CentER

Interdependent individuals:

How aggregation, observation, and persuasion affect economic behavior and judgment

MANWEI LIU

Tilburg University

Interdependent individuals: How aggregation, observation, and persuasion affect economic behavior and judgment

Proefschrift ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. W.B.H.J. van de Donk, in het openbaarte verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de Aula van de Universiteit

op maandag 29 maart 2021 om 13.30 uur door

Manwei Liu,

geboren te Hubei, China

Promotores:

prof. dr. E.C.M. van der Heijden (Tilburg University)
prof. dr. J.J.M. Potters (Tilburg University)

leden promotiecommissie:

prof. dr. S. Gächter (University of Nottingham)
prof. dr. R. Weber (University of Zurich)
prof. dr. S. Suetens (Tilburg University)
dr. B. van Leeuwen (Tilburg University)

Interdependent individuals: How aggregation, observation, and persuasion affect economic behavior and judgment

Manwei Liu

February 2021

Acknowledgements

This dissertation has witnessed some of the most enjoyable and challenging years of my life. It would have missed the enjoyable part had I not been blessed with the love and support of many people.

First and foremost, I wish to thank my supervisors, Eline van der Heijden and Jan Potters, for their incredible guidance and unconditional support both academically and emotionally throughout these years. What they led me to see has fundamentally shaped how I view research and life. Special thanks to Eline, who is the co-author of chapter 2, for her sharp eyes have saved me from making vital or embarrassing mistakes three (maybe more) times. What is more important, Eline has always been a caring, helpful, and trusting person with great energy ever since our first conversation. Jan sets a good example as a passionate researcher and lecturer. His ability of switching between the serious mode and the witty mode so smoothly keeps every meeting alive.

I would also like to express my sincere gratitude to the committee members, Boris van Leeuwen, Sigrid Suetens, and Roberto Weber, for they kindly responded to my needs and shared their insights and expertise in the development of this dissertation whenever I asked for. A big thank you to Simon Gächter for joining the committee and offering his fresh and distinct perspective. All of their feedback has fostered significant improvements in this dissertation.

I am grateful to the Economics Department of Tilburg University, especially to the behavioral and experimental group, for the inspiring and constructive meetings and discussions. The enthusiastic, open, and reciprocal environment is what kept me going when I felt exhausted. Many thanks to Cedric Argenton, Patricio Dalton, Eric van Damme, David Schindler, Florian Sniekers for their valuable suggestions and to Korine Bor and Cecile de Bruijn for their administrative assistance during the job market season. I would have lost half of my confidence and patience without your help.

I am fortunate enough to have encountered many friends and colleagues who have been the best companions I can ever imagine. To Sili Zhang, who is the collaborator of chapter 4: thank you for being a mirror of mine and sharing with me everything we love and hate in life. Thijs Brouwer, Lenka Fiala, Takumin Wang, Tingting Wu, Yadi Yang, Wanqing Zhang: the talks we had about research and life, laughs and cries, hopes and struggles have lighted my burdens and made me more resilient. Thanks to Pavel Čížek, Dorothee Hillrichs, Madina Kurmangaliyeva, Bert Willems, Shuai Yuan, and Yi Zhang, for echoing my love for games (you know which game) which are the main sources of dopamine. Many thanks to my lunch, dinner, and coffee partners: Santiago Bohórquez Correa, Mirthe Boomsma, Sebastian Dengler, Clemens Fiedler, Ruonan Fu, Tao Han, Chen He, Roweno Heijmans, Lei Lei, Shuo Liu, Zihao Liu, Marie le Mouel, Tung Nguyen Huy, Zilong Niu, Laura Capera Romero, Lingbo Shen, Yi Sheng, Chen Sun, Xiaoyu Wang, Oliver Wichert, Mingjia Xie, Xue Xu, Yuanyuan Xu, Yilong Xu, Jierui Yang, Miao Zhang, Xiaoyue Zhang, Sophie Zhou. Food, as well as my Ph.D. career, would never taste the same without you.

I also thank China Scholarship Council for providing financial support for 60 months.

My final and deepest gratitude goes to my parents and other family members for their unfading love.

Thank you all.

Contents

Chapter 1

Introduction

The topic of this dissertation is the interdependence of individuals, specifically, how individuals are influenced by the social environment and social interactions they engage in. In the three chapters of this dissertation, I explore how observation, aggregation, and persuasion affect judgment and economic decision making. All of the three chapters combine behavioral insights with experimental methods.

Through a collective-choice rule, groups can aggregate individual preferences into a collective decision. Chapter 2 explores the role of collective-choice rules in self-governance in a public goods game, and shows that collective-choice rules affect cooperation directly and indirectly. By observing what others do, people extract information from the environment. In chapter 3, I demonstrate that observing a reference action suffices to affect moral judgments. Chapter 4 studies how people deal with information slant and studies the persistent effect of persuasion on judgments and decisions.

Chapter 2 focuses on collective-choice rules used to aggregate individual choices into a group choice. The context of this study is self-governance via institutional design in social dilemmas. In a social dilemma situation, people often appeal to institutions, for example, sanction systems, to sustain cooperation and achieve self-governance. Individual preferences over institutions are aggregated via a certain collective-choice rule into a collective institutional choice. For instance, a referendum is held to decide whether a society adopts a new policy. This study uses an experiment to explore the direct and indirect effects of three different collective-choice rules (majority voting, dictatorship, and rotating dictatorship) on cooperation in a public goods game. Do people behave more pro-socially if institutions are chosen through a democratic rule than if identical institutions are generated by a non-democratic rule? Do collective-choice rules affect the stability of institutions and in turn affect cooperation?

The results show that cooperation levels are higher with the institutions chosen by a dictator than with the same institutions chosen by rotating dictators. Rotating dictatorship tends to produce less stable institutions than majority voting or fixed dictatorship. And the instability of institutions is associated with lower cooperation levels. This study contributes to the strand of literature that examines the significance of collective-choice rules. Apart from adding new evidence to the direct effect of collective-choice rules, we demonstrate that stability matters for the effectiveness of collective decisions. Chapter 3 also extends the vast literature on the endogenous emergence of institutions by allowing for greater freedom in self-governance via institutional design.

Economic literature often sees moral judgments as static and internal: people evaluate an action or a person who chooses that action solely based on how they perceive the nature of the action (for instance, see Abeler et al.,2019; Gneezy et al,2018). Chapter 3 examines whether moral judgments are relative, that is, whether moral judgments about a decision maker depend on a reference action. Such dependence may result from a contrast effect and/or from changes of perceived social norms, which is referred to as a norm-shifting effect. Intuitively, the contrast effect makes a good action look better, while the norm-shifting effect makes a common action look better. The two effects do not always work in the same direction, and thus the reference action may have an asymmetric effect on judgments about moral decision makers and immoral decision makers. I use an experiment to test the predictions.

The results show that people's moral judgments toward a decision maker depend on the reference decision maker's action. Such dependence is indeed asymmetric for judgments about moral and immoral decision makers. In particular, people punish an immoral decision maker more harshly if the reference choice is moral than if it is immoral. A contrast effect, rather than a shift in perceived social norms, accounts for this finding. The punishment moral decision makers inflicted does not vary with the reference choice. Moral judgments do not depend on the perceived descriptive norm. This paper adds to the literature on how previous experience shapes preferences and values and speaks to the growing literature on the determinants of moral decisions.

Chapter 4 studies the persistent effect of information slant. Information slant conveys a particular viewpoint or inclination by choice of words, tone, and framing without necessarily misrepresenting the underlying facts. Sometimes people are still persuaded by information slant despite being aware that it is partial. It is not clear whether and how people can counteract its impact. This chapter examines whether and to what extent we can mitigate the impact of information slant by providing people with an opportunity to acquire more information from both sides of the issue. In a controlled experiment, we choose a concrete issue, the use of GMO mosquitoes in disease control, and construct two versions of slanted articles. The slanted articles are based on the same set of underlying facts but slanted toward either the pro-GMO side or anti-GMO side. Subjects are well informed that the article they receive is slanted toward one side of the issue and toward which side exactly. We elicit subjects' attitudes toward the use of GMO mosquitoes after their exposure to the slanted article. Next, we offer subjects an opportunity to acquire more information from both sides and elicit their attitudes for a second time to see whether and in what direction their attitudes change.

The results show that information slant indeed shifts subjects' attitudes in the direction aligned with the initial article compared to the attitudes of those who have seen both sides. The opportunity to acquire information from both sides only slightly mitigates the impact of information slant, as there is mild evidence that attitudes move closer after information acquisition, whereas exogenously imposed information does not narrow the gap between subjects exposed to different sides. We find that the impact of initial information slant may be persistent: it shifts subjects' attitudes, and the shift in attitudes in turn induces confirmation bias which distorts the way people process subsequent information. This study contributes to the understanding of how people deal with information slant embedded in natural language. It also highlights the role that media bias plays in shaping opinions and preferences.

Chapter 2

The role of collective-choice rules in self-governance via institutional design

2.1 Introduction

Social dilemmas are situations characterized by a conflict between individual interests and collective interests. In such situations, an individual is better off if she/he defects instead of cooperates, but the group is better off if all cooperate than if all defect. In order to sustain cooperation and to achieve self-governance in such social dilemma situations, groups often appeal to institutions. The optimal institutional design, however, is usually not obvious, because it depends on the incentives institutions provide, the group members' characteristics, and the group dynamics. We therefore consider groups' self-governance through institutional design as a problem-solving process. In this process, a group needs a collective-choice rule to aggregate its members' preferences over institutions into a collective decision. In this study, we focus on the instrumental values of collective-choice rules in the problem-solving process. Specifically, we explore whether, and if so how, collective-choice rules affect groups' cooperation and performance in a social dilemma situation.

We use an experiment to simulate a social dilemma situation where groups can choose from a variety of institutions to achieve self-governance. The ba-

This chapter is based on the joint work with Eline van der Heijden.

sic game is a repeated standard public goods game. Every three rounds, group members select institutions as additional rules on top of the basic game. Individual institutional choices are aggregated via a given collective-choice rule into a group institutional choice. Groups then play the public goods game with the chosen institutions for three rounds. We employ a between-subjects design: every group uses only one collective-choice rule throughout the game. Specifically, we implement three collective-choice rules: majority voting, dictatorship (i.e. a single decision maker) and rotating dictatorship (i.e. single decision makers on a rotating basis). Examples of rotating dictatorship are the presidency rotation of the EU council, the presidency rotation of the UN Security Council, etc. These rules are natural candidates to start with because they are simple, representative, and they constitute the basis for more complicated collective-choice rules.

In our social dilemma context, collective-choice rules may have a direct or an indirect effect on cooperation. The direct effect refers to the influence of collective-choice rules on cooperation, conditional on institutions. The indirect effect describes the influence of collective-choice rules on cooperation through the choice of institutions.

Collective-choice rules may *directly* influence cooperation because people have a preference for democratic participation right and value decision right (Bartling et al., 2014; Fehr et al., 2013). Their preference for participation right might be translated into support for the resulting decision (Castore and Murnighan, 1978). For instance, some studies on endogenous institutions find an "endogeneity premium" in the sense that policies chosen by the subjects are more effective than identical policies selected randomly by the computer or imposed exogenously by the experimenter (Arbak and Villeval, 2011; Dal Bó et al., 2010; Casari and Luini, 2009; Kamei, 2016; Markussen et al., 2013; Rivas and Sutter, 2011; Tyran and Feld, 2006). In a similar vein, we explore in this study whether there exists a "democracy premium". That is, given that the institutions are chosen endogenously, do people behave more pro-socially if these institutions are produced by a democratic rule compared to a situation in which the same institutions are generated by a non-democratic rule?

Collective-choice rules may also *indirectly* affect cooperation through their influence on institutions. The indirect effect may come from at least two sources, namely through the types of the institutions chosen and through the stability of the institutional outcomes. As to the types of institutions, we see institutional design as a solution to social dilemma specific to groups. The solution is not necessarily unique and is at each group's discretion. We do not have any supporting theories or evidences to predict the best fitting institutions given a certain collective-choice rule. Nor do we intend to identify the good types of institutions from the bad types. Therefore, we do not form any *ex ante* hypothesis regard this source in this paper.[1] The second source concerns the stability of the institutions. Different collective-choice rules may produce more stable or more unstable institutions over time, and this may affect cooperation. The instability of collective choices, in particular of majority rule outcomes, has been of concern to political scientists and some economists (Haney et al., 1992; Hoffman and Plott, 1983; McKelvey and Ordeshook, 1984; Plott and Levine, 1978; Riker, 1980; Wilson, 1986). And research has shown that political instability can be detrimental to economic performance (Aisen and Veiga, 2013; Alesina et al., 1996; Barro, 1991; Dixit, 2009; Feng, 1997). In our paper, we focus on the indirect effect of collective-choice rules via the (in)stability of institutions rather than via the types of institutions.

A distinct feature of our experiment is that we allow for a rather broad range of institutional design. Existing experimental research on endogenous institutions focuses largely on the emergence of a certain type of institution, for instance, a sanction system. Groups are typically offered a take-it-or-leave-it option: either they implement a particular institution or not implement it (Bischoff, 2007; Ertan et al., 2009; Fehr and Williams, 2013; Guillen et al., 2007; Gürerk et al., 2009; Potters et al., 2005; Sutter et al., 2010). We provide participants with a "menu" of institutions with three items on it: communication, punishment, and reputation. For each item, groups choose whether to implement it or not. This gives them in total eight different combinations of institutions for institutional design. For example, a group can choose communication and punishment, or all of the three institutions. We believe that our menu includes common and crucial institutional features. Communication, punishment, and reputation are among the most intensively studied institutions in economic experiments. Perhaps even more importantly, Rockenbach and Wolff (2016) show that when people can create any institution they prefer, their creations are often limited to a handful of institutions, including communication, punishment, and information feedback.

Another novel characteristic of our design is that we explicitly incorporate institutional costs. Running and maintaining an institution typically entails costs.

[1] One may argue that, since dictators do not need to coordinate with anyone on the institutional choice, they can choose a (pro-social) institution more easily than groups with majority rule. However, there is no guarantee that "good" institutions will be chosen with a better chance. Dictators can make wrong institutional choices as easily as they make right ones. Beside, the dictators are not necessarily pro-social. They can choose punishment to hinder free-riding, or they can make sure free-riding is not punished.

Some institutions have the property of being a public good. Once they are established and made to work, each group member can benefit from increased cooperation regardless of whether she pays for it or not. A second-order social dilemma is thus formed. For example, punishing someone else typically entails costs for the punisher, while other group members can benefit from enforced cooperation without paying any additional cost.[2] Therefore, the successful establishment and use of an effective institution also calls for agreement on who pays for the institutional costs. In our experiment, subjects need to decide collectively on how to divide the costs for each institution (through the given collective-choice rule).

