houser and McCabe (2014) and Camerer et al. (2015) for an overview on this. Additionally, work from cognitive sciences points to humans forming simplified mental representations of tasks to make more efficient use of cognitive resources (Ho et al., 2022). I contribute to this research by shedding light on the cognitive processes involved in decision-making in economic games, showing that the representation of such decision situations seems to be heterogeneous and subject-dependent.

Fourth, this work complements previous studies that use qualitative data in economic research. For example, Xiao and Houser (2005) analyze the con-

[2] See Charness and Fehr (2023) for a recent review.

tent of written messages in an ultimatum game to better understand the role of emotions in punishment. Andre et al. (2022) use written narratives to shed light on perceived causes for inflation, while Ferrario and Stantcheva (2022) discuss employing open-ended survey questions to elicit support for policies. This paper extends the application of qualitative data to economic games to better understand perceptions of economic games and their relevance for experimental research.

Last but not least, this paper is related to work in psychology, sociology, and anthropology on models of the selection of (game) frames (e.g., Eriksson and Strimling, 2014) and spontaneous associations with games (e.g., Yamagishi et al., 2013). The elicitation of such associations is, however, usually done in a closed-ended form based on a set of options. I contribute to this literature by directly measuring mental representations based on open-ended survey questions in a systematic way and demonstrating that the heterogeneity in mental representations is relevant for economic research. Finally, using open-ended instead of closed-ended survey questions avoids priming subjects to particular features of a mental representation. It should therefore not only mitigate experimenter-demand effects but also contribute to eliciting a broader range of associations.

This paper proceeds as follows. I describe the research design, including the experimental design, the method for measuring mental representations, and details on the sample and implementation in Section 2. Section 3 presents the results on the extent of heterogeneity in mental representations, what is driving them, and their relevance for economic research. Section 4 concludes.

2 Research Design

The research design is preregistered at OSF[3] with the following objectives: (1) elicit mental representations in the dictator and trust game based on the measures outlined below, (2) analyze the heterogeneity in mental representations, and (3) study how mental representations are correlated with subject characteristics, depend on exogenous variation in the version of the game itself (framing treatment) and the broader context in which a game is embedded (game order treatment), and moderate the correlation between game and self-reported behavior outside the game ("field behavior").

Addressing these points requires collecting three types of data: behavior in the dictator and trust game, subjects' mental representations of the respective game, and subjects' socio-demographic information as well as self-reported field behavior. I start by describing the design of the experiment which allows to gather these types of data. Afterward, I present and discuss the measure for mental representations in greater detail. Finally, I provide information on the implementation of the experiment and the sample.

2.1 Experimental Design

The experiment consists of three parts. Subjects first participate in the dictator and trust game ("game module"), in which I also elicit the respective mental representation. Thereafter, subjects report their inflation expectations in a separate survey module for another research project by a different researcher. This "inflation module" is included to obfuscate the relationship between the games and the self-reported field behavior, which is elicited, together with subjects' socio-demographic information, in a third and final module ("survey module").[4] The order of the modules is fixed.

[3] While the preregistration itself is still embargoed, it will be available at OSF via Detemple (2023).

[4] At the beginning of the experiment, subjects are informed that the overall study consists of three modules for two separate research projects by two different researchers.

Consider first the economic games. Importantly, while I refer to them as "games" in this paper, they are only referred to as "decision situations" and not games in the experiment. Both the dictator and the trust game involve two player roles: a sender and a receiver. In the dictator game, only the sender has an endowment of 100 points; the receiver does not have any points. The sender can decide how many points to send to the receiver.

In the trust game, however, both the sender and the receiver have an initial endowment of 100 points, i.e., there is no inequality between sender and receiver in the beginning. The sender can decide how many points to send to the receiver. Any points sent are doubled by the experimenter. Thereafter, the receiver can decide how many points from her initial endowment and the points received from the sender to send back. Ideally, the sender would send all points to the receiver, and the receiver would then split all points equally. This, however, requires the sender to "trust" the receiver with all her points in the first place.

The choice of the dictator and trust game as economic games is motivated by including two games that cover both strategic and non-strategic interaction. Moreover, the games exhibit a similar structure, which is easy to explain to subjects and therefore allows to reduce the time subjects need to spend on the online experiment to minimize dropouts. To mitigate potential spill-over effects by announcing the results for one game before playing the other, decisions are matched and results are announced only at the end. This is facilitated by subjects making decisions for each role, i.e., as if they were the sender in the dictator game, the sender in the trust game, and the receiver in the trust game. For the trust game, the strategy method (Selten, 1967) is used: senders can choose between three options — sending 0, 50, or 100 points — and the receiver makes separate decisions for each of these information sets.[5] Additionally, senders in the trust game indicate how much

Moreover, subjects are (truthfully) told that the second module on inflation expectations is about another research project than the game module.

[5] Based on a pilot, the choice set is limited to three options to mitigate survey fatigue,

they think the receiver will send back. At the end of the overall experiment, after all three modules, subjects are matched into pairs, a single game is randomly chosen to be relevant for the payoff, the roles are randomly allocated, and payoffs are computed.[6] Because there are multiple decisions for the receiver in the trust game, a single measure for behavior is constructed for the analyses in Section 3. In line with the literature (e.g., Glaeser et al., 2000; Riedl and Smeets, 2017; Gill et al., 2022), this is done by computing the ratio of the amount sent back by the receiver to the amount the receiver received from the sender (incl. the doubling of the points) and averaging it across the two decisions for which the receiver receives any points from the sender. I refer to this single measure as the "share returned" by the receiver.

Now consider the exogenous variation to provide causal evidence on the drivers of mental representations. The games are implemented in a 2x2 design, with random variation in the order of games (dictator-trust or trust-dictator) and the framing of the games. There is no variation in the order of the sender and receiver decisions within the trust game. The choice of these treatments is based on the literature on game order (Ockenfels and Schier, 2020) and framing effects (Ellingsen et al., 2012; Dreber et al., 2013; Chang, Chen and Krupka, 2019), which, as highlighted above, both seem to indicate that subjects form different mental representations. In the *neutral* framing condition, games were described as a "decision situation", whereas in the *community* framing condition games were referred to as a "community decision situation" throughout the instructions and all decision screens. All instructions for the games and screenshots of the decision screens are reproduced in Appendix A. The framing condition is constant across both games. Both treatments are varied randomly on the individual level and completely

which could result from repeatedly asking very similar variants of the same question for the receiver role in the trust game.

[6] To minimize wait times, the matching is done on the fly when two subjects finish the experiment. If a subject needs to wait longer than five minutes, she is matched with a participant from a previous survey. In this case, her decision only affects her own payoffs. Subjects are informed about this on the wait page.

independently of each other, i.e., there are two independent random draws to set the respective treatment status.

Data from the inflation expectation module is not part of this research project and subjects' expectations are elicited based on Andre et al. (2022). Socio-demographic information and self-reported field behavior are elicited in the last and final module. The selection of which field behavior to include is based on previous findings that use game behavior in the dictator and/or trust game as a preference measurement to predict the respective field behavior. In the literature, the dictator game is often used as an (imperfect) measure for altruism, while sender behavior in the trust game proxies trust and receiver behavior in the same game trustworthiness and positive reciprocity (Levitt and List, 2007).[7]

The selection of field behaviors includes the self-reported altruism score (Galizzi and Navarro-Martinez, 2019), a variant of the General Social Survey trust question (Glaeser et al., 2000), past trusting behavior in the form of lending possessions and/or money to friends (Glaeser et al., 2000), socially responsible investment behavior (Riedl and Smeets, 2017), and working or pursuing a career in the finance sector (Gill et al., 2022).[8] Appendix A provides additional details on eliciting the field behaviors.

[7] Notice that behavior in these games is, by all means, not a perfect measure of the respective preference. Moreover, the literature has not settled on a definition of trust. For the purpose of this paper, trust is treated as a behavioral act of making oneself vulnerable to another person (e.g., in line with Luhmann, 2014). This behavioral act depends on the belief about others' trustworthiness and one's own preferences, e.g., risk preferences but also more fundamental social preferences (Sapienza, Toldra-Simats and Zingales, 2013). However, to simplify language, I will also refer to trust as a preference throughout this paper.

[8] Importantly, the selection of field behaviors is motivated by their relevance for the sample at hand, e.g., contributions to communal reforesting efforts are probably not immediately applicable in the US sample at hand, even though they are of great importance for the sample in Rustagi, Engel and Kosfeld (2010).

2.2 Measurement of Mental Representations

To the best of my knowledge, this is the first paper in economics to directly measure mental representations. Given the lack of existing measures for mental representations, I construct a new measure for them, motivated by the literature on cognitive processes in decision-making, in particular economic research on associative memory (Gennaioli and Shleifer, 2010; Bordalo, Gennaioli and Shleifer, 2020; Bordalo et al., 2021). Subjects are asked about their closest association from everyday life, i.e., "thinking about everyday life, which situation does the sender/receiver role in this decision situation remind you of the most?". Subjects are asked to name the decision situation they feel reminded of and to provide some context in 1-2 sentences such that "somebody who is not you could understand the situation". While the open-ended nature of the survey question makes the subsequent analysis more complex, it does not restrict subjects to a set of options defined by the researcher. This avoids potential priming effects as well as anchoring effects, caused by the (order of) options, and also makes it more difficult to construe some of perceived experimenter-demand.[9]

I also elicit a second, closed-ended measure. Since the only information subjects receive about the decision situation is based on the instruction text, I ask subjects to indicate "which sentence(s) of the instruction did influence how you think about the sender/receiver role in this decision situation?". The results of a small-scale pilot with in-person debriefing interviews revealed that the same sentence is perceived differently by subjects (e.g., some subjects select the unequal endowment in the dictator game because this triggers a feeling of charity, while others select it because they feel entitled to it). I therefore preregistered that this closed-ended measure for mental represen-

[9] Previous work on frame selection in psychology and sociology, for example, asks subjects whether they feel more reminded of a situation involving teamwork or paying taxes (Eriksson and Strimling, 2014). This might signal to subjects that they have to associate the game with either teamwork or taxes. The results in Section 3 show that many associations do not fit into these two categories.

tations will only be used as a secondary, validation measure in the analysis of the results. The same pilot also revealed that eliciting mental representations before behavior triggers more pro-social behavior. Consequently, both the open-ended and closed-ended measures for mental representations are elicited after subjects make their decision.[10] Finally, mental representations are elicited separately for each role for which subjects make a decision, i.e., for the sender role in the dictator game and both the sender and receiver role in the trust game.

Quantitatively analyzing open-ended text responses requires classifying them into a set of categories. In line with previous work using qualitative data (e.g., Andre et al., 2022), this is done by research assistants (RAs) who are blind to the research hypotheses, with the set of categories and codebook developed by the author. After some initial training based on the data from the pilot, three RAs were given the codebook and the text responses in a research-assistant-specific randomized order without any additional data on behavior (either game or field) or treatment status. The three RAs then separately classified each text response. Afterward, results were compared and if RAs disagreed, a majority decision was taken. The cases for which not even two RAs selected the same category were presented again to them in a joint meeting and RAs had to unanimously select a single category, without the author being present.

For the set of categories, I preregistered the following approach: categories into which text responses should be classified are structured based on

[10] The display order of the two question prompts is fixed, with the sentence question coming before the association question. However, all question prompts for mental representations are shown at the same time. The interface is programmed such that subjects are presented with the instructions and the question prompt for the decision as the sender/receiver. After making a decision, the decision prompt and selection are greyed out and the question prompts for mental representations are shown. Importantly, subjects cannot modify their decision even by reloading the website or opening the link again in the same browser. See the screenshots in Section A. In the experiment, I also elicit subjects' main considerations for making a decision. This question is shown last, and, as preregistered, the data are not used in this paper.

the overarching behavioral domains in which the associations from everyday life take place. For example, helping friends or sharing food with co-workers is about interactions with everyday peers, while inserting money into a slot machine lacks any social component and is about gambling. The behavioral domains are explained in greater detail in Table 1, together with an example response fitting that category. Importantly, while the selection of the behavioral domains might seem ad-hoc at first, a machine learning algorithm that identifies clusters of text responses based on the co-occurrence of the same words produces almost the same set of behavioral domains. More details are provided in Appendix A.2.

In addition to these more general behavioral domains, I also preregistered that I will include a very narrowly defined set of categories of situations in which subjects specifically talk about "altruism", "trust", or "reciprocity". These categories are included for the purpose of analyzing whether mental representations moderate the correlation between game and field behavior. These categories only apply if a text response specifically includes the words "altruism", "trust", or "reciprocity" (or some variant of it). Moreover, they also apply if the response talks about a very specific act such as charitable behavior (altruism), personal loan of money/goods to a friend (trust), or repaying/returning a favor (reciprocity). In contrast to the more general categories, e.g., interactions with a friend, which might or might not be motivated by altruism or reciprocity, these narrowly defined preference categories are easier to interpret in the context of analyzing whether the correlation between game and field behavior depends on mental representations.[11]

However, even though a machine learning algorithm identifies a similar set of behavioral domains, the classification still requires some element of

[11] These categories therefore take precedence over the more general categories. For example, loaning the lawn mower to a friend should be assigned to the trust category and not a category about everyday peer interactions. If research assistants were unsure about which category applied, they were asked to select the one for which there was more evidence in favor of, or in case of equal evidence, the first category mentioned.

interpretation (e.g., what about a donation to a panhandler to obtain good karma). Moreover, other researchers might come up with different behavioral domains. To assuage such concerns, I employ an additional approach requiring less interpretation. Before assigning a specific category, RAs were tasked to first classify whether a text response belongs to a "social", "non-social", or "no situation" category. Social situations refer to situations involving an interaction between at least two individuals, whereas non-social situations do not feature any such an interaction (e.g., subjects talk about gambling or investing in the stock market). No situation applies if subjects state that they do not feel reminded of any situation from everyday life.

Table 1: Categories of Mental Representations

Category	**Description/Example**[*]
Social:	
Reciprocity	Any response containing the word "reciprocity" (or some form of it) or describing an explicit act of reciprocity with a proverb, e.g., "repaying/returning the favor", "eye for an eye", "tit for tat", "quid pro quo", ... Example: *tit for tat — someone do good to me, then I do good for him*
Trust	Any response containing the word "trust" (or some form of it) or a personal loan of money/goods from an individual to another individual. Example: *A situation involving trust — If I lend my car to someone and they return it in good shape and full of gas.*
Charity/ Altruism	Any response containing the word "altruism" (or some form of it) or describing an explicit charitable act (e.g., donation to charity). Example: *Giving money to a panhandler — It's similar to a situation in which a stranger with no money came up to me and asked me for money.*

Table 1 Continued from previous page

Category	**Description/Example***
Everyday Peers	Any social act involving helping, sharing, gifting, etc. between two persons who know each other and normally interact as peers (e.g., friends, siblings, co-workers). Example: *giving to friends — Sometimes I have something that I would like to share with friends. I can decide how much to keep for myself and how much to give away.*
Parenting	Any interaction across family generations. Example: *Parents and children — One person has money and they have to decide whether to share it with another person.*
Hierarchical Interaction	Any interaction across hierarchies within an organization (e.g., firm, sports club, etc.), leading to an imbalance in power. Example: *Employee working for a boss — An employee puts in hard work for their boss. In doing so they can sometimes make a large profit, the boss can then decide how to allocate it.*
Tipping	Any act of tipping between a customer and service worker. Example: *tipping a waiter — I am not required to tip a waiter, but it is unfair if I do not. Likewise, not sharing at least something with the receiver is just greed.*
Abstract Social	Any social act involving helping, sharing, splitting, gifting etc. between two persons without any information on the context. Example: *Sharing — When people find something together then they usually share it amongst themselves.*
Social Financial Investment	Any financial investment which involves interaction between two individuals. Example: *An investor — Maybe you give someone money and they use it to start a business and make a profit, so they return money to you as dividends or profit sharing or something.*
Other Social	Any remaining interaction between (at least) two individuals. Example: *Game theory — Usually involves strategic decision making within the context of the game. A real life example will be the friends or foe game show.*

Table 1 Continued from previous page

Category	Description/Example[*]
Non-Social:	
Gambling	Any type of gambling (lottery, slot machine, raffles, ...). Example: *It reminds me of scratch tickets — when you spend you don't know how much you will get in return, Whether you will lose or gain money.*
Financial Investment	Any type of non-social financial investment with a profit motive. Example: *Investing in stock — When you invest in stock, you give over a certain amount of money. The hope is that the stock will ultimately return to you more than what you put in.*
Other Financial	Any non-investment financial act. Example: *loan company — You borrow money from a loan company and you paid them back . Depending on loan , you pay them back with interest.*
Taxes & Government	Any response involving taxes and/or the government. Example: *Government — The government giving the rich tax breaks while not giving the poor people a break.*
Other Non-Social	Any remaining response. Example: *paying at store — paying for your groceries at the store*
No Situation	Any response explicitly stating that subject is not reminded of anything. Example: *There are no similar situations. I wish you had made better suggestions so i could think of something — There is no situation for me to describe. These games are not reality.*

Notes: [*]Original responses, only spelling errors were corrected. Text before the dash is the title subjects gave, text after the dash is the context they provided.

In order to provide additional validation of the human classification by RAs, I also use a large language model (LLM) to classify responses based on the same (but slightly shortened) codebook. All results in Section 3 are qualitatively robust to using the LLM classification instead of the RA classification.[12] Finally, with this paper's main objective to provide first evidence on heterogeneity in mental representations in economic games, I specifically choose to employ this classification of answers based on general behavioral domains to mitigate concerns that the data were interpreted to create (artificial) heterogeneity. However, this decision likely comes at the expense of statistical power for the analyses in Section 3 because the general categories are likely to conflate important features of mental representations (e.g., some subjects specifically mention repeated interactions with their peers, while others say they don't expect anything in return or do not provide any details). Future research may choose more selective coding schemes.

2.3 Prolific Implementation & Sample

The experiment was conducted on Prolific, a survey provider that has already been used in economic research[13], in a US sample representative of the US population in terms of age, ethnicity, and sex with $n = 600$ subjects. Table A.3 in Appendix A.3 provides more details on these socio-demographics.

The experiment was implemented with oTree (Chen, Schonger and Wickens, 2016). Each point in the experiment was worth GBP 0.05, and subjects were paid a participation fee of GBP 6.[14] Subjects had to correctly an-

[12] ChatGPT based on GPT-4 ("Generative Pre-Trained Transformer") by OpenAI was used as the LLM for the classification exercise. GPT-4 has been shown to score in the 96th to 99th percentile in the verbal test of the Graduate Record Examination (OpenAI, 2023). This test is designed to measure the "ability to analyze and evaluate written material and synthesize information obtained from it" among prospective Ph.D. students (GRE, 2023). The codebook was shortened a bit to ensure that it was not too long for the context window of ChatGPT (i.e., the maximum number of characters to which ChatGPT can refer back). Text responses were provided in batches of 30-40 responses in a random order (different from the order for any of the research assistants). After classifying the text responses, a new instance of ChatGPT was initiated for the next batch to ensure that ChatGPT does not classify later batches differently than earlier batches. Results are available upon request.

[13] See Palan and Schitter (2018) for a review of Prolific as a platform for online experiments and, for example, Saccardo and Serra-Garcia (2023) for a recent publication that also uses the Prolific subject pool.

[14] At the time of the experiment, Prolific paid out all payments in GBP, regardless of

swer several comprehension questions before they could participate in each game. For the first of the two games, subjects only had two attempts for each question. If they answered a single comprehension question incorrectly twice, they were screened out and replaced by Prolific. Prolific does not allow screening out subjects in the middle of a study. For the second of the two games, subjects therefore had as many attempts as they needed, but they were still required to answer every comprehension question correctly. Additionally, two attention check questions were included in the survey module of the experiment. Subjects who did not pass both attention checks were also screened out and replaced. The sample size of $n = 600$ and both screening out procedures were part of the preregistration.

Data collection started on June 11th, 2023, with the majority of subjects (98%) participating either on June 11th or 12th. Due to resampling after screening out subjects, the remaining subjects participated in the days until June 18th. On average, subjects took roughly 28 minutes to complete the experiment and earned GBP 10.38 (including the participation fee). The joint ethics committee of Goethe University and the University of Mainz provided IRB approval before the experiment was conducted.

3 Results

This section is structured around three questions. First, does there exist heterogeneity in the mental representations of subjects in the dictator and trust game? Second, which factors influence mental representations and can therefore (partially) explain the heterogeneity, i.e., what are the drivers of mental representations? Third, are mental representations relevant for research using economic games in that mental representations influence behavior in the game and also affect the ability of economic games to predict field behav-

participants' actual location. Moreover, at the time of the experiment, GBP 1 = USD 1.26.

ior? Importantly, results are robust to alternatively using the classification of the open-ended associations by the LLM and also to the second, closed-ended validation measure of mental representations based on which sentences of the instruction were influential in how subjects think about the decision situation.[15]

Before addressing the three questions, however, I briefly comment on the preregistration of the analyses, the potential association between treatments and cognitive skills, and the randomization of treatments. Table D.2 in Appendix D provides a more detailed overview of which analyses and estimation equations were preregistered. However, to summarize, I preregistered to study the extent of heterogeneity in mental representations, how this heterogeneity depends on the order and framing treatment, how it correlates with socio-demographics, and how it moderates the correlation between game and field behavior. I did not preregister to study the correlation between game behavior and mental representations.

As outlined in the previous section, the survey provider only allows to screen out subjects based on a lack of understanding of the first game. While subjects still need to answer all comprehension questions correctly for the second game, they have as many attempts as necessary. Since the trust game is more complex than the dictator game and might therefore be harder to comprehend, playing the trust game first (and the dictator game last) could consequently introduce differential sample selection based on cognitive skills. The exogenous variation in the game order could therefore also introduce variation in cognitive skills, which would affect the interpretation of the treatment effect (i.e., game order *and* different skills). I explore this hypothesis in Appendix C and show that all results on game order effects are robust to controlling for this.

Finally, upon closer inspection of the data, the framing and game order

[15] Appendix C contains the results when using the closed-ended measure. Results for the LLM classification are available upon request.

treatments are very weakly correlated, even though both treatments are assigned independently of each other by two separate random draws. Subjects who are in the community framing treatment are a bit less likely to play the dictator game first. In Appendix C I present evidence that this association is likely to be a statistical coincidence. However, I still include parametric results for the analyses in this section to show that all effects are robust to controlling for the treatment status in the other treatment dimension.

3.1 Heterogeneity in Mental Representations

Figure 1 shows the distribution of the associations from everyday life (cf. Table 1 for the description of the individual categories). In order to avoid that results are driven by outliers and/or categories that are coded very rarely and could therefore be highly subjective, all categories that occur less frequently than 5% are aggregated into the "Other Social" (for social categories) and "Other Non-Social" (for non-social and no-situation categories). The unaggregated distributions are included in Appendix B.1.[16] Figure 1 highlights that, regardless of the game, subjects report substantially different associations. First, consider the distinction between social and non-social associations. Between 11% (dictator game) and 29.5% (sender in the trust game) of subjects feel reminded of situations that do not involve any other person at all. Specifically in the trust game, subjects feel reminded of situations involving playing the lottery or investing in the stock market in more than 10% of all cases. Importantly, the share of subjects who are not reminded of *any* situation from everyday life is negligible at roughly 1% in the dictator game and 2%-3% in the trust game, i.e., almost everyone has some

[16] The results are robust to using a different aggregation threshold of 2.5%, results are available on request. Moreover, notice that this aggregation decreases any variation in mental representation and therefore, if at all, biases the results against the preregistered hypotheses. In terms of increasing statistical power, this would only affect results for the respective "Other"-category. Categories that are not present for any of the three games after this aggregation are omitted from the figures.

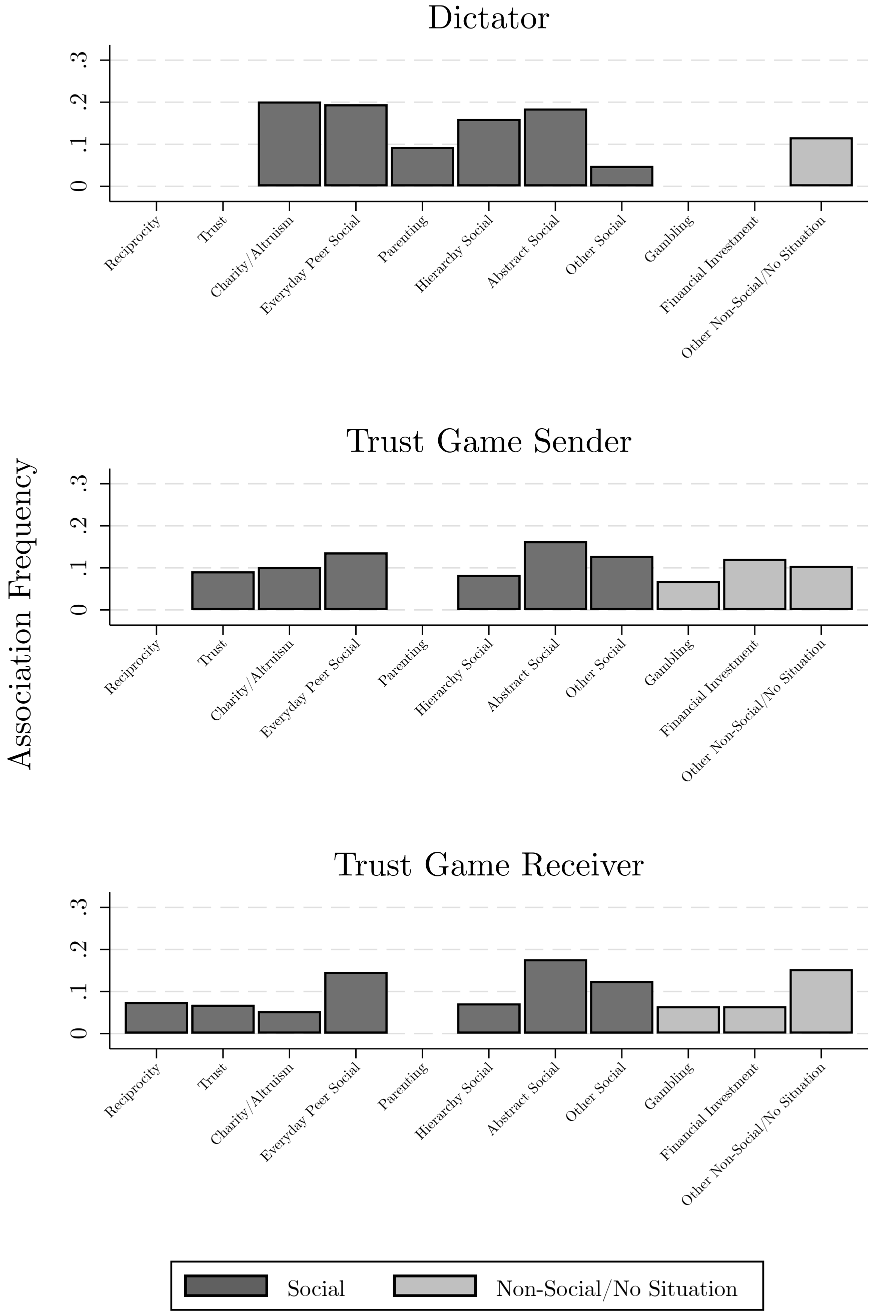


Figure 1: Associations in the Dictator and Trust Game

Notes: Distribution of associations, explanation of categories in Table 1. Categories that occur less frequently than 5% are aggregated into the respective "Other" category.

association.[17]

Second, even within the social categories, there is variation in the associations. For example, subjects feel reminded of situations involving interactions with everyday peers, parenting, interactions across hierarchies within organizations, and other more abstract acts of helping and sharing. Moreover, these associations seem to differ across games. While around 20% of subjects explicitly mention a charitable act in the dictator game, this share drops to 10% (sender) and even further to 5% (receiver) in the trust game. Conversely, subjects are more likely to be reminded of situations involving the word "trust" as the sender in the trust game and, as the receiver in the trust game, "reciprocity".

To quantify the extent of heterogeneity in associations from everyday life, I calculate the generalized variance which is a measure of dispersion for categorical data. It is mathematically equivalent to the probability that two randomly chosen subjects have a different association, i.e., associations belonging to different categories. In the dictator game, the probability is 83.7% and increases even further to 88.2% and 88.1% in the sender and receiver role in the trust game. Moreover, even when looking at the more objective classification of social vs. non-social associations, this probability still is at 20.6% in the dictator game and at roughly 41% in both roles in the trust game. Consequently, there does exist substantial heterogeneity in the associations with everyday life.

How should this heterogeneity in the reported associations be interpreted? Remember that all subjects needed to pass comprehension and attention questions to participate, so the variation is unlikely to be driven by a lack of understanding of the respective decision situation or lack of attention. Indeed, testing for a relationship between the associations and whether a

[17] These subjects are aggregated into the "Other Non-Social" category in Figure 1. While I use the "Other Non-Social/No Situation" label for this category, it is important to bear in mind that this category mostly consists of subjects with an association belonging to the "Other Non-Social" category.

subject answered at least one comprehension question incorrectly or whether a subject missed an attention check does not reveal any statistically significant finding.[18] But associations are self-reported and could therefore just reflect noise without any meaning. A first indicator to the contrary is that associations seem to depend on the game itself. However, this could also be driven by subjects participating in both games and feeling that they should report different, meaningless associations. A sharper test is therefore to look at between-subject variation. If associations did not contain any information about the underlying mental representation, they should not depend on the order of the games, which has been linked to inducing different (mental) decision contexts (Ockenfels and Schier, 2020). Figures 2 and 3 show the distribution of associations depending on which game is played first. In the dictator game, game order leads to a highly statistically significant effect on the associations ($p < 0.0001$ from a χ^2-test). Playing the trust game before the dictator game more than doubles the share of charity associations. This increase is offset by a decrease in associations with parenting and interactions involving hierarchy. The overall share of social associations, however, does not depend on the game order ($p = 0.4653$, χ^2-test). In both the sender and receiver role in the trust game, the evidence is more mixed. While there is no statistically significant effect on the individual categories ($p > 0.2654$, χ^2-test), playing the trust game first increases the share of social associations more generally from roughly 66% to 75% for both the sender and the receiver ($p = 0.0219$ for sender, $p = 0.0237$ for receiver, χ^2-test). In Appendix C I show that game order also affects the selected sentences, with subjects selecting the sentence on the unequal endowment in the dictator game much more frequently when the trust game is played first. The opposite is true for both roles in the trust game. This suggests that the unequal endowment in the dictator game features more prominently in subjects' minds when subjects first experience equal endowments in the trust game, i.e., the contrast to

[18] Details are provided in Appendix C.

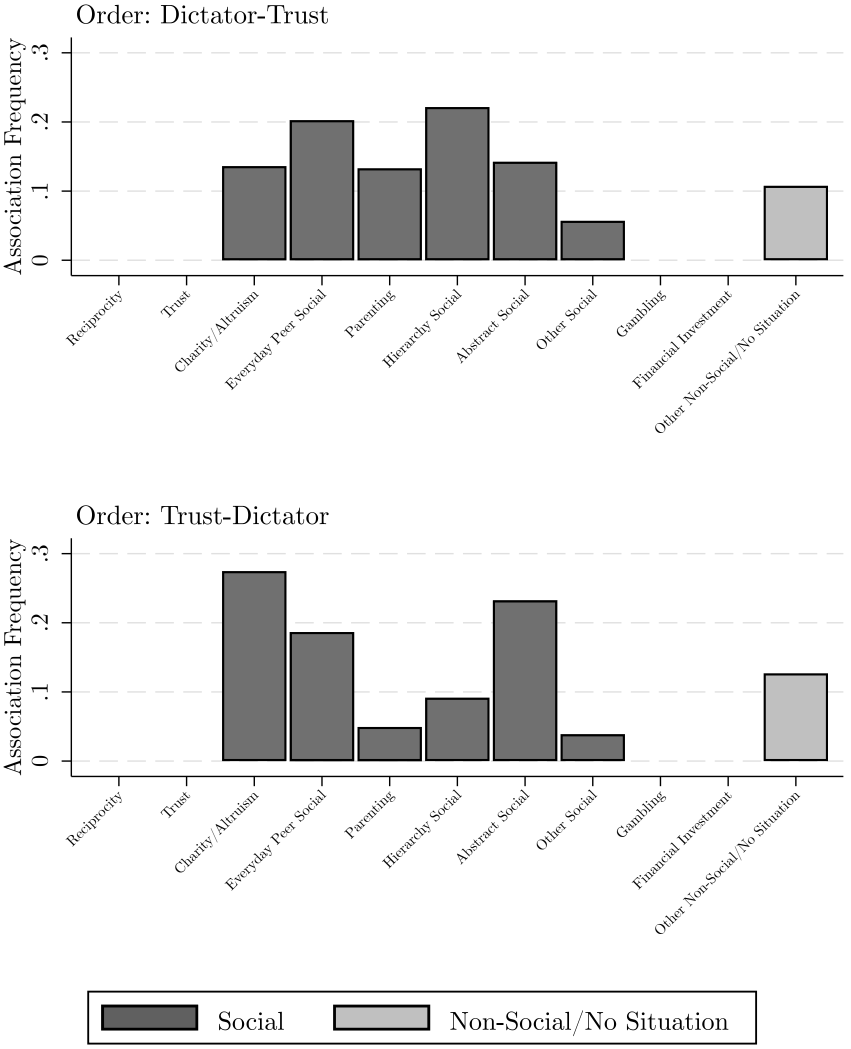
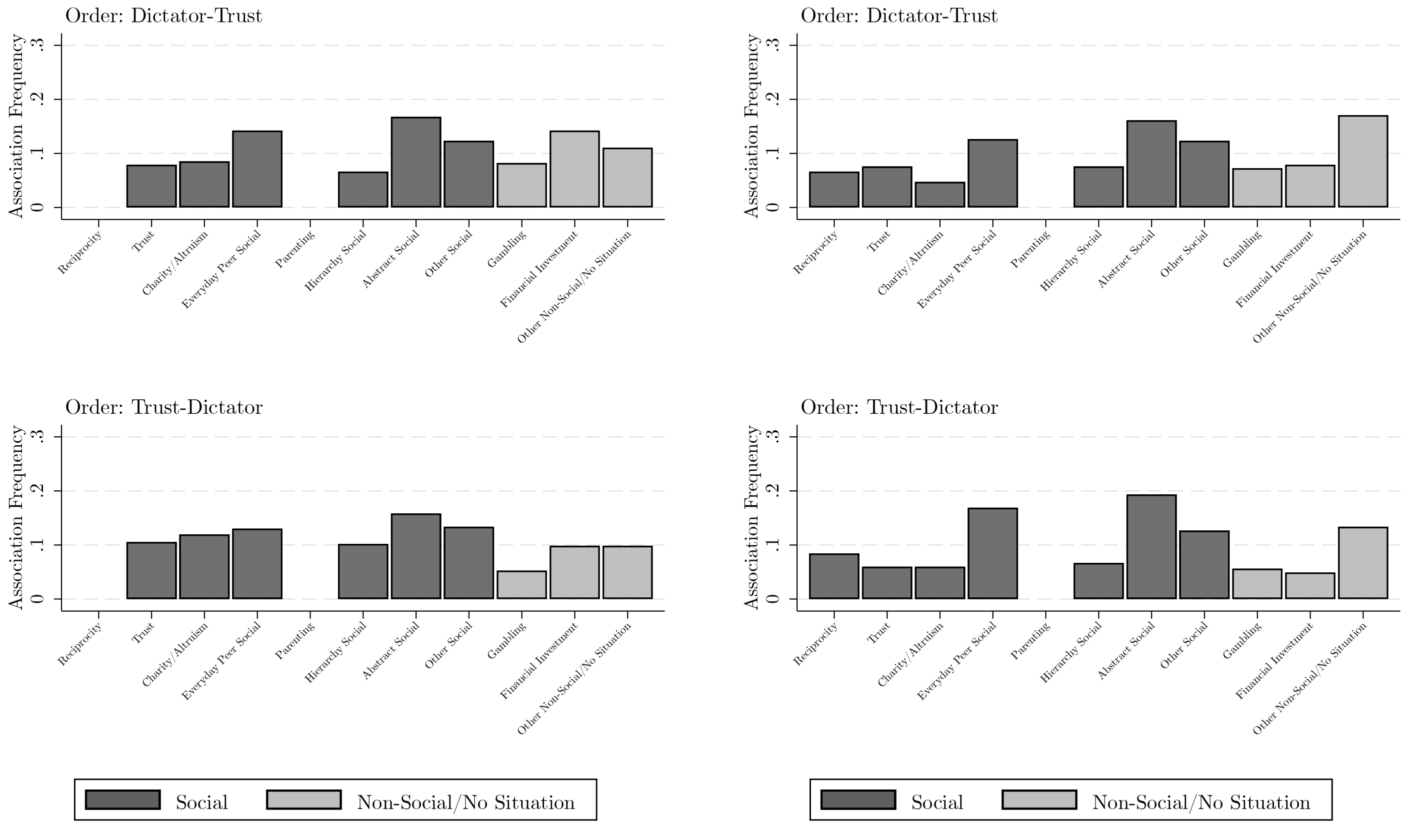


Figure 2: Effect of Game Order on Mental Representations in Dictator Game

Notes: Distribution of associations by game order, explanation of categories in Table 1. Categories that occur less frequently than 5% are aggregated into the respective "Other" category.



(a) Trust Game: Sender

(b) Trust Game: Receiver

Figure 3: Effect of Game Order on Mental Representations in Trust Game

Notes: Distribution of associations by game order, explanation of categories in Table 1. Categories that occur less frequently than 5% are aggregated into the respective "Other" category.

the equal endowment seems to make the unequal endowment in the dictator game much more salient in subjects' minds. This then explains why subjects are more likely to report charity associations in the dictator game when they first play the trust game.

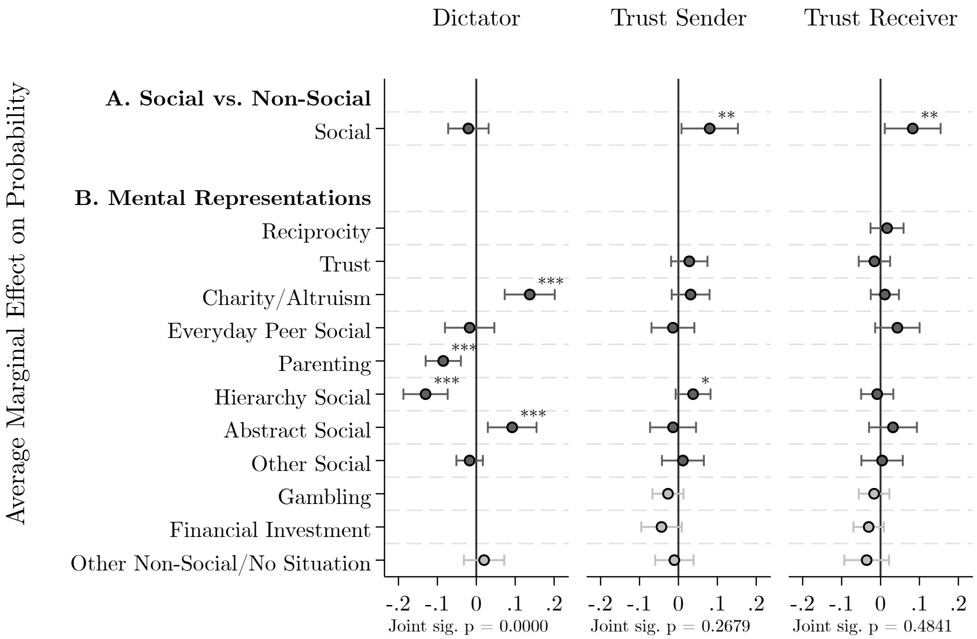


Figure 4: Effect of Playing Trust Game First on Associations across Games

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of playing the trust game first on the different associations, together with a 95% confidence interval. Panel A uses a probit model to regress an indicator for a social association (i.e., involving at least two individuals) on the treatment indicator. Panel B uses a multinomial logit model to regress the individual categories on the treatment indicator. Both models control for the treatment variation in framing. Below each marginal effects plot, the p-value from an F-test of joint significance of all marginal effects in Panel B is reported (i.e., from a test whether all individual marginal effects are jointly zero).

Figure 4 confirms these insights with parametric regressions. For this, Figure 4 plots the marginal effects of playing the trust game first (together with the 95% confidence interval) on the associations across all three player roles (dictator, sender in the trust game, receiver in the trust game), while controlling for the variation in the community framing treatment. In Panel A, a probit regression is used for the distinction between social and non-social associations, while a multinomial logit is used in Panel B for the individual categories.[19] The estimates can therefore be interpreted as the marginal effect of playing the trust game first on the probability of each association category. Additionally, stars indicate the statistical significance of each individual marginal effect, and the p-value from an F-test on joint significance of all marginal effects in Panel B is reported below each plot. The marginal effects confirm the non-parametric insights: playing the trust game first affects the individual associations within the social category for the dictator game, while it shifts associations from non-social to social for both the sender and the receiver in the trust game. Given the statistically significant and plausible effects of between-subject variation in game order on the reported associations, I do not interpret the reported associations as noise. Instead, the reported associations seem to contain (some) information about the underlying mental representation of the decision situation. From now on, I therefore refer to the associations as a measure for mental representations. Result 1 summarizes, highlighting that the control of economic games does not extend to mental representations of the dictator and trust game.

Result 1 *There exists substantial heterogeneity in mental representations in the dictator and trust game. This even holds on the level of classifying associations only based on whether they feature another individual, i.e., whether they entail a social dimension.*

[19] I preregistered that I would use a linear probability model to regress the game order treatment indicator on a set of fixed effects for each association category while controlling for the framing treatment. The results can be found in Appendix C and are identical. Using a probit and multinomial logit model, however, allows to interpret the individual marginal effects more meaningfully as they can be directly interpreted as the change in the probability of a particular association. The same applies to all following parametric regressions that involve associations.

3.2 Drivers of Mental Representations

Another interpretation of the statistically significant effect of game order on mental representations in the dictator and trust game is that the mental representation of a particular game depends on the broader experimental context in which the game is embedded, i.e., on what is happening (immediately) before a game. However, this cannot explain why, for example, there exists heterogeneity within all games which are played first. I therefore explore three additional groups of potential drivers: besides the experimental context in which a game is embedded, the mental representation might depend on the game itself, its implementation, and the subject participating in the game.

First, consider the game itself. Figure 1 already indicates that representations in the dictator game are different from representations in the trust game, e.g., subjects in the dictator game are more likely to perceive the situation as involving charity, while subjects in the trust game also think of "Trust" and "Reciprocity" on the one hand and financial investments on the other hand (probably due to the multiplier). To test this, I use the Stuart-Maxwell-test for paired groups and compare the distribution of mental representations in the dictator game with the respective distributions for the sender and receiver in the trust game. The tests show that mental representations in the dictator game are significantly different from mental representations for both the sender and receiver in the trust game ($p < 0.0001$). Moreover, mental representations are also significantly different across both player roles within the trust game ($p < 0.0001$). Even when comparing the distribution of social vs. non-social mental representations instead of the individual categories, subjects in the dictator game have different (i.e., more social) mental representations than in the trust game ($p < 0.0001$, McNemar's test). However, on this level, mental representations are not significantly different between the sender and receiver in the trust game ($p = 0.5637$).

Second, consider the implementation of a game, i.e., what happens inside

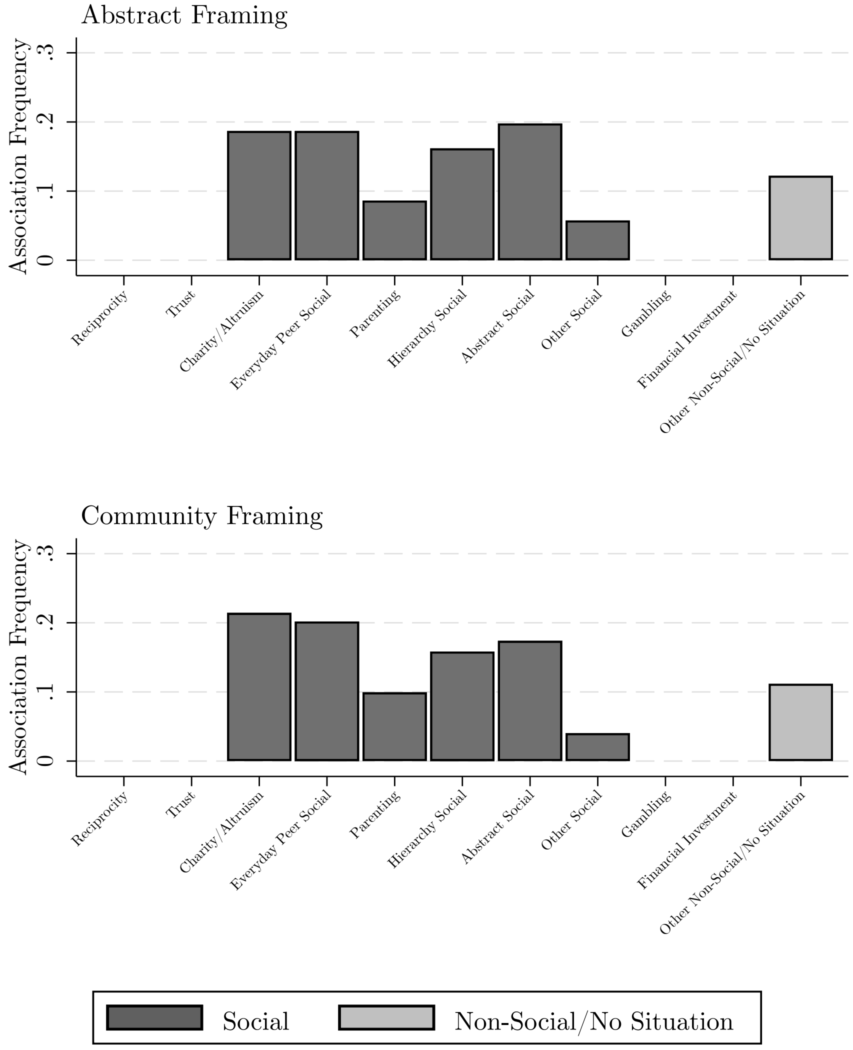
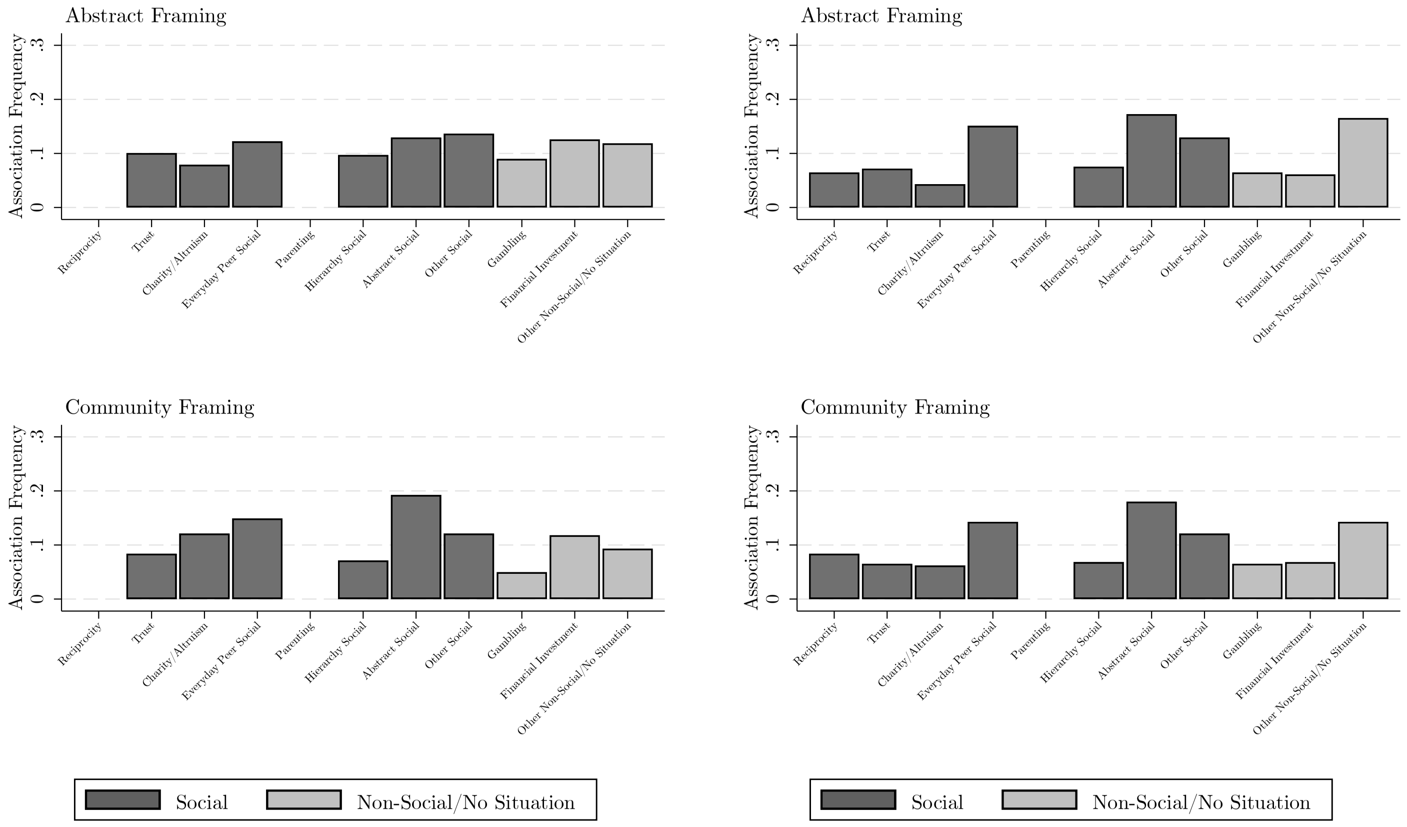


Figure 5: Effect of Framing on Mental Representations in Dictator Game

Notes: Distribution of associations by community framing, explanation of categories in Table 1. Categories that occur less frequently than 5% are aggregated into the respective "Other" category.



(a) Trust Game: Sender

(b) Trust Game: Receiver

Figure 6: Effect of Framing on Mental Representations in Trust Game

Notes: Distribution of associations by community framing, explanation of categories in Table 1. Categories that occur less frequently than 5% are aggregated into the respective "Other" category.

the game. Figures 5 and 6 show the distribution of mental representations across the two framing treatments. Framing the decision situation as a community decision situation does not seem to affect mental representations in the dictator game and for the receiver in the trust game. While the distributions vary a little, this is not statistically significant ($p > 0.8669, \chi^2$-test). Moreover, contrary to the preregistered hypothesis that community framing should induce more social mental representations, the share of social mental representations only increases by 1p.p. in either decision role. However, community framing weakly affects mental representations for the sender in the trust game and increases the share of social mental representations from 66.55% to 73.91% ($p = 0.0485$, χ^2-test). This is driven by a shift from gambling-related mental representations to mental representations featuring acts of helping or sharing. Figure 7 plots the results from parametric regressions and confirms these findings, i.e., describing the decision situation as a community decision situation does not shift mental representations in the dictator game or for the receiver in the trust game. However, it induces slightly more social mental representations for the sender in the trust game. While this constitutes evidence that mental representations, at least for the sender in the trust game, also depend on how a particular game is implemented, more research on other game dimensions is needed to verify whether the weak effect is specific to this framing treatment or applies to game parameters more generally (e.g., what is the effect of varying stake size).

Third, the literature on associative memory and experience effects in finance motivates the hypothesis that the remaining heterogeneity could be caused by different individual experiences, ranging from major shocks that reshaped the neural pathways in the brain (Malmendier, 2021) to having seen memorable images in the news (Enke, Schwerter and Zimmermann, 2020). Briefly notice that having played a game before the current game does constitute a (very recent) event that does influence mental representations in

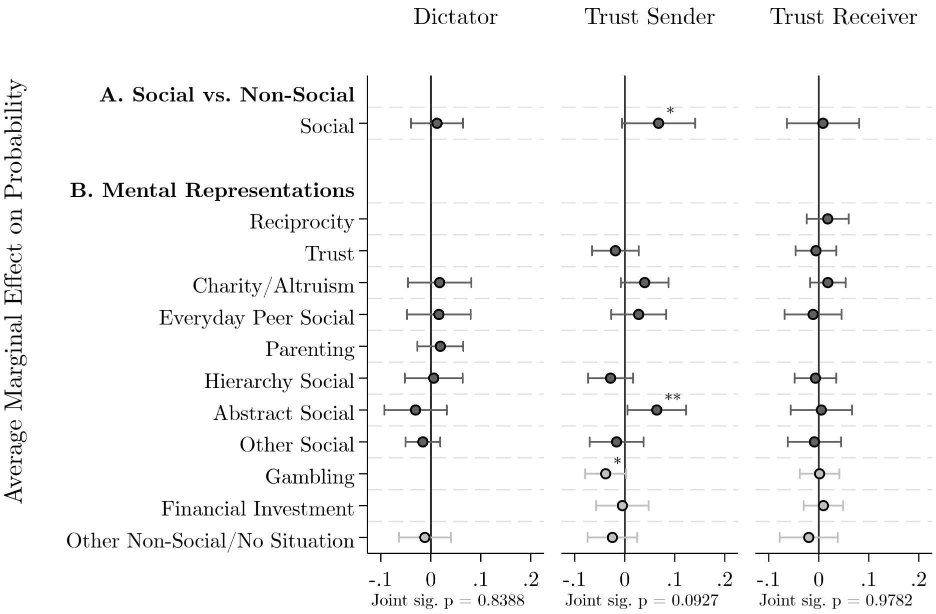


Figure 7: Marginal Effect of Community Framing on Mental Representations

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of community framing on the different associations, together with a 95% confidence interval. Panel A uses a probit model to regress an indicator for a social association (i.e., involving at least two individuals) on the treatment indicator. Panel B uses a multinomial logit model to regress the individual categories on the treatment indicator. Both models control for the treatment variation in game order. Below each marginal effects plot, the p-value from an F-test of joint significance of all marginal effects in Panel B is reported (i.e., from a test whether all individual marginal effects are jointly zero).

this experiment (cf. game order effects). Turning to personal experiences, I do not have data on previous shocks or the main activities of subjects on the day(s) leading up to the experiment.[20] Still, socio-demographic information for which this sample is representative in the US, i.e., age, ethnicity, and sex, can be used as (noisy) proxies for such influential events.

Figures 8, 9, and 10 plot the coefficients from parametric regressions of mental representations on these socio-demographics. Figures with the respective distributions are included in Appendix B.2.

Starting with the effect of age, being older than the median age shifts mental representations from perceiving the dictator game as an interaction involving hierarchies in an organization to parenting and more abstract acts of helping and sharing. This seems plausible given that being a parent and having a more senior position in an organization (i.e., where upward hierarchy is less salient) are likely to correlate positively with age. However, being older than the median age does not generally induce more social mental representations in the dictator game. Considering the trust game, older subjects are generally more likely to perceive the sender role as a social situation and are also more likely to perceive it as a charity situation or an abstract act of helping or sharing. The evidence for the receiver role is less clear, with older people perceiving the receiver role more likely as an abstract act of helping or sharing.

Regarding ethnicity, there is limited variation in ethnicity to begin with (77% of the sample are white). Moreover, there is no clear evidence that being white affects mental representations. When it comes to the effect of being female, there is a strong and persistent effect across all player roles, with women being substantially more likely to perceive each game as an abstract situation involving helping or sharing compared to more concrete interactions. Being female does not lead to generally more social mental

[20] An interesting avenue for future research is to have subjects write a diary in the days before an experiment and analyze how mental representations correlate with the daily events leading up to the experiment.

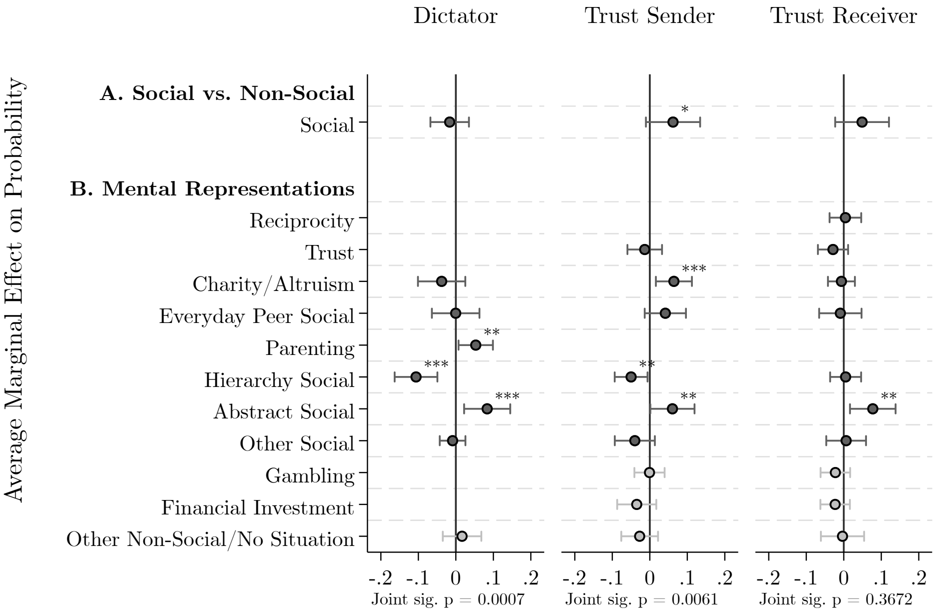


Figure 8: Marginal Effect of Being Above Median Age on Mental Representations

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of being above median age on the different associations, together with a 95% confidence interval. Panel A uses a probit model to regress an indicator for a social association (i.e., involving at least two individuals) on an indicator for being above median age. Panel B uses a multinomial logit model to regress the individual categories on an indicator for being above median age. Both models control for the treatment variation in game order and framing. Below each marginal effects plot, the p-value from an F-test of joint significance of all marginal effects in panel B is reported (i.e., from a test whether all individual marginal effects are jointly zero).

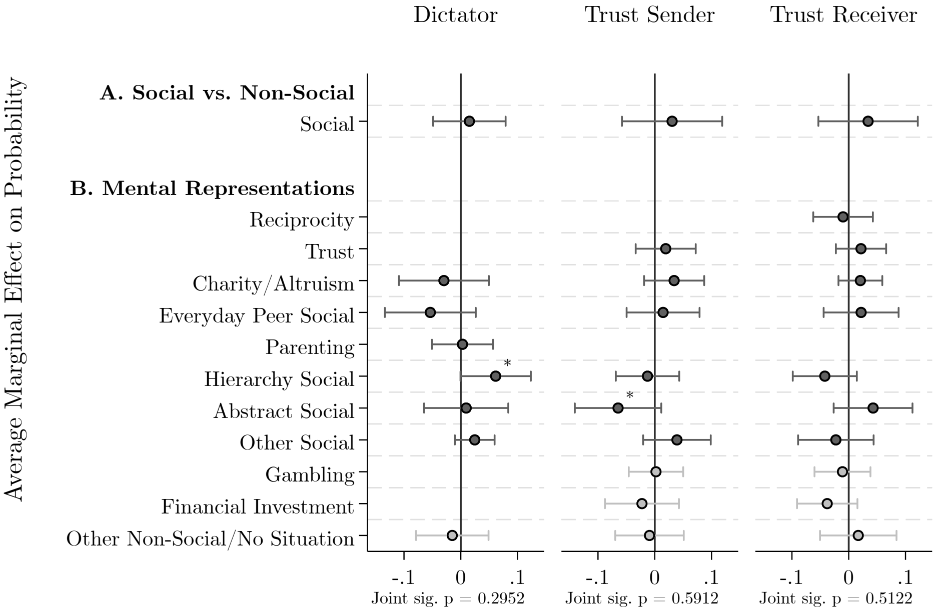


Figure 9: Marginal Effect of Being White on Mental Representations

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of being white on the different associations, together with a 95% confidence interval. Panel A uses a probit model to regress an indicator for a social association (i.e., involving at least two individuals) on an indicator for being white. Panel B uses a multinomial logit model to regress the individual categories on an indicator for being white. Both models control for the treatment variation in game order and framing. Below each marginal effects plot, the p-value from an F-test of joint significance of all marginal effects in panel B is reported (i.e., from a test whether all individual marginal effects are jointly zero).

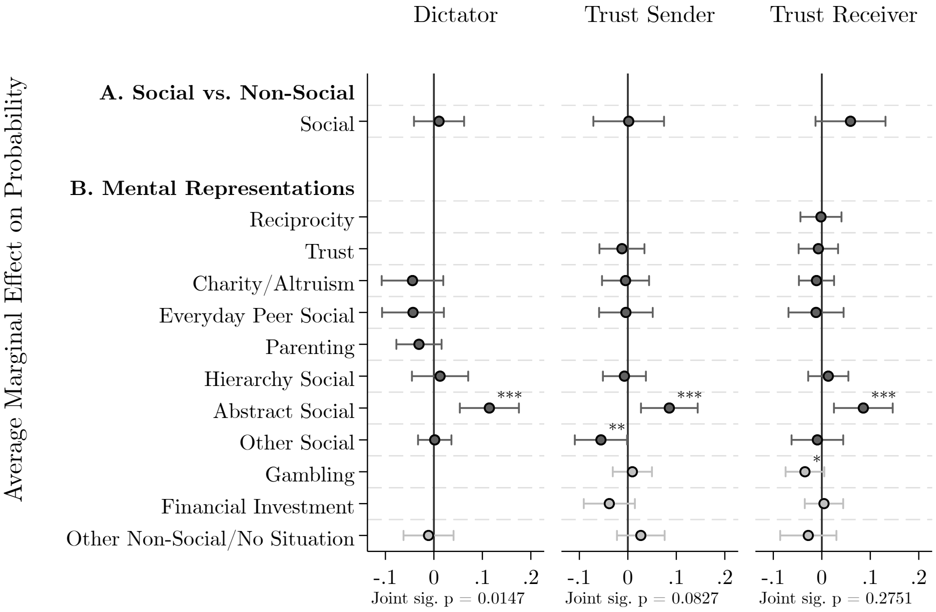


Figure 10: Marginal Effect of Being Female on Mental Representations

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of being female on the different associations, together with a 95% confidence interval. Panel A uses a probit model to regress an indicator for a social association (i.e., involving at least two individuals) on an indicator for being female. Panel B uses a multinomial logit model to regress the individual categories on an indicator for being female. Both models control for the treatment variation in game order and framing. Below each marginal effects plot, the p-value from an F-test of joint significance of all marginal effects in panel B is reported (i.e., from a test whether all individual marginal effects are jointly zero).

representations, however.

Summing up, the heterogeneity in mental representations seems to be driven by a combination of game order effects, framing effects (for the sender in the trust game), and socio-demographic information, in particular age. While I interpret this as evidence that mental representations in economic games are driven by the broader (experimental) context in which the game is embedded, the game itself as well as its implementation (e.g., framing), and the subject herself, more research with better data — in particular on past experiences of subjects — is clearly needed to explore this further. However, Result 2 highlights that the drivers of mental representations of economic games are (partly) outside the control of experiments.

Result 2 *Mental representations in a game depend on the general type of game (dictator vs. trust game), the experimental context in which the game is embedded (i.e., game order), and the subject herself. Framing the game as a community decision situation does not systematically shift mental representations in the dictator and trust game.*

Finally, it is interesting to see that across most of these dimensions, the effects in the dictator game all happen within the social dimension, i.e., mental representations are shifted from one social category to another. In the trust game, however, the shift mainly happens from non-social to social categories (and vice versa). Moreover, effect sizes are generally smaller for the trust game. This could suggest that while the dictator game is more clearly recognizable as a social situation than the trust game, the specific mental representation of the dictator game is also more malleable and dependent on (external) cues.

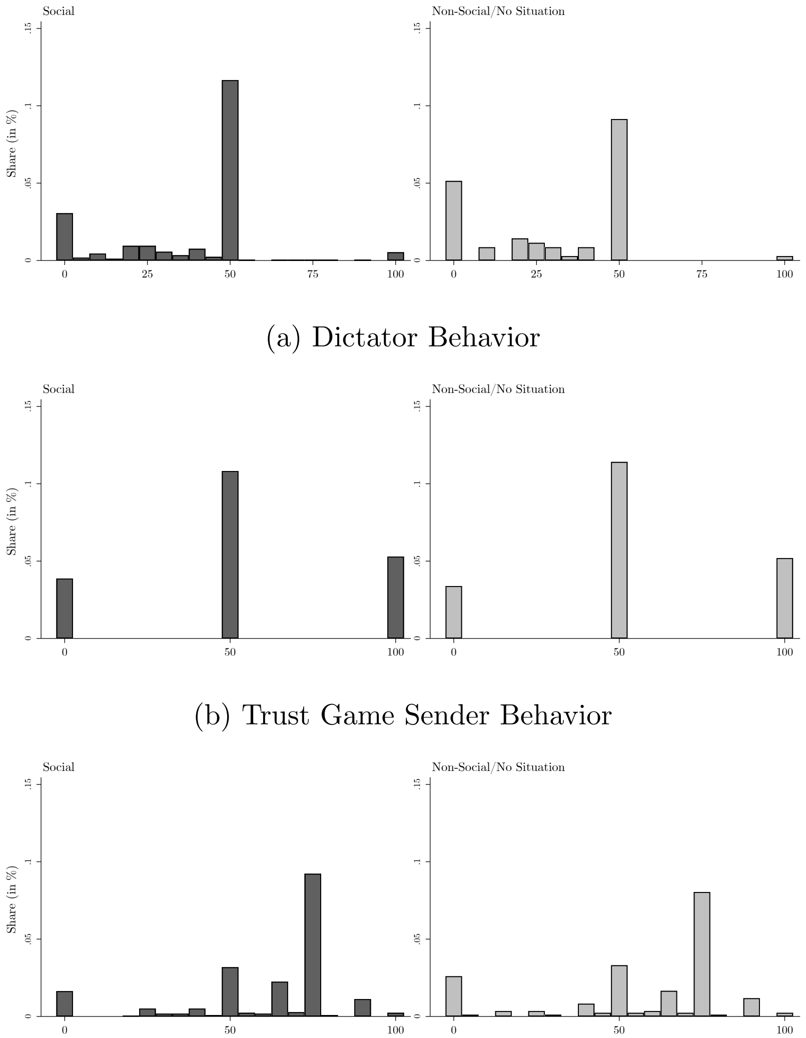
3.3 Relevance of Mental Representations for Economic Research

Results 1 and 2 indicate that there is heterogeneity in mental representations and also explore different drivers of this heterogeneity. However, it remains to be seen whether mental representations are also relevant for economic research beyond shedding light on how subjects perceive economic games. To demonstrate said relevance, remember that economic games are often used to either causally test hypotheses on human behavior by introducing exogenous variation in the game environment or infer preferences from behavior in these economic games.

Consider first research applications that use economic games to better understand (drivers of) human behavior by exposing subjects in economic games to different treatment conditions. Suppose that mental representations were linked to behavior in economic games. Result 2 shows that mental representations react to changes in the game and in the experimental context in which the game is embedded. Accounting for mental representations would therefore allow to (more precisely) pin down channels of treatment effects, in addition to contributing to a better understanding of human behavior and the associated cognitive processes themselves. In the following, I therefore analyze to which extent mental representations are correlated with behavior in economic games.[21] Behavior in the dictator game and as the sender in the trust game are parameterized as the amount sent to the receiver (with the amount in the trust game being discrete choices), while behavior as the receiver in the trust game is parameterized as the share returned (cf. Section 2.1).

Figure 11 shows the distribution of choices in each role of the dictator and

[21] To fully explore the potential of mental representations in this context, it is of course important to put more structure on mental representations, move beyond correlations, and put this idea to the test by measuring mental representations in a variety of different treatment conditions.



(a) Dictator Behavior

(b) Trust Game Sender Behavior

(c) Trust Receiver Behavior

Figure 11: Mental Representations and Behavior

Notes: Distribution of behavior in the dictator and trust game (sender and receiver), split by whether subjects report a social or non-social representation. Behavior in the dictator game and as the sender in the trust game is parameterized as the amount sent. Behavior as the receiver in the trust game is parameterized as the average share returned across all decisions. More details in Section 2.1.

trust game, split across social and non-social mental representations.[22] Figure 11 shows that subjects give more in the dictator game when they have a social mental representation. The average amount sent increases from 31.06 to 38.56 points ($p = 0.0062$, Mann-Whitney-U-test), i.e., an increase of 25%. While there is no effect on the amount sent in the trust game ($p = 0.7400$, χ^2-test), the share returned by the receiver increases when subjects have a social mental representation from 55.86% to 61.65%, but this increase is not statistically significant at conventional levels ($p = 0.1121$, Mann-Whitney-U-test). Figure 12 confirms these insights with parametric regressions and also provides results for how the individual mental representations influence behavior.[23] For this, Figure 12 plots the marginal effects of regressions of game behavior on mental representations. In Panel A, this is done with the social vs. non-social distinction of mental representations, while Panel B reports results when using individual mental representations. The reference category is set to "Charity/Altruism" for the dictator game, "Trust" for the sender in the trust game, and "Reciprocity" for the receiver in the trust game. Coefficients can therefore be interpreted as the marginal effect on behavior relative to the reference category or, in the case of the sender in the trust game, the marginal effect on the choice probability. The analyses reveal that there is heterogeneity in behavior across the individual mental representations for the dictator game and the sender in the trust game: testing whether all individual coefficients are zero, i.e., that there is no heterogeneity in behavior across mental representations, yields $p < 0.0001$ for the dictator game and $p = 0.0051$ for the outcome of sending 50 points in the trust game. For choosing to send 100 points, mental representations are individually but not jointly significant. Testing whether mental representations are jointly

[22] If receivers in the trust game return more than the points they received, the share is larger than 100%. This is the case for four subjects. For illustrative purposes, Figure 11 omits these four subjects.

[23] Figures B.8, B.9, and B.10 in Appendix C provide the distributions of behavior for each individual mental representation.

significant for any of the decisions of the sender in the trust game reveals a p-value of $p = 0.0660$.

How should one interpret the association between mental representations and behavior? First, consider the direction of the effects. Setting the reference category to a narrowly defined category helps interpret the direction, compared to, for example, setting it to “Other Social” which is a mixture of different representations. For example, consider on the one hand that subjects give substantially more in the dictator game for any social mental representation, compared to the reference category of “Charity/Altruism”. In contrast to “Charity/Altruism”, the other mental representations are more likely to feature an element of repeated interaction in real life. One hypothesis to rationalize these effects could be that subjects subconsciously consider benefits from future interaction through their mental representation even though the game is not repeated. Consider, on the other hand, the sender in the trust game. Having any mental representation different from “Trust” increases the likelihood of sending 50 points to the receiver. Conversely, it decreases the likelihood of sending no or all points. This seems to indicate that when subjects perceive the game to be about trust, they either do not trust at all or entrust everything to the receiver.

Second, consider that subjects report their mental representations after they make a decision in the respective game role. Mental representations could therefore be caused by behavior, with subjects, for example, justifying their behavior by selecting a fitting mental representation for their behavior. While I cannot rule this out completely with the current experimental design, no clear association exists between prosocial behavior and social mental representations. If subjects, for example, wanted to maximize their income in the dictator game, they should keep everything and report a non-social association to justify their behavior. This is not the case, with subjects sending as many points with a charity mental representation as a non-social mental representation.