Our main findings are as follows. First, we do not find evidence of a "democracy premium". On the contrary, fixed dictatorship has a positive direct effect on contribution behavior and earnings compared with rotating dictatorship. Second, institutional choices generated by rotating dictators are significantly more unstable than those generated via majority voting or fixed dictatorship. Third, the instability of institutional choices is associated with lower cooperation levels. Finally, overall, groups in the fixed dictatorship treatment achieve higher cooperation levels, spend less on institutions, and hence earn more than those in rotating dictatorship.

To our knowledge, this paper is the first to investigate both the direct and indirect effects of collective-choice rules on cooperation in a social dilemma environment. We contribute to the discussion over the instrumental value of collective-choice rules. We provide two perspectives: participation right and stability. People might value the participation right ensured by democratic collective-choice rules, but our findings suggest that such rules do not necessarily bring better performance nor lead to a more cooperative community. Our results also imply that the stability of collective decisions may be important and therefore deserves to be included when discussing the impact of collective-choice rules on group decision-making. These findings have implications for the organization of small-scale group decision-making.

2.2 Related literature

Our paper is related to two strands of literature. The first strand of literature examines the endogenous emergence of a certain type of institution. In most exper-

[2] Related literature finds that people suffer as much from higher order social dilemmas as from first-order social dilemma (Sigmund et al., 2010; Zhang et al., 2014). Fehr and Gächter (2002) claim that free riding can only be solved if sufficient agents are willing to "altruistically" provide the second-order public good.

imental studies, subjects are offered an exogenously given institution, and they can choose to implement this institution or not (Bischoff, 2007; Ertan et al., 2009; Fehr and Williams, 2013; Guillen et al., 2007; Gürerk et al., 2009; Markussen et al., 2013; Potters et al., 2005; Sutter et al., 2010). For example, in Guillen et al. (2007), after having experienced a central sanction mechanism for several rounds, groups can choose whether to keep it; they all choose to remove the sanction system. Potters et al. (2005) let two players decide unanimously whether to apply a sequential structure in a voluntary contribution mechanism. Groups predominantly choose to move sequentially. Some studies have expanded the institution set to examine variations of a certain type of institution. For instance, Sutter et al. (2010) provide subjects with three institutions to choose from: standard voluntary contribution mechanism (VCM), VCM with punishment, and VCM with reward. They find that groups prefer the reward option, although punishment turns out to be more effective in sustaining cooperation. Fehr and Williams (2013) offer four communities different forms of a punishment system to choose from, including one without punishment. They find that a centralized punishment system dominates and almost eliminated free-riding.

An expansion of this strand of the research compares the effectiveness of endogenous institutions with the same institutions exogenously imposed by the experimenters (Arbak and Villeval, 2011; Dal Bó et al., 2010; Casari and Luini, 2009; Kamei, 2016; Markussen et al., 2013; Rivas and Sutter, 2011; Tyran and Feld, 2006). Findings from these studies provide the basis for our conjecture about the direct effect of collective-choice rules.

We contribute to this strand of literature by pushing endogenous institutions to the direction of institutional design. Compared with existing literature, we give subjects more freedom in choosing and designing institutions and allow subjects to repeatedly change their institutional choices. We choose this design not only because institutional design and its evolution are interesting in themselves, but also because the collective decision to be made–choosing the right set of institutions–varies with timing and the group composition and thus it is difficult and conflict-invoking. Collective-choice rules likely play a bigger role when there are disagreements. It is for a similar reason that the institutions in our design are costly. The study of Rockenbach and Wolff (2016) mostly resembles such an idea of institutional design. In their study, subjects can create any rules for a public good game , and they can repeatedly improve their design. The authors then decompose the subjects' rule sets into rule components and classify these components along different categories. They find that subjects usually combine two or more rule components, rather than use exclusively a single category of institution. In particular, designers often implement vertical communication and punishment and abstain from giving feedback on contributions. A critical difference between our setting and theirs is that we have more control on the collective decisions process by providing a preset menu of institutions. The institutions we offer are the ones that were chosen most frequently by subjects in the paper by Rockenbach and Wolff (2016), communication, punishment and information.

The second strand of literature investigates the difference between collective-choice rules from various perspectives. Some papers view collective-choice rules as information aggregation devices and look at how these rules influence the accuracy of group decisions (Nitzan and Paroush, 1982; Shapley and Grofman, 1984; Sorkin et al., 2001). For instance, Shapley and Grofman (1984) shows theoretically that collective-choice rules affect the likelihood of a correct choice and suggests that simple majority rule may be near optimal for large groups. Sorkin et al. (2001) compares super majority, simple majority, and unanimity from a signal detection theory perspective. They find that simple majority performs the best and unanimity is the worst. Other studies use collective-choice rules as devices to resolve conflicts of individual distributional preferences (Buchanan and Tullock, 1962; Dougherty et al., 2014; Walker et al., 2000). For example, Walker et al. (2000) studies how majority voting and unanimity affect the allocation proposals and efficiency gains in a common pool resources game. Their results suggest that unanimity rules lead to higher efficiency in symmetric groups with complete information.

We contribute to the research on collective-choice rules by adding another perspectives. We see collective-choice rules as problem-solving devices that involve more than information and preference aggregation. In our study, subjects are placed in a relatively rich decision environment, in which they interact with each other under strategic uncertainty. The complex but arguably more realistic setting enables us to see how collective-choice rules affect behavior and performance from new angles.

2.3 Experimental design and procedure

2.3.1 The Institutional Design Game

We use what we call an “Institutional Design Game” (IDG) to simulate a social dilemma environment where groups can design their own institutions. The main body of the IDG is a standard public goods game. The standard game can be modified by using institutions which affect the rules of the game. Groups choose institutions via a collective-choice rule. Therefore, a group's choices of institutions pin down a personalized version of a modified public goods game for the group.

Many things can be varied in the IDG, depending on the research interest. Since our main aim is to examine whether, and to what extent, collective-choice rules affect cooperation behavior in such a social dilemma situation, we fix the collective-choice rule for each group. The details on the basic game, the available institutions, and how the collective-choice rules work are provided in the following subsections. In the last subsection we formulate the hypotheses.

Standard public goods game

The main body of the IDG is a standard three person public goods game widely used in studies into social dilemmas (Ledyard, 1994). The public goods game is repeated for 18 rounds, which are divided into six terms of three rounds. In every round, group members are assigned ID numbers as their temporary identity; ID numbers change every round. The three participants form a fixed group (partner matching). We use subscript $i \in 1, 2, 3$ to denote individual level decisions and subscript j to denote group level decisions or outcomes. In every round, all group members receive an endowment $e = 20$ points, and they decide simultaneously how much to invest in a public project. Every point invested into the project yields a return $\alpha = 0.5$ to each group member. Points that are kept in private remain unchanged. The payoff function of player i in round t if he chooses to invest x_{it} into the public project is then:

$$\pi_{it} = 20 - x_{it} + 0.5 * \sum_{k}^{3} x_{kt}$$

The marginal per capita return $\alpha = 0.5$ ensures that there is a conflict between group interest and individual interest since $0.5 < 1$ while $3 * 0.5 > 1$. That is, there are incentives to achieve a higher contribution level; full cooperation gives every group member 30 while fully selfish group members earn 20. After contributions have been made, group members receive information on every group member's contribution and payoff in the current round. However, without reputation, the contribution and payoff information is shown in match with the temporary IDs in the current round. With reputation, the contribution and payoff information is shown together with IDs fixed for the current term (see the next section for more details).

Institutions

In the experiment, several institutions can be established to extend the rules of the basic game. At the beginning of each term, groups can choose the institution(s) to be implemented in that term. We provide a menu with three available institutions: communication, reputation, and punishment. For each institution, groups choose whether to implement it or not. This gives subjects the freedom to combine two or more institutions. For instance, a group can choose communication and reputation, or communication and punishment. It is also possible to implement none of the institutions. Hence, there are in total eight possible institutional choices. To avoid connotation, in the experiment, these institutions are referred to as mechanism A, B and C, respectively. For each group, the individual choices are aggregated to determine if an institution will be implemented in a term. How individual choices are aggregated depends on the collective choice rule, i.e. on the treatment.

One property of institutions is that they are typically costly to establish and/or to maintain. To capture this characteristic in the experiment, the establishment of an institution in a group costs 3 points per round, and these costs will be shared equally among group members in each of the three rounds of the term. On top of that, every time the function of communication/punishment is executed, a variable cost arises, as will be described in more detail below. Variable costs can be shared equally among group members or burdened by the individual who initiates it, depending on the cost-sharing rules chosen by the group members. Table 1 summarizes the fixed costs and variable costs for each institution. This table was also presented in the instructions. We will now elaborate on the details of the institutions.

With communication (mechanism A), group members can simultaneously send messages to a "blackboard" that every group member can see before making a contribution decision. The content of the message is semi-specified, and participants are required to fill in the blank before they send the message. Every message they send entails a variable cost of 1 point. They can choose one or more messages from the following set:

- "I propose we invest ____(integer number from 0 to 20) points in the project."
- "I propose we use mechanism A/B/C/none."

- “I propose we share the variable costs of mechanism A/C/none.”

Reputation (mechanism B) makes a person’s (term) history tractable and known to all group members. With reputation, the temporary IDs assigned to group members are fixed for this term; group member 1 remains group member 1 throughout the term. And subjects see each group member’s record of contribution and payoff in this term, after making contributions. They can thus see six numbers (the contribution and payoff of each group member) in the first round of this term and 18 numbers in the last round of this term. Unlike the other two institutions, reputation does not incur any variable cost. Once it is established at fixed costs of 3 per round, it keeps working for the entire term without any further variable costs.

The last institution, punishment (mechanism C), enables group members to directly reduce others’ payoffs at some costs after they have seen every member’s contribution and payoffs. The fine-to-fee ratio is 3 in this experiment, following the standard practice in public good experiments (Nikiforakis and Normann, 2008). That is, it costs one point to reduce another individual’s payoff by 3 points. To rule out anti-social punishment behavior, subjects can only punish those who contribute less than them.[3]

Subjects are informed of all of the above information in the instructions they receive (see Appendix B).

Table 2.1: Overview of costs (in points) per treatment

Institution	Fixed Cost per round	Variable Cost per use
Communication	3	1 per message
Punishment	3	1/3 per unit of deduction
Reputation	3	-

Treatments: Collective-choice rules

Our experiment consists of three treatments, which only differ in the collective-choice rules employed to translate individual preferences into one collective decision: majority voting (MV), fixed dictatorship (FD), and rotating dictatorship (RD). The rules of the basic game and the institutions explained in previous sections apply to all treatments.

[3] Anti-social punishment widely exists when punishment is possible (Fehr and Gachter, 2000; Herrmann et al., 2008). We forbid antisocial punishment in order to keep the function of the “punishment” institution straightforward and to avoid inducing more complicated motives and interactions.

At the beginning of each term, group members choose institutions and the corresponding cost-sharing rules of each institution individually and simultaneously. For each item (or institution) on the menu, they answer two questions: (1) whether they want this institution to be established with fixed costs that will be shared by all group members and (2) whether they would like to share the potential variable cost of that institution, if it is put into use.

In the MV treatment, whether an institution will be established depends on whether the majority of the group votes yes or no concerning question (1). If the majority votes yes, the cost-sharing rule will be also determined by majority rule according to their answers to question (2).

In the FD treatment, one of the group members will be randomly chosen as the decision maker (fixed dictator) and remains the decision maker throughout the entire game. In the very first round, all group members are asked to answer questions (1) and (2) without knowing whether they are the decision maker or not. After the voting phase in the first round, each subject knows whether he is assigned as the decision maker or not. Nevertheless, to minimize procedural differences between the treatments, non-dictators are still required to privately indicate their preferences over institutions at the beginning of each term, even though their choices do not affect the institutional outcome.

In the RD treatment, participants take turns to be the decision maker (rotating dictator). The order is randomly determined by the computer. All group members answer questions (1) and (2) in all six terms without knowing whether they are going to be the decision maker in that term. Only after the voting phase will they be informed of whether their choices have been selected and implemented or not. There are six terms, and hence six institutional choices. Subjects know that only two times their decision(s) will actually be implemented as the collective-choice. If a group member has been selected twice as the decision maker, she can therefore conclude that her institutional choice would not matter anymore in later terms.

2.3.2 Procedure

The experiment was conducted at CentERlab at Tilburg University, and it was programmed using zTree (Fischbacher, 2007). Participants were recruited via the UvTlab system of Tilburg University. We employed a between-subjects design. Participants could only participate in one experimental session, and in each session, only one treatment (i.e. collective-choice rule) was conducted. We implemented in total six sessions, with two sessions per treatment.

We start by reading aloud instructions. Subjects are first introduced to the standard public good game as described in section 3.1.1. A test of their understanding of the public good game is performed before entering the next stage of instruction. Then we introduce the available institutions as described in section 3.1.2 and the collective-choice rule used in their collective-decision process. Subjects can ask questions, which are answered privately. Finally, the experiment starts without any practice round. The complete procedure of the game is presented in Appendix A.

In each treatment, subjects are randomly matched in fixed groups of three (partner matching). There are in total 18 rounds, which are divided into six terms of three rounds each. In rounds 1, 4, 7, 10, 13, 16, groups play a full Institutional Design Game consisting of two phases: a collective decision phase and a simultaneous contribution phase. In all other rounds participants skip the collective decision phase and only make contribution decisions. Every round participants are randomly (re)assigned an ID number 1, 2 or 3 (except for the case where they have established reputation institution).

In the collective decision phase, group members choose which institutions they would like to implement in the next three rounds and the corresponding cost-sharing rule individually and simultaneously. Their choices are aggregated into group decisions via the collective-choice rule assigned to the group. At the end of this collective decision phase, the group decisions are displayed on the screen and hence known to all group members. Individual votes are not revealed.

Then follows the second phase of simultaneous contribution. Participants play a standard public game with the institutions that they have chosen themselves. The first stage of the contribution phase is the communication stage. Only groups that have chosen the communication institution are eligible to participate in this stage. Subjects of eligible groups can choose and fill in the message(s) that they would like to convey to his/her group members. Not sending any message is also possible. Messages from each member are shown on the group's blackboard. In the second stage, participants play the standard public good game and make their contribution decisions simultaneously. In the succeeding stage, individual contribution and payoffs of this round are shown to all group members. The fourth stage is the punishment stage. If the punishment institution is established in a group, participants indicate whom they want to punish and by how many points. These decisions are made individually and simultaneously. In the final stage, punishment and final payoffs of each group member in this round are shown on the screen. Groups that have chosen reputation can identify contributions and payoffs of the group member in a term; without reputation this information cannot be linked to individual group members (only to random IDs).

The monetary payoff of each participant in the experiment is determined by the sum of individual tokens earned in all 18 rounds plus a show-up fee of 3 euros. 100 tokens are translated into 3 euros.

In total, 114 subjects participated in the experiment (39 in MV, 36 in FD, and 39 in RD), of whom 43% are women. In the questionnaire conducted at the end of the experiment, around 74% of the subjects claimed that they had participated in a "similar" experiment. This suggests that these participants generally have a good understanding of the game, or at least they believe so. Subjects earned on average 16.9 euros in about 1 hour 15 minutes.