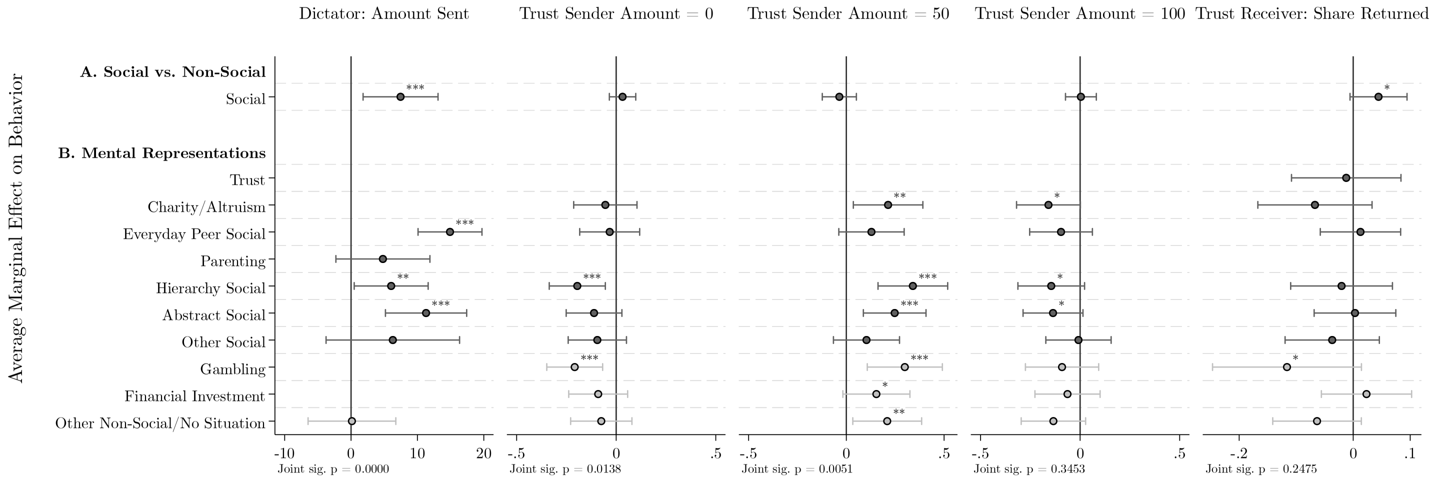


Figure 12: Mental Representations and Behavior

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of different mental representations on behavior in the respective game. For the dictator and the receiver in the trust game, a variable for the amount sent (dictator) or share returned (trust game receiver) is regressed on an indicator for a social mental representation (Panel A) and, separately on indicator variables for each mental representation (Panel B). For the sender's behavior in the trust game, a multinomial logit is used with 0 as the base outcome. For the sender in the trust game, the coefficients can therefore be interpreted as the change in the probability that the respective amount of points is chosen. The reference groups "Non-Social/No Situation", "Charity/Altruism" (dictator), "Trust" (sender in the trust game), and "Reciprocity" (receiver in the trust game) are omitted in the respective model. As preregistered all regressions control for variation in the game order and framing treatment. Below each marginal effects plot, the p-value from an F-test of joint significance of all marginal effects in Panel B is reported (i.e., from a test whether all individual marginal effects from mental representations are jointly zero).

Summing up, mental representations seem to be associated with statistically significantly different behavior in the dictator game but also for both roles in the trust game — albeit more modestly. More work is needed to assess whether mental representations influence behavior causally. However, the current evidence already suggests that accounting for mental representations can help to better understand the drivers of behavior in economic games and, in future research, pin down the channels of treatment effects. In Appendix B.4, I provide one exemplary use case for this by arguing that accounting for mental representations gives rise to a new explanation why framing might not affect behavior (Ellingsen et al., 2012; Dreber et al., 2013): framing treatments might simply not shift the mental representation, i.e., the "frame", sufficiently strongly. Framing treatments should therefore be understood as intention-to-treat treatments, with not every subject "complying" with the intended mental representation.

Besides aiding in better understanding (drivers of) behavior by exposing subjects to different treatment conditions, economic games such as the dictator and trust game are also frequently used to elicit a measure for social preferences. As spelled out above, due to the controlled nature of economic games, heterogeneity in preferences can be inferred from heterogeneity in behavior under certain assumptions. Treating behavior in economic games as a measure of preferences implies that behavior in economic games should be associated with behavior outside the game ("field behavior") that is influenced by the same preference that game behavior is supposed to measure. Previous research, however, shows that this is not always the case, with meta-studies finding no statistically significant effect across studies (e.g., Galizzi and Navarro-Martinez, 2019). Charness and Fehr (2023) discuss why this might be the case and point to a variety of pitfalls in previous studies measuring preferences based on game behavior, chief among which is that not every economic game — or rather its implementation — is on average a clean measure of a particular preference. Result 1 and 2 indicate a related reason why game behavior might not measure preferences: subjects have a mental representation of the game in which the preference to be measured is unlikely to play a major role.[24] For example, even in the community framing treatment, subjects perceive the sender decision in the trust game — often used as a measure for trust — as related to investments into the stock market ("Financial Investment"), while some subjects in the dictator game — often used (and criticized) as a measure for altruism — relate the decision to interactions which likely involve repeated elements (e.g., interactions with "Everyday Peers").

To test this, I study whether mental representations uncover heterogeneity in the relationship between self-reported behavior from the field and behavior in the dictator and trust game. The selection of field behavior is based on previous research: Galizzi and Navarro-Martinez (2019) find a positive but weak relationship between dictator game giving and an index of self-reported altruistic behavior (e.g., frequency of (blood) donations and altruistic acts like helping a stranger). Glaeser et al. (2000) report a positive association between behavior in the trust game and the General Social Survey trust question ("Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?"), but only for the receiver and not the sender. Glaeser et al. (2000), however, do find a positive relationship between past lending behavior to friends and sender behavior in the trust game. Riedl and Smeets (2017) highlight that social preferences — as measured through second-mover behavior in the trust game — predict the likelihood that an investor holds socially responsible equity but are negatively (but statistically not significantly) associated with the share of socially responsible investments in the overall portfolio. Finally, Gill et al. (2022) link receiver behavior in the trust game to aspiring to and working in

[24] Result 2 shows that mental representations also depend on the game and its implementation. Mental representations could therefore also serve as a test for what a "good" implementation is.

the finance section among university students. They interpret this finding as students who aspire to work in the finance industry being less trustworthy.

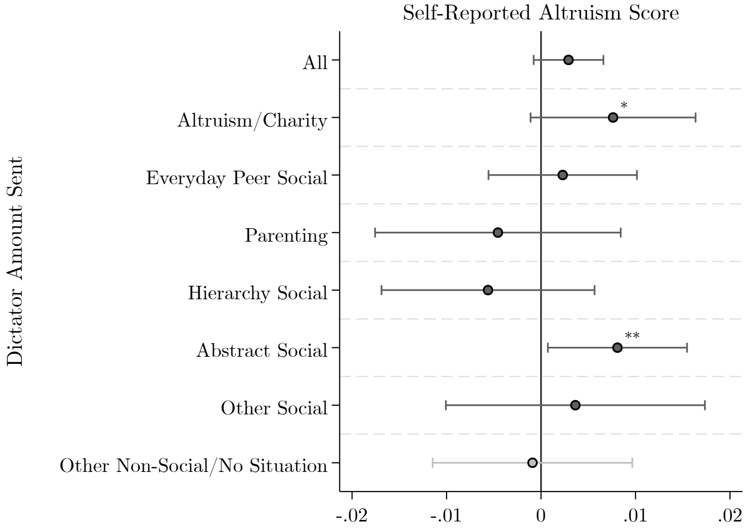


Figure 13: Dictator Game and Field Behavior

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of behavior as the dictator on field behavior. The first row indicates the coefficient in the overall sample, while the following rows indicate the coefficient for the sample split by the mental representations of the respective game role. Linear regressions are used for all outcomes except SRI, which uses a probit model. See Table A.3 for more details on the elicitation of each field behavior. Effects for SRA are in units of standard deviations.

This experiment takes place in a general population sample and I have to rely on self-reported data. I therefore use survey questions to match the respective field behavior as closely as possible. More details on the respective field behaviors and how they are elicited in this study are provided in Table A.3 in Appendix A.4.

Figures 13, 14, and 15 plot the results of regressing each field behavior on the respective game behavior, i.e., self-reported altruistic behavior ("SRA", effect in standard deviations) on the amount sent by the dictator, General Social Survey trust question ("GSS", effect in standard deviations) and self-reported lending behavior to friends ("lending", binary outcome) on

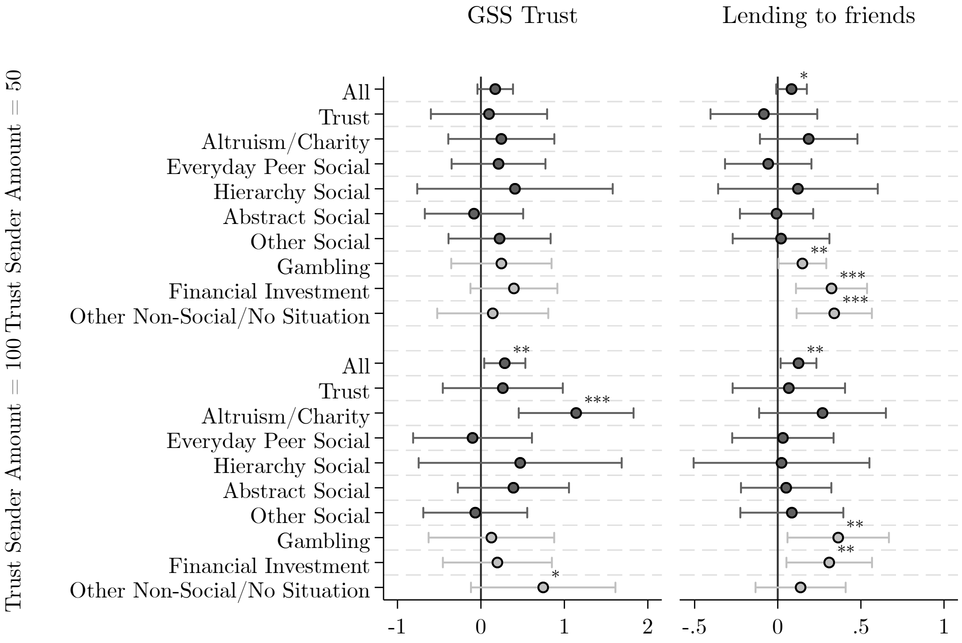


Figure 14: Sender in Trust Game and Field Behavior

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of behavior as the sender in the trust game on field behavior. The first row indicates the coefficient in the overall sample, while the following rows indicate the coefficient for the sample split by the mental representations of the respective game role. Linear regressions are used for all outcomes except SRI, which uses a probit model. See Table A.3 for more details on the elicitation of each field behavior. Effects for GSS are in units of standard deviations; effects on lending as change in the probability of lending more than once per year.

the amount sent in the trust game[25], and, finally, GSS, pursuing a career (or working) in the finance industry ("finance", binary outcome), socially responsible investment behavior ("SRI", binary outcome), and a hypothetical donation to a non-profit organization which offsets CO_2 emissions ("Atmosfair", share donated) on the average share of the amount received which is sent back by the receiver in the trust game. The first row always lists the average marginal effect in the overall sample[26], while the subsequent rows report the estimate in the subsample for each mental representation separately. Importantly, only 36 subjects indicate that they pursue a career (or work) in the finance industry in the overall sample. The results for this field behavior should therefore be interpreted with caution.[27] While I focus on the heterogeneity in the relationship between game and field behavior across the different mental representations in this section, I provide more details on how the findings in the overall sample compare to the original results in Appendix D.[28]

Notice at first that I find evidence of heterogeneity across the different mental representations for all field outcomes. Moreover, the heterogeneity is partially in line with what one would expect, treating game behavior as a measurement for altruism (dictator game), trust (sender in the trust game),

[25] In line with the original estimation strategy and this paper's interpretation of trust as a "behavioral act", I do not control for the belief of senders about what the receiver is going to do. However, results are robust to also controlling for the sender's belief. Results are available upon request.

[26] The analyses for SRI and pursuing a career in finance are based on a subsample with $n = 451$ (SRI; excluding subjects who do not want to invest their money at all) and $n = 540$ (finance; excluding subjects who are permanently unemployed or are not working because they take care of their home or family or indicate "other" as occupation).

[27] Among subjects with a mental representation related to hierarchy and financial investments, no subject indicates a preference for working in the finance industry.

[28] In short, in the overall sample, I do not replicate the original effect for SRA, GSS and trust game sender behavior, pursuing a career in finance, and the Atmosfair donation. This is probably driven by the different sample in which I operate (e.g., general population vs. student sample in Glaeser et al. (2000) and Gill et al. (2022)) and the modifications in the elicitation of field behaviors (e.g., using a hypothetical donation to Atmosfair instead of the actual equity in socially responsible investments).

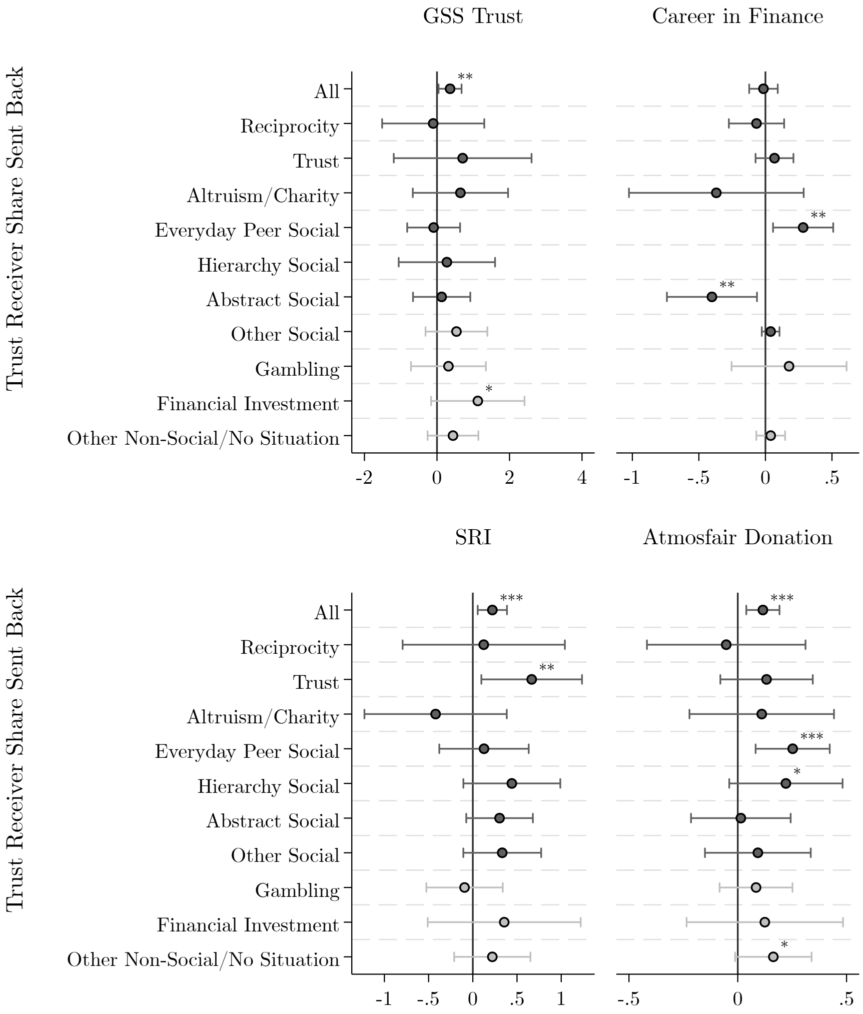


Figure 15: Receiver in Trust Game and Field Behavior

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of behavior as the receiver in the trust game on field behavior. The first row indicates the coefficient in the overall sample, while the following rows indicate the coefficient for the sample split by the mental representations of the respective game role. Linear regressions are used for all outcomes except SRI, which uses a probit model. See Table A.3 for more details on the elicitation of each field behavior. Effects for GSS are in units of standard deviations, effect on finance (SRI) as the change in probability of working in finance (indicating paying special considerations to sustainability when investing). Effects on Atmosfair as the change in the share donated to Atmosfair. No subjects in "Hierarchical Interaction" and "Financial Investment" pursue a career or work in the finance industry.

and trustworthiness/reciprocity (receiver in the trust game). For example, dictator game giving is only positively associated with outside altruistic behavior for subjects who have in mind a charity representation or a more abstract mental representation that involves an act of sharing, giving, or helping. It is, however, not significantly associated with altruistic field behavior among subjects who liken the dictator game to parenting or (likely repeated) interactions with friends. Similarly, receiver behavior in the trust game is negatively associated with (aspiring to) an occupation in the finance industry for subjects with a mental representation involving an abstract act of sharing, giving, or helping. Moreover, the positive effect of receiver behavior in the trust game on socially responsible investment is driven by social mental representations, in particular related to trust.

However, this exploratory analysis also reveals some unexpected insights which should be tested more rigorously in future research. For example, trust game sender behavior is positively associated with the GSS trust question for subjects who have a “Charity/Altruism” representation in mind. This could be an artifact of the GSS trust question, which asks about trust in *strangers*. Moreover, working or pursuing a career in finance is — contrary to the original findings in Gill et al. (2022) — positively associated with subjects who think of everyday peer interactions and not at all associated with people who liken the game to trust. Potentially, it is not the least trustworthy who are pursuing careers in finance, but those who possess sufficient social capital to thrive in repeated interactions with peers but are not willing to sacrifice their own well-being for others, as demonstrated by the negative association among subjects who perceive the decision as an abstract act of helping and sharing. Moreover, there is also evidence of a statistically significant relationship between game and field behaviors among subjects with non-social mental representations. These findings all occur for the trust game, particularly for the sender in the trust game. This could be driven by risk preferences, which should influence gambling and investment decisions but also trusting behavior due to the associated uncertainty.

It is important to bear in mind that analyzing the relationship between game and field behavior within each mental representation does not only lead to substantially smaller sample sizes for each individual analysis (in particular for the finance outcome) but also, necessarily, to different sample sizes for different mental representations. This could contribute to some of the findings (in particular the lack thereof). However, keep in mind that the shares of individual categories are somewhat evenly spread for the trust game and very comparable within social categories for the dictator game (except for "Other Social" and "Parenting"), i.e., power is broadly comparable for these mental representations. Finally, the results presented so far could also reflect spurious correlations because subjects self-select into different mental representations. In Appendix C, I exploit the exogenous variation in the order of the games, which shifts mental representations, to provide evidence that speaks against spurious results.[29]

Taken together, while some of the insights are in line with economic theory, others raise new questions. However, the main takeaway from this, despite the preregistration still quite exploratory, exercise is that mental representations help uncover heterogeneity in the relationship between game and field behavior.

Result 3 *Mental representations correlate with game behavior and help uncover heterogeneity in the extent to which game behavior predicts field behavior in this experiment.*

Result 3 summarizes the findings regarding the relevance of mental representations for research using economic games. How subjects perceive an economic game correlates with their behavior in the game, thereby potentially

[29] Remember that while the effect of game order on mental representations is statistically significant, it shifts mental representations across the distribution. Considering the heterogeneity uncovered within social mental representations, future research may use more targeted treatments to shift mental representations more precisely to specific categories.

allowing researchers to learn more about the cognitive processes involved in human behavior, while also shedding light on the channels through which exogenous treatments operate (or fail to operate, cf. Section B.4). Furthermore, mental representations help uncover heterogeneity in the relationship between game and field behavior, which is broadly in line with economic theory. Taken together, the evidence presented in this paper indicates that heterogeneity in mental representations — or, putting it differently, heterogeneity in how people think about economic games — contains relevant information to advance research in economics.

4 Conclusion

Economic games in which humans interact in stylized and (more or less) abstract decision situations are an important method in economic research. Among other purposes, these games are used to better understand human behavior, for example by exposing subjects to different treatment conditions, but also to measure preferences based on behavior in these games.

This is the first paper in the economic literature to explicitly measure the mental representations of two economic games. I demonstrate that the controlled environment of these economic games does not extend to how people think about them. Mental representations are heterogeneous, even to the extent that when confronted with a game involving interaction with another subject, some subjects have a representation in mind that does not include any other person. Importantly, this is not driven by a lack of attention or comprehension of these games. The probability that two subjects report different mental representations within the same game ranges from 83.7% to 88.2% and still is between 20% and 40% when just considering whether the representation features another individual or not. Moreover, I show that mental representations seem to depend not only on the game itself but also on the broader experimental context and socio-demographic characteristics of the subject, with the latter probably picking up on different (life) experiences outside the game. Finally, this paper also demonstrates that accounting for mental representations helps to better understand game behavior and also allows to derive more precise preference measurements. Putting it more generally, heterogeneity in how people think about economic games seems to be informative about heterogeneity in game behavior and the ability of games to capture preferences.

These results are promising when considering research with economic games more broadly. A (not yet existent) theory that integrates mental representations (and their drivers), preferences, beliefs, and behavior could provide a unifying explanation for the inconclusive evidence on the relationship between game and field behavior (Galizzi and Navarro-Martinez, 2019) and the effect of framing on game behavior (Ellingsen et al., 2012; Dreber et al., 2013; Chang, Chen and Krupka, 2019). Moreover, if how games are perceived changes over time and/or is driven by the experiences subjects make in the days before an experiment, heterogeneity in mental representations across time — both on the individual and group level — could provide an explanation for the lack of stability in game behavior across time (Chuang and Schechter, 2015) and replication failures in experimental research (Camerer et al., 2016). Last but not least, mental representations can provide a sharper test to pin down treatment effects and could be partially responsible for the correlation of game behavior with socio-demographics (e.g., Chapman et al., 2023), variation across cultures (Henrich et al., 2001; Falk et al., 2018), and reaction to major shocks such as wars (Bauer et al., 2016), which might — at least in the short term — shift mental representations before shifting preferences.

Clearly, more work is needed before such a theory of mental representations can exist. First, this experiment should be repeated in different samples, ideally with more precise data on previous major and minor shocks and experiences, and with larger sample sizes to mitigate concerns regarding the lack of statistical power when analyzing different subgroups of mental representations separately. Extending the analysis to different samples will also shed light on whether the effects documented here are driven by the sample at hand, i.e., while the sample is representative of the US population in terms of age, ethnicity, and sex, it is also very homogeneous in that it only contains subjects who signed up at a survey provider to earn money. More heterogeneous samples might reveal more heterogeneous mental representations. Moreover, future research should also study different economic games and economic experiments without any interaction between subjects more generally, in particular because the analysis of mental representations of experiments becomes more intricate if subjects also form beliefs about the mental representations of other subjects. Second, more precisely targeted treatments to induce selected mental representations should be designed and implemented to provide causal evidence on the relevance of mental representations for research with economic games (driving behavior, explaining treatment effects, causing correlations between game behavior and subject characteristics/experiences or even cultures, and moderating the ability of game behavior to predict field behavior). While the treatments employed in this paper already provide insights into some of these use cases, they are not precise enough to induce a, potentially use-case-specific, targeted shift in mental representations. Finally, although most results are robust to a more objective classification procedure of the text responses (social vs. non-social) and having a large language model instead of research assistants classify the text responses, different coding schemes could be explored to demonstrate that this particular categorization does not drive the results. More selective coding schemes could also increase statistical power for the analyses of how mental representations are correlated with behavior and moderate the relationship between game and field behavior, albeit at the cost of objectivity, since the aggregation based on general behavioral domains probably conflates different important features of mental representations (e.g., some subjects mention the desire to match previous presents in gift-giving, while others do not). Ideally, after sufficiently much open-ended text data has been collected, semi-closed-ended measures can be developed (e.g., subjects selecting from a range of categories with the set of options based on an initial short open-ended text response).

However, despite these shortcomings, this first evidence on mental representations of economic games highlights that heterogeneity in how people think about economic games contains relevant information for research in economics. Future research can build on the insights in this paper to integrate mental representations in a framework that brings together heterogeneity in behavior, beliefs, preferences, and mental representations of economic games.

References

Alekseev, Aleksandr, Gary Charness, and Uri Gneezy. 2017. "Experimental Methods: When and Why Contextual Instructions Are Important." *Journal of Economic Behavior & Organization*, 134: 48–59.

Ambrus, Attila, and Ben Greiner. 2012. "Imperfect Public Monitoring with Costly Punishment: An Experimental Study." *American Economic Review*, 102(7): 3317–3332.

Andre, Peter, Ingar Haaland, Christopher Roth, and Johannes Wohlfart. 2022. "Narratives about the Macroeconomy." ECONtribute ECONtribute Discussion Paper No. 127. Available at `https://www.econtribute.de/RePEc/ajk/ajkdps/ECONtribute_127_2021.pdf` (accessed 2023-09-25).

Ash, Elliott, and Stephen Hansen. 2023. "Text Algorithms in Economics." *Annual Review of Economics*, 15: 659–688.

Bauer, Michal, Christopher Blattman, Julie Chytilová, Joseph Henrich, Edward Miguel, and Tamar Mitts. 2016. "Can War Foster Cooperation?" *Journal of Economic Perspectives*, 30(3): 249–274.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, Frederik Schwerter, and Andrei Shleifer. 2021. "Memory and Representativeness." *Psychological Review*, 128(1): 71–85.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2020. "Memory, Attention, and Choice." *The Quarterly Journal of Economics*, 135(3): 1399–1442.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten,

Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science*, 351(6280): 1433–1436.

Camerer, Colin F, Jonathan Cohen, Ernst Fehr, Paul Glimcher, and David Laibson. 2015. "Neuroeconomics." In *The Handbook of Experimental Economics*. Vol. 2. Princeton University Press.

Cason, Timothy N., and Charles R. Plott. 2014. "Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing." *Journal of Political Economy*, 122(6): 1235–1270.

Castillo, Daniel, François Bousquet, Marco A. Janssen, Kobchai Worrapimphong, and Juan Camillo Cardenas. 2011. "Context Matters to Explain Field Experiments: Results from Colombian and Thai Fishing Villages." *Ecological Economics*, 70(9): 1609–1620.

Chang, Daphne, Roy Chen, and Erin Krupka. 2019. "Rhetoric Matters: A Social Norms Explanation for the Anomaly of Framing." *Games and Economic Behavior*, 116: 158–178.

Chapman, Jonathan, Mark Dean, Pietro Ortoleva, Erik Snowberg, and Colin Camerer. 2023. "Econographics." *Journal of Political Economy Microeconomics*, 1(1): 115–161.

Charness, Gary, and Ernst Fehr. 2023. "Social Preferences: Fundamental Characteristics and Economic Consequences." CESifo CESifo Working Paper No. 10488, Munich. Available at `https://www.cesifo.org/DocDL/cesifo1_wp10488.pdf` (accessed 2023-09-25).

Charness, Gary, Thomas Garcia, Theo Offerman, and Marie Claire Villeval. 2020. "Do Measures of Risk Attitude in the Laboratory Predict

Behavior under Risk in and Outside of the Laboratory?" *Journal of Risk and Uncertainty*, 60: 99–123.

Charness, Gary, Uri Gneezy, and Alex Imas. 2013. "Experimental Methods: Eliciting Risk Preferences." *Journal of Economic Behavior & Organization*, 87: 43–51.

Chen, Daniel L., Martin Schonger, and Chris Wickens. 2016. "oTree—An Open-Source Platform for Laboratory, Online, and Field Experiments." *Journal of Behavioral and Experimental Finance*, 9: 88–97.

Chuang, Yating, and Laura Schechter. 2015. "Stability of Experimental and Survey Measures of Risk, Time, and Social Preferences: A Review and Some New Results." *Journal of Development Economics*, 117: 151–170.

Cohen, Jonathan, Keith Marzilli Ericson, David Laibson, and John Myles White. 2020. "Measuring Time Preferences." *Journal of Economic Literature*, 58(2): 299–347.

Detemple, Julian. 2023. "Mental Representations in Economic Games." OSF (still under embargo).

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. "Individual Risk Attitudes: Measurement, Determinants, And Behavioral Consequences." *Journal of the European Economic Association*, 9(3): 522–550.

Dreber, Anna, Tore Ellingsen, Magnus Johannesson, and David G. Rand. 2013. "Do People Care about Social Context? Framing Effects in Dictator Games." *Experimental Economics*, 16: 349–371.

Duffy, John, and Daniela Puzzello. 2014. "Gift Exchange versus Monetary Exchange: Theory and Evidence." *American Economic Review*, 104(6): 1735–1776.

Ellingsen, Tore, Magnus Johannesson, Johanna Mollerstrom, and Sara Munkhammar. 2012. "Social Framing Effects: Preferences or Beliefs?" *Games and Economic Behavior*, 76(1): 117–130.

Engel, Christoph, and David G. Rand. 2014. "What Does "Clean" Really Mean? The Implicit Framing of Decontextualized Experiments." *Economics Letters*, 122(3): 386–389.

Enke, Benjamin, Frederik Schwerter, and Florian Zimmermann. 2020. "Associative Memory and Belief Formation." National Bureau of Economic Research NBER Working Paper w26664, Cambridge, MA. Available at `http://www.nber.org/papers/w26664.pdf` (accessed 2023-09-25).

Eriksson, Kimmo, and Pontus Strimling. 2014. "Spontaneous Associations and Label Framing Have Similar Effects in the Public Goods Game." *Judgment and Decision Making*, 9(5): 360–372.

Falk, A., and J. J. Heckman. 2009. "Lab Experiments Are a Major Source of Knowledge in the Social Sciences." *Science*, 326(5952): 535–538.

Falk, Armin, and Michael Kosfeld. 2006. "The Hidden Costs of Control." *American Economic Review*, 96(5): 1611–1630.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. "Global Evidence on Economic Preferences." *The Quarterly Journal of Economics*, 133(4): 1645–1692.

Falkinger, Josef, Ernst Fehr, Simon Gachter, and Rudolf Winter-Ebmer. 2000. "A Simple Mechanism for the Efficient Provision of Public Goods: Experimental Evidence." *American Economic Review*, 90(1): 247–264.

Ferrario, Beatrice, and Stefanie Stantcheva. 2022. "Eliciting People's First-Order Concerns: Text Analysis of Open-Ended Survey Questions." *AEA Papers and Proceedings*, 112: 163–169.

Fischbacher, Urs, Simon Gächter, and Ernst Fehr. 2001. "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment." *Economics Letters*, 71(3): 397–404.

Gächter, Simon, Felix Kölle, and Simone Quercia. 2022. "Preferences and Perceptions in Provision and Maintenance Public Goods." *Games and Economic Behavior*, 135: 338–355.

Galizzi, Matteo M., and Daniel Navarro-Martinez. 2019. "On the External Validity of Social Preference Games: A Systematic Lab-Field Study." *Management Science*, 65(3): 976–1002.

Gennaioli, N., and A. Shleifer. 2010. "What Comes to Mind." *The Quarterly Journal of Economics*, 125(4): 1399–1433.

Gill, Andrej, Matthias Heinz, Heiner Schumacher, and Matthias Sutter. 2022. "Social Preferences of Young Professionals and the Financial Industry." *Management Science*, 69(7): 3905–3919.

Glaeser, Edward L., David I. Laibson, Jose A. Scheinkman, and Christine L. Soutter. 2000. "Measuring Trust." *Quarterly Journal of Economics*, 115(3): 811–846.

Gneezy, U., and A. Imas. 2017. "Lab in the Field: Measuring Preferences in the Wild." In *Handbook of Economic Field Experiments*. Vol. 1, 439–464. Elsevier.

GRE. 2023. "GRE General Test Verbal Reasoning Overview." `https://www.ets.org/gre/test-takers/general-test/prepare/content/verbal-reasoning.html` (accessed on 2023-09-05).

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. 2001. “In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies.” *American Economic Review*, 91(2): 73–78.

Ho, Mark K., David Abel, Carlos G. Correa, Michael L. Littman, Jonathan D. Cohen, and Thomas L. Griffiths. 2022. “People Construct Simplified Mental Representations to Plan.” *Nature*, 606(7912): 129–136.

Houser, Daniel, and Kevin McCabe. 2014. “Experimental Economics and Experimental Game Theory.” In *Neuroeconomics*. . Second Edition ed., 19–34. Elsevier.

Kahneman, Daniel, and Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision under Risk.” *Econometrica*, 47(2): 263–292.

Levitt, Steven D, and John A List. 2007. “What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?” *Journal of Economic Perspectives*, 21(2): 153–174.

Luhmann, Niklas. 2014. *Vertrauen: ein Mechanismus der Reduktion sozialer Komplexität. UTB*. 5. Aufl ed., Konstanz:UVK Verlagsgesellschaft.

Malmendier, Ulrike. 2021. “Experience Effects in Finance: Foundations, Applications, and Future Directions.” *Review of Finance*, 25(5): 1339–1363.

Naar, Nicole. 2020. “Gaming Anthropology: The Problem of External Validity and the Challenge of Interpreting Experimental Games.” *American Anthropologist*, 122(4): 784–798.

Ockenfels, Axel, and Uta K. Schier. 2020. “Games as Frames.” *Journal of Economic Behavior & Organization*, 172: 97–106.

OpenAI. 2023. "GPT-4 Technical Report." Available at arXiv: `https://arxiv.org/abs/2303.08774` (accessed 2023-09-25).

Palan, Stefan, and Christian Schitter. 2018. "Prolific.Ac—A Subject Pool for Online Experiments." *Journal of Behavioral and Experimental Finance*, 17: 22–27.

Riedl, Arno, and Paul Smeets. 2017. "Why Do Investors Hold Socially Responsible Mutual Funds?" *The Journal of Finance*, 72(6): 2505–2550.

Rustagi, D., S. Engel, and M. Kosfeld. 2010. "Conditional Cooperation and Costly Monitoring Explain Success in Forest Commons Management." *Science*, 330(6006): 961–965.

Saccardo, Silvia, and Marta Serra-Garcia. 2023. "Enabling or Limiting Cognitive Flexibility? Evidence of Demand for Moral Commitment." *American Economic Review*, 113(2): 396–429.

Sapienza, Paola, Anna Toldra-Simats, and Luigi Zingales. 2013. "Understanding Trust." *The Economic Journal*, 123(573): 1313–1332.

Selten, Reinhard. 1967. "Die Strategiemethode Zur Erforschung Des Eingeschränkt Rationalen Verhaltens Im Rahmen Eines Oligopolexperimentes." In *Beiträge Zur Experimentellen Wirtschaftsforschung.* , ed. H. Sauermann, 136–168. Tübingen:J.C.B. Mohr (Paul Siebeck).

Viceisza, Angelino C. G. 2016. "Creating A Lab In The Field: Economics Experiments For Policymaking." *Journal of Economic Surveys*, 30(5): 835–854.

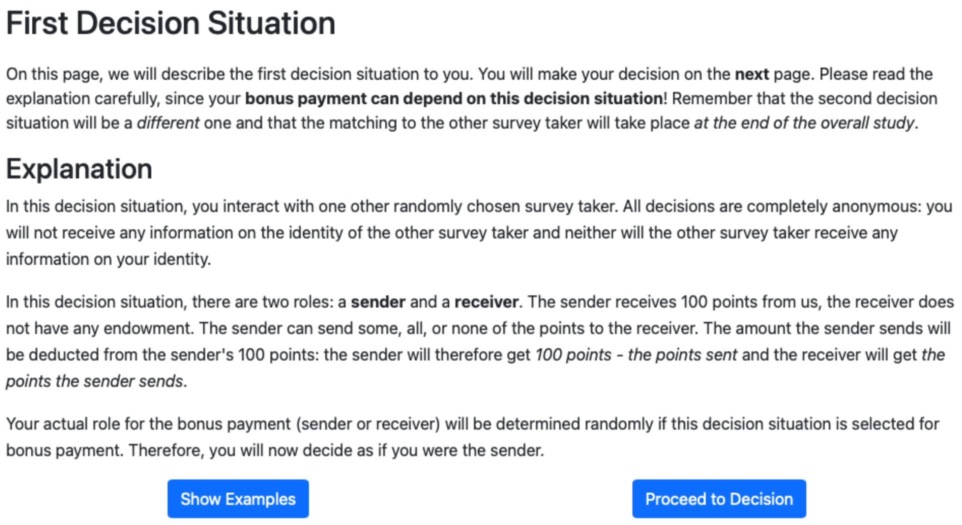
Xiao, Erte, and Daniel Houser. 2005. "Emotion Expression in Human Punishment Behavior." *Proceedings of the National Academy of Sciences*, 102(20): 7398–7401.

Yamagishi, Toshio, Nobuhiro Mifune, Yang Li, Mizuho Shinada, Hirofumi Hashimoto, Yutaka Horita, Arisa Miura, Keigo Inukai, Shigehito Tanida, Toko Kiyonari, Haruto Takagishi, and Dora Simunovic. 2013. "Is Behavioral Pro-Sociality Game-Specific? Pro-social Preference and Expectations of pro-Sociality." *Organizational Behavior and Human Decision Processes*, 120(2): 260–271.

A Research Design: Additional Details

A.1 Instructions & Screenshots of Decision Screens

Table A.1 provides the instructions in the dictator and trust game in text form. Figures A.1 to A.9 provide screenshots of the decision screens for the first part of the experiment with the dictator game coming first.



First Decision Situation

On this page, we will describe the first decision situation to you. You will make your decision on the **next** page. Please read the explanation carefully, since your **bonus payment can depend on this decision situation**! Remember that the second decision situation will be a *different* one and that the matching to the other survey taker will take place *at the end of the overall study*.

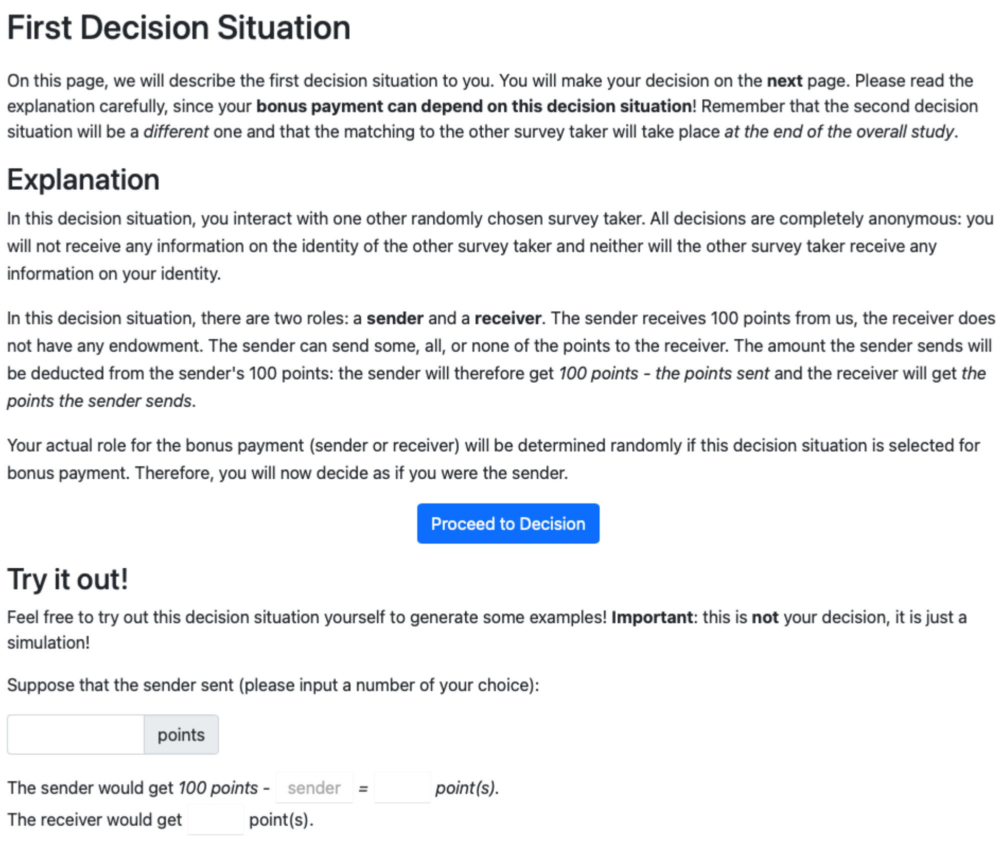
Explanation

In this decision situation, you interact with one other randomly chosen survey taker. All decisions are completely anonymous: you will not receive any information on the identity of the other survey taker and neither will the other survey taker receive any information on your identity.

In this decision situation, there are two roles: a **sender** and a **receiver**. The sender receives 100 points from us, the receiver does not have any endowment. The sender can send some, all, or none of the points to the receiver. The amount the sender sends will be deducted from the sender's 100 points: the sender will therefore get *100 points - the points sent* and the receiver will get *the points the sender sends*.

Your actual role for the bonus payment (sender or receiver) will be determined randomly if this decision situation is selected for bonus payment. Therefore, you will now decide as if you were the sender.

Figure A.1: Instructions Dictator Game



First Decision Situation

On this page, we will describe the first decision situation to you. You will make your decision on the **next** page. Please read the explanation carefully, since your **bonus payment can depend on this decision situation**! Remember that the second decision situation will be a *different* one and that the matching to the other survey taker will take place *at the end of the overall study*.

Explanation

In this decision situation, you interact with one other randomly chosen survey taker. All decisions are completely anonymous: you will not receive any information on the identity of the other survey taker and neither will the other survey taker receive any information on your identity.

In this decision situation, there are two roles: a **sender** and a **receiver**. The sender receives 100 points from us, the receiver does not have any endowment. The sender can send some, all, or none of the points to the receiver. The amount the sender sends will be deducted from the sender's 100 points: the sender will therefore get *100 points - the points sent* and the receiver will get *the points the sender sends*.

Your actual role for the bonus payment (sender or receiver) will be determined randomly if this decision situation is selected for bonus payment. Therefore, you will now decide as if you were the sender.

Try it out!

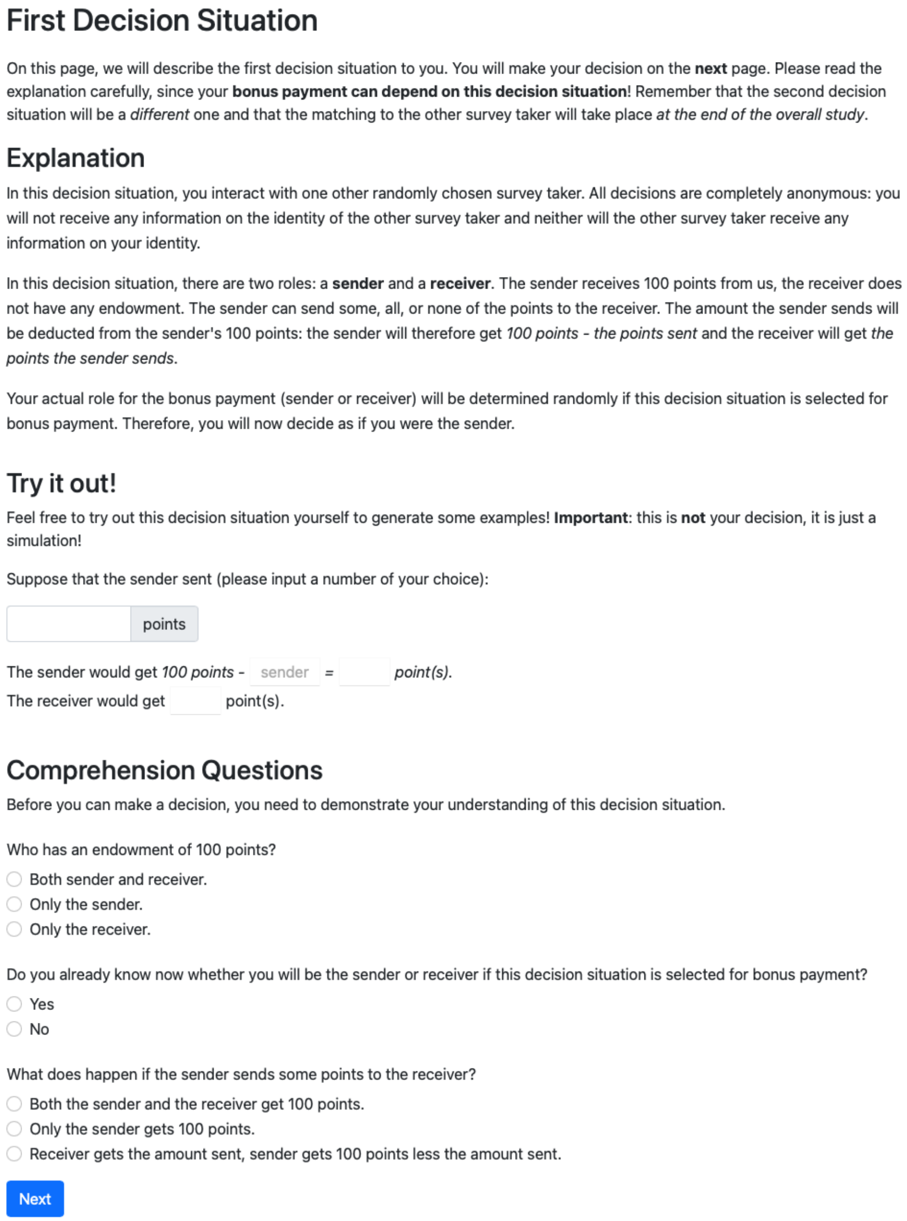
Feel free to try out this decision situation yourself to generate some examples! **Important**: this is **not** your decision, it is just a simulation!

Suppose that the sender sent (please input a number of your choice):

The sender would get *100 points -* sender = *point(s).*

The receiver would get point(s).

Figure A.2: Instructions Dictator Game after Clicking on “Show Examples”



First Decision Situation

On this page, we will describe the first decision situation to you. You will make your decision on the **next** page. Please read the explanation carefully, since your **bonus payment can depend on this decision situation**! Remember that the second decision situation will be a *different* one and that the matching to the other survey taker will take place *at the end of the overall study*.

Explanation

In this decision situation, you interact with one other randomly chosen survey taker. All decisions are completely anonymous: you will not receive any information on the identity of the other survey taker and neither will the other survey taker receive any information on your identity.

In this decision situation, there are two roles: a **sender** and a **receiver**. The sender receives 100 points from us, the receiver does not have any endowment. The sender can send some, all, or none of the points to the receiver. The amount the sender sends will be deducted from the sender's 100 points: the sender will therefore get *100 points - the points sent* and the receiver will get *the points the sender sends*.

Your actual role for the bonus payment (sender or receiver) will be determined randomly if this decision situation is selected for bonus payment. Therefore, you will now decide as if you were the sender.

Try it out!

Feel free to try out this decision situation yourself to generate some examples! **Important**: this is **not** your decision, it is just a simulation!

Suppose that the sender sent (please input a number of your choice):

The sender would get *100 points -* sender = *point(s).*

The receiver would get point(s).

Comprehension Questions

Before you can make a decision, you need to demonstrate your understanding of this decision situation.

Who has an endowment of 100 points?

○ Both sender and receiver.
○ Only the sender.
○ Only the receiver.

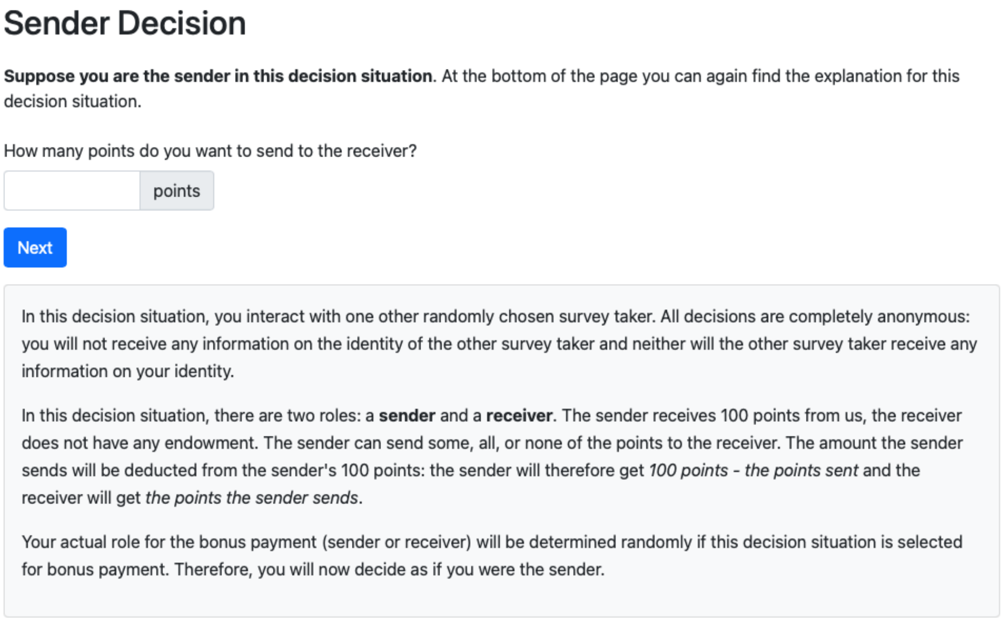
Do you already know now whether you will be the sender or receiver if this decision situation is selected for bonus payment?

○ Yes
○ No

What does happen if the sender sends some points to the receiver?

○ Both the sender and the receiver get 100 points.
○ Only the sender gets 100 points.
○ Receiver gets the amount sent, sender gets 100 points less the amount sent.

Figure A.3: Comprehension Questions Dictator Game



Sender Decision

Suppose you are the sender in this decision situation. At the bottom of the page you can again find the explanation for this decision situation.

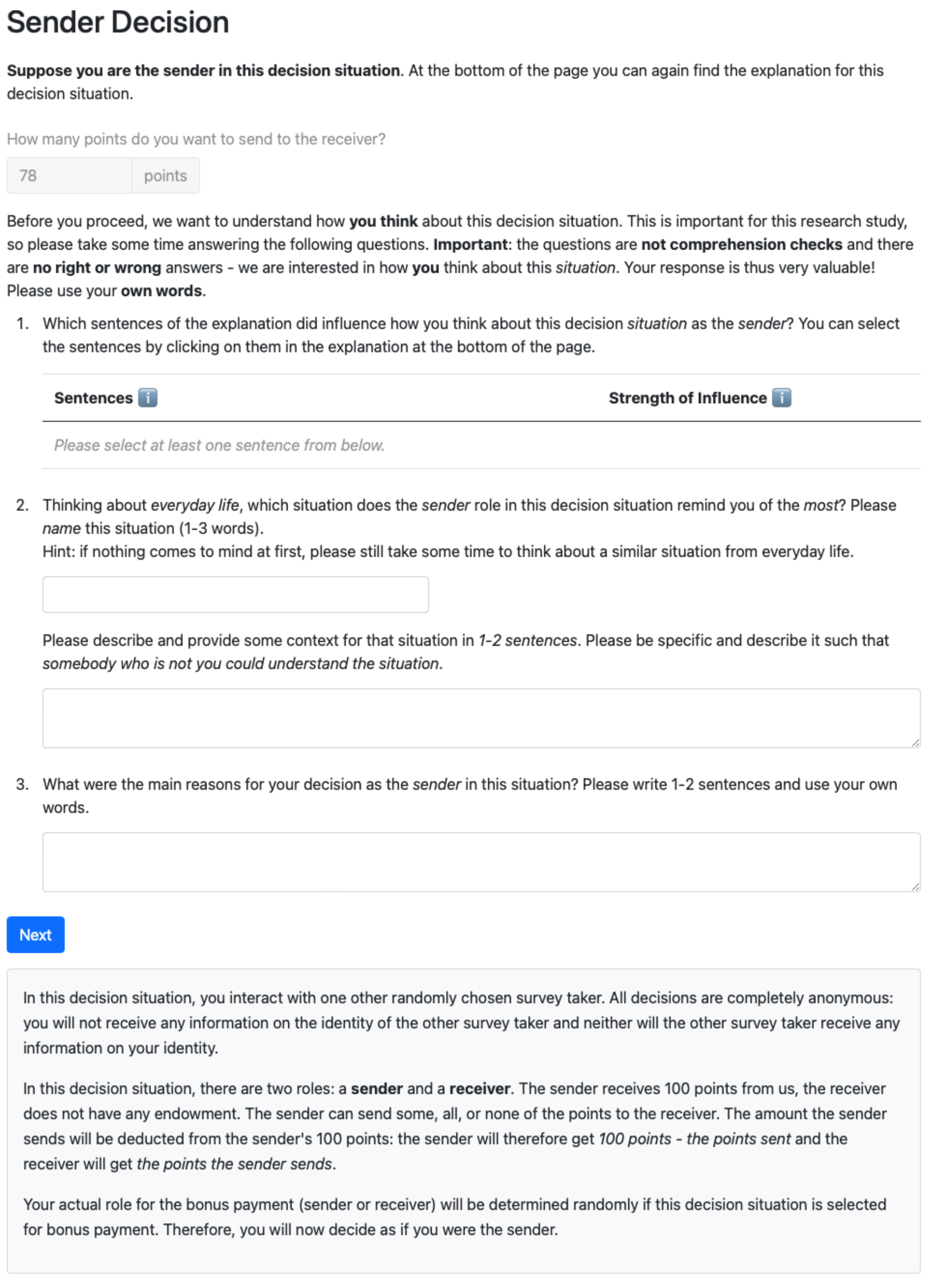
How many points do you want to send to the receiver?

In this decision situation, you interact with one other randomly chosen survey taker. All decisions are completely anonymous: you will not receive any information on the identity of the other survey taker and neither will the other survey taker receive any information on your identity.

In this decision situation, there are two roles: a **sender** and a **receiver**. The sender receives 100 points from us, the receiver does not have any endowment. The sender can send some, all, or none of the points to the receiver. The amount the sender sends will be deducted from the sender's 100 points: the sender will therefore get *100 points - the points sent* and the receiver will get *the points the sender sends*.

Your actual role for the bonus payment (sender or receiver) will be determined randomly if this decision situation is selected for bonus payment. Therefore, you will now decide as if you were the sender.

Figure A.4: Decision Screen Dictator Game



Sender Decision

Suppose you are the sender in this decision situation. At the bottom of the page you can again find the explanation for this decision situation.

How many points do you want to send to the receiver?

78 points

Before you proceed, we want to understand how **you think** about this decision situation. This is important for this research study, so please take some time answering the following questions. **Important**: the questions are **not comprehension checks** and there are **no right or wrong** answers - we are interested in how **you** think about this *situation*. Your response is thus very valuable! Please use your **own words**.

1. Which sentences of the explanation did influence how you think about this decision *situation* as the *sender*? You can select the sentences by clicking on them in the explanation at the bottom of the page.

Sentences ℹ	**Strength of Influence** ℹ
Please select at least one sentence from below.	

2. Thinking about *everyday life*, which situation does the *sender* role in this decision situation remind you of the *most*? Please *name* this situation (1-3 words).
Hint: if nothing comes to mind at first, please still take some time to think about a similar situation from everyday life.

   Please describe and provide some context for that situation in *1-2 sentences*. Please be specific and describe it such that *somebody who is not you could understand the situation*.

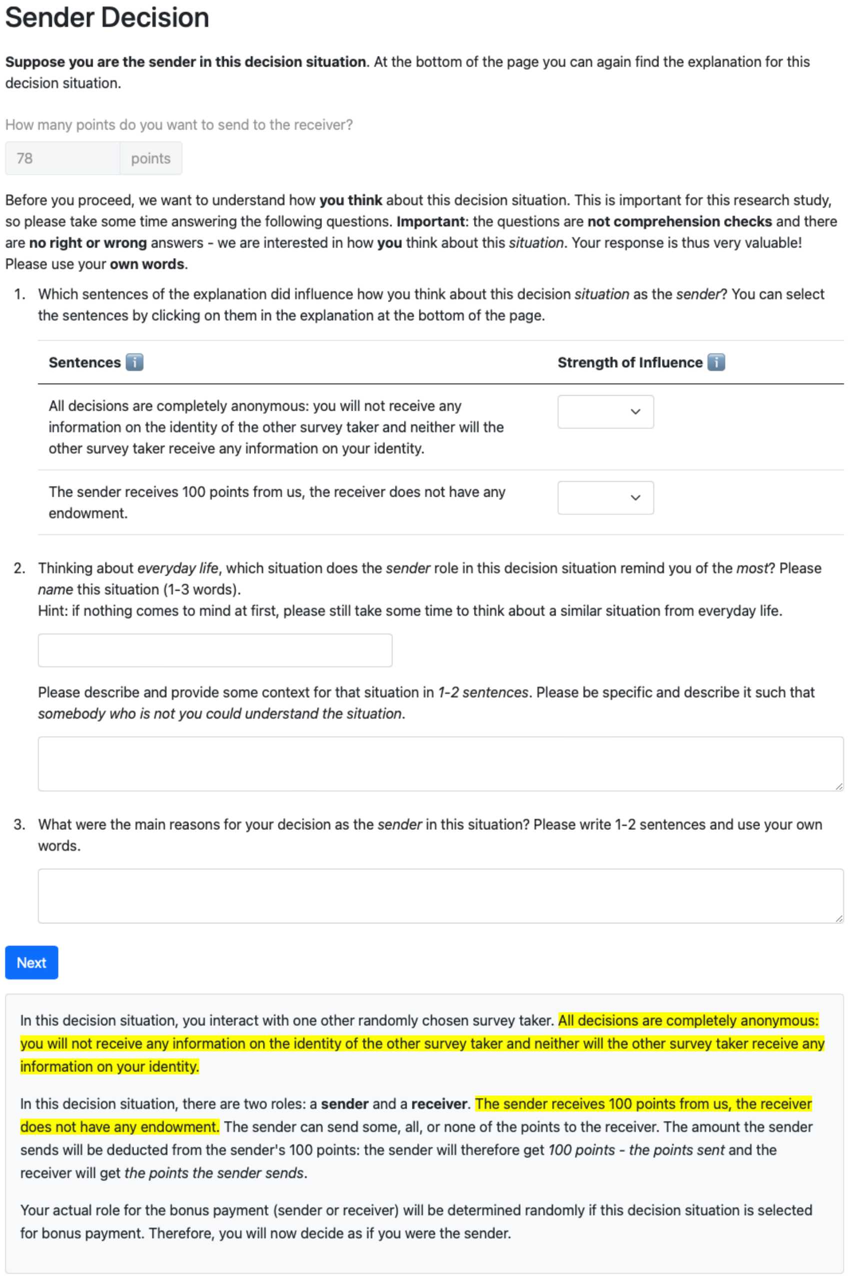
3. What were the main reasons for your decision as the *sender* in this situation? Please write 1-2 sentences and use your own words.

In this decision situation, you interact with one other randomly chosen survey taker. All decisions are completely anonymous: you will not receive any information on the identity of the other survey taker and neither will the other survey taker receive any information on your identity.

In this decision situation, there are two roles: a **sender** and a **receiver**. The sender receives 100 points from us, the receiver does not have any endowment. The sender can send some, all, or none of the points to the receiver. The amount the sender sends will be deducted from the sender's 100 points: the sender will therefore get *100 points - the points sent* and the receiver will get *the points the sender sends*.

Your actual role for the bonus payment (sender or receiver) will be determined randomly if this decision situation is selected for bonus payment. Therefore, you will now decide as if you were the sender.

Figure A.5: Elicitation of Mental Representations after Decision in Dictator Game



Sender Decision

Suppose you are the sender in this decision situation. At the bottom of the page you can again find the explanation for this decision situation.

How many points do you want to send to the receiver?

78 points

Before you proceed, we want to understand how **you think** about this decision situation. This is important for this research study, so please take some time answering the following questions. **Important**: the questions are **not comprehension checks** and there are **no right or wrong** answers - we are interested in how **you** think about this *situation*. Your response is thus very valuable! Please use your **own words**.

1. Which sentences of the explanation did influence how you think about this decision *situation* as the *sender*? You can select the sentences by clicking on them in the explanation at the bottom of the page.

Sentences ⓘ	**Strength of Influence** ⓘ
All decisions are completely anonymous: you will not receive any information on the identity of the other survey taker and neither will the other survey taker receive any information on your identity.	
The sender receives 100 points from us, the receiver does not have any endowment.	

2. Thinking about *everyday life*, which situation does the *sender* role in this decision situation remind you of the *most*? Please *name* this situation (1-3 words).
Hint: if nothing comes to mind at first, please still take some time to think about a similar situation from everyday life.

Please describe and provide some context for that situation in *1-2 sentences*. Please be specific and describe it such that *somebody who is not you could understand the situation*.

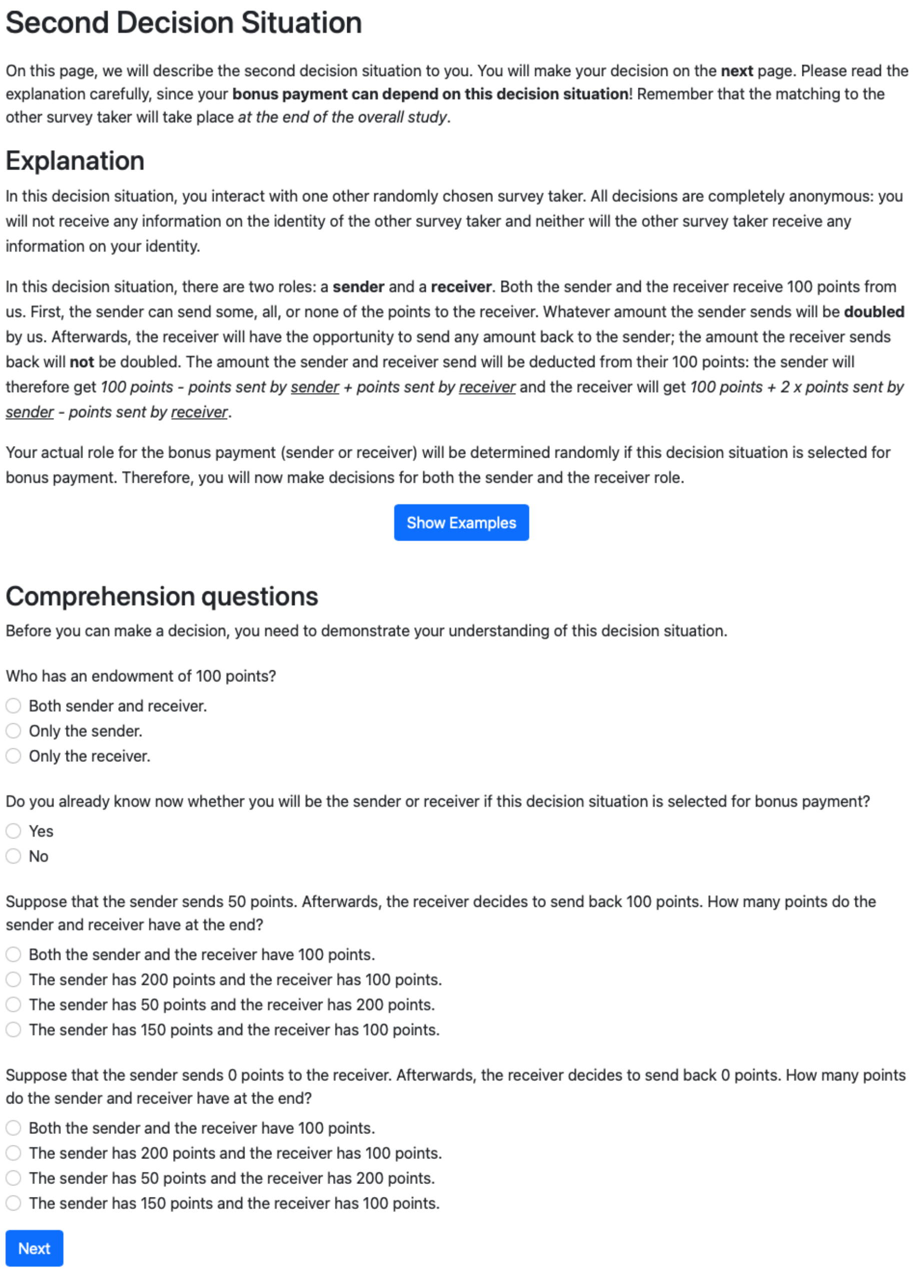
3. What were the main reasons for your decision as the *sender* in this situation? Please write 1-2 sentences and use your own words.

In this decision situation, you interact with one other randomly chosen survey taker. All decisions are completely anonymous: you will not receive any information on the identity of the other survey taker and neither will the other survey taker receive any information on your identity.

In this decision situation, there are two roles: a **sender** and a **receiver**. The sender receives 100 points from us, the receiver does not have any endowment. The sender can send some, all, or none of the points to the receiver. The amount the sender sends will be deducted from the sender's 100 points: the sender will therefore get *100 points - the points sent* and the receiver will get *the points the sender sends*.

Your actual role for the bonus payment (sender or receiver) will be determined randomly if this decision situation is selected for bonus payment. Therefore, you will now decide as if you were the sender.

Figure A.6: Selecting Sentences after Decision in Dictator Game



Second Decision Situation

On this page, we will describe the second decision situation to you. You will make your decision on the **next** page. Please read the explanation carefully, since your **bonus payment can depend on this decision situation**! Remember that the matching to the other survey taker will take place *at the end of the overall study*.

Explanation

In this decision situation, you interact with one other randomly chosen survey taker. All decisions are completely anonymous: you will not receive any information on the identity of the other survey taker and neither will the other survey taker receive any information on your identity.

In this decision situation, there are two roles: a **sender** and a **receiver**. Both the sender and the receiver receive 100 points from us. First, the sender can send some, all, or none of the points to the receiver. Whatever amount the sender sends will be **doubled** by us. Afterwards, the receiver will have the opportunity to send any amount back to the sender; the amount the receiver sends back will **not** be doubled. The amount the sender and receiver send will be deducted from their 100 points: the sender will therefore get *100 points - points sent by sender + points sent by receiver* and the receiver will get *100 points + 2 x points sent by sender - points sent by receiver*.

Your actual role for the bonus payment (sender or receiver) will be determined randomly if this decision situation is selected for bonus payment. Therefore, you will now make decisions for both the sender and the receiver role.

Comprehension questions

Before you can make a decision, you need to demonstrate your understanding of this decision situation.

Who has an endowment of 100 points?

○ Both sender and receiver.
○ Only the sender.
○ Only the receiver.

Do you already know now whether you will be the sender or receiver if this decision situation is selected for bonus payment?

○ Yes
○ No

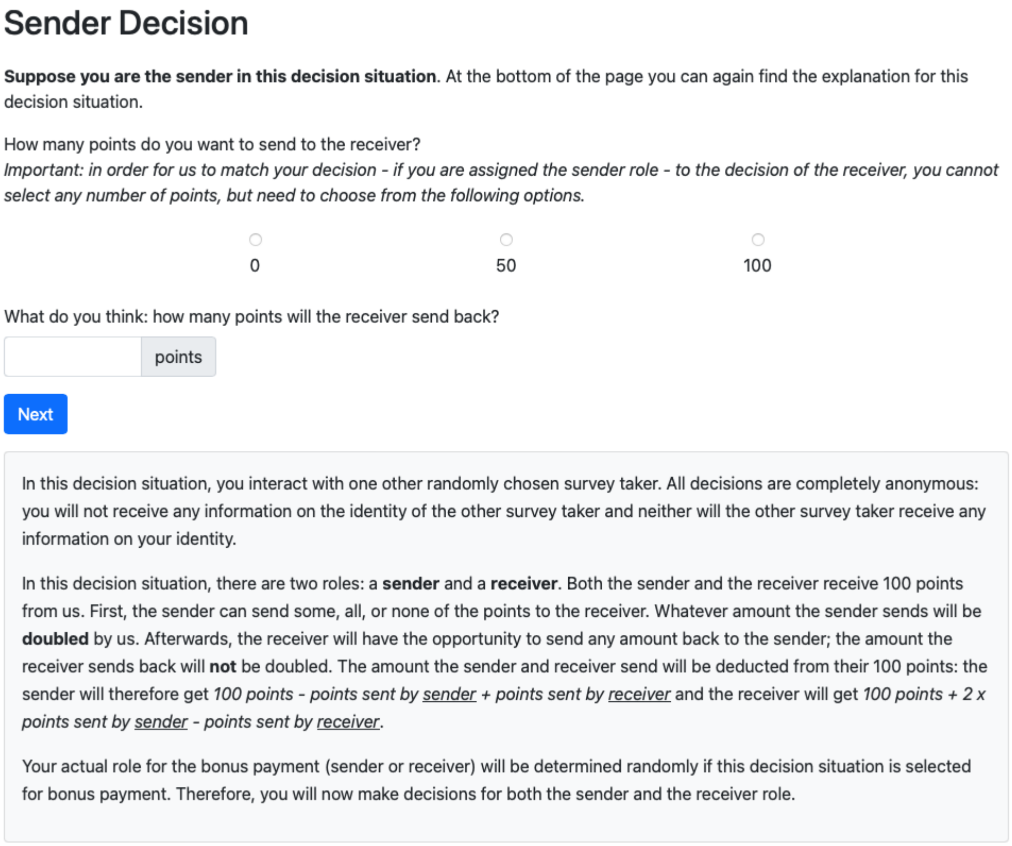
Suppose that the sender sends 50 points. Afterwards, the receiver decides to send back 100 points. How many points do the sender and receiver have at the end?

○ Both the sender and the receiver have 100 points.
○ The sender has 200 points and the receiver has 100 points.
○ The sender has 50 points and the receiver has 200 points.
○ The sender has 150 points and the receiver has 100 points.

Suppose that the sender sends 0 points to the receiver. Afterwards, the receiver decides to send back 0 points. How many points do the sender and receiver have at the end?

○ Both the sender and the receiver have 100 points.
○ The sender has 200 points and the receiver has 100 points.
○ The sender has 50 points and the receiver has 200 points.
○ The sender has 150 points and the receiver has 100 points.

Figure A.7: Comprehension Questions Trust Game



Sender Decision

Suppose you are the sender in this decision situation. At the bottom of the page you can again find the explanation for this decision situation.

How many points do you want to send to the receiver?
Important: in order for us to match your decision - if you are assigned the sender role - to the decision of the receiver, you cannot select any number of points, but need to choose from the following options.