2.3.3 Hypotheses

Our first hypothesis is that collective-choice rules have a direct effect on contribution behavior. The direct effect is reflected in the difference in contribution behavior by collective-choice rules conditional on having the same institutions. This hypothesis builds on people's preference for democratic participation right. Participation right provides a feeling of inclusion and lends legitimacy to the collective decisions. Being able to participate and decide may increase group members' mental attachment to the selected institutions and hence enhance compliance with the resulting institutions (Arbak and Villeval, 2011; Dal Bó et al., 2010; Casari and Luini, 2009; Castore and Murnighan, 1978; Frey et al., 2004). In our setting, majority voting guarantees every group member participation right in all six terms. With rotating dictatorship, group members share participation rights equally over all terms, but each group member has decision power only 1/3 of the time. Fixed dictatorship assigns all the decision power to only one of the three group members, which likely undermines the legitimacy of the selected institutions. While majority voting is obviously democratic, rotating dictatorship is arguably democratic, depending on one's understanding of "democracy". Based on these arguments, we formulate the first hypothesis as follows.

Hypothesis 2.1. *(Direct effect). Given the same endogenous institution combination, contributions to the public good are greater with majority voting than with rotating dictatorship, and are greater with rotating dictatorship than with fixed dictatorship.*

Our second hypothesis consists of two parts. First, we hypothesize that collective-choice rules affect the stability of institutional outcomes. Second, the stability of institutional outcomes in turn affects cooperation behavior. As to the first part, intuitively, rotating dictatorship is expected to produce the least stable institutional outcomes. Regarding majority voting, a number of studies use controlled experiments to test the stability of majority rule. They have shown that the resulting collective decisions under the majority voting framework are not necessarily predictable and may depend on details of the voting procedure, for example, agenda setting, or pre-communication (Haney et al., 1992; Hoffman and Plott, 1983; McKelvey and Ordeshook, 1984; Plott and Levine, 1978; Wilson, 1986). Following these results, we hypothesize that institutional outcomes produced via a fixed dictatorial collective-choice rule are more stable than those produced via majority voting.

Hypothesis 2.2a. *Institutional choices generated under fixed dictatorship are the most stable among the three collective-choice rules, and those generated under rotating dictatorship are the least stable.*

The second part of the second hypothesis is based on extensive empirical research on the relationship between political stability and economic performance (Aisen and Veiga, 2013; Alesina et al., 1996; Barro, 1991; Dixit, 2009; Feng, 1997). Literature sometimes finds instability to be the cause of poor economic performance and sometimes the consequence of it. We therefore expect to observe a positive correlation between the institutional stability and cooperation in our setting.

Hypothesis 2.2b. *The stability of institutional choices is positively correlated with cooperation.*

Collective-choice rules may also affect the type of institutions groups choose. As explained in section 1, we cannot formulate a testable hypothesis about the relationship between collective-choice rules and the type of institutions groups prefer. Nor can we say anything about the effects of a certain combination of collective-choice rule and institutional design on cooperation behavior. Therefore, we remain open to any result that surfaces from our data.

2.4 Results

We first present some general results on subjects' behavior in the public goods game, including the performance, contributions and institutional choices. Bearing these results in mind, we will then proceed to investigate the direct and indirect effects of the collective-choice rules.

2.4.1 General results

Performance

The first result we are interested in is how well groups perform in self-governance, or their "performance". Performance is measured by the earnings of an individual or a group in the game: subtracting the costs related to the establishment and usage of institutions from the total profits made in the public goods game. Performance is the materialized individual or social welfare. For a group, performance is also a net measure of cooperation, broadly considered. It increases when group members contribute more to the public good, and it decreases when the group achieves that cooperation level at greater costs.

As a benchmark, if a group never establishes any institution and all group members contribute nothing to the public good, each individual group member will earn 360 experimental points over the 18 rounds of the game. Full contributions without any costs of institutions would yield maximum individual earnings of 540 points. Figure 2.1 presents the distributions of individual accumulated earnings (in experimental points, one observation per individual) in each treatment, where the vertical lines are drawn at the mean value. Over all treatments subjects earn on average 483.2 points, which is an increase of more than 34% compared to the benchmark earnings of 360 points.

The average earnings of subjects are highest in the FD treatment and lowest in the RD treatment. As can bee seen in Figure 2.1, the variation in individual earnings is smaller in the FD treatment than in the other two treatments. Using independent observations at the group level, Kolmogorov-Smirnov tests show that average group earnings in the FD treatment are significantly higher than in the RD treatment ($p = 0.05$, $n_1 = 12$, $n_2 = 13$), but not significantly higher than in the MV treatment ($p = 0.21$, $n_1 = 12$, $n_2 = 13$). Earnings in the MV and RD treatment do not differ significantly from each other ($p = 0.58$, $n_1 = 13$, $n_2 = 13$). Figure 10 in Appendix C presents the earnings in the three treatments over rounds.

Result 2.1. *Total earnings with fixed dictatorship are significantly higher than with rotating dictatorship.*

As group performance depends on how much a group contributes to the pub-

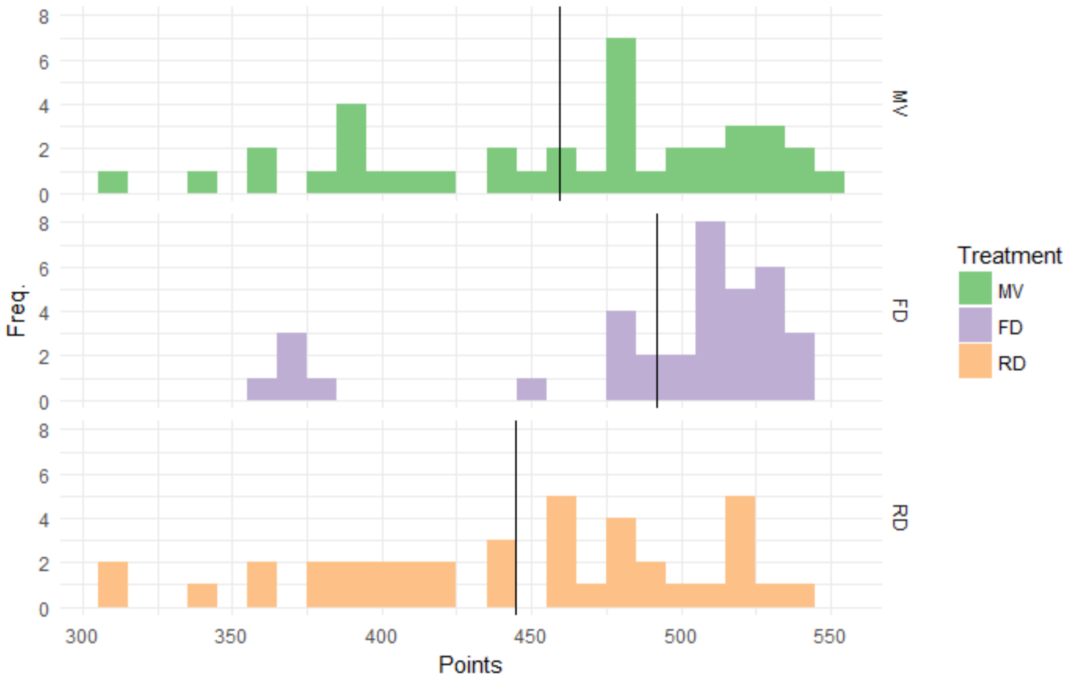


Figure 2.1: Distribution of individual earnings

lic good and how intensively it uses the institutions a group's total earnings can be decomposed into two parts: the profits from the public and private good, and the costs associated with the usage of institutions. Figure 2.2 and Table 6 (in Appendix C) give more detailed information on the two parts. In Figure 2.2, the darker and lighter dots are located at the mean of per round group contributions and group institution-related costs across terms for the three treatments, respectively. The box-and-whisker plots show the distribution of group earnings in each treatment across terms.

Figure 2.2 suggests that contributions to the public good play a decisive part in determining the group performance. Institution-related costs, including the fixed and variable costs of institutions and the loss from punishment, are relatively small in magnitude. However, both parts contribute to the differences between treatments.

In the subsequent sections, we look more closely into contribution behavior and institutional choices.

Contribution behavior

For each treatment, the development of the average group contributions over 18 rounds is shown in Figure 2.3. In all three treatments of our institutional design game, groups manage to sustain cooperation at a rather high level. Per round

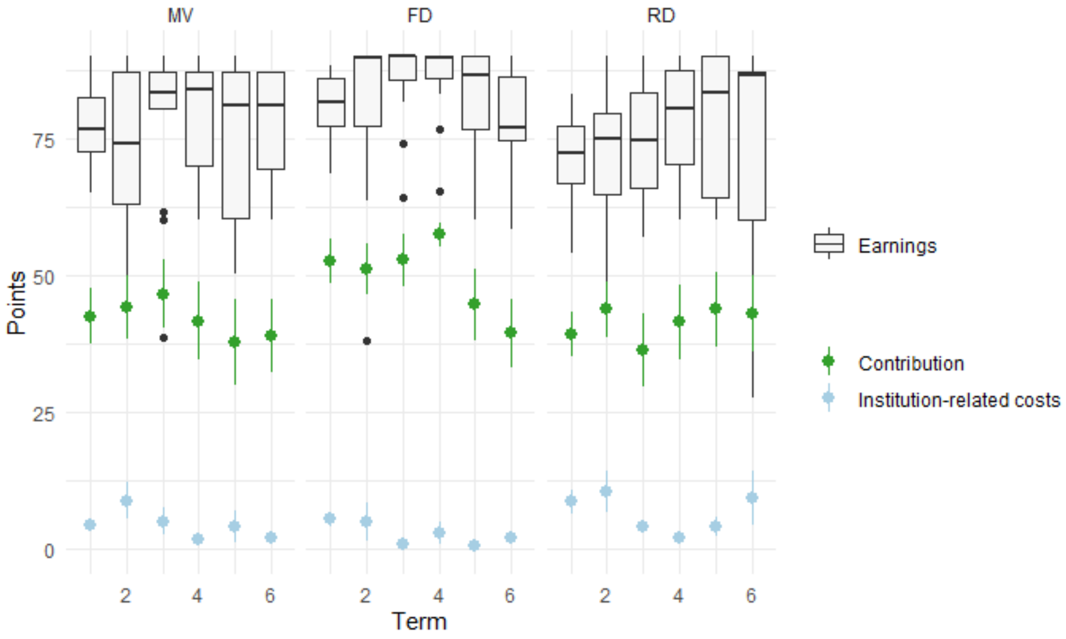


Figure 2.2: Per round group earnings, group contributions, and group institution-related costs over terms.
Note: The tails of the dots cover one standard deviation. The bold line in the box represents the median earnings. The end of the box shows the first and the third quartiles. The vertical extreme lines show the highest and lowest earnings excluding outliers.

group contributions to the public good are between 28 and 59 points, with an average group contribution of 44.18 points, which is approximately 73.6% of the group endowment.

Contributions to the public good do not demonstrate a declining trend over time. This result is not in line with the stylized fact of declining cooperation found in repeated public goods experiments (Ledyard, 1994; Fischbacher and Gachter, 2010), but such a pattern is typically observed in public good games with punishment. In the FD treatment, cooperation even seems to be climbing up until round 12. In the very last term, we observe a sharp falling of group contributions in the FD and MV treatment, which suggests that cooperation in previous rounds is more of a strategic play rather than purely driven by social preferences.

Interestingly, we observe a declining trend within terms in the MV and RD treatment. In these two treatments subjects contribute significantly less in the third round of a term than in the first round of that same term (both treatments $p < 0.01$, $n = 13$); we may call this the “draining” effect. In contrast, they contribute significantly more upon entering a new term, compared with their contributions in the last round of the previous term (both treatments $p < 0.01$,

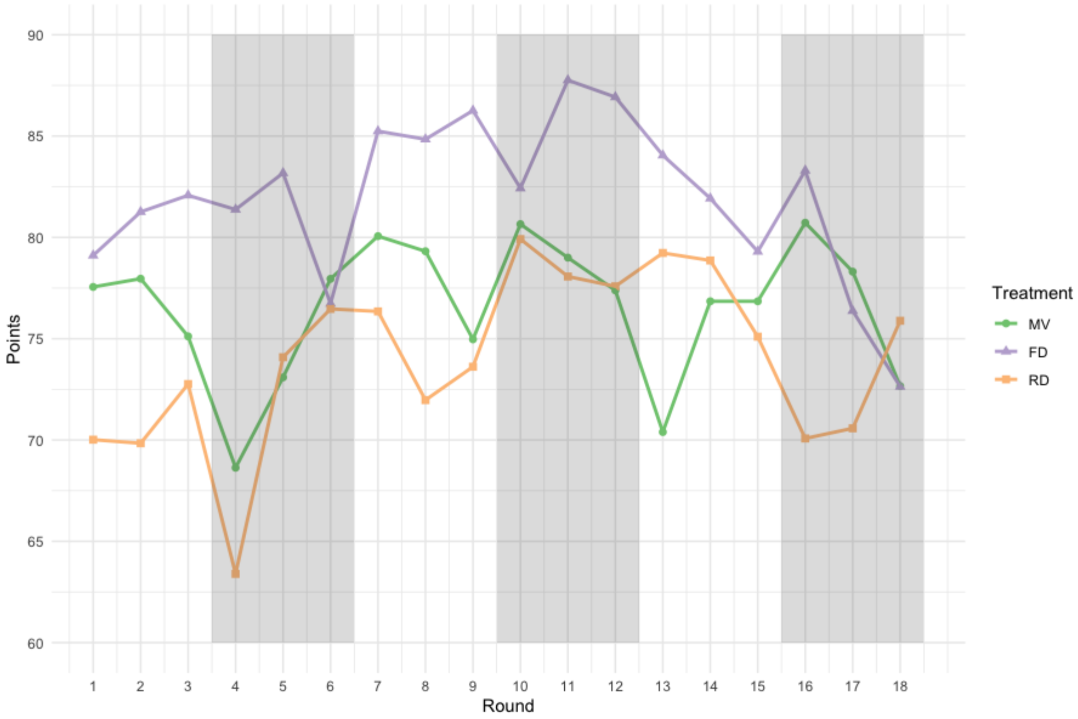


Figure 2.3: Group contributions over rounds

$n = 13$); we call this the "boosting" effect.[4] Furthermore, we find that a change in the institutions between two consecutive terms is positively correlated with the boosting effect and the draining effect. That is, a new set of institutions gives a boost to cooperation, but at the same time cooperation declines more than if the institutions do not change. This suggests a correlation between the stability of institutions and the stability of cooperation behavior. We will consider this in more detail in Section 4.1.2.