○ 0 ○ 50 ○ 100

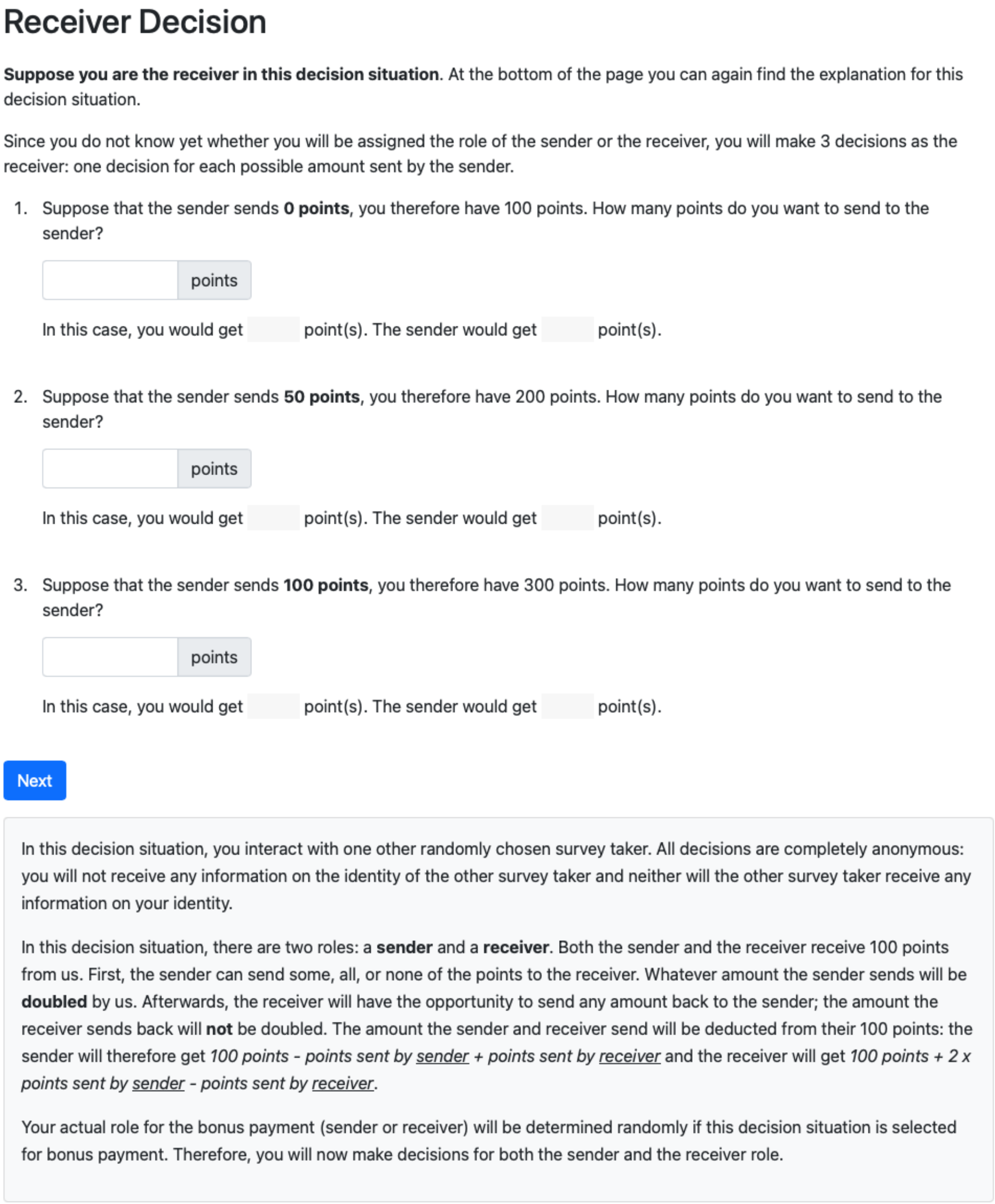
What do you think: how many points will the receiver send back?

In this decision situation, you interact with one other randomly chosen survey taker. All decisions are completely anonymous: you will not receive any information on the identity of the other survey taker and neither will the other survey taker receive any information on your identity.

In this decision situation, there are two roles: a **sender** and a **receiver**. Both the sender and the receiver receive 100 points from us. First, the sender can send some, all, or none of the points to the receiver. Whatever amount the sender sends will be **doubled** by us. Afterwards, the receiver will have the opportunity to send any amount back to the sender; the amount the receiver sends back will **not** be doubled. The amount the sender and receiver send will be deducted from their 100 points: the sender will therefore get *100 points - points sent by sender + points sent by receiver* and the receiver will get *100 points + 2 x points sent by sender - points sent by receiver*.

Your actual role for the bonus payment (sender or receiver) will be determined randomly if this decision situation is selected for bonus payment. Therefore, you will now make decisions for both the sender and the receiver role.

Figure A.8: Decision as Sender in Trust Game



Receiver Decision

Suppose you are the receiver in this decision situation. At the bottom of the page you can again find the explanation for this decision situation.

Since you do not know yet whether you will be assigned the role of the sender or the receiver, you will make 3 decisions as the receiver: one decision for each possible amount sent by the sender.

1. Suppose that the sender sends **0 points**, you therefore have 100 points. How many points do you want to send to the sender?

   In this case, you would get point(s). The sender would get point(s).

2. Suppose that the sender sends **50 points**, you therefore have 200 points. How many points do you want to send to the sender?

   In this case, you would get point(s). The sender would get point(s).

3. Suppose that the sender sends **100 points**, you therefore have 300 points. How many points do you want to send to the sender?

   In this case, you would get point(s). The sender would get point(s).

In this decision situation, you interact with one other randomly chosen survey taker. All decisions are completely anonymous: you will not receive any information on the identity of the other survey taker and neither will the other survey taker receive any information on your identity.

In this decision situation, there are two roles: a **sender** and a **receiver**. Both the sender and the receiver receive 100 points from us. First, the sender can send some, all, or none of the points to the receiver. Whatever amount the sender sends will be **doubled** by us. Afterwards, the receiver will have the opportunity to send any amount back to the sender; the amount the receiver sends back will **not** be doubled. The amount the sender and receiver send will be deducted from their 100 points: the sender will therefore get *100 points - points sent by sender + points sent by receiver* and the receiver will get *100 points + 2 x points sent by sender - points sent by receiver*.

Your actual role for the bonus payment (sender or receiver) will be determined randomly if this decision situation is selected for bonus payment. Therefore, you will now make decisions for both the sender and the receiver role.

Figure A.9: Decision as Receiver in Trust Game

Table A.1: Instructions in Text Form

<table>
<tr><th>Sentence Label</th><th>Text Dictator Game</th><th>Text Trust Game</th></tr>
<tr><td>Random Interaction†</td><td colspan="2">In this [community] decision situation, you interact with one other randomly chosen survey taker.</td></tr>
<tr><td>Anonymity†</td><td colspan="2">All decisions are completely anonymous: you will not receive any information on the identity of the other survey taker and neither will the other survey taker receive any information on your identity.</td></tr>
<tr><td>Two Roles†</td><td colspan="2">In this [community] decision situation, there are two roles: a sender and a receiver.</td></tr>
<tr><td>Endowment</td><td>The sender receives 100 points from us, the receiver does not have any endowment.</td><td>Both the sender and the receiver receive 100 points from us.</td></tr>
<tr><td>Sender action space</td><td>The sender can send some, all, or none of the points to the receiver.</td><td>First, the sender can send some, all, or none of the points to the receiver.</td></tr>
<tr><td>Multiplier</td><td></td><td>Whatever amount the sender sends will be doubled by us.</td></tr>
<tr><td>Receiver action space</td><td></td><td>Afterwards, the receiver will have the opportunity to send any amount back to the sender; the amount the receiver sends back will not be doubled.</td></tr>
<tr><td>Payoffs</td><td>The amount the sender sends will be deducted from the sender's 100 points: the sender will therefore get <i>100 points - the points sent</i> and the receiver will get <i>the points the sender sends.</i></td><td>The amount the sender and receiver send will be deducted from their 100 points: the sender will therefore get <i>100 points - points sent by sender + points sent by receive</i> and the receiver will get <i>100 points + 2 × points sent by sender - points sent by receiver.</i></td></tr>
</table>

Table A.1 Continued from previous page

Sentence Label	**Text Dictator Game**	**Text Trust Game**
Role unknown ex-ante	Your actual role for the bonus payment (sender or receiver) will be determined randomly if this [community] decision situation is selected for bonus payment. Therefore, you will now decide as if you were the sender.	Your actual role for the bonus payment (sender or receiver) will be determined randomly if this [community] decision situation is selected for bonus payment. Therefore, you will now make decisions for both the sender and the receiver role.

Notes: † Identical instructions for dictator and trust game.

A.2 Results from ML-Algorithm to Identify Clusters

While the selection of the categories (cf. Table 1) can seem ad-hoc at first, similar clusters are identified by a machine learning algorithm, applied to the qualitative text responses.

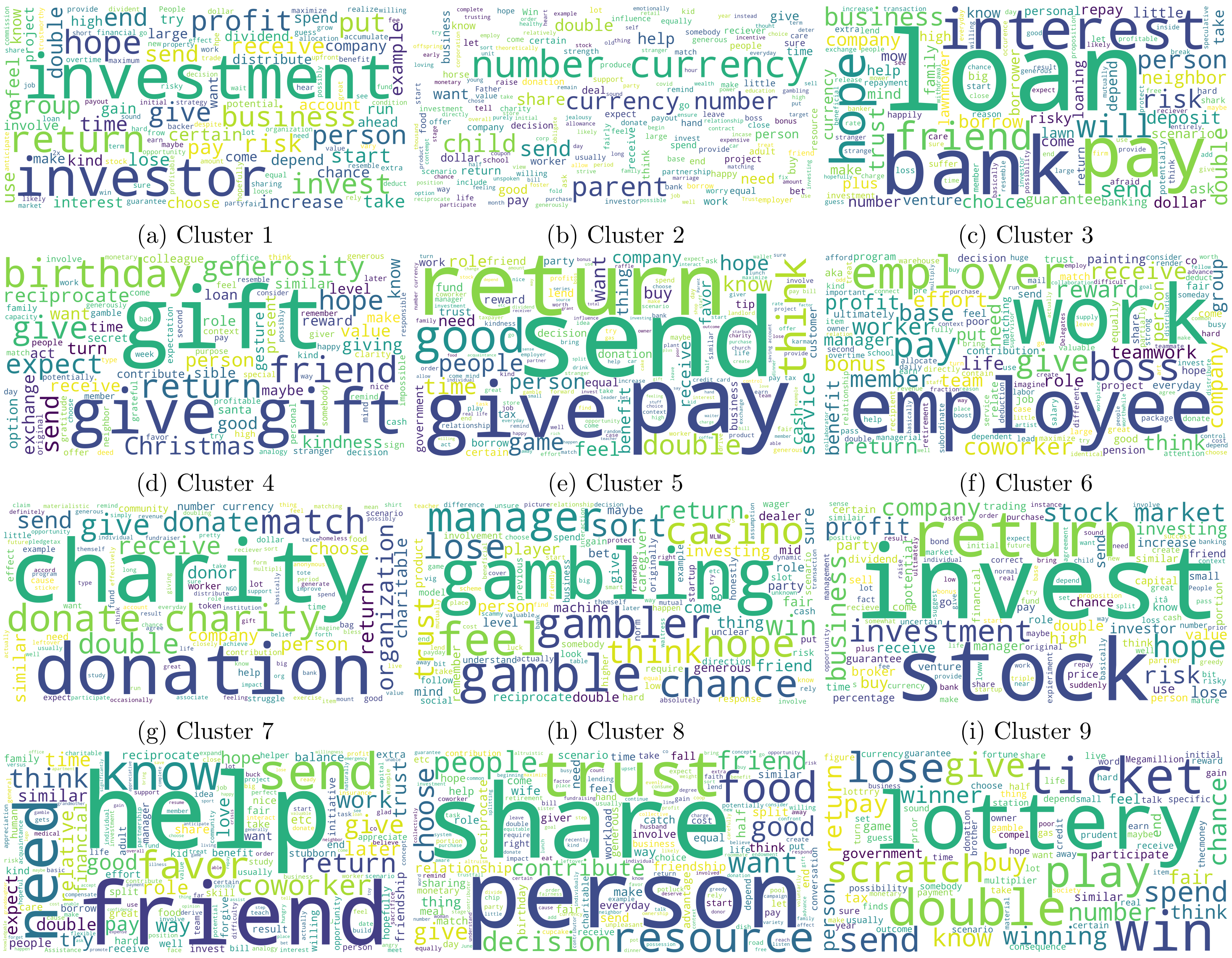
For this, the chosen name of the decision situation from everyday life is merged with the context subjects provided. Then, a term-frequency-inverse-document-frequency algorithm (tf-idf) transforms each text into a vector representation based on the tf-idf-score of each word. See Ash and Hansen (2023) for a discussion of the tf-idf algorithm (and other approaches) to analyze open text in economic research. The tf-idf score of each word is computed as the product of term frequency (the ratio of how many times a word appears in a text response to the total number of words in that response) and inverse-document-frequency (the logarithm of the ratio of the total number of responses to the number of responses in which the word occurs). Consequently, words that are very common and occur in every text response receive a very low score, even if they occur very frequently in a single text response. Only words that occur very frequently in a particular response but not in many other text responses, i.e., which seem to contain information about the unique meaning of that response, are assigned a high score. To ensure that similar words such as "(to) invest" and "investing" can be identified as the same word across different text responses, text responses are cleaned before applying the tf-idf algorithm. This involves the removal of stop words (i.e., words that are common but do not contain lots of information such as "situation", "reminds", and "points") and lemmatizing each word, i.e., each word is transformed to its root (e.g., "investing" becomes "invest").

Afterward, a K-Means algorithm is used to identify clusters of common responses based on the tf-idf-scores. To provide an example of the output of the ML clustering exercise, Figure A.10 contains an illustration of the most common words within each cluster of text responses for the sender in the trust game for 12 clusters. There is some variation in the output depending on the number of clusters.

First, consider that also the “objective” ML algorithm identifies clusters based on the behavioral domain in which the associations from everyday life take place. For example, cluster 1 contains responses in which people talk about investments that involve the role of an investor and potentially another “person”. Cluster 2 is more difficult to interpret but features the words parent and child besides mentioning numbers and currencies (any number and currency symbol are replaced by the words “number” and “currency” in the pre-processing steps). Cluster 3 mentions financial transactions other than investments, in particular bank loans. Clusters 4 and 10 contain situations related to gift-giving or helping among friends. Cluster 6 involves interactions between employer and employee. Cluster 7 picks up on charity donations, while clusters 8 and 12 feature elements of gambling. Cluster 9 contains associations related to financial investments in the stock market. Finally, cluster 11 is about sharing something with an (abstract) person but also trust.

Second, some of the clusters contain conceptual overlap (e.g., giving gifts and helping friends; cf. clusters 4 and 10). Consequently, the output from the ML clustering exercise is fine-tuned by the author. This involves having a common set of categories across all game roles, e.g., financial investment occurs almost exclusively in the trust game, but to have the same starting point it should also be a category in the dictator game. Moreover, I attach labels to the clusters identified, distinguish between acts of helping/sharing among peers and abstract people, add the respective “other” categories, and enrich the set of categories with the “preference”-categories.

Summing up, while the decision on the categories might seem ad-hoc at first, even an objective ML clustering algorithm based on the co-occurrence of words identifies a similar set of clusters, which is then fine-tuned to suit the needs of this study.



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

(d) Cluster 4 (e) Cluster 5 (f) Cluster 6

(g) Cluster 7 (h) Cluster 8 (i) Cluster 9

(j) Cluster 10 (k) Cluster 11 (l) Cluster 12

Figure A.10: Clusters Identified by K-Means Algorithm for Sender in Trust Game

A.3 Details on the Sample

Table A.2 reports socio-demographic information on the sample with $n = 600$.

A.4 Eliciting Field Behaviors

Table A.3 provides an overview of the different types of field behaviors elicited and how the elicitation in this paper differs from the elicitation in the original paper. While Galizzi and Navarro-Martinez (2019) (SRA) and Riedl and Smeets (2017) (SRI) regress field behavior on the respective game behavior, Glaeser et al. (2000) (GSS, lending) and Gill et al. (2022) (finance) regress game behavior on field behavior. To harmonize the interpretation of the coefficients, I always regress field behavior on game behavior. In line with the original estimation models, I use linear regressions for all field behaviors except the binary SRI-indicator for which I use a probit model. The results for lending and finance are robust to using a probit model instead of a linear probability model. In the main part, I omit all control variables because I cannot replicate the original results for most outcomes (cf. Appendix D).

	Mean	Std.	Data source
Age	45.81	15.56	Prolific
Ethnicity			Prolific
Asian	0.06		
Black	0.13		
White	0.77		
Mixed	0.02		
Other	0.02		
Sex			Prolific
Female	0.51		
Male	0.49		
Household Annual Income (in USD)			Self-reported
Less than 10,000	0.04		
10,000-24,999	0.09		
25,000-49,999	0.27		
50,000-74,999	0.19		
75,000-99,999	0.15		
More than 100,000	0.22		
Prefer not to say	0.04		
Occupation			Self-reported
Working full time now	0.52		
Working part time now	0.17		
Temporarily laid off	0.04		
Unemployed	0.01		
Retired	0.07		
Permanently disabled	0.03		
Taking care of home or family	0.03		
Student	0.05		
Other	0.05		

Notes: Overview of socio-demographic data collected as part of the study. Std = standard deviation, only computed for age. All other variables are indicator variables. Data source refers to whether the data are collected by Prolific or self-reported by subjects.

Table A.2: Sample Demographics

Table A.3: Elicitation of Field Behaviors

Name	**Abbrv.**	**Ref.**	**Sample**	**Outcome & Elicitation**
Self-Reported Altruism Scale	SRA	Galizzi and Navarro-Martinez (2019)	UK students	**Original:** 20-item questionnaire (summed up for the index) on frequency of altruistic acts. **This paper:** shortened 5-item questionnaire (money/goods donation to charity, sending help to individual, helping a stranger in need, donating blood, volunteering for charity). SRA score is standardized.
General Social Survey Trust Question	GSS	Glaeser et al. (2000)	US students	**Original:** answer to the question "Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?", most likely on 11-item Likert scale. **This paper:** 5-item Likert scale with 0 = "You can't be too careful", 5 = "Most people can be trusted", no labels in between. GSS response is standardized.
Past trusting behavior	lending	Glaeser et al. (2000)	US students	**Original:** index based on a multi-item questionnaire on past trusting behavior (frequency of lending money to friends, frequency of lending personal possessions to friends, intentionally leaving the door open). **This paper:** single question on frequency of lending money or personal possessions to friends; for power reasons aggregated to binary indicator whether happens "more than once per year".

Table A.3 Continued from previous page

Name	Abbrv.	Reference	Sample	Elicitation
Occupation in finance	finance	Gill et al. (2022)	German students	**Original:** indicating an interest in pursuing a career in the finance industry on Likert scale ("To what extent can you imagine working in the following industries in the future?") from one ("certainly not") to seven ("definitively") and self-reported current job. **This paper:** single question on "In which industry do you aspire to work (students)/did you work (retired)/do you work (employed)?" with binary indicator whether subjects select "finance". Collapse all in a single indicator. Subjects who are permanently unemployed or are not working because they take care of home or family or indicate "other" as occupation are excluded from the analysis.
Socially Responsible Investment	SRI	Riedl and Smeets (2017)	Dutch Investors	**Original:** actual investment behavior, parameterized in two ways. First, binary indicator of whether investors hold SRI equity. Second, share of total equity invested in SRI equity. **This paper:** for binary indicator, use the answer to the question which option best describes how subjects would like to invest their money with options: "I would like my money to be invested in a way that contributes to sustainability." (indicator for SRI), "I would like my money to be invested without giving special consideration to sustainability criteria.', "I do not want to invest my money at all.", "I have no money to save or invest.". Subjects who indicate "I do not want to invest my money at all." or "I have no money to save or invest." are excluded from the analysis (i.e., SRI indicator is set to missing). Instead of the actual share invested, the amount hypothetically donated to NGO working on reducing CO_2 emissions, parameterized as the share of the hypothetical endowment of USD 450.

B Additional Analyses & Figures

B.1 Full Distributions of Associations

Figure B.1 plots the distribution of associations before aggregating every social (non-social) category with less than 5% into the "Other Social" ("Other Non-Social") category. Associations belonging to "No Situation" are also aggregated into the "Other Non-Social" category. This ensures that any statistical findings are not driven by outliers (e.g., categories that are assigned very infrequently and might therefore be more subjective). A black line indicates this aggregation threshold of 5%. All results in Section 3 are qualitatively robust to using a lower aggregation threshold of 2.5%.

B.2 Mental Representations across Age, Ethnicity, and Sex

Figures B.2 and B.3 compare the distributions of mental representation for subjects above the median age with subjects below the median age. They complement the parametric analysis contained in Figure 8.

Figures B.4 and B.5 compare the distributions of mental representation for subjects who are white with subjects who are not white. They complement the parametric analysis contained in Figure 9.

Figures B.6 and B.7 compare the distributions of mental representation for subjects who are female with subjects who are not female. They complement the parametric analysis contained in Figure 10.

B.3 Individual Mental Representations and Behavior

Figures B.8, B.9, and B.10 plot the distribution of behavior as the dictator, the sender in the trust game, and the receiver in the trust game for each individual mental representation separately. In line with the parametric analyses

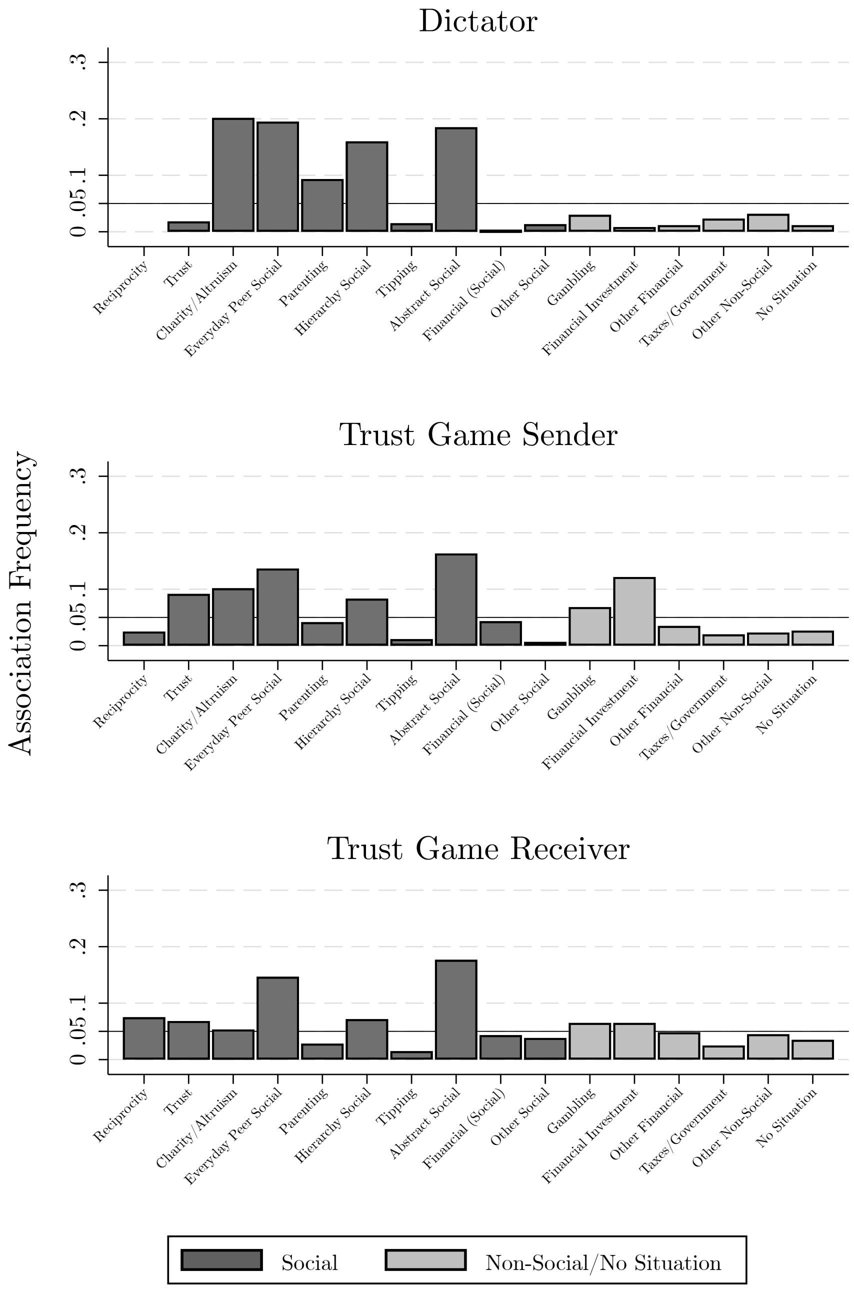


Figure B.1: Associations in Economic Games (Full Distribution)

Notes: Full distribution of associations, explanation of categories in Table 1. The black horizontal line indicates the aggregation threshold of 5%.

contained in Figure 12, the figures highlight that there does exist heterogeneity in behavior across the individual mental representations and even within the class of social mental representations.

Dictator behavior is parameterized as the absolute amount sent to the other subject by the dictator. Sender behavior in the trust game is the absolute amount sent to the receiver in the trust game. In line with the literature (e.g., Gill et al., 2022), receiver behavior in the trust game is parameterized as the average number of points sent back relative to the points received (i.e., amount of points sent × multiplier), or putting it differently, the average share of the points received which are sent back across the two relevant information sets (sender sending 50 points and sender sending 100 points). This share can be more than 100% if subjects decide to send back even more than they received (receivers also have an endowment). The number of receivers who do so is negligible ($n = 4$). These subjects are omitted in Figure B.10 for illustrative purposes.

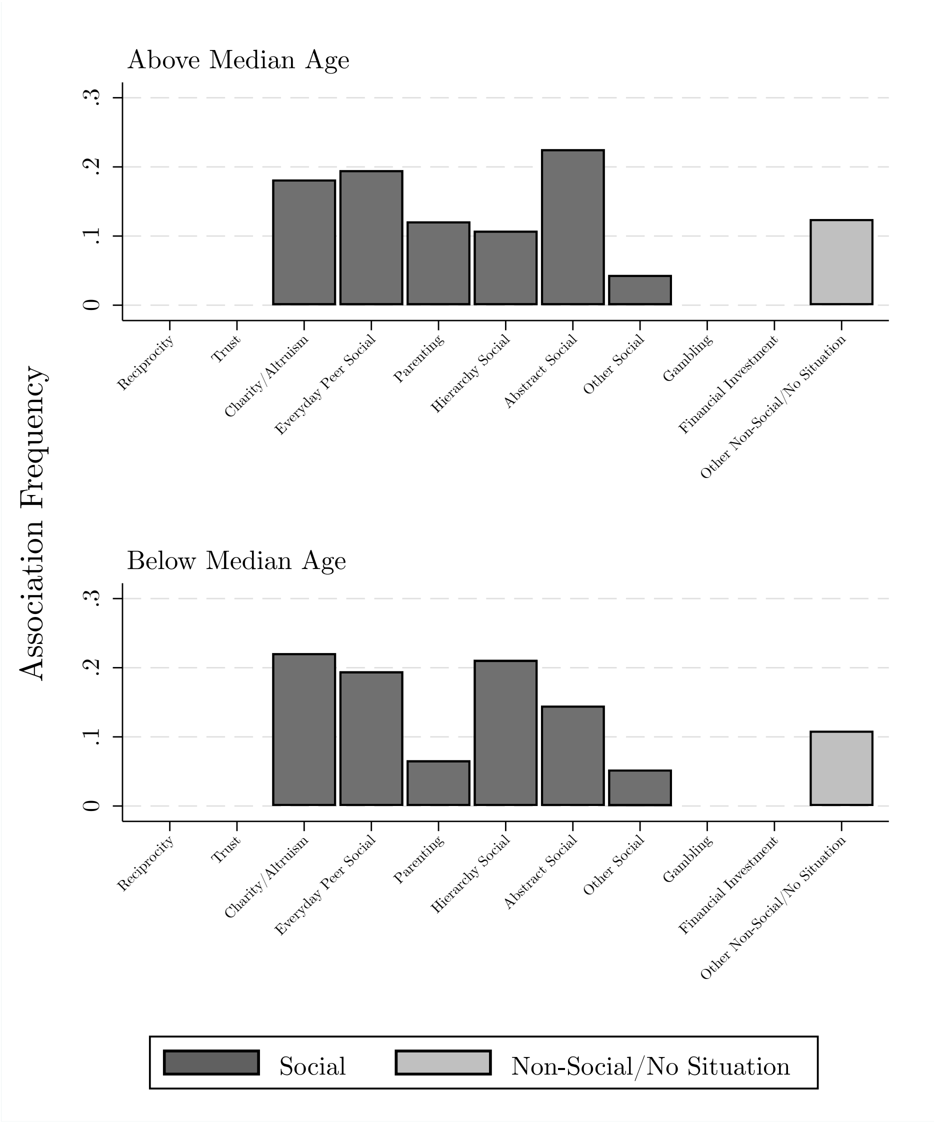
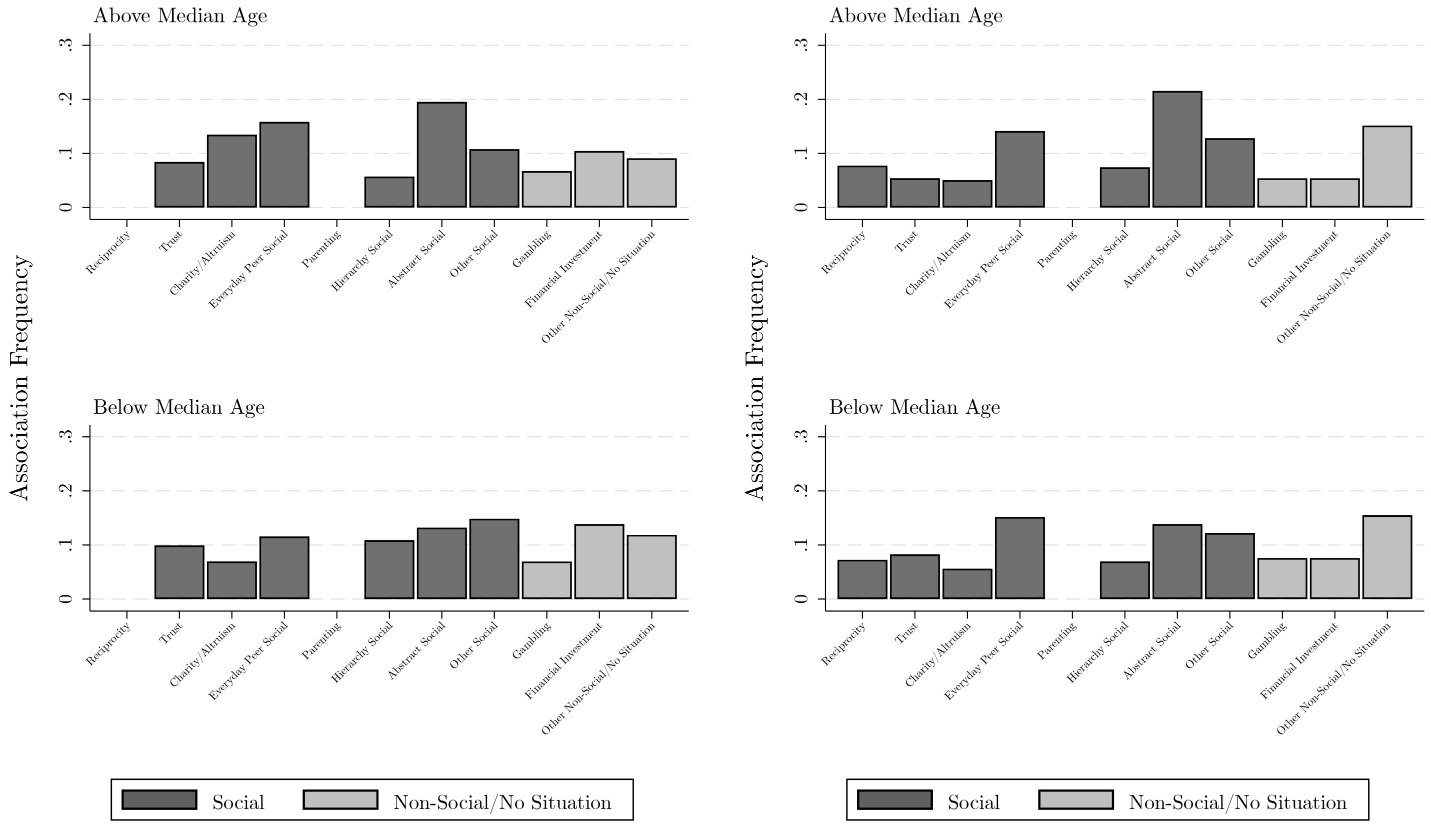


Figure B.2: Mental Representations Across Age in Dictator Game

Notes: Distribution of associations by age, explanation of categories in Table 1. Categories that occur less frequently than 5% are aggregated into the respective "Other" category.



(a) Trust Game: Sender

(b) Trust Game: Receiver

Figure B.3: Mental Representations Across Age in Trust Game

Notes: Distribution of associations by age, explanation of categories in Table 1. Categories that occur less frequently than 5% are aggregated into the respective "Other" category.

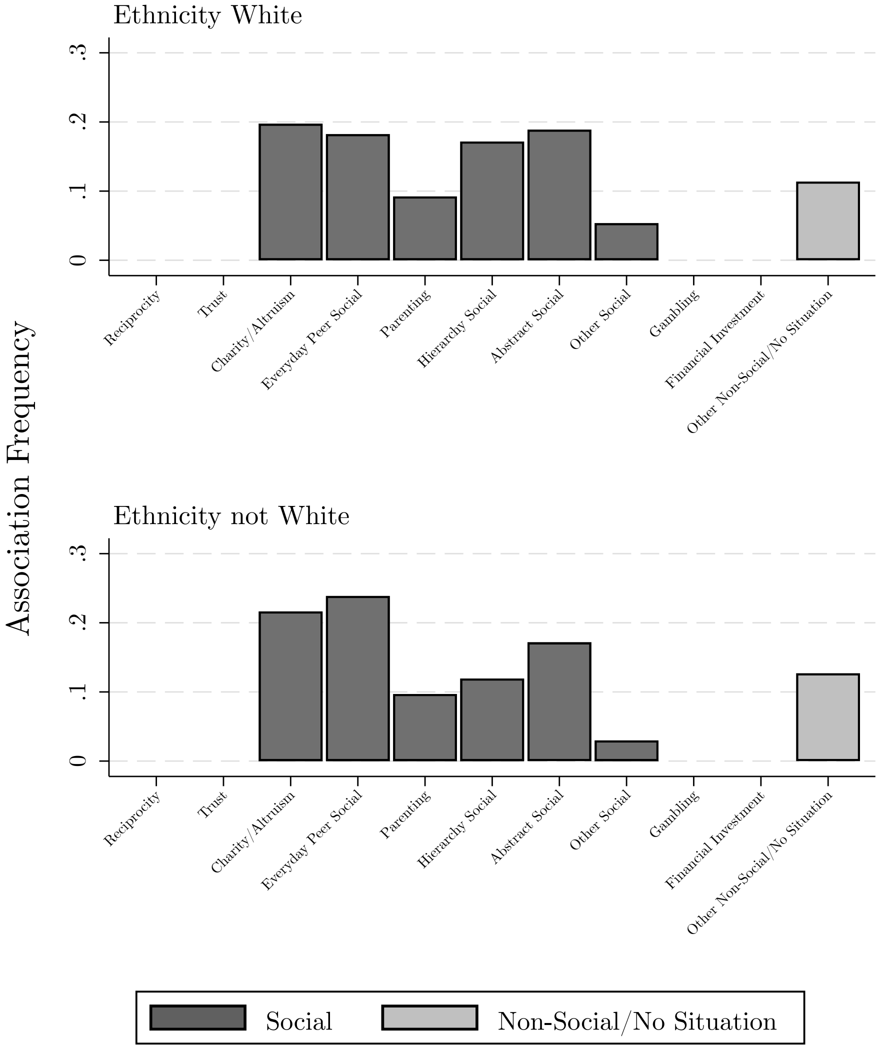
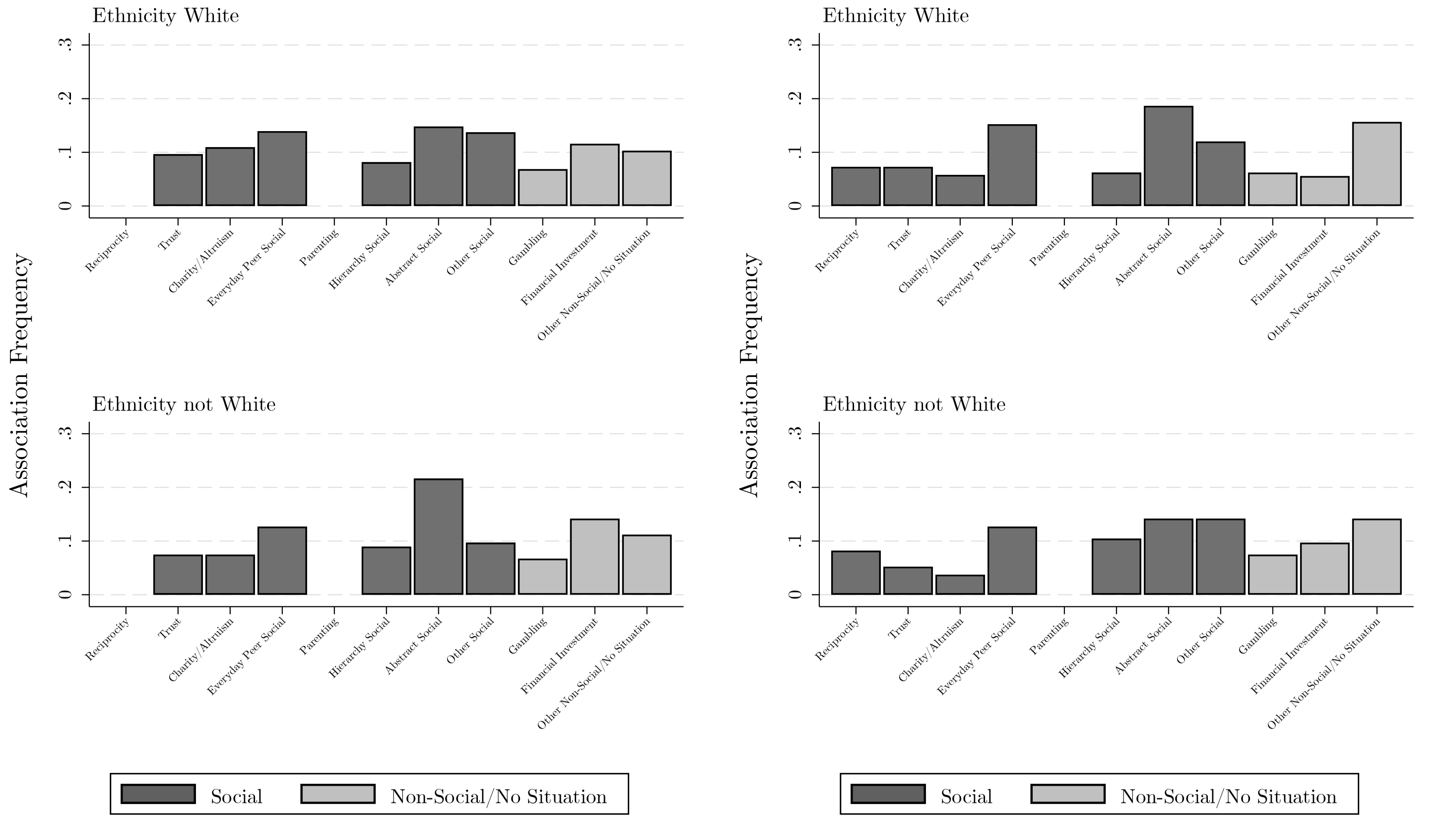


Figure B.4: Mental Representations Across Ethnicity in Dictator Game

Notes: Distribution of associations by ethnicity, explanation of categories in Table 1. Categories that occur less frequently than 5% are aggregated into the respective "Other" category.