Turning to the differences in cooperation across treatments, Figure 2.3 shows that groups in the FD treatment constantly and consistently have higher contribution levels than MV and RD groups with the exception of some of the very last rounds. This can also be seen from Table 7 in Appendix C, which shows the average contributions per term and treatment as well as results of comparisons between treatments. Average group contributions in the FD treatment are about 20% higher than in the other two treatments. Although the differences are substantial in most terms, only the differences between the two treatments with dictatorship are significant in several terms while the differences between

[4] This boosting effect resembles the so-called restart effect (Andreoni, 1988; Cookson, 2000). In our experiment the (new) term was not unexpectedly announced, however, but announced ex ante. Furthermore, it is not a pure restart effect as the rules in the new term may be different, depending on the institution combination selected.

contribution with majority voting and with fixed dictatorship are not.

Result 2.2. *In early terms and over all six terms, groups with fixed dictatorship contribute significantly more than those with rotating dictatorship. Groups with majority voting contribute less than FD and more than RD, but neither difference is statistically significant.*

Institutional choices

We will first look at the overall institutional choices in all six terms. The frequency of each institution combination chosen is displayed in Figure 2.4. Of in total 228 group institutional choices (38 groups x 6 terms), 51.32% are "no institution". The most frequently selected institutions are (only) punishment (14.9%) and (only) communication (11.4%). Then it follows with the combination of punishment and communication (8.3%) and the combination of all three institutions (6.6%). Participants show little interest in reputation, neither alone nor combined with other institutions. The reason could be that reputation in this experiment is rather weak while still costly.[5]

If we look at the institutional choices over time, groups seem to prefer communication in the beginning of the game and then turn to punishment when they approach the end of the game; see Figure 2.5 which shows choices of institutions in the first term (top panel) and last term (bottom panel). One explanation could be that in early rounds, communication may be a useful and powerful tool to build up group morale and mutual trust, while in later rounds the credible threat offered via punishment is more useful. The patterns observed in Figure 2.5 offer some support for such a "carrots first, then sticks" conjecture: between terms 1 and 6 the percentages of groups using communication, either alone or combined with other institutions, decreases strongly (from 66% to 15.8%) while the use of punishment increases (from 26.3% to 45%). The willingness to share the variable costs of both institutions shows a similar development. This also suggests that subjects are (only) willing to pay for the institutions they vote for.

When comparing the use of institutions across treatments, it turns out FD groups use institutions less frequently (24 times) than MV groups (40 times) and RD groups (47 times). Less institution usage means lower institution-related costs, which contributes to the better performance of FD groups, as was shown in the previous section.

[5] Rockenbach and Wolff (2016) find that institution designers tend to provide aggregate contribution information, and leave individual information vague. Our findings do not conflict theirs.

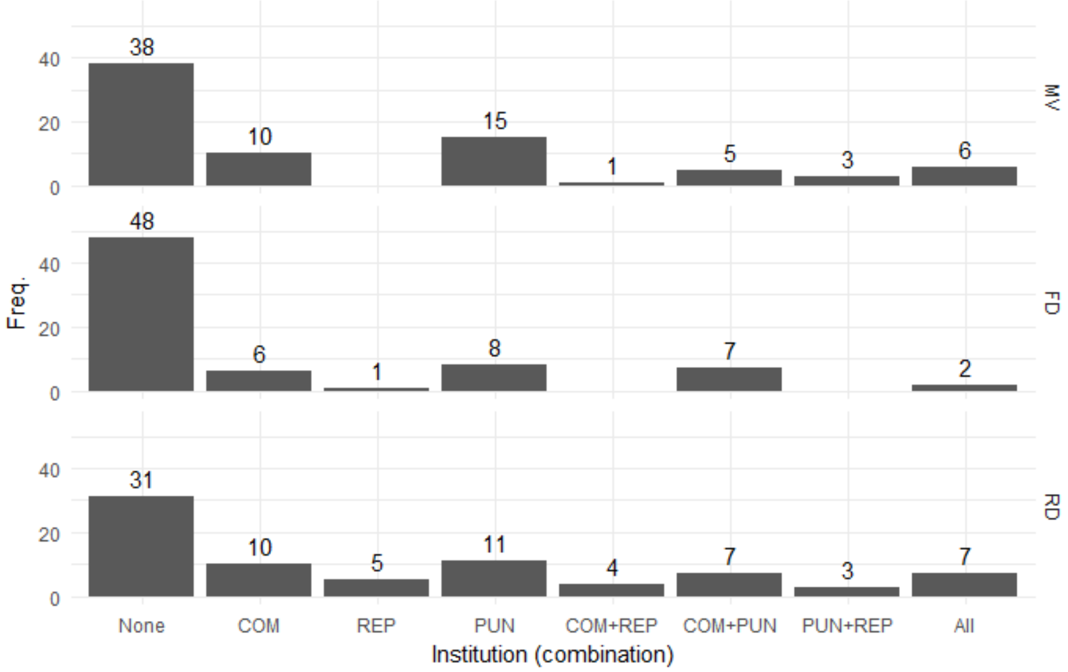


Figure 2.4: Group institutional choices over all six terms

Do collective-choice rules also affect the type of institutions chosen by groups? We compare the first round institutional choices across treatments since they are made individually and simultaneously before any form of interaction. The individual institutional choices in the first round are presented in Figure 11, Appendix C. There is no evidence that initial preferences over institutions differ given different collective-choice rules. Also individual preferences over cost-sharing rules, i.e whether group members are willing to share the variable costs of institutions, do not differ significantly across treatments. Therefore, we find no evidence showing that different collective-choice rules might induce different preferences over institutions.

2.4.2 Direct and indirect effects of collective-choice rules

We now test our hypotheses regarding the direct and indirect effects of collective-choice rules.

Direct effect

The direct effect of collective-choice rules is the influence of these rules on contribution behavior conditional on institutions. To test whether there is difference in contributions between treatments, given the same institution, we run a regres-

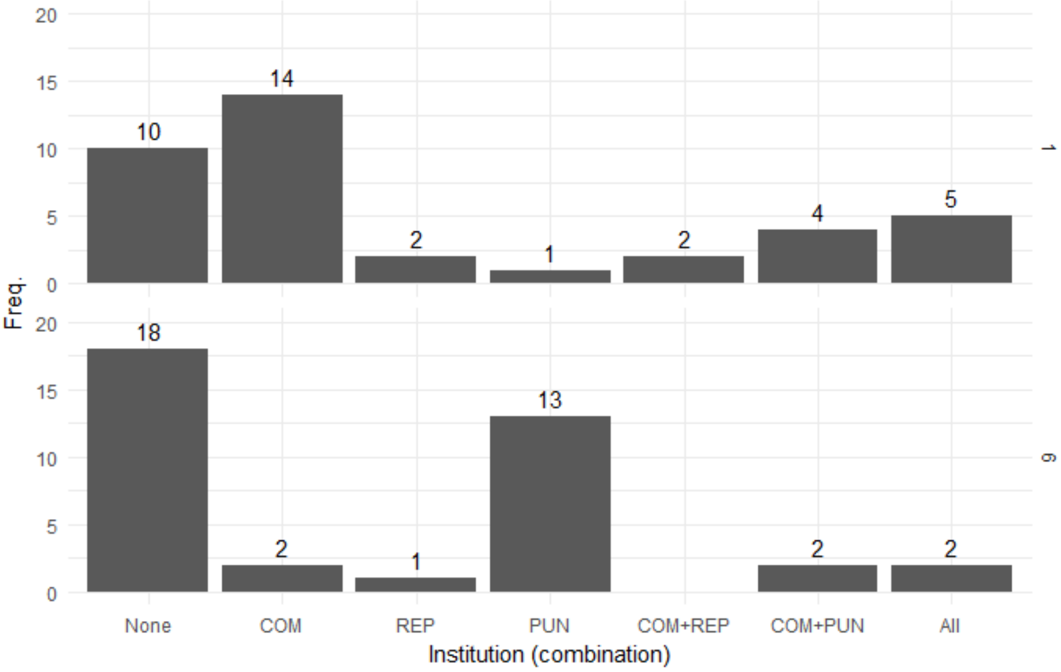


Figure 2.5: Group institutional choices in term 1 and term 6

sion on individual contribution behavior in the first round.[6]

We estimate three specifications using a censored tobit model. The first specification takes the form

$$x_{ij1} = \alpha + \beta_1 MV + \beta_2 RD + \epsilon_{i1} \tag{2.1}$$

where x_{ij1} is the contribution to the public good of individual i in group j, round 1. MV and RD are dummy variables indicating whether the individual is in treatment MV and whether the individual is in treatment RD respectively. The benchmark treatment in the regression is FD. The second specification is identical to the first one, except that it also takes into consideration the institutional choices of a group in round 1

$$x_{ij1} = \alpha + \beta_1 MV + \beta_2 RD + \delta_1 COM_{j1} + \delta_2 PUN_{j1} + \delta_3 REP_{j1} + \epsilon_{i1} \tag{2.2}$$

where COM_{j1}, PUN_{j1}, and REP_{j1} represent the establishment of the 3 available institutions in group j, term 1. Our last specification controls for more institutional features and some individual characteristics, in addition to the second

[6] Using only first-round data is a rather strict but clean test, as it rules out path dependency which may occur when using data from all rounds. Tobit is used because a substantial fraction of contributions is at the boundaries, 0 or 20.

specification

$$\begin{aligned} x_{ij1} =& \alpha + \beta_1 MV + \beta_2 RD + \delta_1 COM_{j1} + \delta_2 PUN_{j1} \\ & + \delta_3 REP_{j1} + \sigma_1 Share\_COM_{j1} + \sigma_2 Share\_PUN_{j1} \\ & + \eta_1 gender_i + \eta_2 experience_i + \epsilon_{i1} \end{aligned} \tag{2.3}$$

where Share_COM$_{j1}$ and Share_PUN$_{j1}$ denote whether the variable costs of communication and punishment are shared among group members, respectively. $experience_i$ is a dummy variable denoting whether the individual has experience in similar experiments, and gender is a dummy variable taking value 1 for females.

Results are reported in Table 4.1. The results show that subjects in treatment FD contribute more than in the other two treatments and significantly so in treatment RD, in line with Result 2.2. This positive effect of fixed dictatorship on contributions persists in all model specifications, so even after controlling for institutions. We also find that the establishment of communication raises the contribution level significantly. Of the other variables only gender has a significant effect; women contribute significantly more than men in the first round.

Result 2.3. *Fixed dictatorship has an immediate and significantly positive direct effect on contributions compared with rotating dictatorship. Fixed dictatorship also has a positive direct effect on contributions compared with majority voting but it is not statistically significant. Contributions with majority voting do not differ from contributions with rotating dictators significantly.*

Our first hypothesis is thus not supported. We do not find a positive direct effect of democratic collective-choice rules on cooperation. On the contrary, the direct effect, if any, works in the opposite direction as predicted, with highest contributions with fixed dictatorship.

Indirect effect

To test our second hypothesis, we need a measure of the stability of groups' institutional choices over time. Formally, let $\mathbf{p} = (p_{COM}, p_{REP}, p_{PUN})$ be an ordered triple in a 3-dimensional Euclidean space, with each component $p \in \{0, 1\}$ representing the individual or group choice on the establishment of one institution. For example, $\mathbf{p_{jt}} = (1, 0, 1)$ indicates that the institutional choice of group j in term t is communication and punishment. We then compute the linear distance between the institutional choices in two consecutive terms and denote it as the

Table 2.2: Determinants of contribution in 1st round

	Dependent variable:		
	Contribution		
	(1)	(2)	(3)
MV	-5.964	-3.720	-3.262
	(5.269)	(4.716)	(4.633)
RD	-13.183^{**}	-12.446^{***}	-10.694^{**}
	(5.213)	(4.788)	(4.657)
Communication		16.538^{***}	14.563^{***}
		(4.028)	(4.865)
Punishment		0.339	1.527
		(4.557)	(5.759)
Reputation		6.579	6.065
		(4.815)	(4.956)
Share_COM			2.431
			(4.957)
Share_PUN			-1.907
			(5.860)
Female			7.581^{**}
			(3.727)
Experience			2.902
			(3.840)
logSigma	2.889^{***}	2.714^{***}	2.676^{***}
	(0.148)	(0.145)	(0.145)
Constant	31.821^{***}	17.981^{***}	12.094^{**}
	(4.548)	(4.172)	(5.235)
Observations	114	114	114
Akaike Inf. Crit.	402.159	382.923	385.595
Bayesian Inf. Crit.	413.104	402.077	415.694
Note:			*p<0.1; **p<0.05; ***p<0.01

group *instability index* $index_{jt} = \|\mathbf{p_{j,t-1}}, \mathbf{p_{j,t}}\|$. The overall instability of group j's institutional choices over time is defined by adding up the instability index from term 2 to term 6: $instability_j = \sum_{t \in 2,\ldots 6} \|\mathbf{p_{j,t-1}}, \mathbf{p_{j,t}}\|$.[7]

To examine the stability of institutions we first present for each treatment a complete and detailed picture of how institutions evolve over time, see Figures 2.6, 2.7, and 2.8. On theses graphs, each sub-plot shows the institutional choices of one group over six terms. Note that although the combinations on the vertical axis are not completely ordered, broadly speaking higher points correspond to more institutions. The color of dots denotes the instability index of the institutional choice in a term. A red dot suggests that institutions have changed a lot from previous term while a blue dot signals stability.

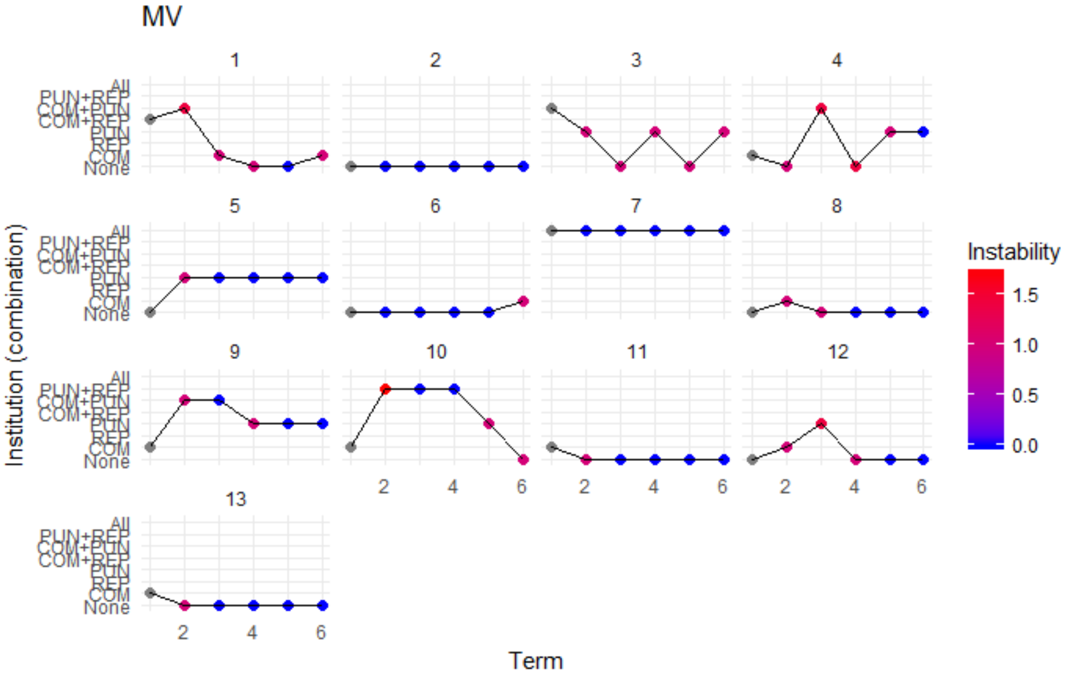


Figure 2.6: Evolution of institutions in treatment MV

Visual inspection shows that institutional choices of the FD groups are the most "stable" as the graph of treatment FD looks more blue. 7 out of 12 FD groups converge to no institutions immediately after the second term. In sharp contrast, the RD groups seem to be struggling with designing their (best) institutions. There is frequent and seemingly random switching among different institution combinations. What happens in the MV groups is somewhere in between the case of the FD and RD groups. The results of Mann-Whitney U tests

[7] Because all individuals are asked to indicate their institutional choices in every term, we could compute an instability index at the individual level, but we focus on group (in)stability.