(a) Trust Game: Sender

(b) Trust Game: Receiver

Figure B.5: Mental Representations Across Ethnicity in Trust Game

Notes: Distribution of associations by ethnicity, explanation of categories in Table 1. Categories that occur less frequently than 5% are aggregated into the respective "Other" category.

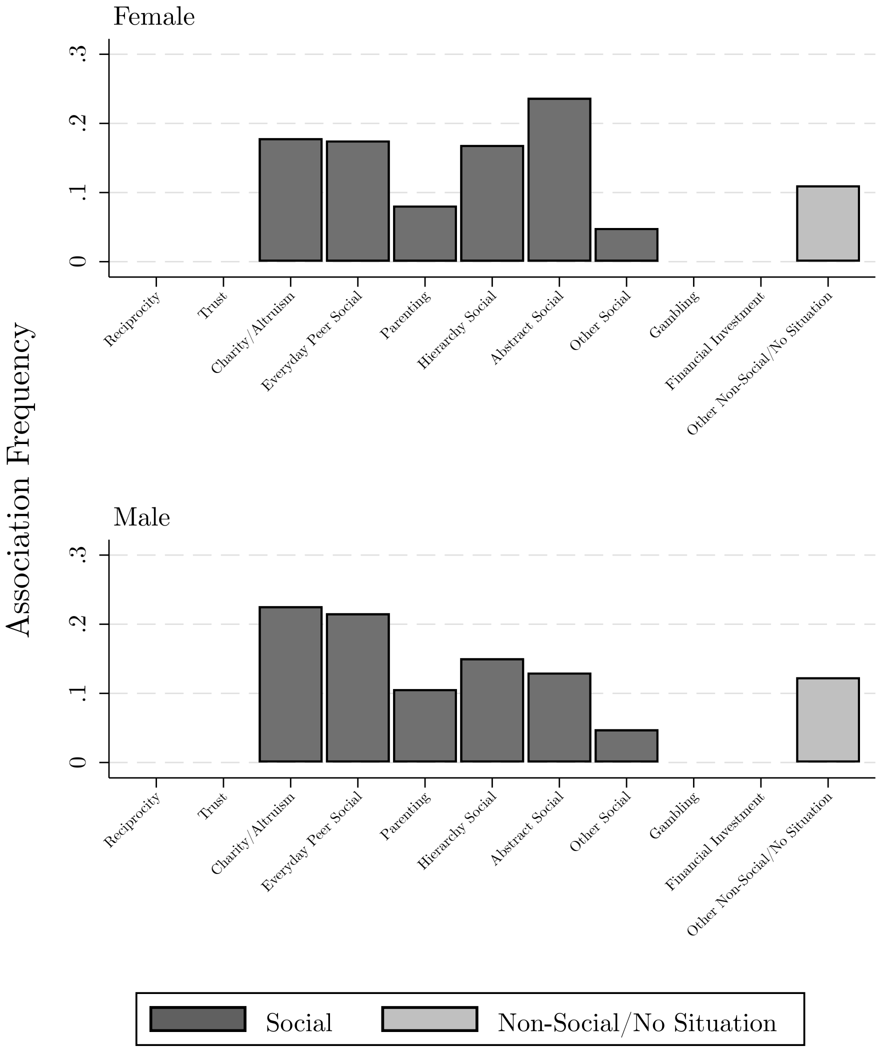
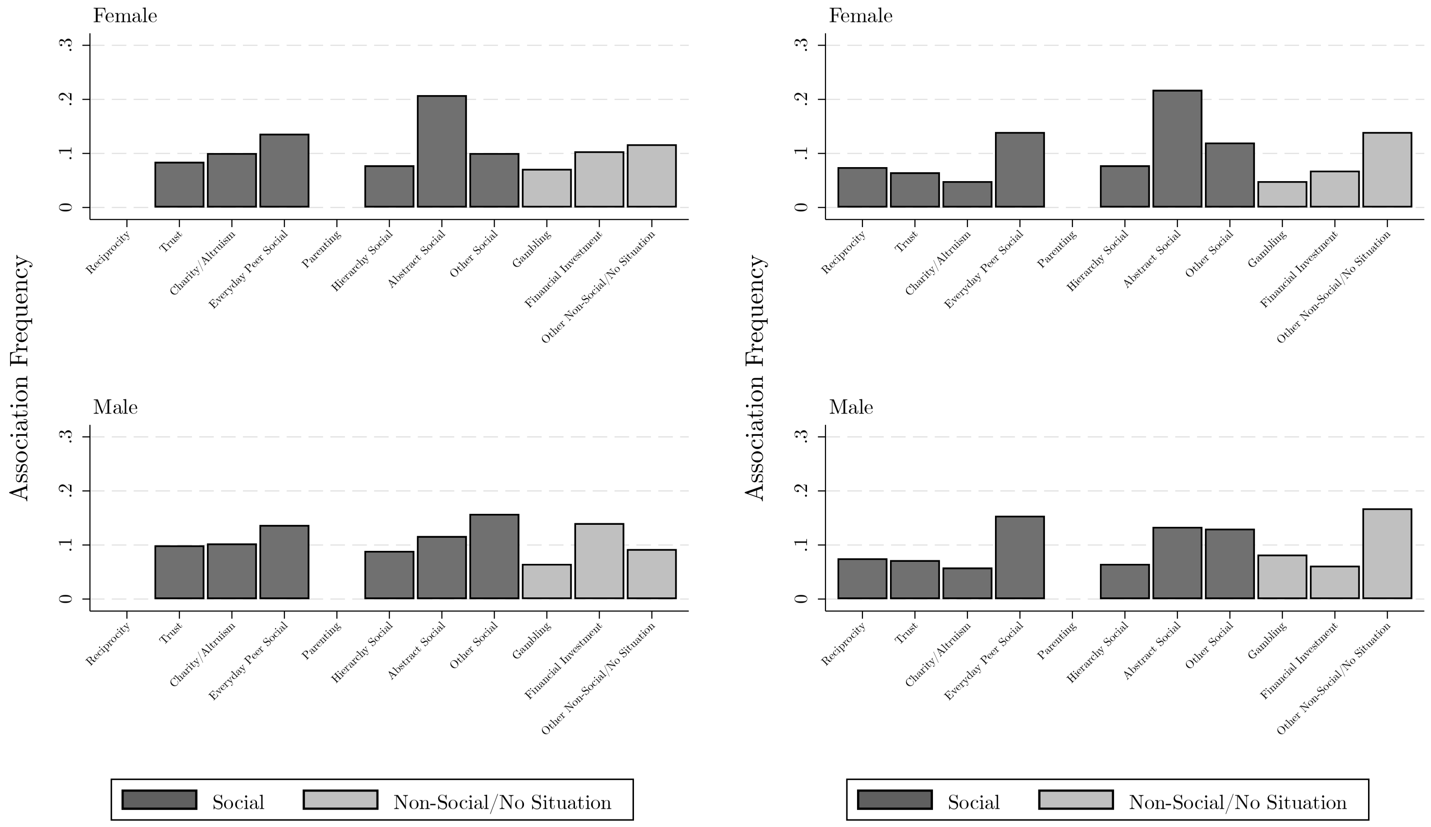


Figure B.6: Mental Representations Across Sex in Dictator Game

Notes: Distribution of associations by sex, explanation of categories in Table 1. Categories that occur less frequently than 5% are aggregated into the respective "Other" category.



(a) Trust Game: Sender

(b) Trust Game: Receiver

Figure B.7: Mental Representations Across Sex in Trust Game

Notes: Distribution of associations by sex, explanation of categories in Table 1. Categories that occur less frequently than 5% are aggregated into the respective "Other" category.

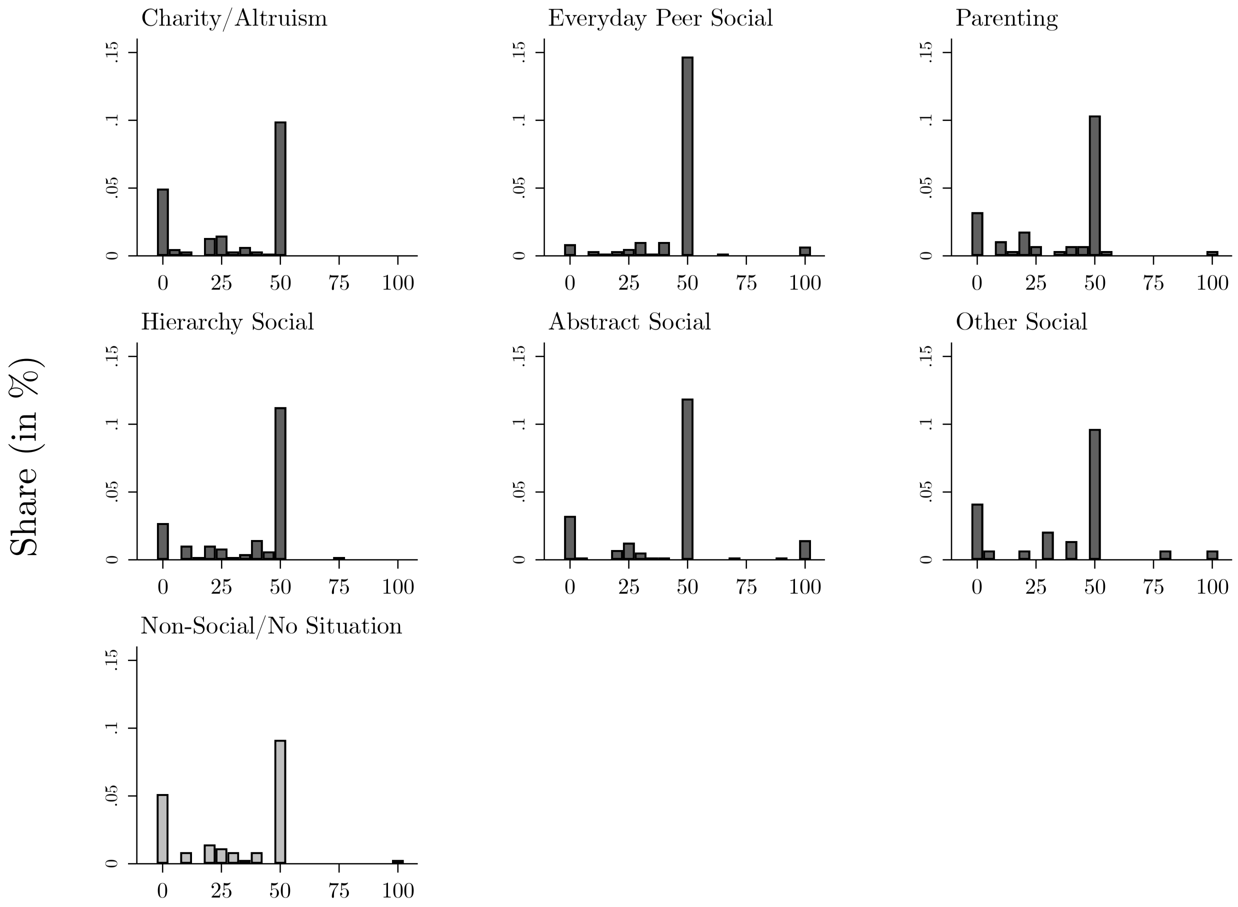


Figure B.8: Mental Representations and Behavior as Dictator

Notes: Distribution of behavior in the dictator game for each individual mental representation. Behavior in the dictator game is parameterized as the amount sent. Darker shades indicate a social mental representation.

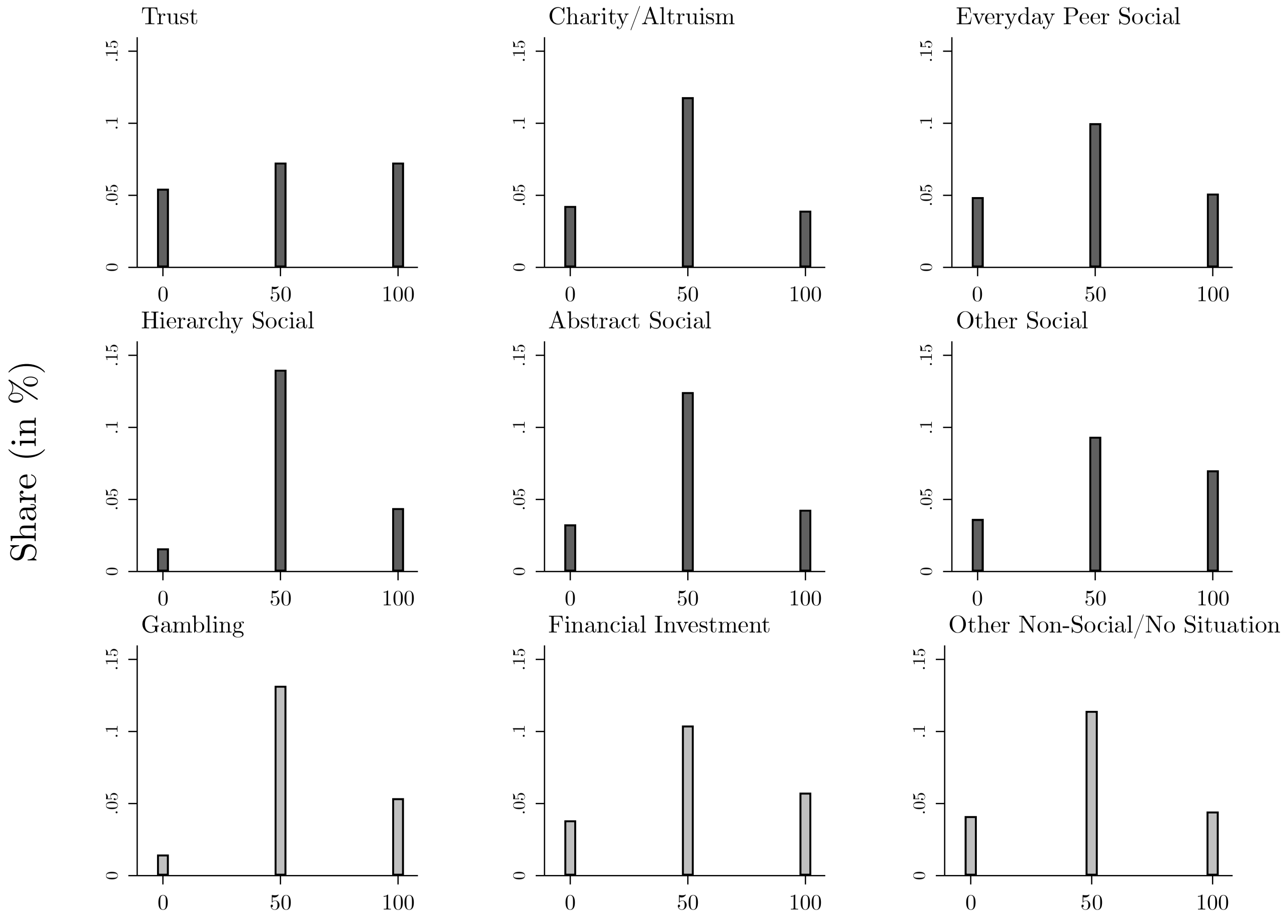


Figure B.9: Mental Representations and Behavior as Sender in Trust Game

Notes: Distribution of behavior as the sender in the trust game for each individual mental representation. Behavior as the sender in the trust game is parameterized as the amount sent. Darker shades indicate a social mental representation.

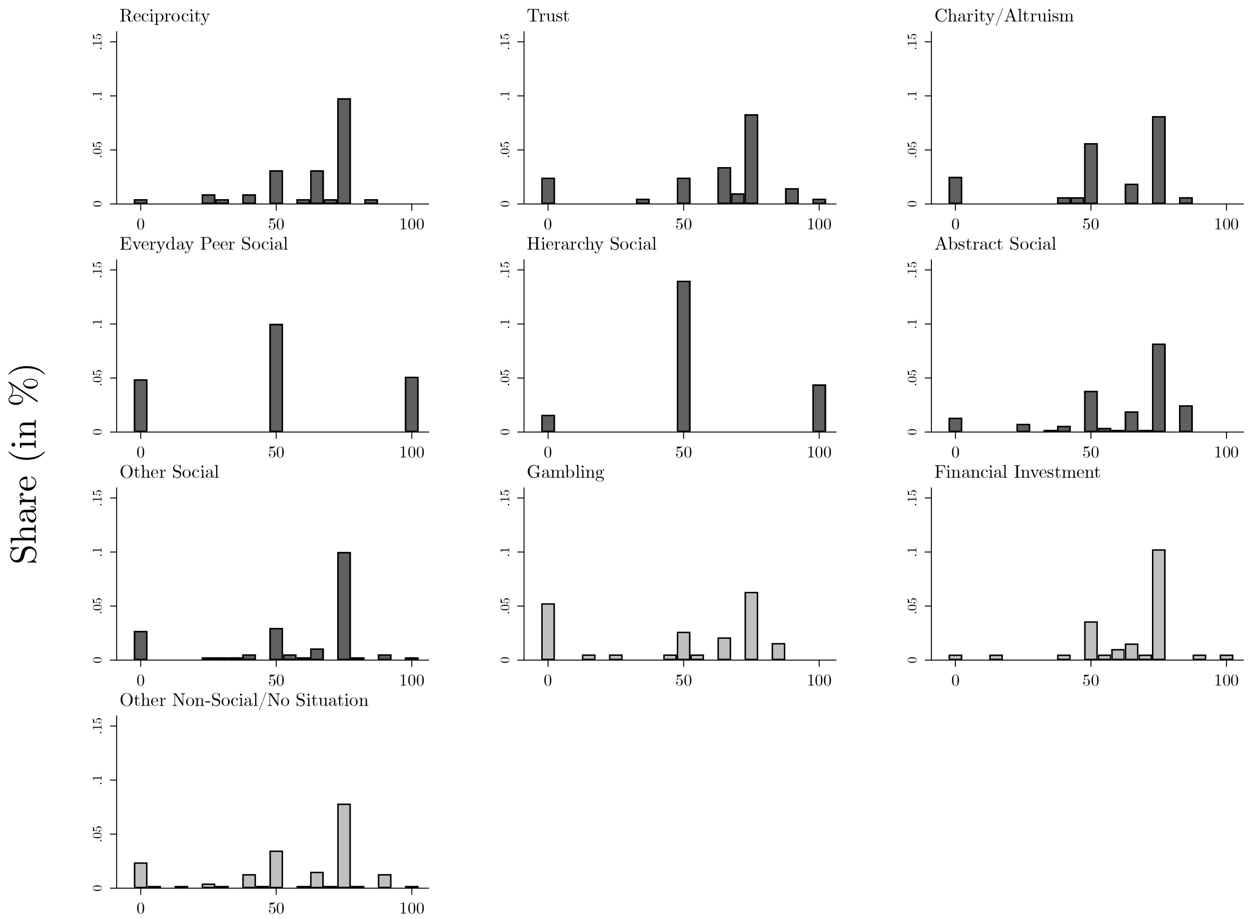


Figure B.10: Mental Representations and Behavior as Receiver in Trust Game

Notes: Distribution of behavior as the receiver in the trust game for each individual mental representation. Behavior as the receiver in the trust game is parameterized as the average share returned across all decisions. Subjects who return more than they receive are omitted for illustrative purposes. More details in Section 2.1. Darker shades indicate a social mental representation.

B.4 Interpreting the Effect of Framing

The literature on framing effects in economic games finds that community framing does not seem to affect behavior in the dictator game and in a sequentially played prisoner's dilemma (Ellingsen et al., 2012; Dreber et al., 2013). The authors argue that this is because framing acts as a coordination device and does not shift preferences themselves. Taking into account mental representations, however, gives rise to another hypothesis: a framing treatment should be understood as an *intention-to-treat* treatment, with some subjects not complying with the framing by not adapting their (non-social) mental representation. I will provide evidence in favor of this hypothesis.

First, consider Figure B.11 and Table B.1. Figure B.11 plots the distribution of behavior for all decision roles in the dictator and (sequentially played) trust game across the different framing treatments. Table B.1 provides accompanying parametric tests by regressing behavior in each decision role on an indicator for the framing treatment while controlling for variation in the game order. Dictator behavior is parameterized as the absolute amount sent to the other subject by the dictator. Sender behavior in the trust game is the absolute amount sent to the receiver in the trust game. Receiver behavior in the trust game is parameterized as the average number of points sent back relative to the points received (i.e., amount of points sent × multiplier), or putting it differently, the average share of the points received which are sent back across the two relevant information sets (sender sending 50 points and sender sending 100 points). This share can be more than 100% if subjects decide to send back even more than they received (receivers also have an additional endowment). The number of receivers who do so is negligible ($n = 4$). These subjects are omitted in Figure B.11 for illustrative purposes but are included in the parametric analyses. Figure B.11 and Table B.1 show that framing the respective game as a "community decision situation" instead of simply a "decision situation" does not affect behavior at all for the dictator and the receiver in the trust game. For the sender in the trust game, there seems to be a small increase in the likelihood of choosing to send 50 points, but this increase is not statistically significant. Controlling for the belief of the sender and analyzing whether beliefs are affected by the framing treatment does not yield any statistically significant findings either.[30]

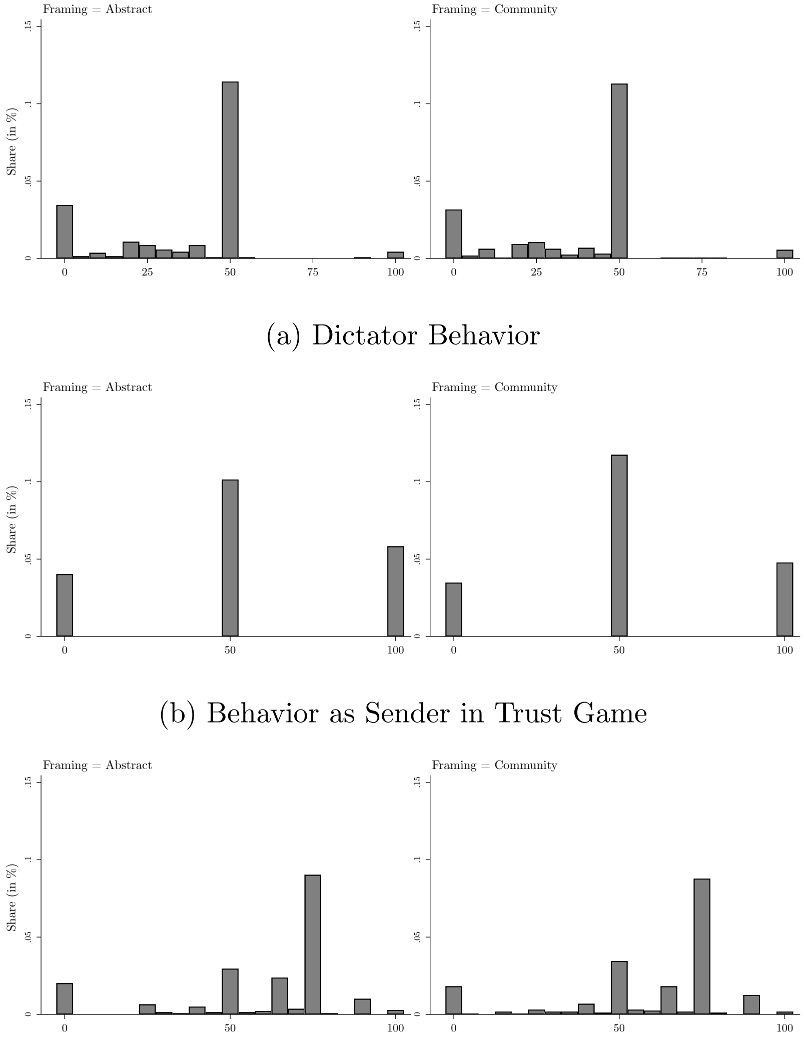
	Dictator Game	Trust Game		
	Amount Sent	Amount Sent		Share Returned
		0 vs 50	0 vs. 100	
Framing = Community Situation	0.818	0.274	-0.087	0.000
	(1.814)	(0.220)	(0.249)	(0.021)
Control for game order	Yes	Yes	Yes	Yes
(Pseudo-)R^2	0.00	0.01		0.00
Obs.	600	600		600

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Column (1) regresses the absolute amount sent in the dictator game on a set of treatment indicators. Columns (2) and (3) report the estimates from a multinomial logit of the choice options (sending 0, 50, or 100 points) as the sender in the trust game, using 0 points as the base outcome. Column (4) regresses the average share returned as the receiver in the trust game on the same set of treatment indicators. Robust standard errors are used for the sender in the dictator game and the receiver in the trust game.

Table B.1: Framing Effect on Behavior

Result 2 provides an explanation to rationalize the lack of framing effects on behavior. Clearly, framing treatments are meant to change the "frame" of subjects, or using the terminology of this paper, the mental representation of the game. Result 2 shows that this is not the case. Only for the sender in the trust game, the framing treatment increases the likelihood of having a social representation by less than 10 p.p., but as Figures 5 and 6 show, despite the community framing, a substantial share of subjects still have mental representations related to gambling, financial investment, or other non-social situations. Moreover, these effects are robust to controlling for attention, cognitive skills, and lack of understanding of the game: Figure B.12 replicates the parametric analyses from Result 2 while also controlling for general lack of attention, lack of understanding of the game, and cognitive skills. There is no statistically significant effect of framing on any of the

[30] Results available upon request.



(a) Dictator Behavior

(b) Behavior as Sender in Trust Game

(c) Behavior as Receiver in Trust Game

Figure B.11: Framing Effect on Game Behavior

Notes: Distribution of behavior in the dictator and trust game (sender and receiver), split by framing treatment status. Behavior in the dictator game and as the sender in the trust game is parameterized as the amount sent. Behavior as the receiver in the trust game is parameterized as the average share returned across all decisions (subjects who return more than they receive are omitted for illustrative purposes). More details in Section 2.1.

non-social categories and the overall increase in the likelihood of a social mental representation does not become stronger, i.e., subjects with a non-social mental representation in the community framing treatment condition are not just generally inattentive or have lower cognitive skills.

Summing up, accounting for mental representations gives rise to a new explanation as to why framing treatments might not affect behavior: they are simply not shifting mental representations sufficiently strongly. Framing treatments should therefore be understood as intention-to-treat treatments, with some subjects not complying with the treatment.

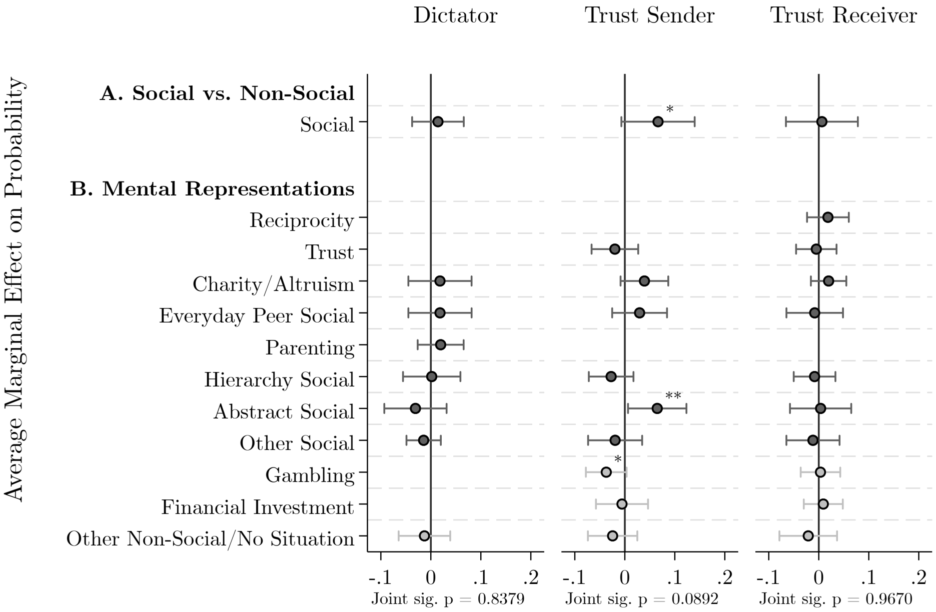


Figure B.12: Marginal Effect of Community Framing on Mental Representations Controlling for Game Understanding, Attention, and Cognitive Skills

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of community framing on the different associations, together with a 95% confidence interval. Panel A uses a probit model to regress an indicator for a social association (i.e., involving at least two individuals) on the treatment indicator. Panel B uses a multinomial logit model to regress the individual categories on the treatment indicator. Both models control for the treatment variation in game order and, additionally, the number of failed attention checks in the third part of the experiment, whether a subject answered at least one comprehension question for the respective game incorrectly, and results from a cognitive reflection test. Below each marginal effects plot, the p-value from an F-test of joint significance of all marginal effects in Panel B is reported (i.e., from a test whether all individual marginal effects are jointly zero).

C Robustness of Main Results

C.1 Treatment Interpretation and Assignment

Figure C.1 shows the results of a multinomial regression of the results of a cognitive reflection test (CRT) on indicators for treatment variation in the game order (playing the trust game first) and community framing. The CRT consists of two common questions to measure cognitive ability (e.g., Chapman et al., 2023): "In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how many days would it take for the patch to cover half of the lake?" and "A bat and a ball cost USD 1.10 in total. The bat costs USD 1.00 more than the ball. How much does the ball cost (in USD)?". The coefficients can be interpreted as the change in the probability of obtaining a particular score in the cognitive reflection test.

While playing the trust game first seems to induce a statistically significant effect on answering at least one CRT question correctly, the effect is small in magnitude. Moreover, a joint test of whether the effect on any of the three possible test outcomes is significant yields $p = 0.1501$. However, to ensure that estimates for playing the trust game first only reflect the effect of game order and do not pick up on potential differences in cognitive skills among subjects, below (cf. Figure C.2) I show that the effect of game order on mental representations does not change when one additionally controls for cognitive skills.

Furthermore, upon closer inspection of the data, playing the trust game first is weakly positively associated with being exposed to the community framing treatment ($p = 0.0826$ from a χ^2-test). The Spearman correlation coefficient is 0.0709 ($p = 0.0828$). Treatments were assigned independently of each other by the computer, so the small positive association can be either a statistical coincidence or driven by correlated sample selection across the two treatments. Sample selection can occur because subjects are screened

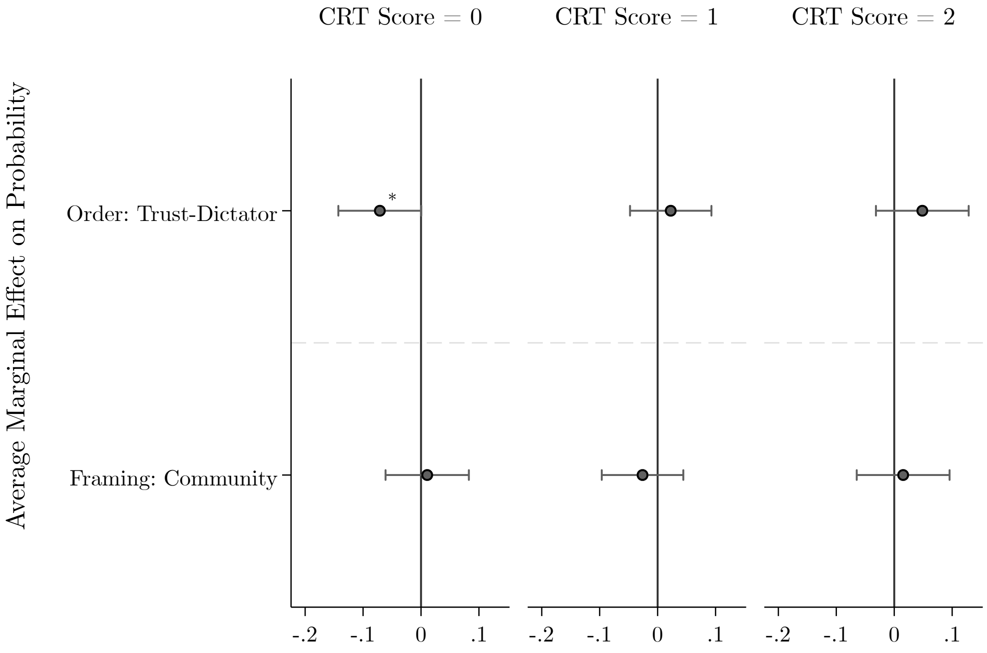


Figure C.1: Treatment Variation and Cognitive Ability

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of community framing and game order on the performance in a cognitive reflection test, together with a 95% confidence interval. The cognitive reflection test consists of two common questions to measure cognitive ability (e.g., Chapman et al., 2023): "In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how many days would it take for the patch to cover half of the lake?" and "A bat and a ball cost USD 1.10 in total. The bat costs USD 1.00 more than the ball. How much does the ball cost (in USD)?".

out of the study if they cannot correctly answer comprehension questions related to the games. This screening out is enforced by Prolific for the first game that subjects play. However, treatments need to be assigned before the comprehension questions. If the association between both treatments was not a statistical coincidence, both treatments should therefore be associated with higher cognitive skills. Figure C.1 shows that while the evidence is mixed for the game order treatment, the framing treatment is not associated with higher cognitive skills of subjects. I therefore interpret the weakly positive association between both treatments as a statistical coincidence. However, in Section 3 I always show parametric results that include treatment indicators for both treatments to control for the association between them.