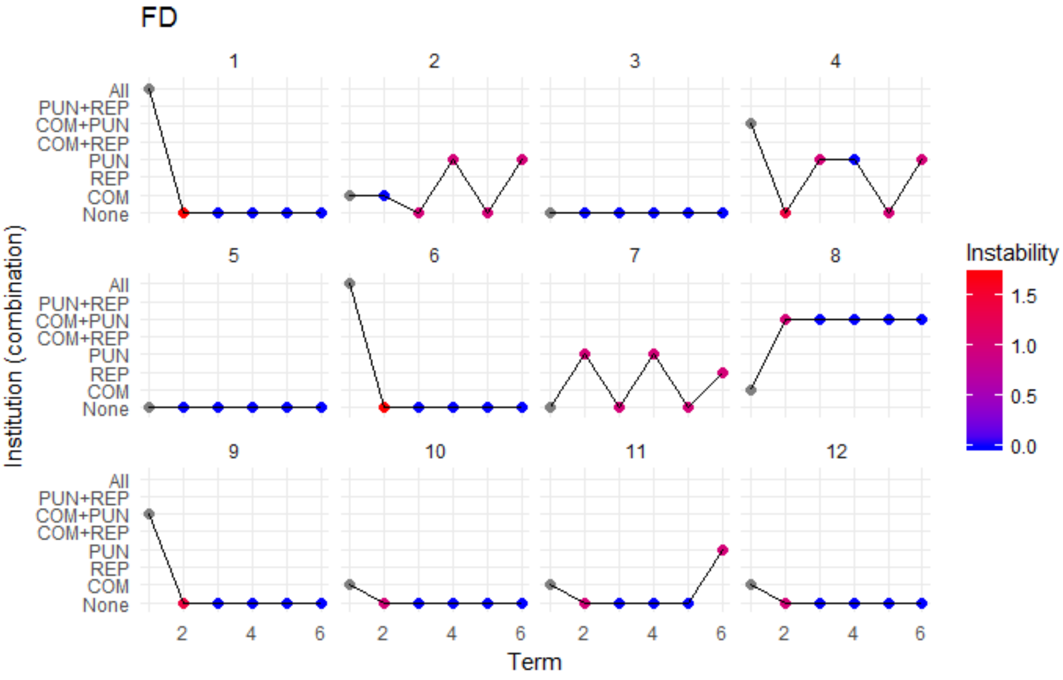


Figure 2.7: Evolution of institutions in treatment FD

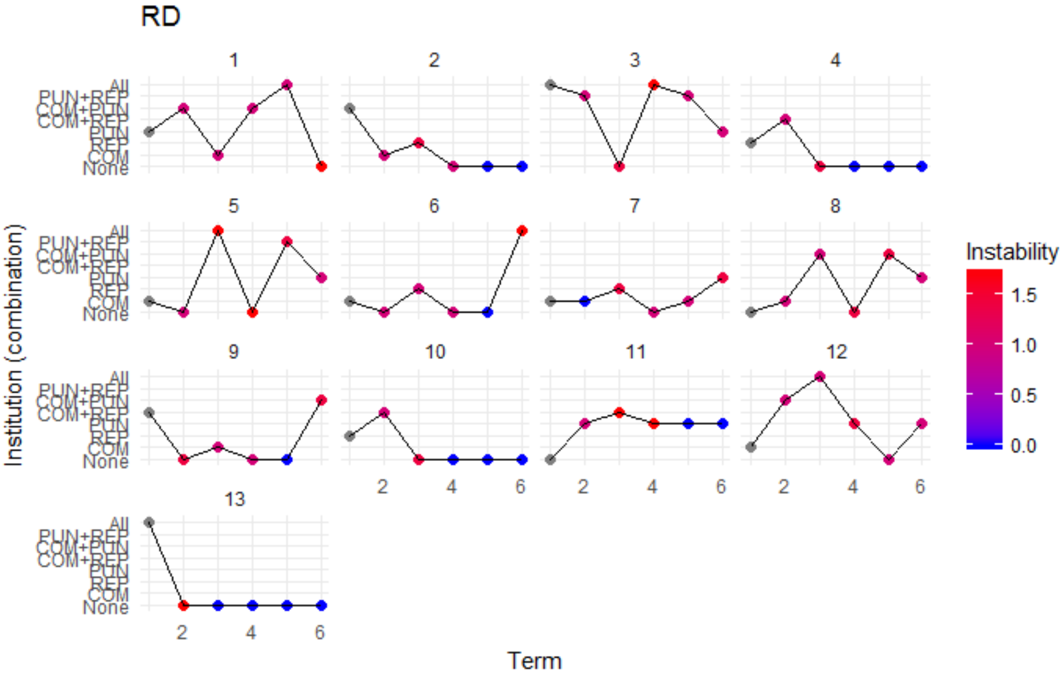


Figure 2.8: Evolution of institutions in treatment RD

on the instability index confirm that institutional choices in FD and MV groups are indeed significantly more stable than in RD groups (both differences $p < 0.01$, $n = 12$ for FD, $n = 13$ for MV and RD) while the difference between FD and MV groups is not significant ($p = 0.74$).

However, simple comparisons of the overall instability across treatments may suffer from endogeneity problems. For example, among the many factors that may influence a group's institutional design, cooperation history is an important one. A group might decide to try a new set of institutions because its members are currently not contributing enough. At the same time, low cooperation might be traced back to the direct effect of collective-choice rules in the very beginning. Therefore, instability could stem from the direct effect of collective-choice rules, rather than the inherent instability of collective-choice rules.

Ideally, we want to control for all other factors which are possibly correlated with collective-choice rules and which could affect institutional choices, such as previous cooperation history and group morale, and only vary the collective-choice rule. One way to achieve this is to construct counterfactual group institutional choices. A counterfactual group institutional choice is the institutional choice a group would have selected if its members' preferences were aggregated via a different collective-choice rule. We can construct these counterfactual choices by exploiting the individual institutional choices of subjects in treatment MV. Because all the decisions in this treatment are made in an incentivized manner, the individual institutional choices can be aggregated *ex post* in whichever way we want. We use each individual's six institutional choices to compute a counterfactual FD instability index, as if they were the dictators. This gives us three counterfactual FD indices for each group, one per subject. As to the counterfactual RD index, we first generate all possible permutations of random dictators for a group according to our rotating rule specified in section 3. We compute the counterfactual RD instability index in each permutation and then take the average of the counterfactual instability index for all permutations. The counterfactual instability indices constructed in this way allow us to do paired difference test on the stability of institutions across treatments while keeping other things constant.

The results of these exercises provide strong evidence showing that institutional choices produced by rotating dictatorship are significantly less stable than by fixed dictatorship (Wilcoxon signed-rank test, $p < 0.01$, $n = 39$) and significantly less stable than by majority voting (Wilcoxon signed-rank test, $p < 0.01$, $n = 13$). But there is no evidence suggesting any difference in stability between majority voting and fixed dictatorship (Wilcoxon signed-rank test, $p = 0.844$, $n = 39$).

Result 2.4. *Institutional choices generated by rotating dictators are significantly less stable than those generated via majority voting or fixed dictatorship. The stability of institutional choices does not differ between majority voting and fixed dictatorship.*

Next, we explore the relationship between instability and cooperation (Hypothesis 2.2b). We estimate the following specification for all groups using a linear model[8] and using data from all rounds

$$\begin{aligned} X_{jt} =& \alpha + \lambda instability_{jt} + \gamma X_{j,t-1} + \theta_t + \delta_1 COM_{jt} + \delta_2 PUN_{jt} + \delta_3 REP_{jt} \\ &+ \beta_1 MV + \beta_2 RD + \epsilon_{jt} \end{aligned} \tag{2.4}$$

where X_{jt} is the per round contributions of group j averaged over the three rounds in term t. $instability_{jt}$ is the instability index which measures how much the institutional choice(s) of group j change from term $t-1$ to term t, as defined previously. COM_{jt}, PUN_{jt}, and REP_{jt} stand for the establishment of the three institutions in group j, term t. θ_t is term fixed effects, capturing the trend of cooperation over time. And MV and RD are dummy variables indicating the collective-choice rule group j is using.

Table 2.3 presents the results. The instability of institutional choice has a negative effect on group contributions. The magnitude of this effect is relatively small but significant. The establishment of punishment significantly increases group contributions by 21.6 points, and reputation decreases group contributions by 6.2 points. The results combined with the group institutional choices presented in the previous section justify our argument that institutional design can be seen as a problem-solving process: Groups try to find out which institutions work best in terms of promoting cooperation. Institutions that work (punishment) are chosen more frequently, while institutions that don't work lose their places. The results also show that group contributions are highly correlated with contributions in the last term, with a coefficient of about 0.66.[9]

Result 2.5. *The instability of institutional choice decreases subsequent group contributions.*

[8] The dependent variable here is group contributions, which are never censored. Therefore, we use a linear model instead of a censored tobit model.

[9] The results do not depend on the specifications; the signs and significance of variables are very similar if group contributions in the last term are not included.

Table 2.3: Instability and contributions

	Dependent variable:
	Group Contributions
Instability	-3.876^{**}
	(1.963)
Group contributions in the last term	0.657^{***}
	(0.051)
Term	-2.175^{***}
	(0.785)
Communication	2.014
	(2.814)
Punishment	21.623^{***}
	(2.414)
Reputation	-6.159^{*}
	(3.418)
MV	-4.826^{*}
	(2.588)
RD	-0.773
	(2.781)
Constant	20.510^{***}
	(4.938)
Observations	190
Adjusted R^2	0.601
Note:	*p<0.1; **p<0.05; ***p<0.01

Our second hypothesis is thus partially supported. Rotating dictatorship produces unstable institutional choices, compared with majority voting and fixed dictatorship. Instability is associated with a lower cooperation level.

2.5 Discussions: What drives the direct effect

Result 2.3 offers no support for Hypothesis 2.1, which stated that democratic participation right would increase compliance with the selected institutions and results in higher contributions. In this section, we try to explore why this may be the case. Among the many potential explanations we focus on two, namely communication and leadership. The first conjecture points to the use and in particular the *content* of communication as a coordination device. The messages subjects exchange in the communication phase may account for the difference between FD and RD groups. A second conjecture is that the dictators in the FD treatment act (more) as *de facto* leaders. They might take their responsibilities in promoting group cooperation by sending messages and making larger contributions than non-dictators. As a result, the whole group benefits from having a leading figure, even though they cannot identify which member is the leader. We find some evidence supporting the first conjecture and evidence against the second.

2.5.1 The usage and the content of communication

In section 4, we have shown that in the first round communication increases contributions to the public good. Now we consider in more detail the usage and the content of the messages in this round.

The first three rows of Table 2.4 give a summary of the usage of communication across treatments: the percentage of groups that have established communication, the percentage of individuals who send any message at all, and the number of messages sent sorted by content. Fixed dictators choose to establish communication more frequently and send more messages than groups in the MV and RD treatments. Remember that the content of a messages is semi-specified. It can be a proposal for certain amount of contribution to the public good, a proposal for the establishment of certain institution combinations, or a proposal for a certain cost-sharing rule. Table 2.4 also shows that the types of messages sent are similar across treatments.

Table 2.4: Usage of communication in round 1

	MV	FD	RD
% groups using communication	61.5	75	61.5
% individuals sending any message	51.3	58.3	51.3
NO. messages sent	39	38	33
message about contribution	18	19	20
message about institution	13	10	7
message about sharing-rule	8	9	6

What might matter more than the number of messages sent is the specific content of the message, and the proposals for contribution in particular. Table 2.5 presents the frequency of the proposed amount of contribution in round 1. We immediately notice that in the RD treatment less people propose to contribute the full endowment to the public good. Results from Fisher's exact tests confirm that the proposals for contribution are significantly different between MV and RD groups ($p = 0.08$, $n = 38$), and between FD and RD groups ($p = 0.04$, $n = 39$).

Previous findings from economic experiments demonstrate that a large fraction of the population can be categorized as conditional cooperator (Fischbacher

Table 2.5: Frequency of individual proposed contribution in round 1

proposed amount	MV	FD	RD
0	0	0	1
3	0	1	0
5	0	0	1
8	0	1	0
10	1	0	5
15	0	1	1
20	17	16	12
N	18	19	20

et al., 2001; Fischbacher and Gachter, 2010; Gächter and Renner, 2018; Neugebauer et al., 2009; Thöni and Volk, 2018): they cooperate only if they expect others to cooperate. Seeing selfish signals, group members may lower their expectations of how much others would contribute, and this may result in a lower group cooperation level. Results of a regression of actual contributions on the minimal proposed amount of contribution in a group (not reported here) suggest that a one point decrease in the minimal proposed amount of contribution decreases actual contributions significantly, by 1.07 point on average ($p < 0.01$). The fact that group members in the RD treatment observe less (very) positive signals may explain, at least partly, why RD groups cooperate at a lower level.

2.5.2 The leadership effect

To examine whether some sort of leadership effect (Moxnes and Van der Heijden, 2003; Gächter and Renner, 2018; Güth et al., 2007) exists and whether it is stronger in the FD treatment than in the RD treatment, we compare the behaviors of dictators and non-dictators in all rounds with communication.

First, dictators are more willing to send messages to their group members than non-dictators when communication is possible. This finding also holds for both FD and RD sub-samples separately. But the dictators in the FD treatment are not more willing to communicate than those in the RD treatment. Second, dictators in both FD and RD treatments propose a higher amount of contribution than non-dictators do, but the contribution proposals by the dictators in the FD treatment and in the RD treatment are indistinguishable. Finally, dictators behave similarly as non-dictators in terms of actual contributions to the public good in treatments FD and RD. Taken together, this implies that there indeed seems to be a “leadership” effect in both treatments with dictatorship, albeit weak. The dictators are more willing to communicate and propose larger amount of contributions but are not actually contributing more than non-dictators. More importantly, we find no evidence that this leadership effect is stronger in the FD treatment than in the RD treatment. Therefore, the higher cooperation level observed in the FD treatment compared with the RD treatment cannot be attributed to a (stronger) leadership effect in the first treatment.

2.6 Conclusion

This paper studies the role of collective-choice rules in a problem-solving decision environment where groups can exploit the possibility of institutional design to achieve self-governance. We test if collective-choice rules directly affect cooperation after controlling for institutions and if collective-choice rules indirectly affect cooperation through the stability of institutional choices. Specifically, we study majority voting, dictatorship, and rotating dictatorship. The Institutional Design Game we use is based on a standard public goods game. Groups repeatedly select institutions they would like to implement via a given collective-choice rule and play the self-tailored public goods game with the chosen institutions. Institutions are costly.

Our main findings are: (1) Cooperation level is not higher when the institutions are chosen via a democratic rule than when the same institutions are chosen via a non-democratic rule. On the contrary, groups with a fixed dictator cooperate at a higher level than those with rotating dictators, conditional on institutions. (2) Institutional choices chosen by majority voting or by a fixed dictator are more stable over time than those chosen by rotating dictators. (3) The instability of institutions is associated with lower cooperation level.

Previous literature has established an "endogeneity premium" (Dal Bó et al., 2010), namely that institutions chosen by the subjects themselves are more effective than the same institutions imposed on them. The "endogeneity premium" is widely interpreted as the merit of democratic participation in group decision-making. But it could also be that endogenous institutions work better because they are chosen by in-group members whose payoffs are at stake. We exclude this possibility by comparing institutions that are all chosen by subjects themselves but through democratic or non-democratic rules. We do not find evidence of a "democracy premium" (institutions chosen via democratic rules are more effective than chosen via non-democratic rules). Our results do not refute the findings of an "endogeneity premium", but suggest further investigation into the underlying mechanisms through which democratic participation right is transmitted to higher cooperation.

Our second finding concerns the stability of collective decisions in a complex problem-solving setting where the consequences of these decisions are uncertain and individual preferences are not necessarily fixed. In the presence of this complexity, one does not expect widespread agreement on the optimal collective choice. Therefore, collective choice is likely exposed to greater variance when the decision rights are more dispersed. We show that indeed decisions are most unstable when they are made by individuals on a rotating basis. Yet we detect no significant difference in the stability of institutional choices between majority voting and fixed dictatorship. Future work could look at whether the difference in stability persists as the size of group grows and as the complexity of decision environment increases. Finally, in our environment, unstable institutions is not constructive to cooperation.

Taken together, our results indicate that certain collective-choice rules may matter for the effectiveness of collective decisions. It helps to think about the value of democratic participation right, the cost of instability, and the possible tradeoff between the two before a group sets the collective-choice rule for making collective decisions.

Of course, there are several questions remaining to be answered. For example, concerning the direct effect, it is a bit surprising that individuals in the treatment of rotating dictatorship send more selfish signals than those in other treatments. It could be a coincidence given the relatively small sample. Alternatively, one may suspect that subjects with rotating dictators anticipate the instability of their institutional environment and that the anticipation of instability already hurts their willingness to fully cooperate. As to the indirect effect, our setting suffers from endogeneity issue as many empirical studies do. If we want to learn more about the causal relationship between stability and cooperation, we need an experimental design which is more tailored to this purpose. Finally, the direct and indirect effects observed here may serve as starting points for studying and better understanding the role of collective-choice rules. It is not clear, for example, whether the impact of collective-choice rules will be stronger or weaker, as the decision environment becomes more complex. All these aspects may be examined in further research.

Appendix A. Experiment Procedure

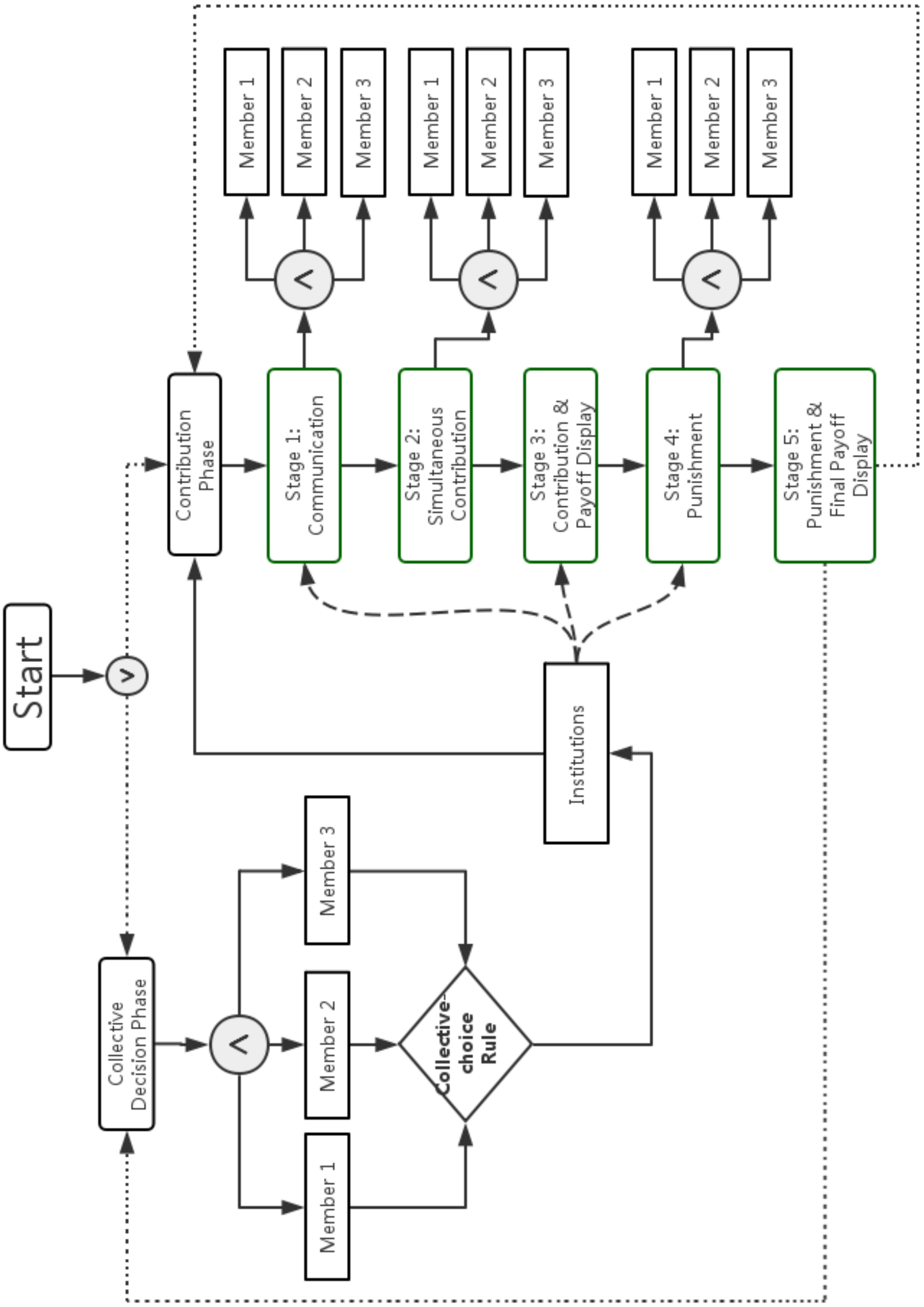


Figure 9: Complete experimental procedure

Appendix B. Instructions

Welcome and thanks for your participation. You receive €3 for having shown up on time. If you read these instructions carefully, you can earn more. Your earnings will be paid out to you in cash immediately after the experiment.

It is strictly forbidden to communicate with the other participants during the experiment. If you have any questions or concerns, please raise your hand. We will answer your questions individually.

The experiment consists of 18 rounds. It is divided into 6 terms. Each consists of 3 rounds. You will interact with two other participants. The three of you form a group that will remain the same in all 18 rounds. You will never know which of the other participants are in your group.

The sum of the 18 round payoffs will determine your final earnings. 100 points in the experiment are converted to 3 Euros: 100 points = €3.

Basic Game

This is the game that you play every round. Each participant will be (re)assigned as either Member 1, Member 2, or Member 3. This identity changes every round.

In each round, each of you receive 20 points. In the following, we shall refer to this amount as your endowment. **Your task is to decide how much of your endowment you want to invest in a project.** After investment, the rest of your endowment is kept for yourself.

One POINT, no matter invested by whom, in the project gives 0.5 POINT to each member in your group. Therefore how much you earn from the project depends on both your own investment and your group members' investment. To calculate how much you earn from the project:

1. Investments into the project from the three group members are added up as the total group investment.

2. Each member in this group receives half amount of the total group investment.

The amount that you keep for yourself remains the same. Your profit is therefore the sum of the following two parts:

Profit = 0.5 * total group investment + the amount that you keep for yourself

You and your group members make this decision at the same time without knowing others' choices. The individual investments and profits of all group members will be shown to you at the end of each round.

Now please look at the screen and answer a few questions. These questions aim at making sure that you understand the game. They do not influence your earnings. If you have any questions, please raise your hand.

Possible Extensions

As mentioned, the experiment will consist of 6 terms of 3 rounds. At the beginning of each term, you are able to change the rules of the basic game by using one or more of the following three Mechanisms:

Mechanism A: It is used before investment. You are able to exchange messages with your group members before making your investment. There are at most three pieces of message you can send, which will be listed below.

Mechanism B: It is used after investment, when investments and profits are shown to you. You are able to see not only the investments and profits in the current round, but also the history of investments and profits of your group members in the previous rounds starting from this term. Your identity (Member 1, Member 2, or Member 3) will not change during this term.

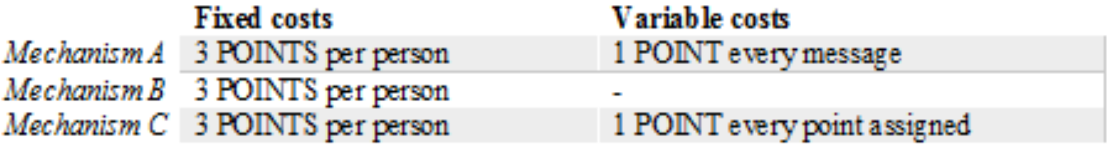
Mechanism C: It is used after seeing your investments and profits. You are able to assign points to other group members to reduce their earnings. Each point assigned to a group member will reduce the earnings of this group member by 3 points. You can only assign points to those who invest less than you.

To be able to use these Mechanisms, you will have to pay for the fixed costs as a group. Each Mechanism costs each member of the group 3 points in this term. That is, each of you pay 1 point per round in the term in which the Mechanism is used.

Additionally, Mechanism A and C might incur variable costs, which depend on your actual usage. Within Mechanism A, the variable cost of each piece of message is 1 point. Within Mechanism C, the variable cost of each assigned point is 1 point. For example, if Member 1 assigns 1 point to Member 2, the variable cost of such action is 1 point, whereas the earnings of Member 2 will be reduced by 3 points.

Your group can choose to share the variable costs of the Mechanism A and C equally among all group members or not. If your group chooses to share the variables costs, then no matter who sends a message or who assigns points to others, the variable costs of such actions will be shared equally by all group members. Otherwise the costs of sending a message or assigning points to others will only be burdened by the member who takes the action.

The following table gives you a summary of the costs of each Mechanism.



	Fixed costs	Variable costs
Mechanism A	3 POINTS per person	1 POINT every message
Mechanism B	3 POINTS per person	-
Mechanism C	3 POINTS per person	1 POINT every point assigned

Your task is to decide:

At the beginning of each term

1. With regard to each of the three Mechanisms, would you like it to be used in your group? (Yes/No)

2. With regard to Mechanism A and Mechanism C, would you like the variable costs to be shared equally by all group members? (Yes/No)

In each round

3. (If applicable) Do you have any messages to send? If yes, what would you send?

There are three pieces of message to choose from. You need to fill in the blank before sending. You can choose more than one of them:

- "I propose that we invest ___(integer number from 0 to 20) points in the project."
- "I propose that we later use Mechanism ___ (A/B/C/none)."
- "I propose that we later share the variable cost of Mechanism ___ (A/C/none)."

4. (If applicable) How many points you want to assign to those who invest less than you?

[*FOR MV ONLY*] Whether a Mechanism will be actually used depends on whether a majority in your group agree to use it. That is, if two or more group members vote yes for a Mechanism, it will be used in the following term (three rounds) and all group members will pay for its fixed costs.

For Mechanism A and C, whether the variable costs would be shared equally by all group members depends on whether a majority in your group agree to it.

[*FOR FD ONLY*] One of the group members will be randomly assigned the role of decision maker throughout the 6 terms. Whether a Mechanism will be actually used and whether the variable costs will be shared equally totally depend on the decision maker's choice. You are informed of your role after all group members have made their choices regarding the Mechanisms in term 1.

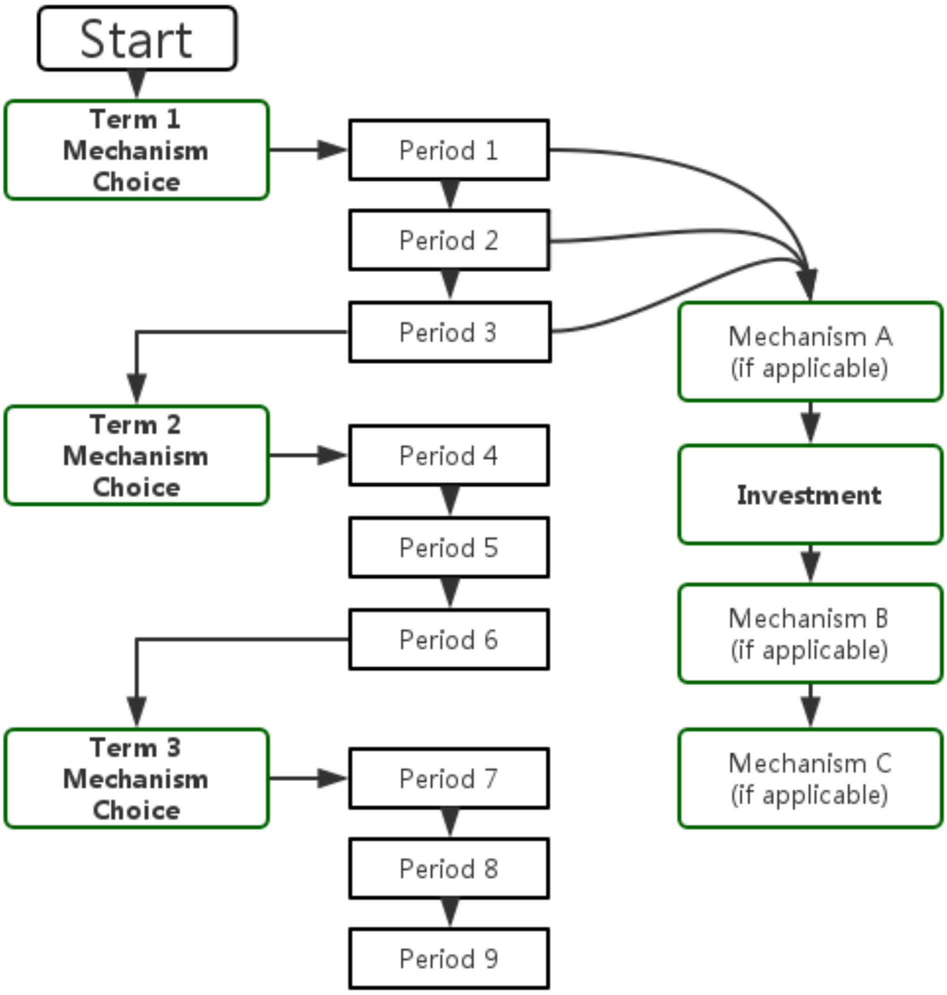
If you are not the decision maker, you are still able to indicate your preferences although your choices will not affect the outcome.