C.2 Result 1: Heterogeneity in Mental Representations

I first provide the robustness analyses based on using the associations from everyday life as a measure for mental representations. Afterward, I show the results based on which sentences from the instructions subjects selected as influential.

Measuring Mental Representations with Associations

In the main analysis, I analyze the effect of game order on associations (i.e., mental representations) with a multinomial logit to facilitate easier interpretation of the coefficients. Results are robust to additionally controlling for performance in a cognitive reflection test (to ensure that the indicator for playing the trust game first only captures the effect of game order; cf. Figure C.2) and using a linear probability model to regress an indicator for game order on a set of indicator variables for each mental representations, while also controlling for the framing treatment and cognitive skills (cf. Table C.1).

Subjects need to pass comprehension and attention checks in order to par-

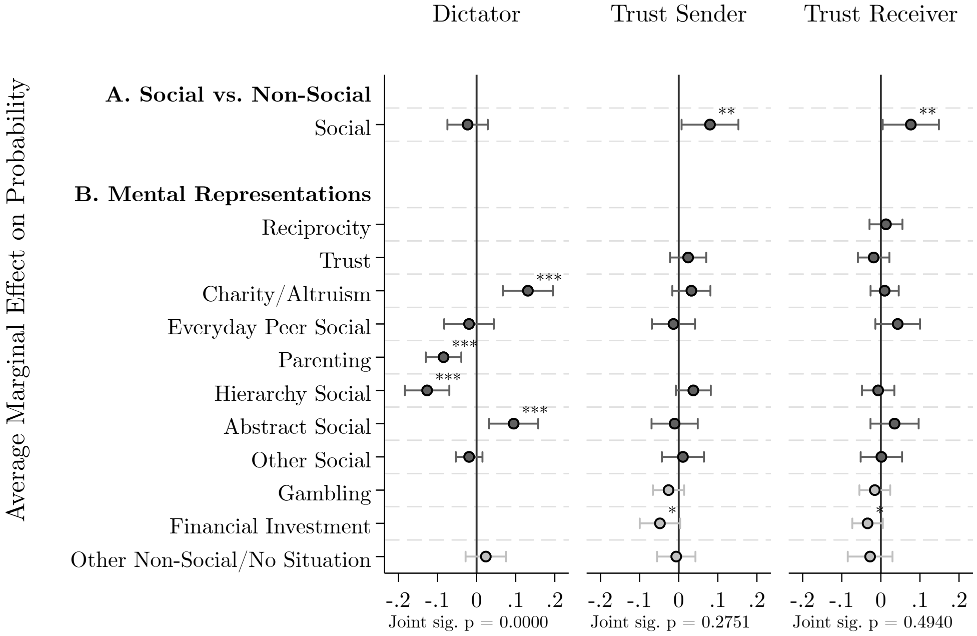


Figure C.2: Effect of Playing Trust Game First on Associations across Games and Controlling for Cognitive Skills

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of playing the trust game first on the different associations, together with a 95% confidence interval. Panel A uses a probit model to regress an indicator for a social association (i.e., involving at least two individuals) on the treatment indicator. Panel B uses a multinomial logit model to regress the individual categories on the treatment indicator. Both models control for the treatment variation in framing as well as cognitive skills by subjects, as measured through two standard cognitive reflection test questions. Below each marginal effects plot, the p-value from an F-test of joint significance of all marginal effects in Panel B is reported (i.e., from a test whether all individual marginal effects are jointly zero).

	Order: Dictator-Trust					
	Dictator		Trust Sender		Trust Receiver	
	(1)	(2)	(3)	(4)	(5)	(6)
Mental Representations						
- Reciprocity					-0.040	
					(0.094)	
- Trust			-0.045		0.069	
			(0.088)		(0.097)	
- Charity/Altruism	-0.258**		-0.059		-0.042	
	(0.103)		(0.086)		(0.106)	
- Everyday Peer Social	-0.071		0.045		-0.071	
	(0.104)		(0.080)		(0.080)	
- Parenting	0.129					
	(0.110)					
- Hierarchy Social	0.102		-0.091		0.026	
	(0.104)		(0.092)		(0.097)	
- Abstract Social	-0.220**		0.036		-0.047	
	(0.104)		(0.077)		(0.076)	
- Gambling			0.114		0.061	
			(0.094)		(0.098)	
- Financial Investment			0.118		0.133	
			(0.081)		(0.097)	
- Other Non-Social/No Situation	-0.142		0.036		0.046	
	(0.111)		(0.085)		(0.078)	
Aggregated Representations						
- Social		0.057		-0.096**		-0.094**
		(0.064)		(0.044)		(0.045)
Control for framing	Yes	Yes	Yes	Yes	Yes	Yes
Control for cognitive ability	Yes	Yes	Yes	Yes	Yes	Yes
(Joint) sig. associations (p-value)	0.0000	0.3734	0.2799	0.0296	0.5009	0.0352
R^2	0.09	0.01	0.03	0.02	0.02	0.02
Obs.	600	600	600	600	600	600

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Results from regressions of game order indicator on mental representations. Columns (1), (3), and (5) report results from a regression on indicators for the individual mental representations. "Other Social" is the reference category. Columns (2), (4), (6) report results from a regression on an indicator for having a social vs. non-social mental representation, with "Other Non-Social/No Situation" as the reference category. Explanation of individual categories in Table 1. "(Joint) sig. associations (p-value)" refers to the p-value from an F-test for (joint) significance of the indicator(s) for the respective mental representation(s).

Table C.1: Effect of Game Order on Mental Representations

ticipate in the experiment. However, it could be the case that this only filters out subjects without *any* understanding and/or attention. Consequently, mental representations could still be associated with a lack of understanding or attention. Non-parametric tests for both the social vs. non-social distinction and on the distribution of the individual mental representations show that this is not the case ($p > 0.1354$, χ^2-test). Parametric regressions controlling for treatment variation confirm this.[31]

Cognitive skills, however, are associated with different individual mental representations. In particular, cognitive skills seem to induce a greater emphasis on interactions involving hierarchy in the dictator game and a smaller emphasis on mental representations featuring "Charity/Altruism", "Trust", and "Reciprocity" — even when controlling for game understanding and attention. Moreover, cognitive skills are associated with a greater focus on "Other Non-Social/No Situation" for the receiver in the trust game.

Summing up, mental representations are not driven by a lack of game understanding or attention, i.e., by a lack of engagement with the games. However, mental representations do seem to be associated with cognitive skills, but cognitive skills probably also correlate with different experiences in life. This effect should therefore not be interpreted as inducing artificial variation in the mental representations, but, instead, be considered as another dimension through which socio-demographics can explain variation in mental representations.

Measuring Mental Representations with Influential Sentences

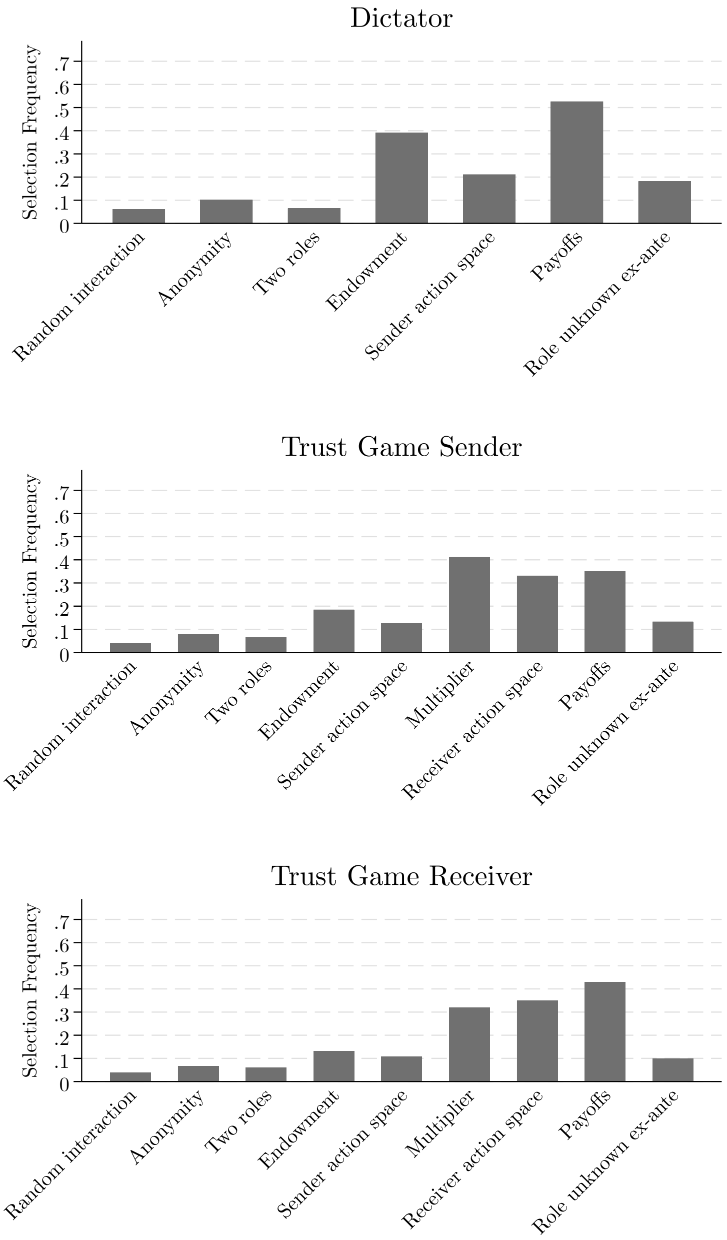
Figure C.3 plots the distribution of which sentences are selected by subjects as influencing how they think about the decision situation. Additionally, influence weights are indicated by subjects. All robustness analyses build, for now, on the unweighted selection. Table A.1 in Appendix A provides details on the sentences together with their labels used in Figure C.3.

[31] Results available upon request.

Subjects have to select at least one sentence but can select multiple sentences. Consequently, the selection frequency of a particular sentence is not necessarily independent of the selection frequency of another sentence, and the selection frequencies do not sum up to 100%. This means that it is not possible to directly compute a measure of dispersion from Figure C.3. Despite that, Figure C.3 highlights that — at least qualitatively — there is variation in the sentences that are selected, mirroring the heterogeneity in the associations, e.g., notice how the selection of the "endowment" sentence in the dictator game coincides with the frequency of the "Charity/Altruism" category for associations from everyday life (and the respective lower frequency of "endowment" and "Charity/Altruism" in the trust game).

Turning to the effect of exogenously varying the game order, Figures C.4 and C.5 provide the distribution of selected sentences across the different game orders. Similar to the analyses in the main part, game order affects which sentences are selected in all games, i.e., the effect of game order on mental representations is not an artifact of the open-ended nature of the elicitation method. While the same sentence can be selected for different reasons and interpretation is therefore difficult, the clear emphasis on sentences related to the (unequal) endowment and payoffs in the dictator game stands out. Similar to the emphasis on associations related to "Altruism/Charity", when the trust game is played before the dictator game, subjects select the sentence on the unequal endowment much more frequently in the dictator game. For the sender in the trust game, the emphasis on the sentence related to the (now equal) endowment is mirrored (and also to some extent for the receiver in the trust game): it is selected much more frequently when subjects are first exposed to the unequal endowment in the dictator game before experiencing the equal endowment in the trust game.

Table C.2 confirms these insights with parametric tests, i.e., the heterogeneity in associations and the effect of game order on them is mirrored by a highly statistically significant effect of game order on which sentences are



Dictator

Trust Game Sender

Trust Game Receiver

Figure C.3: Instructions Selected in Economic Games

Notes: Distribution of which sentences are selected as influential in how subjects think about the game. Subjects can select multiple sentences. Table A.1 provides more details on the content of the sentences.

selected.

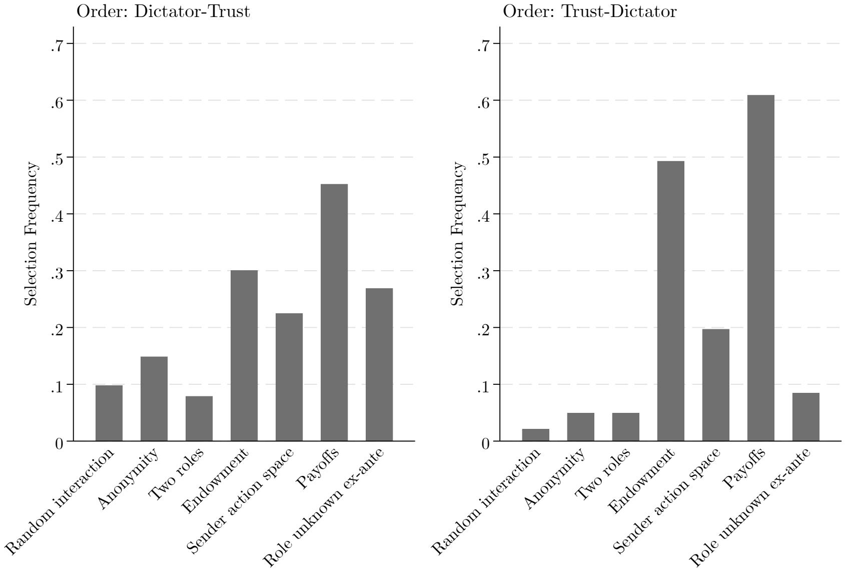
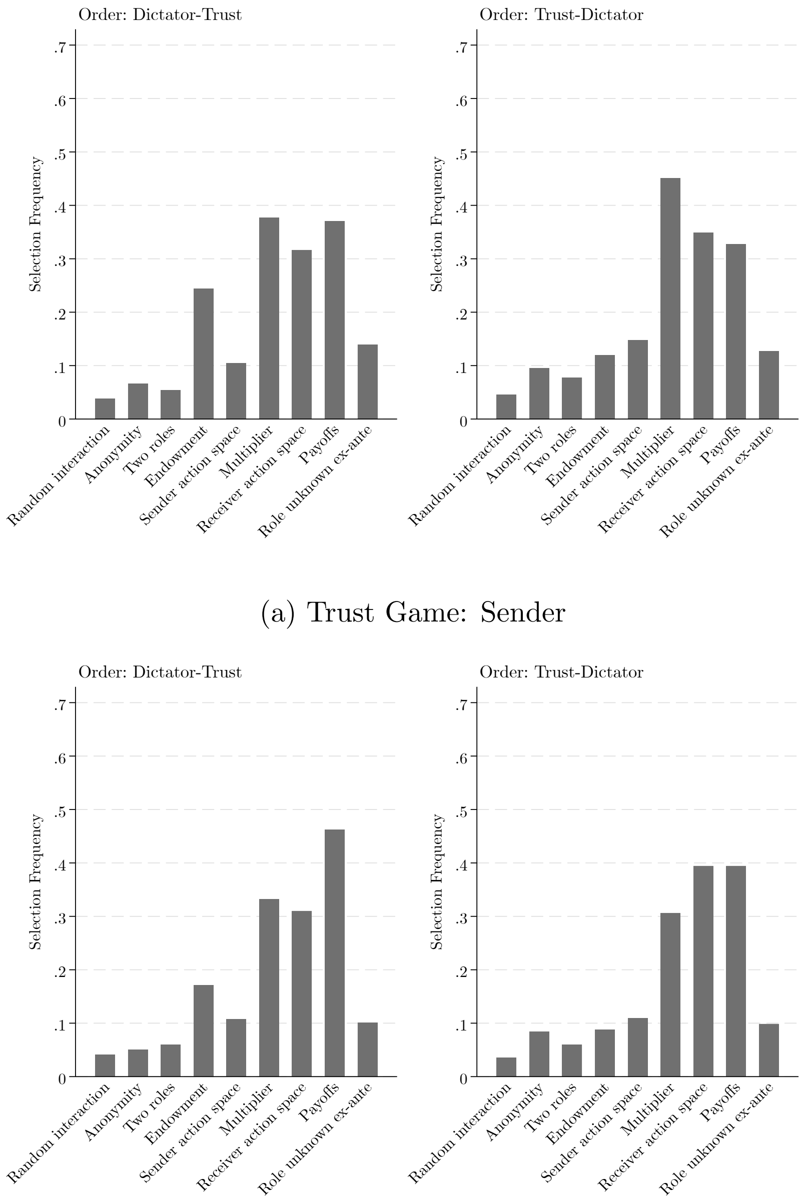


Figure C.4: Effect of Game Order on Sentences in Dictator Game

Notes: Distribution of which sentences are selected as influential in how subjects think about the game, split by game order. Subjects can select multiple sentences. Table A.1 provides more details on the content of the sentences.



(a) Trust Game: Sender

(b) Trust Game: Receiver

Figure C.5: Effect of Game Order on Sentences in Trust Game

Notes: Distribution of which sentences are selected as influential in how subjects think about the game, split by game order. Subjects can select multiple sentences. Table A.1 provides more details on the content of the sentences.

	Order: Trust-Dictator		
	Dictator	Trust Sender	Trust Receiver
	(1)	(2)	(3)
Selected Sentences			
- Anonymity	-0.190***	0.116	0.145*
	(0.059)	(0.075)	(0.080)
- Two roles	-0.051	0.136	0.021
	(0.074)	(0.087)	(0.089)
- Endowment	0.190***	-0.236***	-0.184***
	(0.041)	(0.049)	(0.058)
- Sender action space	-0.064	0.127**	-0.007
	(0.047)	(0.062)	(0.069)
- Multiplier		0.069	-0.032
		(0.042)	(0.046)
- Receiver action space		0.027	0.090**
		(0.043)	(0.045)
- Payoffs	0.132***	-0.035	-0.066
	(0.040)	(0.044)	(0.044)
- Role unknown ex-ante	-0.218***	-0.017	0.005
	(0.048)	(0.062)	(0.069)
Control for framing	Yes	Yes	Yes
Control for cognitive ability	Yes	Yes	Yes
Joint sig. sentences (p-value)	0.0000	0.0000	0.0078
R^2	0.13	0.06	0.04
Obs.	600	600	600

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Results from regressions of game order indicator on which sentences are selected as influential in how subjects think about the game. Columns (1), (2), and (3) report results from a regression on indicators for the individual sentences. "Random interaction" is the reference category. Table A.1 provides more details on the content of the sentences. "Joint sig. sentences (p-value)" refers to the p-value from an F-test for joint significance of the indicators for the sentences.

Table C.2: Effect of Game Order on Selected Sentences

C.3 Result 2: Drivers of Mental Representations

Again, I first provide the robustness analyses based on using the associations from everyday life as a measure for mental representations.

Measuring Mental Representations with Associations

Result 2 builds on parametric analyses that use a multinomial logit model to estimate the influence of the framing treatment, being above median age, being white, and being female on the mental representations in each game. Tables C.3, C.4, C.5, and C.6 confirm these findings, using linear probability models by regressing an indicator for framing and socio-demographic information (above median age, white, and female) on a set of indicators for the mental representations. In particular, notice how community framing again only affects mental representations of the sender in the trust game.

	Framing = Community Situation					
	Dictator		Trust Sender		Trust Receiver	
	(1)	(2)	(3)	(4)	(5)	(6)
Mental Representations						
- Reciprocity					0.076	
					(0.094)	
- Trust			-0.019		-0.003	
			(0.088)		(0.098)	
- Charity/Altruism	0.102		0.128		0.101	
	(0.105)		(0.084)		(0.105)	
- Everyday Peer Social	0.102		0.082		-0.002	
	(0.105)		(0.079)		(0.080)	
- Parenting	0.133					
	(0.115)					
- Hierarchy Social	0.091		-0.052		-0.006	
	(0.107)		(0.092)		(0.097)	
- Abstract Social	0.040		0.129^{*}		0.024	
	(0.106)		(0.075)		(0.076)	
- Gambling			-0.108		0.023	
			(0.096)		(0.099)	
- Financial Investment			0.022		0.052	
			(0.082)		(0.099)	
- Other Non-Social/No Situation	0.056		-0.027		-0.015	
	(0.111)		(0.085)		(0.078)	
Social vs. Non-Social						
- Social		0.029		0.082^{*}		0.011
		(0.063)		(0.045)		(0.045)
Control for game order	Yes	Yes	Yes	Yes	Yes	Yes
(Joint) sig. associations (p-value)	0.8431	0.6498	0.0944	0.0693	0.9788	0.8050
R^2	0.01	0.01	0.03	0.01	0.01	0.01
Obs.	600	600	600	600	600	600

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Results from regressions of framing indicator on mental representations. Columns (1), (3), and (5) report results from a regression on indicators for the individual mental representations. "Other Social" is the reference category. Columns (2), (4), (6) report results from a regression on an indicator for having a social vs. non-social mental representation, with "Other Non-Social/No Situation" as the reference category. Explanation of individual categories in Table 1. "(Joint) sig. associations (p-value)" refers to the p-value from an F-test for (joint) significance of the indicator(s) for the respective mental representation(s).

Table C.3: Effect of Framing on Mental Representations

	Age = Above Median Age					
	Dictator		Trust Sender		Trust Receiver	
	(1)	(2)	(3)	(4)	(5)	(6)
Mental Representations						
- Reciprocity					0.002 (0.094)	
- Trust			0.040 (0.088)		-0.118 (0.096)	
- Charity/Altruism	0.002 (0.105)		0.238^{***} (0.084)		-0.041 (0.108)	
- Everyday Peer Social	0.045 (0.104)		0.155^{*} (0.079)		-0.028 (0.079)	
- Parenting	0.184 (0.113)					
- Hierarchy Social	-0.123 (0.105)		-0.073 (0.088)		0.004 (0.096)	
- Abstract Social	0.161 (0.104)		0.173^{**} (0.076)		0.097 (0.076)	
- Gambling			0.072 (0.098)		-0.099 (0.099)	
- Financial Investment			0.007 (0.082)		-0.102 (0.099)	
- Other Non-Social/No Situation	0.082 (0.111)		0.013 (0.085)		-0.019 (0.078)	
Social vs. Non-Social						
- Social		-0.040 (0.064)		0.076^{*} (0.045)		0.060 (0.045)
Control for game order	Yes	Yes	Yes	Yes	Yes	Yes
Control for framing	Yes	Yes	Yes	Yes	Yes	Yes
(Joint) sig. associations (p-value)	0.0006	0.5292	0.0063	0.0916	0.3669	0.1838
R^2	0.04	0.00	0.04	0.01	0.02	0.01
Obs.	600	600	600	600	600	600

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Results from regressions of above-median-age indicator on mental representations. Columns (1), (3), and (5) report results from a regression on indicators for the individual mental representations. “Other Social” is the reference category. Columns (2), (4), (6) report results from a regression on an indicator for having a social vs. non-social mental representation, with “Other Non-Social/No Situation” as the reference category. Explanation of individual categories in Table 1. “(Joint) sig. associations (p-value)” refers to the p-value from an F-test for (joint) significance of the indicator(s) for the respective mental representation(s).

Table C.4: Mental Representations and Age

	Ethnicity = White					
	Dictator		Trust Sender		Trust Receiver	
	(1)	(2)	(3)	(4)	(5)	(6)
Mental Representations						
- Reciprocity					0.009 (0.081)	
- Trust			-0.016 (0.068)		0.086 (0.078)	
- Charity/Altruism	-0.118 (0.077)		0.005 (0.064)		0.098 (0.081)	
- Everyday Peer Social	-0.138* (0.078)		-0.034 (0.063)		0.057 (0.065)	
- Parenting	-0.081 (0.086)					
- Hierarchy Social	-0.018 (0.076)		-0.077 (0.075)		-0.070 (0.088)	
- Abstract Social	-0.083 (0.077)		-0.122* (0.063)		0.073 (0.063)	
- Gambling			-0.047 (0.079)		0.001 (0.087)	
- Financial Investment			-0.085 (0.067)		-0.072 (0.093)	
- Other Non-Social/No Situation	-0.113 (0.084)		-0.068 (0.070)		0.050 (0.066)	
Social vs. Non-Social						
- Social		0.026 (0.055)		0.025 (0.039)		0.028 (0.039)
Control for game order	Yes	Yes	Yes	Yes	Yes	Yes
Control for framing	Yes	Yes	Yes	Yes	Yes	Yes
(Joint) sig. associations (p-value)	0.3278	0.6387	0.6051	0.5190	0.5355	0.4788
R^2	0.02	0.01	0.02	0.01	0.02	0.01
Obs.	600	600	600	600	600	600

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Results from regressions of being white on mental representations. Columns (1), (3), and (5) report results from a regression on indicators for the individual mental representations. "Other Social" is the reference category. Columns (2), (4), (6) report results from a regression on an indicator for having a social vs. non-social mental representation, with "Other Non-Social/No Situation" as the reference category. Explanation of individual categories in Table 1. "(Joint) sig. associations (p-value)" refers to the p-value from an F-test for (joint) significance of the indicator(s) for the respective mental representation(s).

Table C.5: Mental Representations and Ethnicity

			Sex = Female			
	Dictator		Trust Sender		Trust Receiver	
	(1)	(2)	(3)	(4)	(5)	(6)
Mental Representations						
- Reciprocity					0.012	
					(0.096)	
- Trust			0.074		-0.008	
			(0.088)		(0.097)	
- Charity/Altruism	-0.061		0.096		-0.033	
	(0.102)		(0.085)		(0.105)	
- Everyday Peer Social	-0.063		0.100		-0.002	
	(0.102)		(0.079)		(0.079)	
- Parenting	-0.092					
	(0.114)					
- Hierarchy Social	0.009		0.085		0.064	
	(0.105)		(0.090)		(0.095)	
- Abstract Social	0.147		0.237***		0.138*	
	(0.102)		(0.075)		(0.074)	
- Gambling			0.139		-0.114	
			(0.097)		(0.097)	
- Financial Investment			0.030		0.034	
			(0.081)		(0.099)	
- Other Non-Social/No Situation	-0.031		0.170**		-0.027	
	(0.109)		(0.085)		(0.078)	
Social vs. Non-Social						
- Social		0.027		0.003		0.071
		(0.063)		(0.045)		(0.045)
Control for game order	Yes	Yes	Yes	Yes	Yes	Yes
Control for framing	Yes	Yes	Yes	Yes	Yes	Yes
(Joint) sig. associations (p-value)	0.0119	0.6749	0.0826	0.9529	0.2552	0.1150
R^2	0.04	0.01	0.03	0.01	0.03	0.02
Obs.	600	600	600	600	600	600

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Results from regressions of being female on mental representations. Columns (1), (3), and (5) report results from a regression on indicators for the individual mental representations. "Other Social" is the reference category. Columns (2), (4), (6) report results from a regression on an indicator for having a social vs. non-social mental representation, with "Other Non-Social/No Situation" as the reference category. Explanation of individual categories in Table 1. "(Joint) sig. associations (p-value)" refers to the p-value from an F-test for (joint) significance of the indicator(s) for the respective mental representation(s).

Table C.6: Mental Representations and Sex

Measuring Mental Representations with Influential Sentences

Turning to the sentences that are selected as influential by subjects, Figures C.6 and C.7 plot the distribution of mental representations across the two framing treatment conditions. Similar to the (lack of an) effect of community framing on mental representations when measured through associations from everyday life, there is almost no change in which sentences are selected in the dictator game and for the receiver in the trust game. However, notice that, again, for the sender in the trust game, community framing seems to influence which sentences are selected. Table C.7 confirms this qualitatively, but the effect on sentences in the trust game is not jointly statistically significant ($p = 0.1516$).

Turning to how the selected sentences depend on age, ethnicity, and sex, Tables C.8, C.9, and C.10 provide parametric analyses for how these sociodemographics correlate with age, ethnicity, and sex. Notice that the results are again very similar to the analyses of how age, ethnicity, and sex are related to associations from everyday life. In particular, being above the median age influences which sentences are selected most frequently.

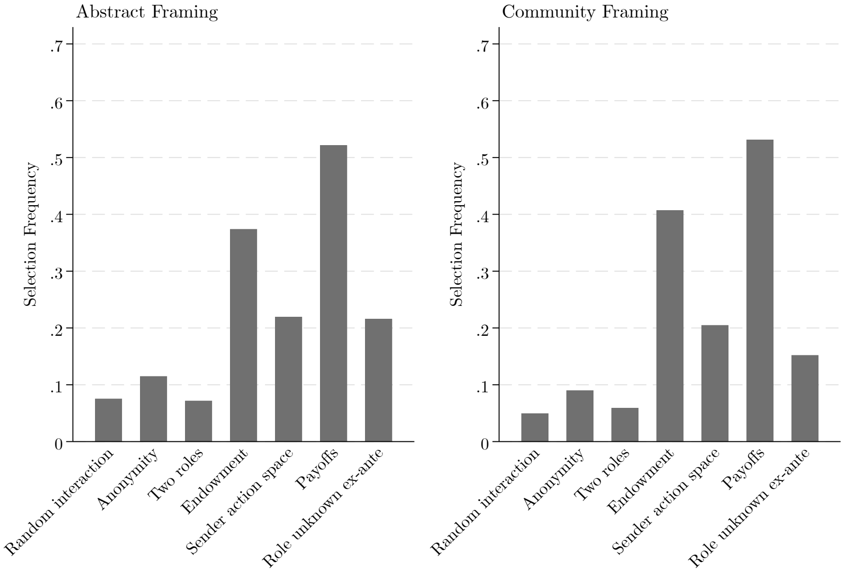
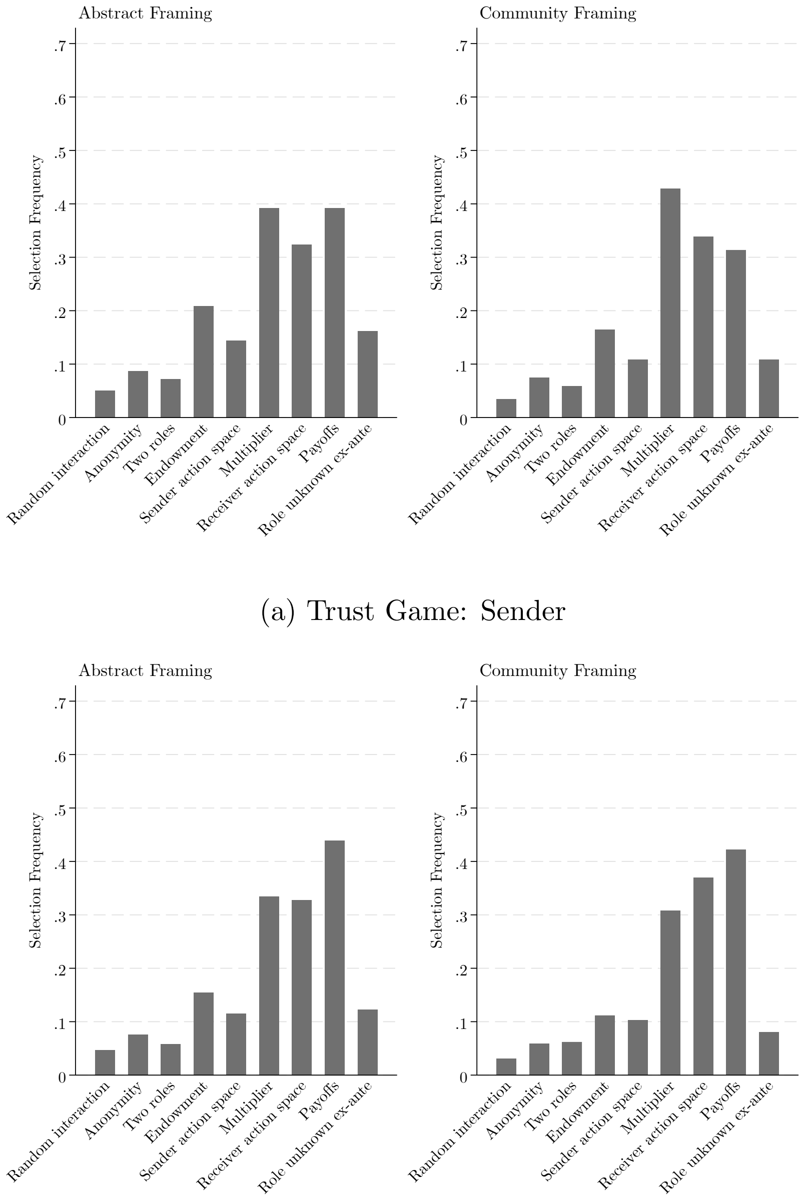


Figure C.6: Effect of Framing on Sentences in Dictator Game

Notes: Distribution of which sentences are selected as influential in how subjects think about the game, split by framing treatment. Subjects can select multiple sentences. Table A.1 provides more details on the content of the sentences.



(a) Trust Game: Sender

(b) Trust Game: Receiver

Figure C.7: Effect of Framing on Sentences in Trust Game

Notes: Distribution of which sentences are selected as influential in how subjects think about the game, split by framing treatment. Subjects can select multiple sentences. Table A.1 provides more details on the content of the sentences.

	Framing = Community Situation		
	Dictator	Trust Sender	Trust Receiver
	(1)	(2)	(3)
Selected Sentences			
- Anonymity	-0.042	-0.016	-0.067
	(0.067)	(0.073)	(0.082)
- Two roles	-0.029	-0.018	0.073
	(0.083)	(0.080)	(0.082)
- Endowment	0.011	-0.046	-0.079
	(0.044)	(0.055)	(0.062)
- Sender action space	-0.028	-0.075	-0.026
	(0.050)	(0.062)	(0.066)
- Multiplier		0.011	-0.029
		(0.043)	(0.046)
- Receiver action space		0.014	0.036
		(0.044)	(0.045)
- Payoffs	-0.016	-0.080*	-0.019
	(0.043)	(0.043)	(0.043)
- Role unknown ex-ante	-0.090	-0.104*	-0.119*
	(0.057)	(0.061)	(0.068)
Control for game order	Yes	Yes	Yes
Joint sig. sentences (p-value)	0.6862	0.1516	0.4405
R^2	0.01	0.02	0.02
Obs.	600	600	600

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Results from regressions of framing treatment indicator on which sentences are selected as influential in how subjects think about the game. Columns (1), (2), and (3) report results from a regression on indicators for the individual sentences. "Random interaction" is the reference category. Table A.1 provides more details on the content of the sentences. "Joint sig. sentences (p-value)" refers to the p-value from an F-test for joint significance of the indicators for the sentences.

Table C.7: Effect of Framing on Selected Sentences

	Age = Above Median Age		
	Dictator	Trust Sender	Trust Receiver
	(1)	(2)	(3)
Selected Sentences			
- Anonymity	0.035	-0.053	-0.052
	(0.069)	(0.073)	(0.084)
- Two roles	0.101	0.086	0.147*
	(0.081)	(0.082)	(0.087)
- Endowment	-0.042	-0.104*	0.003
	(0.044)	(0.054)	(0.062)
- Sender action space	-0.019	0.128**	-0.067
	(0.050)	(0.062)	(0.069)
- Multiplier		-0.111***	-0.016
		(0.043)	(0.046)
- Receiver action space		-0.082*	-0.092**
		(0.043)	(0.045)
- Payoffs	-0.135***	-0.137***	-0.089**
	(0.043)	(0.043)	(0.044)
- Role unknown ex-ante	-0.098*	-0.101*	-0.159**
	(0.055)	(0.058)	(0.065)
Control for game order	Yes	Yes	Yes
Control for framing	Yes	Yes	Yes
Joint sig. sentences (p-value)	0.0356	0.0001	0.0700
R^2	0.02	0.05	0.02
Obs.	600	600	600

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Results from regressions of above-median-age indicator on which sentences are selected as influential in how subjects think about the game. Columns (1), (2), and (3) report results from a regression on indicators for the individual sentences. “Random interaction” is the reference category. Table A.1 provides more details on the content of the sentences. “Joint sig. sentences (p-value)” refers to the p-value from an F-test for joint significance of the indicators for the sentences.