[*FOR RD ONLY*] In each term, one group member will be randomly assigned the role of decision maker. Whether a Mechanism will be actually used and whether the variable costs will be shared equally totally depend on the decision maker's choice. Each of the group members will become the decision maker exactly twice in the six terms. You are informed of your role only after all group members have made their choices regarding the Mechanisms.

If you are not the decision maker, you are still able to indicate your preferences although your choices will not influence the outcome.

How is the game played?

The procedure of the game is shown in the following picture. (Only the first three terms are presented).


Start
Term 1 Mechanism Choice
Period 1
Period 2
Period 3
Term 2 Mechanism Choice
Period 4
Period 5
Period 6
Term 3 Mechanism Choice
Period 7
Period 8
Period 9
Mechanism A (if applicable)
Investment
Mechanism B (if applicable)
Mechanism C (if applicable)

Appendix C. Additional results

Performance over time

We present the average group performance over time of different treatments in Figure 10. The benchmark group earnings are 60 points per round. Figure 10 shows that on average, groups in all treatments manage to earn more than the benchmark in all rounds. In the FD treatment, the average group performance follows a hump-shape pattern. In the MV and RD treatments, there are ups and downs over time with no obvious trend. Groups in treatment FD on average outperform those in treatments MV and RD most of the time.

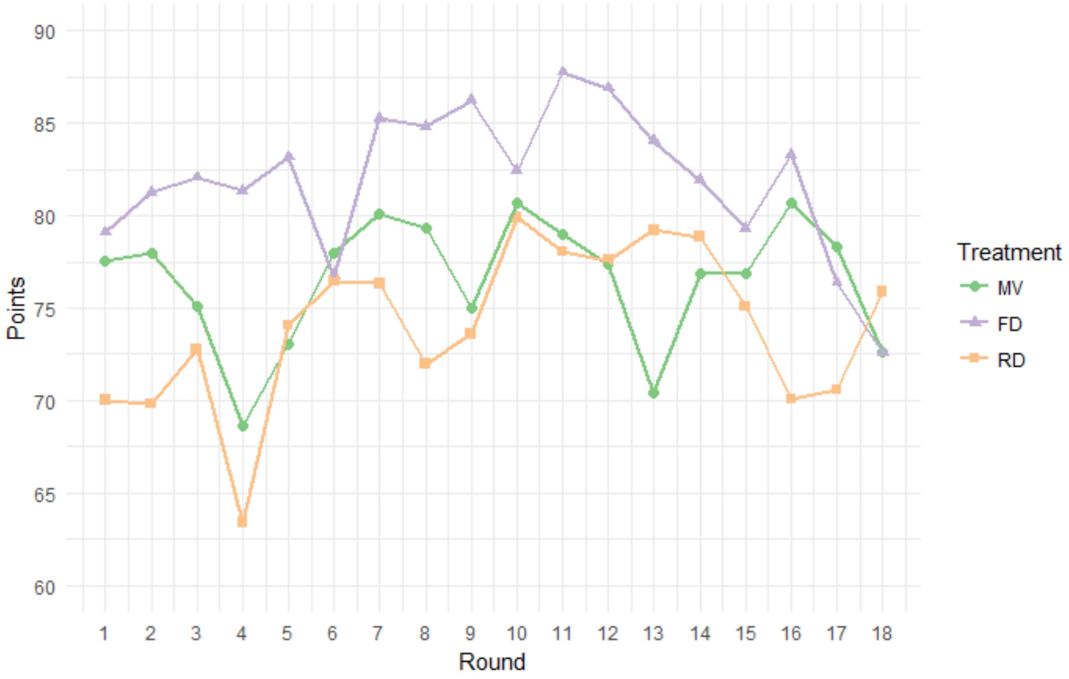


Figure 10: Group earnings over rounds

Decomposition of group performance

Table 6 shows that FD groups on average earn 7.83 more points than RD groups, of which 4.19 points come from the profit side and 3.64 come from the cost side.

Average per term group contributions

Table 7 shows the average per term group contributions. Subjects in the FD treatment contribute more than those in the RD treatment in earlier terms. The

Table 6: Decomposition of group performance

	Earnings			Costs			Profits		
	MV	FD	RD	MV	FD	RD	MV	FD	RD
Term 1	81.28	86.36	79.65	4.41	5.55	8.78	76.88	80.81	70.87
Term 2	82.14	85.57	81.88	8.92	5.14	10.57	73.23	80.43	71.32
Term 3	83.29	86.39	78.18	5.18	0.94	4.20	78.12	85.45	73.98
Term 4	80.86	88.69	80.73	1.85	2.99	2.21	79.01	85.70	78.52
Term 5	78.87	82.33	81.88	4.18	0.58	4.15	74.69	81.76	77.73
Term 6	79.49	79.71	81.56	2.26	2.28	9.38	77.23	77.44	72.18
Avg.	80.99	84.84	80.65	4.46	2.91	6.55	76.53	81.93	74.10

contributions of FD groups are also consistently higher than those in treatment MV, but the differences are not statistically significant.

Table 7: Group contributions over terms

	Group contributions		
	MV	FD	RD
Term 1	42.56	$52.72^{\ddagger}$	$39.31^{\ddagger}$
Term 2	44.28	51.14	43.77
Term 3	46.59	$52.78^{\dagger}$	$36.36^{\dagger}$
Term 4	41.72	$57.39^{\dagger}$	$41.46^{\dagger}$
Term 5	37.74	44.67	43.77
Term 6	38.97	39.42	43.13
Avg.	41.98	$49.69^{\ddagger}$	$41.30^{\ddagger}$

Notes. Two-sided Mann-Whitney U tests. † Significant difference ($p < 0.1$) between FD and RD. ‡ Significant difference ($p < 0.05$) between FD and RD.

First round individual institutional choices

As shown in Figure 11, the initial preference of individuals over institutions does not differ across treatments ($p = 0.412$, $n = 114$). It suggests that collective-choice rules do not affect the type of institutions people choose to use.

References

Aisen, A. and Veiga, F. J. (2013). How does political instability affect economic growth? *European Journal of Political Economy*, 29:151–167.

Alesina, A., Özler, S., Roubini, N., and Swagel, P. (1996). Political instability and economic growth. *Journal of Economic Growth*, 1(2):189–211.

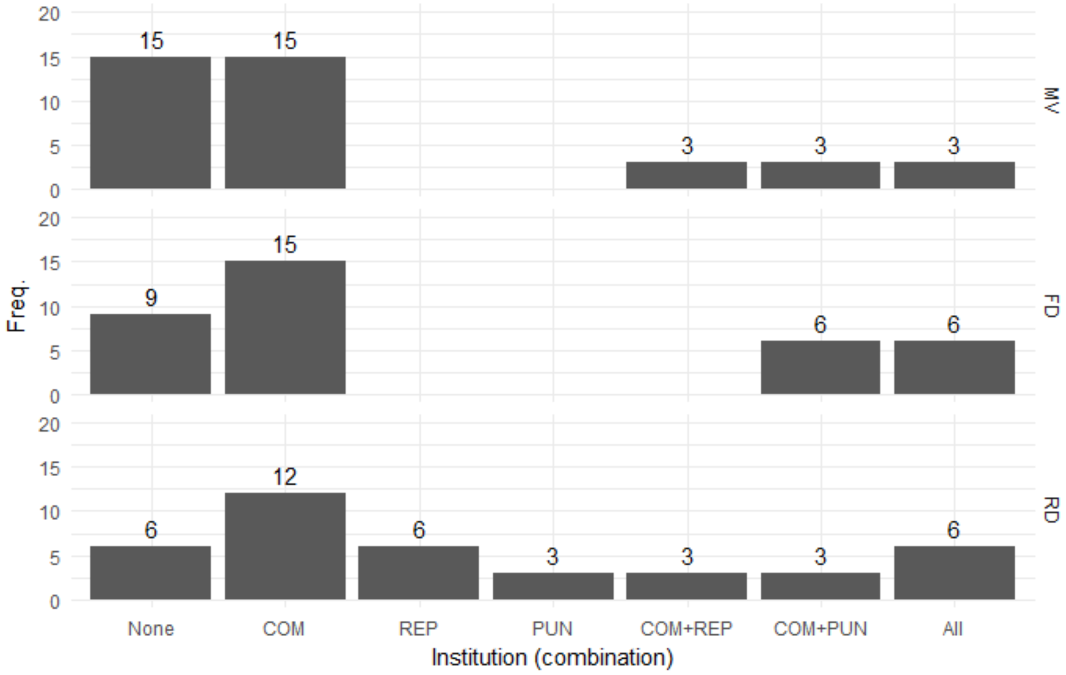


Figure 11: First round individual institutional choices

Andreoni, J. (1988). Why free ride?: Strategies and learning in public goods experiments. *Journal of Public Economics*, 37(3):291–304.

Arbak, E. and Villeval, M. C. (2011). Endogenous leadership selection and influence.

Barro, R. J. (1991). Economic growth in a cross section of countries. *The Quarterly Journal of Economics*, 106(2):407–443.

Bartling, B., Fehr, E., and Herz, H. (2014). The intrinsic value of decision rights. *Econometrica*, 82(6):2005–2039.

Bischoff, I. (2007). Institutional choice versus communication in social dilemmas—an experimental approach. *Journal of Economic Behavior & Organization*, 62(1):20–36.

Buchanan, J. M. and Tullock, G. (1962). *The calculus of consent*, volume 3. University of Michigan press Ann Arbor.

Casari, M. and Luini, L. (2009). Cooperation under alternative punishment institutions: An experiment. *Journal of Economic Behavior & Organization*, 71(2):273–282.

Castore, C. H. and Murnighan, J. K. (1978). Determinants of support for group decisions. *Organizational Behavior and Human Performance*, 22(1):75–92.

Cookson, R. (2000). Framing effects in public goods experiments. *Experimental Economics*, 3(1):55–79.

Dal Bó, P., Foster, A., and Putterman, L. (2010). Institutions and behavior: Experimental evidence on the effects of democracy. *American Economic Review*, 100(5):2205–2229.

Dixit, A. (2009). Governance institutions and economic activity. *American Economic Review*, 99(1):5–24.

Dougherty, K., Pitts, B., Moeller, J., and Ragan, R. (2014). An experimental study of the efficiency of unanimity rule and majority rule. *Public Choice*, 158(3-4):359–382.

Ertan, A., Page, T., and Putterman, L. (2009). Who to punish? individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, 53(5):495–511.

Fehr, E. and Gachter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994.

Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868):137.

Fehr, E., Herz, H., and Wilkening, T. (2013). The lure of authority: Motivation and incentive effects of power. *American Economic Review*, 103(4):1325–59.

Fehr, E. and Williams, T. (2013). Endogenous emergence of institutions to sustain cooperation. Technical report, Working Paper.

Feng, Y. (1997). Democracy, political stability and economic growth. *British Journal of Political Science*, 27(3):391–418.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Fischbacher, U. and Gachter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, 100(1):541–56.

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, 71(3):397–404.

Frey, B. S., Benz, M., and Stutzer, A. (2004). Introducing procedural utility: Not only what, but also how matters. *Journal of Institutional and Theoretical Economics*, 160(3):377–401.

Gächter, S. and Renner, E. (2018). Leaders as role models and 'belief managers' in social dilemmas. *Journal of Economic Behavior & Organization*, 154:321–334.

Guillen, P., Schwieren, C., and Staffiero, G. (2007). Why feed the leviathan? *Public Choice*, 130(1-2):115–128.

Gürerk, Ö., Irlenbusch, B., and Rockenbach, B. (2009). Voting with feet: Community choice in social dilemmas.

Güth, W., Levati, M. V., Sutter, M., and Van Der Heijden, E. (2007). Leading by example with and without exclusion power in voluntary contribution experiments. *Journal of Public Economics*, 91(5):1023–1042.

Haney, P. J., Herzberg, R. Q., and Wilson, R. K. (1992). Advice and consent: unitary actors, advisory models, and experimental tests. *Journal of Conflict Resolution*, 36(4):603–633.

Herrmann, B., Thöni, C., and Gächter, S. (2008). Antisocial punishment across societies. *Science*, 319(5868):1362–1367.

Hoffman, E. and Plott, C. R. (1983). Pre-meeting discussions and the possibility of coalition-breaking procedures in majority rule committees. *Public Choice*, 40(1):21–39.

Kamei, K. (2016). Democracy and resilient pro-social behavioral change: an experimental study. *Social Choice and Welfare*, 47(2):359–378.

Ledyard, J. O. (1994). Public goods: A survey of experimental research.

Markussen, T., Putterman, L., and Tyran, J.-R. (2013). Self-organization for collective action: An experimental study of voting on sanction regimes. *The Review of Economic Studies*, 81(1):301–324.

McKelvey, R. D. and Ordeshook, P. C. (1984). An experimental study of the effects of procedural rules on committee behavior. *The Journal of Politics*, 46(1):182–205.

Moxnes, E. and Van der Heijden, E. (2003). The effect of leadership in a public bad experiment. *Journal of Conflict Resolution*, 47(6):773–795.

Neugebauer, T., Perote, J., Schmidt, U., and Loos, M. (2009). Selfish-biased conditional cooperation: On the decline of contributions in repeated public goods experiments. *Journal of Economic Psychology*, 30(1):52–60.

Nikiforakis, N. and Normann, H.-T. (2008). A comparative statics analysis of punishment in public-good experiments. *Experimental Economics*, 11(4):358–369.

Nitzan, S. and Paroush, J. (1982). Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, pages 289–297.

Plott, C. R. and Levine, M. E. (1978). A model of agenda influence on committee decisions. *The American Economic Review*, 68(1):146–160.