Table C.8: Selected Instructions and Age

	Ethnicity = White		
	Dictator	Trust Sender	Trust Receiver
	(1)	(2)	(3)
Selected Sentences			
- Anonymity	-0.024	0.036	0.117^{*}
	(0.059)	(0.062)	(0.062)
- Two roles	-0.008	-0.137^{*}	-0.135
	(0.072)	(0.081)	(0.083)
- Endowment	0.101^{***}	-0.062	-0.006
	(0.036)	(0.048)	(0.053)
- Sender action space	0.008	-0.077	-0.167^{***}
	(0.041)	(0.057)	(0.065)
- Multiplier		0.032	0.029
		(0.036)	(0.039)
- Receiver action space		0.027	-0.001
		(0.036)	(0.038)
- Payoffs	-0.001	0.019	0.021
	(0.036)	(0.037)	(0.038)
- Role unknown ex-ante	-0.000	0.029	-0.037
	(0.050)	(0.050)	(0.061)
Control for game order	Yes	Yes	Yes
Control for framing	Yes	Yes	Yes
Joint sig. sentences (p-value)	0.1587	0.3383	0.1068
R^2	0.02	0.02	0.03
Obs.	600	600	600

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Results from regressions of being white on which sentences are selected as influential in how subjects think about the game. Columns (1), (2), and (3) report results from a regression on indicators for the individual sentences. "Random interaction" is the reference category. Table A.1 provides more details on the content of the sentences. "Joint sig. sentences (p-value)" refers to the p-value from an F-test for joint significance of the indicators for the sentences.

Table C.9: Selected Instructions and Ethnicity

	Sex = Female		
	Dictator	Trust Sender	Trust Receiver
	(1)	(2)	(3)
Selected Sentences			
- Anonymity	-0.016	-0.023	-0.054
	(0.068)	(0.077)	(0.081)
- Two roles	-0.140*	-0.057	-0.057
	(0.082)	(0.085)	(0.090)
- Endowment	0.079*	0.097*	0.080
	(0.044)	(0.053)	(0.062)
- Sender action space	0.017	-0.019	-0.165**
	(0.050)	(0.063)	(0.065)
- Multiplier		0.004	-0.022
		(0.043)	(0.046)
- Receiver action space		-0.031	0.045
		(0.044)	(0.045)
- Payoffs	-0.063	-0.047	-0.001
	(0.043)	(0.045)	(0.044)
- Role unknown ex-ante	-0.069	0.051	0.027
	(0.056)	(0.059)	(0.069)
Control for game order	Yes	Yes	Yes
Control for framing	Yes	Yes	Yes
Joint sig. sentences (p-value)	0.0678	0.5769	0.1992
R^2	0.03	0.02	0.03
Obs.	600	600	600

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Results from regressions of being female on which sentences are selected as influential in how subjects think about the game. Columns (1), (2), and (3) report results from a regression on indicators for the individual sentences. “Random interaction” is the reference category. Table A.1 provides more details on the content of the sentences. “Joint sig. sentences (p-value)” refers to the p-value from an F-test for joint significance of the indicators for the sentences.

Table C.10: Selected Instructions and Sex

C.4 Result 3: Relevance of Mental Representations for Economic Research

Measuring Mental Representations with Associations

Consider that the results on how mental representations allow to uncover heterogeneity in the relationship between game and field behavior could also be spurious. I therefore exploit the exogenous variation in game order — which does affect mental representations for all games contrary to the framing treatment — and interact game behavior with an indicator for playing the trust game first, while controlling for variation in the framing treatment. Except for the relationship between SRI and receiver behavior in the trust game, all estimations are based on linear regressions with robust standard errors. A probit model is used for the relationship between SRI and trust game receiver behavior.

Figures C.8 and C.9 plot the results from these analyses. They first plot the estimate for the behavior in the respective game and, afterward, the linear combination of the behavior and the interaction terms of behavior and playing the trust game first (i.e., the sum of both coefficients). Consequently, the coefficient for the behavior itself can be interpreted as the estimate among subjects who play the *dictator* game first, while the estimate for the linear combination is the effect among the subjects who play the *trust* game first. For interpreting the results, it is important to bear three things in mind. First, playing the *trust* game first induces, on average, more social mental representations in the trust game, while inducing more mental representations related to "Charity/Altruism" and "Abstract Social" at the expense of fewer mental representations related to "Parenting" and "Hierarchy Social" in the dictator game. Second, these effects, while statistically significant, are not large. Third, since both treatments are weakly correlated, playing the *trust* game first will also pick up some of the effects of community framing. Taken together, statistical power will be small and the estimate for the interaction term (and thus the linear combination of behavior itself and the interaction term) will reflect the correlation between game and field behavior averaged across the mental representations which are (weakly) induced by playing the trust game first (and being exposed to community framing). Again, the marginal effects on SRA and GSS are in units of standard deviations, while the estimates for lending, finance, and SRI are the marginal effects on the probability that subjects lend money/possessions to friends more than once per year, pursue a career in finance, and pay special considerations to sustainability when investing. For Atmosfair, the estimates are the marginal effects on the share of the hypothetical donation to Atmosfair. The analyses for SRI and pursuing a career in finance are based on a subsample with $n = 451$ (SRI; excluding subjects who do not want to invest their money at all) and $n = 540$ (finance; excluding subjects who are permanently unemployed or are not working because they take care of their home or family or indicate "other" as occupation).

Figures C.8 and C.9 highlight that there is some evidence of statistically significant heterogeneity in the relationship between game and field behavior depending on the game order. Moreover, this heterogeneity is generally in line with taking the parametric results on the relationship between game and field behavior in each mental representation (cf. Figures 13, 14, and 15) and averaging them across the mental representations which are induced by playing the trust game first. For example, while not statistically significant and small in magnitude, the relationship between dictator game giving and SRA seems to be stronger among subjects who play the trust game first, i.e., for whom the "Altruism/Charity" and "Abstract Social" mental representations are weakly induced (remember that this in terms of standard deviations of SRA). For the sender in the trust game, game behavior predicts the answer to the GSS trust question more strongly for subjects who play the trust game first, i.e., for whom a social mental representation is induced (also notice that the effect is stronger for sending everything to the receiver which again

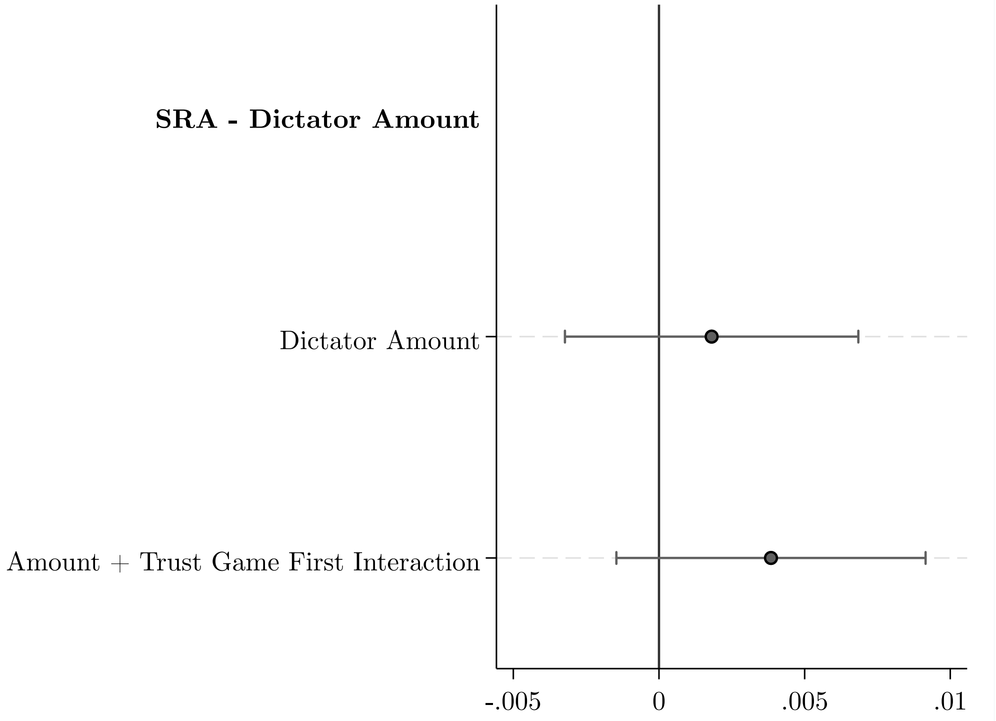
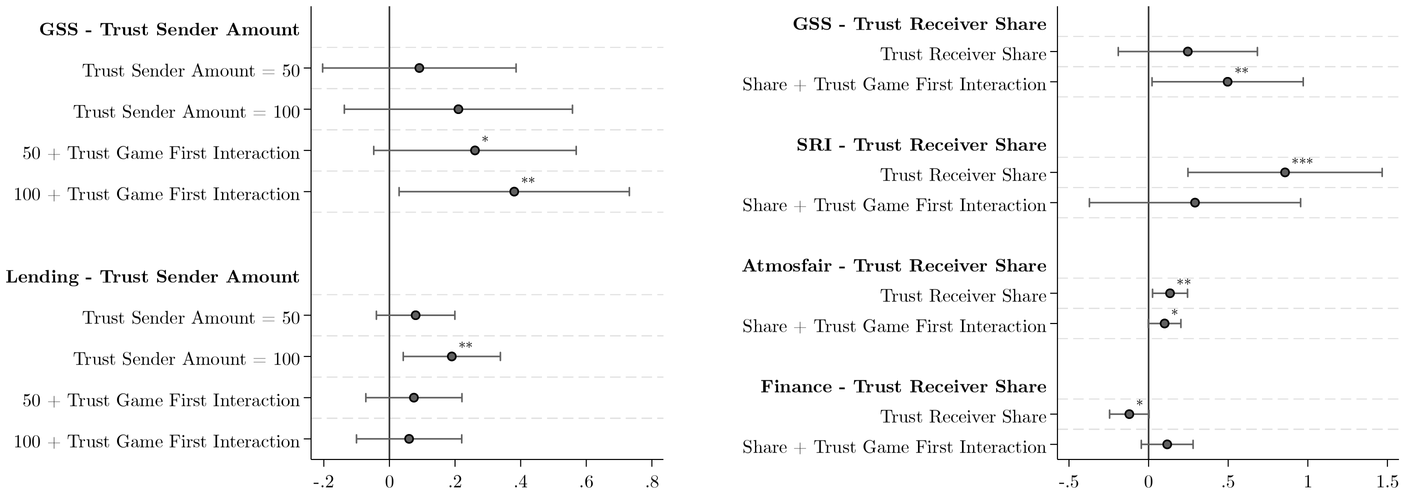


Figure C.8: Treatment Variation: Relationship between Dictator Giving and Field Behavior

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of behavior as the dictator on SRA, interacted with an indicator for playing the trust game first, in units of standard deviations. The second (first) row reports the effect among subjects who did (not) play the trust game first. See Table A.3 for more details on the elicitation of SRA.



(a) Sender in the Trust Game

(b) Receiver in the Trust Game

Figure C.9: Treatment Variation: Relationship between Behavior in Trust Gamer and Field Behavior

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of behavior in the trust game on the respective field behavior (GSS, lending, SRI, Atmosfair, finance), interacted with an indicator for playing the trust game first. For GSS, the effects are in units of standard deviations. For the sender in the trust game, the third and fourth (first and second) rows report the effect among subjects who did (not) play the trust game first. For the receiver in the trust game, the second (first) row reports the effect among subjects who did (not) play the trust game first. See Table A.3 for more details on the elicitation of each field behavior.

reflects the results from Figure 14). The opposite is true for self-reported lending behavior, whose relationship with sender behavior in the trust game is driven by subjects with a non-social mental representation (cf. Figure 14). Finally, the results for the receiver in the trust game are not as clear, which reflects that, as Figure 15 shows, both social and non-social mental representations seem to drive the relationship between game and field behavior — except for pursuing a career or working in the finance sector where two different social mental representation point to relationships into opposite directions.

Summing up, while more work is clearly needed to establish that mental representations causally affect the relationship between game and field behavior, the evidence presented so far indicates that the results from Figures 13, 14, and 15 are not (entirely) spurious.

Measuring Mental Representations with Influential Sentences

Consider that mental representations when measured based on associations from everyday life are correlated with game behavior. Table C.11 provides parametric regressions to show that this also holds when using the selected sentences from the instruction text. Table C.11 uses a linear regression for behavior as the dictator and receiver in the trust game (with robust standard errors) and a multinomial logit for the amount sent as the sender in the trust game and regresses the respective behavior on a set of indicators for each sentence. As the individual estimates and the tests for joint significance show, selecting different sentences as influential in how subjects think about the game is associated with significantly different behavior.

As preregistered, I do not use the sentence-based measure to identify heterogeneity in the relationship between game and field behavior. This choice is based on the notion that different subjects select the same sentence for different reasons (as debriefing interviews in a pilot showed). Therefore, the sentence-based measure is even less precise and more difficult to interpret than using the association-based measure.

	Dictator Game	Trust Game		
	Amount Sent	Amount Sent		Share Returned
		0 vs 50	0 vs. 100	
Selected Sentences				
- Anonymity	1.024	-0.061	0.556	-0.022
	(3.073)	(0.440)	(0.478)	(0.044)
- Two roles	1.683	1.972***	1.666**	-0.041
	(3.936)	(0.757)	(0.825)	(0.045)
- Endowment	5.519***	-0.899***	-1.408***	0.069**
	(1.923)	(0.284)	(0.373)	(0.027)
- Sender action space	-0.592	0.039	-0.850*	-0.038
	(2.138)	(0.354)	(0.463)	(0.035)
- Multiplier		2.117***	2.735***	0.078***
		(0.332)	(0.363)	(0.022)
- Receiver action space		-0.436*	-0.296	0.022
		(0.252)	(0.291)	(0.023)
- Payoffs	-1.918	0.347	0.724**	0.052**
	(1.863)	(0.258)	(0.297)	(0.022)
- Role unknown ex-ante	-0.280	0.568	0.917**	-0.002
	(2.612)	(0.380)	(0.418)	(0.037)
Control for game order	Yes	Yes	Yes	Yes
Control for framing	Yes	Yes	Yes	Yes
Joint sig. sentences (p-value)	0.0623	0.0000		0.0015
(Pseudo-)R^2	0.02	0.10		0.04
Obs.	600	600		600

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Column (1) regresses the absolute amount sent in the dictator game on a set of indicators for each sentence and treatment indicators. Columns (2) and (3) report the estimates from a multinomial logit of the choice options (sending 0, 50, or 100 points) as the sender in the trust game, using 0 points as the base outcome. Column (4) regresses the average share returned as the receiver in the trust game on the same set of treatment indicators. Robust standard errors are used for the sender in the dictator game and the receiver in the trust game. The same reference category is used across all regressions: selecting the sentence on "Random interaction".

Table C.11: Instructions Selected and Behavior

D Preregistration

The preregistration is available at Detemple (2023) (currently still under embargo and to be treated confidentially). The following two questions are preregistered: (1) Is there variation in the mental representation of subjects in the dictator and trust game and what is driving this variation? (2) Do mental representations influence the correlation between game behavior and outside behavior?

These two preregistered questions are split into the three results of this paper. Additionally, I analyze to which extent mental representations correlate with game behavior. In terms of methodology, the classification of the open-ended survey questions on associations from everyday life, the emphasis on associations as the primary measure, the experimental design, and the selection of the field behaviors are all preregistered. There are some deviations from the preregistered estimation equations in the main part; Table D.2 provides more details and where the preregistered analyses are located.

The main difference between the preregistered approach and the analyses contained in this paper is that I do not use the original estimation equation for analyzing the relationship between game and field behavior and also look at subsamples based on the individual mental representations instead of interacting game behavior with an indicator for whether subjects have the "right" mental representation. This is motivated by the failure to replicate half of the original findings in this experiment, the great heterogeneity even within social mental representations, and maximizing power for the estimations within the subsample of each mental representation. Due to lack of data on the standard errors in some of the papers, Table D.1 compares the original results with my "replication" estimation results qualitatively.

Finally, Figures D.1 to D.4 provide the results from the preregistered approach to interact game behavior with an indicator for whether subjects have the "right" mental fit in order to detect heterogeneity in the relationship between game and field behavior. All analyses are based on the (original) estimation equations as outlined in Table D.1. Three different mental fit indicators are used in each figure: first, subjects must have a social mental representation. Second, subjects must have a mental representation that is identical to the preference that the game is supposed to measure, i.e., “Charity/Altruism” for the dictator game, “Trust” for the sender in the trust game, and “Trust” *or* “Reciprocity” for the receiver in the trust game. Third, game order is used as an intention-to-treat treatment for the “right” mental fit to provide more causal evidence. Notice that these analyses are just included for completeness since they were preregistered. Because I cannot replicate half of the original findings with the original estimation equation, I focus on simple regressions without additional control variables. The results for this exercise are contained in Section 3.

Table D.1: Comparison to Original Findings

Paper	**Replicates**	**Game behavior**	**Field behavior**	**Other controls**	**Original result**	**This paper***
Galizzi and Navarro-Martinez (2019)	No	Sender in dictator game	SRA	Behavior in UG, TG†, and PGG	Positive effect, significant at 10% level.	95%-CI: $[-.0046, 0.0039]$ with $p = 0.876$.
Glaeser et al. (2000)	No	Sender in trust game	GSS	male†, white†, freshman, lost mail, only child	Positive, but not significant effect	95%-CI: $[-0.0123, 0.4216]$ with $p = 0.065$ for sending 50 instead of 0, $[0.0481, 0.5510]$ with $p = 0.020$ for sending 100 instead of 0.
Glaeser et al. (2000)	Yes	Sender in trust game	Lending	male†, white†, freshman, lost mail, only child	Positive, significant at 10% level	95%-CI: $[-0.0413, 0.1439]$ with $p = 0.277$ for sending 50 instead of 0, $[0.0073, 0.2226]$ with $p = 0.036$ for selecting 100 instead of 0.
Glaeser et al. (2000)	Yes	Receiver in trust game	GSS	male†, white†, freshman, lost mail, only child	Positive, significant at 10% level	95%-CI: $[0.0210, 0.6543]$ with $p = 0.037$.

Table D.1 Continued from previous page

Paper	Replicates	Game behavior	Field behavior	Other controls	Original result	This paper[*]
Riedl and Smeets (2017)	Yes	Receiver in trust game	Holding SRI	investment behavior, university degree[†], risk preferences, female[†], age[†], income[†]	Positive, significant at 1% level	95%-CI: $[0.1232, 1.0386]$ with $p = 0.013$.
Riedl and Smeets (2017)s	No	Receiver in trust game	Share invested in SRI	investment behavior, university degree[†], risk preferences, female[†], age[†], income[†]	Negative, not significant	95%-CI: $[0.0477, 0.2004]$ with $p = 0.001$.
Gill et al. (2022)	No	Receiver in trust game	Career in finance	age[†], gender[†], cognitive skills[†]	Negative, significant at 1% level	95%-CI: $[-0.1209, 0.0880]$ with $p = 0.757$.

[*] With same specification (regressing field behavior on game behavior, controls as indicated) except for Glaeser et al. (2000) and Gill et al. (2022). For them, I use linear probability models regressing field behavior on game behavior, instead of the other way around, to have a common estimation strategy across all field behaviors (robust to using a probit model instead).
[†] Also included as a control in this section. Controls are omitted in the main part of the paper.
See Table A.3 for an overview of field behaviors. UG = ultimatum game, TG = trust game, PGG = public goods game.

Table D.2: Overview of Preregistered Analyses

Result	**Analysis**	**Pre-registered**	**Notes**
Result 1	Effect of game order on associations	Yes	Instead of LPM of treatment indicator on associations, use mlogit for easier interpretation. Preregistered analysis in Table C.1.
Result 1	Effect of game order on sentences	Yes	See Table C.2.
Result 1	Effect of inattention, game understanding, and cognitive skills on associations	Yes (exploratory)	See results in Appendix C.
Result 2	Comparison of association distributions across games	Yes	Did not preregister the dispersion measure.
Result 2	Effect of framing on associations	Yes	Instead of LPM of treatment indicator on associations, use mlogit for easier interpretation. Preregistered analysis in Table C.3.
Result 2	Effect of framing on sentences	Yes	See Table C.7.
Result 2	Effect of age, ethnicity, sex on associations	Yes (exploratory)	Instead of LPM of binary socio-demographic indicator on associations, use mlogit for easier interpretation. Preregistered analyses in Table C.4, C.5, and C.6.
Result 2	Effect of age, ethnicity, sex on sentences	Yes (exploratory)	See Tables C.8, C.9, and C.10.
Result 3	Correlation of behavior with associations	No	
Result 3	Correlation of behavior with sentences	No	

Table D.2 Continued from previous page

Result	**Analysis**	**Pre-registered**	**Notes**
Result 3	Influence of associations on relationship between game and field behavior.	Yes	• Due to failure to replicate findings with estimation strategy from original papers (cf. Table D.1), use simple regression (linear regression, probit) of field behavior on respective game behavior in each mental representation separately. See Figures D.1 to D.4 for preregistered approach with original estimation equation (incl. controls) and interacting behavior with an indicator variable for the fit of mental representations. • In line with the preregistration, do not use framing treatment as an inducement treatment for mental representations because framing treatment does not shift associations. Instead, use game order, see Figures C.8 and C.9 for plain regressions and Figures D.1 to D.4 for the original estimation equations. • Do not separately study SRA based on monetary question items, since none of the SRA-related questions ended up being exclusively about money, i.e., all questions ask about money and/or goods/efforts to decrease the number of questions.

LPM = linear probability model, mlogit = multinomial logit, SRA = self-reported altruism score.

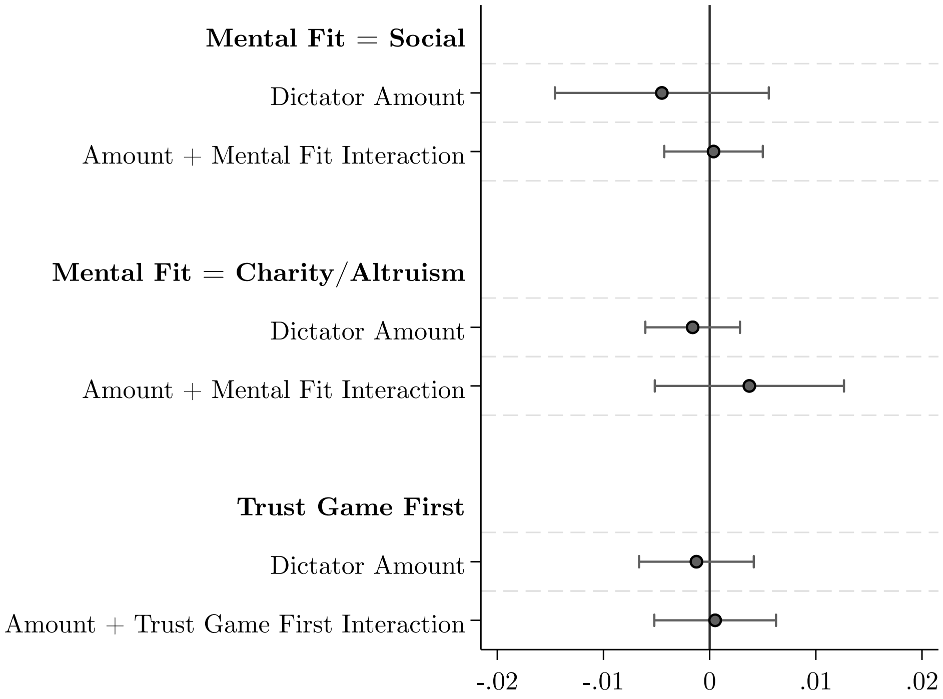
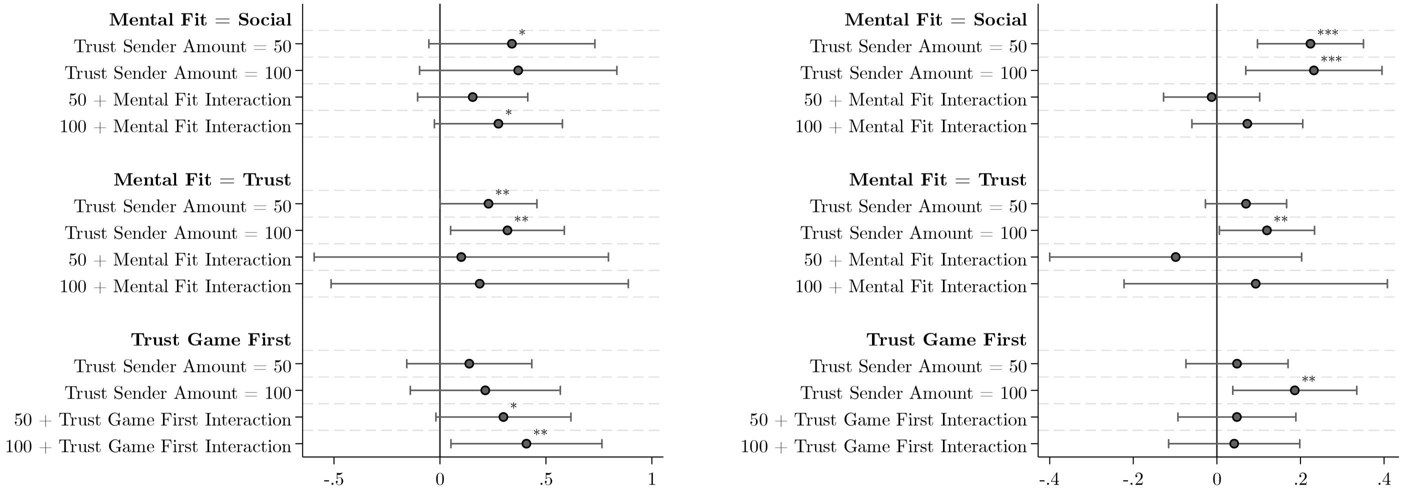


Figure D.1: Heterogeneity in Relationship between SRA and Dictator Amount Sent

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of behavior in the dictator game on SRA, interacted with different indicators, in units of standard deviations. In the first panel, game behavior is interacted with an indicator for a social mental representation. In the second panel, game behavior is interacted with an indicator for a "Charity/Altruism" mental representation. In the third panel, game behavior is interacted with an indicator for playing the trust game first. Within each panel, the second (first) row reports the effect among subjects who did (not) have a social mental representation (first panel), "Charity/Altruism" mental representation (second panel), or play the trust game first (third panel). See Table A.3 for more details on the elicitation of SRA. Table D.1 contains details on the estimation strategy.

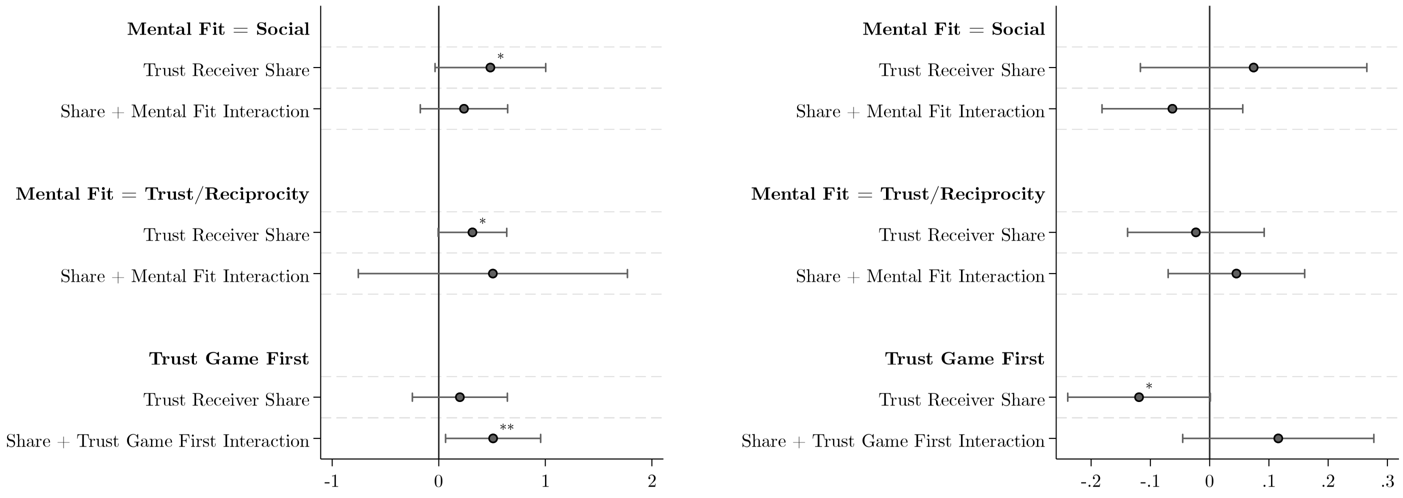


(a) GSS Trust and Trust Game Sender Amount

(b) Lending and Trust Game Sender Amount

Figure D.2: Heterogeneity in Relationship between Game and Field Behavior for Trust Game Sender

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of behavior as the sender in the trust game on GSS (left) and lending (right), interacted with different indicators. Effects for GSS are in units of standard deviations. For each outcome, in the first panel, game behavior is interacted with an indicator for a social mental representation. In the second panel, game behavior is interacted with an indicator for a "Trust" mental representation. In the third panel, game behavior is interacted with an indicator for playing the trust game first. Within each panel, the third and fourth (first and second) row report the effect among subjects who did (not) have a social mental representation (first panel), "Trust" mental representation (second panel), or play the trust game first (third panel). See Table A.3 for more details on the elicitation of GSS and lending. Table D.1 contains details on the estimation strategy.

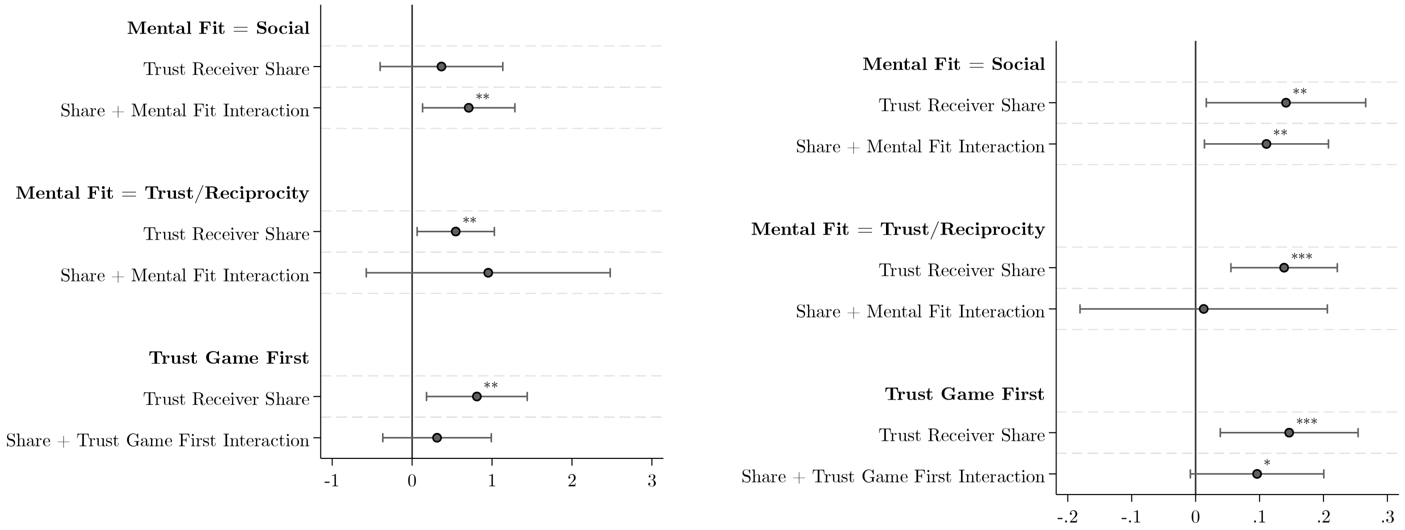


(a) GSS Trust and Trust Receiver Share

(b) Finance Career and Trust Receiver Share

Figure D.3: Heterogeneity in Relationship between Game and Field Behavior for Trust Game Receiver (1/2)

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of behavior as the receiver in the trust game on GSS (left) and finance (right), interacted with different indicators. Effects for GSS are in units of standard deviations. For each outcome, in the first panel, game behavior is interacted with an indicator for a social mental representation. In the second panel, game behavior is interacted with an indicator for a "Trust" or "Reciprocity" mental representation. In the third panel, game behavior is interacted with an indicator for playing the trust game first. Within each panel, the second (first) row reports the effect among subjects who did (not) have a social mental representation (first panel), "'Trust"/"Reciprocity" mental representation (second panel), or play the trust game first (third panel). See Table A.3 for more details on the elicitation of GSS and finance. Table D.1 contains details on the estimation strategy.



(a) Socially Responsible Investment and Trust Receiver Share

(b) Atmosfair Donation and Trust Receiver Share

Figure D.4: Heterogeneity in Relationship between Game and Field Behavior for Trust Game Receiver (2/2)

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This figure reports the average marginal effects of behavior as the receiver in the trust game on SRI (left) and Atmosfair (right), interacted with different indicators. For each outcome, in the first panel, game behavior is interacted with an indicator for a social mental representation. In the second panel, game behavior is interacted with an indicator for a "Trust" or "Reciprocity" mental representation. In the third panel, game behavior is interacted with an indicator for playing the trust game first. Within each panel, the second (first) row reports the effect among subjects who did (not) have a social mental representation (first panel), "'Trust"/"Reciprocity" mental representation (second panel), or play the trust game first (third panel). See Table A.3 for more details on the elicitation of SRI and Atmosfair. Table D.1 contains details on the estimation strategy.

Recent Issues

No. 421	Claes Bäckman, Patrick Moran, Peter van Santen	Mortgage Design, Repayment Schedules, and Household Borrowing
No. 420	Konstantin Bräuer, Andreas Hackethal, Tobin Hanspal, Samuel M. Hartzmark	Educating Investors about Dividends
No. 419	Andreas Hackethal, Philip Schnorpfeil, Michael Weber	Inflation and Trading
No. 418	Ioannis Asimakopoulos, Tobias Tröger	Reform of the CMDI Framework – Driving Off With the Breaks On
No. 417	Thiago Fauvrelle, Max Riedel, Mathias Skrutkowski	Collateral Pledgeability and Asset Manager Portfolio Choices During Redemption Waves
No. 416	Olga Balakina, Claes Bäckman, Anastasiia Parakhoniak	Beyond Connectivity: Stock Market Participation in a Network
No. 415	Monica Billio, Aoife Fitzpatrick, Carmelo Latino, Loriana Pelizzon	Unpacking the ESG Ratings: Does One Size Fit All?
No. 414	Peter Andre, Teodora Boneva, Felix Chopra, Armin Falk	Misperceived Social Norms and Willingness to Act Against Climate Change
No. 413	Errikos Melissinos	Real Term Premia in Consumption-Based Models
No. 412	Andreas Hackethal, Thomas Pauls	Conform to the Norm. Peer Information and Sustainable Investments
No. 411	Peter Andre, Marc Kaufmann, Botond Köszegi	Understanding Markets with Socially Responsible Consumers
No. 410	Carmelo Latino	Surfing the Green Wave: What's in a "Green" Name Change?
No. 409	Kevin Bauer, Yan Chen, Florian Hett, Michael Kosfeld	Group Identity and Belief Formation: A Decomposition of Political Polarization
No. 408	Gianluca Anese, Marco Corazza, Michele Costola, Loriana Pelizzon	Impact of Public News Sentiment on Stock Market Index Return and Volatility
No. 407	Michele Costola, Matteo Iacopini, Casper Wichers	Bayesian SAR Model with Stochastic Volatility and Multiple Time-Varying Weights