Potters, J., Sefton, M., and Vesterlund, L. (2005). After you—endogenous sequencing in voluntary contribution games. *Journal of Public Economics*, 89(8):1399–1419.

Riker, W. H. (1980). Implications from the disequilibrium of majority rule for the study of institutions. *American Political Science Review*, 74(2):432–446.

Rivas, M. F. and Sutter, M. (2011). The benefits of voluntary leadership in experimental public goods games. *Economics Letters*, 112(2):176–178.

Rockenbach, B. and Wolff, I. (2016). Designing institutions for social dilemmas. *German Economic Review*, 17(3):316–336.

Shapley, L. and Grofman, B. (1984). Optimizing group judgmental accuracy in the presence of interdependencies. *Public Choice*, 43(3):329–343.

Sigmund, K., De Silva, H., Traulsen, A., and Hauert, C. (2010). Social learning promotes institutions for governing the commons. *Nature*, 466(7308):861–863.

Sorkin, R. D., Hays, C. J., and West, R. (2001). Signal-detection analysis of group decision making. *Psychological review*, 108(1):183.

Sutter, M., Haigner, S., and Kocher, M. G. (2010). Choosing the carrot or the stick? endogenous institutional choice in social dilemma situations. *The Review of Economic Studies*, 77(4):1540–1566.

Thöni, C. and Volk, S. (2018). Conditional cooperation: Review and refinement. *Economics Letters*, 171:37–40.

Tyran, J.-R. and Feld, L. P. (2006). Achieving compliance when legal sanctions are non-deterrent. *The Scandinavian Journal of Economics*, 108(1):135–156.

Walker, J. M., Gardner, R., Herr, A., and Ostrom, E. (2000). Collective choice in the commons: Experimental results on proposed allocation rules and votes. *The Economic Journal*, 110(460):212–234.

Wilson, R. K. (1986). Forward and backward agenda procedures: committee experiments on structurally induced equilibrium. *The Journal of Politics*, 48(2):390–409.

Zhang, B., Li, C., De Silva, H., Bednarik, P., and Sigmund, K. (2014). The evolution of sanctioning institutions: an experimental approach to the social contract. *Experimental Economics*, 17(2):285–303.

Chapter 3

The relativity of moral judgments

3.1 Introduction

People are constantly making judgments as to whether an action, a motive, or a person is kind or unkind. They evaluate the kindness of the available options and derive utility from doing good and disutility from doing bad beyond monetary payoffs (Andreoni, 1990; Crumpler and Grossman, 2008). They also evaluate the kindness of whom they interact with and reciprocate based on those moral judgments in social interactions (Fehr and Fischbacher, 2004; Henrich et al., 2006; Nowak and Sigmund, 2005).

Economic literature often sees moral judgments as static and internal: one evaluates an action or a person who chooses that action solely based on how she perceives the nature of the action. Thus, moral judgments are invariant to other external factors, such as the actions of other people in the same decision environment [1]. For instance, in Abeler et al. (2019); Gneezy et al. (2018), moral judgments about a person are modeled as a function of whether this person lies or not. This one-to-one relation may not hold true in a social environment in which more than one decision maker is involved.

This paper examines whether moral judgments are relative, that is, whether moral judgments about a decision maker depend on a reference action. Such

[1] An exception is Cubitt et al. (2011), in which they show that moral judgments of a free rider depend strongly on others' behaviour and how people perceive the nature and the intensity of the behavior responds to features of the whole situation, not just the consequences of the judged behavior.

dependence may result from a contrast effect and/or from changes of perceived social norms, which is later referred to as a norm-shifting effect. The contrast effect, which has been found in various domains (Conway Dato-on and Dahlstrom, 2003; Hartzmark and Shue, 2018; Kenrick and Gutierres, 1980; Pepitone and DiNubile, 1976), makes a moral action seems more moral and an immoral action more immoral when they are in contrast than when they are judged in isolation. The norm-shifting effect proposes that others' actions affect moral judgments by shifting the perceptions about social norms. By observing a subset of decision makers, people sample and make inference about what most people do, i.e. the descriptive norm, and even about how much the relevant group approves of the actions, i.e. the injunctive norm (Bicchieri et al., 2020; Eriksson et al., 2015; Kelley, 1971; Lindström et al., 2018; McGraw, 1985; Trafimow et al., 2001; Welch et al., 2005). People's perceived social norms change as they observe different behavior, which may in turn alter how they judge the decision maker. Put simply, the contrast effect makes a good action look better, while the norm-shifting effect makes a common action look better.

The framework of this study suggests that, a bit unexpectedly, the contrast effect and the norm-shifting effect of a reference action do not always work in the same direction, and thus the reference action may have asymmetric effect on judgments about moral decision makers and immoral decision makers. As a simply example of the two effects working against each other, a moral action seems better when it is contrasted by an immoral action, but on the other hand, it is worse because it's less common. The overall effect of a reference action on moral judgments is therefore unknown. As an example of the two effects working in alignment, an immoral action is worse when it is contrasted by a moral action and even worse because it is less common. In this case, we can predict the direction of the effect of a reference action.

In the experiment of this study, subjects are assigned one of the two roles in the experiment: a decision maker or an evaluator. Decision makers choose between two options in a binary dictator game simultaneously and independently: either they share ten euros equally between themselves and their paired subjects or they take seven euros for themselves and leave two euros to the paired subject. Evaluators are fully informed of the decision environment. They observe the choices of two random decision makers and judge one of the decision makers; the other decision maker's choice is naturally the reference point. I use two measures of moral judgment: a behavioral measure and an explicit measure. After observing two choices, evaluators immediately allocate a number of real-effort tasks[2] to the decision maker without incurring any cost. Assigning the real-effort to the decision maker can be regarded as a form of punishment because it is meaningless, boring, and time-consuming. Doing these tasks benefits neither the decision makers nor the experimenter. Therefore, the number of real-effort task assigned to the decision maker is used as a behavioral measure of moral judgments.[3] Additionally, evaluators are asked to judge the decision maker along eight personality traits, which gives me the explicit measure of moral judgment. To separate the contrast effect and the norm-shifting effect, I elicit the evaluators' beliefs about the descriptive norm and the injunctive norm in an incentive compatible manner before and after they observe decision makers' choices.

I find that moral judgments indeed depend on the reference action and such dependence is asymmetric for judgments about moral and immoral decision makers. In particular, immoral decision makers are punished more harshly if the reference choice is moral than if it is immoral due to a contrast effect. The punishment moral decision makers inflicted does not vary with the reference choice. Moral judgments do not depend on the perceived descriptive norm.

This paper adds to the literature on how previous experience shapes preferences and values. Kahneman et al. (1986) use hypothetical questions to show that previous transactions serve as reference points and people judge the fairness of prices by the standard of reference points. They believe the psychological basis of such phenomenon is that people adapt to what they observe and eventually accept any stable status. Similarly, Gächter and Schulz (2016) use cross-societal experimental data to demonstrate that intrinsic honesty is stronger in countries where rule violations is less prevalent. Roth and Wohlfart (2018) find in administrative data that those who have experienced a high level of inequality in their lifetime hold different fairness views and are thus less in favor of income redistribution. Their finding suggests that people adapt their fairness views to what they experience and observe over time. Herz and Taubinsky (2017) confirms the role of previous transactions as reference points in an experiment. In their study, subjects were first exposed to market games with different market structures, and are therefore used to high or low prices. They find that experiences in the market games change subjects' perceptions of what is the fair price in the

[2] This task is "counting zeroes", a variation of the counting task in Abeler et al. (2011).

[3] I do not use monetary punishment because deducting payoffs from decision makers may be driven by preferences for redistribution. For example, since the expected payoffs of those who choose the (7,2) scheme is higher than those who choose the (5,5) scheme, one may argue that evaluators use the monetary punishment to equalize the decision makers' expected payoffs, which could be irrelevant to moral judgments about the decision maker.

subsequent ultimatum games. They attribute their finding to a contrast effect. These studies provide evidence for the relativity of fairness judgments, but they do not agree on the underlying mechanisms. I contribute to this strand of literature by examining both the contrast effect suggested by Herz and Taubinsky (2017) and the norm-shifting effect suggested by Kahneman et al. (1986) in the same setting. I show that the relativity of moral judgments, i.e., its dependence on a reference point, can be explained by a contrast effect. I do not find evidence for the norm-shifting effect at least in the setting with minimal observational experience.

This study also speaks to the growing literature on the determinants of moral decisions (Abeler et al., 2019; Benabou and Tirole, 2006; Bénabou et al., 2018; Ellingsen and Johannesson, 2008; Gneezy et al., 2018; Grossman and Van Der Weele, 2017; Khalmetski and Sliwka, 2019) especially in a social context. Economists increasingly recognize the concern for image as one of the driving factors of moral decisions. Image is essential observers' moral judgments about the decision maker. The current study shows that a single reference choice suffices to have a sizable impact on one's moral image and thus has spillover effects on other decision makers' judgments and behaviors. Thus, seemingly independent individual decision making may actually be interdependent. This finding provides a simple alternative explanation to the frequent observation that people tend to behave more immorally in groups than in isolation. Immoral behavior can be contagious due to a contrast effect when there is no diffusion of responsibility (Darley and Latané, 1968; Mynatt and Sherman, 1975; Mathes and Kahn, 1975) or "replacement excuses" (Bartling et al., 2014; Falk and Szech, 2013).

The next section lays out a conceptual framework to formulate the hypotheses. I describe the experimental design in section 3, present the results in Section 4, and conclude in Section 5.

3.2 Conceptual framework and Hypotheses

In this section, I present a conceptual framework to illustrate how a reference action affects moral judgments through a contrast effect and a norm-shifting effect.

Consider a decision environment with $A = \{0, 1\}$ the set of actions available to decision makers. An evaluator uses a function ϕ to judge whether a decision maker j who chooses action a_j is good or bad, after she observes a profile of actions denoted as (a_j, a_{-j}).

$$\phi(a_j) = \phi(v(a_j|a_{-j}, A), u(a_j|a_{-j}, A)) \tag{3.1}$$

The evaluation function takes a single value which can be normalized to $[-1, 1]$. The first component of this function v captures people's instinctive, automatic and hard-wired moral sense. The second component u incorporates learning from and conforming to social norms.[4]

The distinct feature of this framework is that the evaluation of decision maker j is conditional on a profile of actions (a_j, a_{-j}) the evaluator observes. Consider the simplest case in which the evaluator observes the actions of two decision makers and judges decision maker j. Her judgment about decision maker j depends not only on j's action a_j, but also on the other decision maker's choice a_{-j} which is later referred to as the reference action.

Suppose that the presence of the reference action a_{-j} affects $v(.)$ through a contrast effect and/or $u(.)$ through a norm-shifting effect. Notice that both $v(.)$ and$u(.)$ use the choice set A as an input, as the morality of an action depends on available alternatives. Since I do not explore how alternatives affect moral judgments, the choice set remains constant and known to all in this study. Therefore, A will be omitted from the evaluation function in what follows.

Contrast effect The contrast effect is ubiquitous in cognition, perception, and judgments.[5] It describes the phenomenon that the perception of a stimulus is enhanced or diminished by another stimulus of less or greater value. The contrast effect often results in perceptions that go against reasoning. For example, the famous *Ebbinghaus' Tichtener Circles*, as in Appendix B, uses the contrast effect to create two identical circles that look drastically different: one of the circles is surrounded by smaller circles and the other one is contrasted by bigger circles.

The contrast effect may work in the domain of moral judgments in a similar way. Given a choice set $A = \{0, 1\}$, people rank the actions in terms of morality based on their internal moral sense. Suppose $a = 1$ is the moral choice and $a = 0$ is the immoral choice. People always judge the moral choice to be better than the immoral one, regardless of the reference choice: $v(a_j = 1|a_j) > v(a_j = 0|a'_j)$. This unconditional ranking can be thought of as absolute moral judgments.

[4] There is no consensus over the origins of morality or the hard-wired/learned dichotomy. Here I follow Haidt (2012) and assume that morality is a combination of innateness and social learning.

[5] See Conway Dato-on and Dahlstrom (2003); Hartzmark and Shue (2018); Kahneman et al. (1986); Kenrick and Gutierres (1980); Simonsohn (2006); Simonsohn and Loewenstein (2006); Pepitone and DiNubile (1976).

With different reference actions, the perception of an action being moral would be enhanced if the reference action is of less value, i.e. immoral. Likewise, an immoral action appears even more immoral if contrasted by a moral reference action. More formally, the contrast effect implies that

$$v(a_j = 1|a_{-j} = 0) > v(a_j = 1|a_{-j} = 1) \tag{3.2}$$

$$v(a_j = 0|a_{-j} = 0) > v(a_j = 0|a_{-j} = 1) \tag{3.3}$$

Norm-shifting effect Bandura and Walters (1977) proposes that among the many social factors that determine moral judgments, there are the base rate of occurrence and the degree of norm variation, which are often phrased as the two dimensions of social norms in economic literature. One dimension, corresponding to the base rate of occurrence, is the descriptive norm, i.e. how much a behavior is exhibited. The other dimension is the injunctive norm, i.e. how much a society approves of a behavior. Ever since the first experiments conducted by Sherif (1937) and Asch (1955), a large literature has shown that people tend to conform to social norms.[6] As people's perception of social norms changes with what they observe, their evaluation of a behavior may change in accordance with how social norms judge that behavior.

Note that in a very specific and non-recurring environment, people often do not know exactly the descriptive norm or the injunctive norm. In what follows, when I refer to social norms, I mean perceived social norms, or people's beliefs about social norms. Define the descriptive norm as the proportion of decision makers who choose an action $a, a \in A$, denoted as $\theta(a) \in [0,1]$. In a decision environment with only two possible actions, $\theta(a = 0) + \theta(a = 1) = 1$. Denote $N(a)$ as the injunctive norms over the action a, i.e. how much a society approves of this behavior. $N(a = 1)$ and $N(a = 0)$ may or may not be negatively correlated: a society can approve one action while disapproving the other, or approve or disapprove both actions. The second component of the evaluation function $u(.)$ can be written as

$$u(a_j|\theta(a_j|a_{-j}), N(a_j|a_{-j})), \tag{3.4}$$

[6] For more literature on conformity, see for example Allcott (2011); Allen and Wilder (1977); Bicchieri (2016); Bicchieri et al. (2020); Cialdini et al. (1990); Cialdini and Goldstein (2004); Gerber and Rogers (2009); Keizer et al. (2008); Nolan et al. (2008); Prentice and Miller (1993); Schultz et al. (2007).

where u is increasing in both $\theta(a_j)$ and $N(a_j)$.

The connection between the reference action and the beliefs about the descriptive norm is obvious. Rational evaluators would adapt their beliefs about the descriptive norm to the frequency of a certain behavior they actually observe. Fixing the prior and a_j, we immediately have: $\theta(a_j = 1|a_{-j} = 1) > \theta(a_j = 1|a_{-j} = 0)$, and $\theta(a_j = 0|a_{-j} = 1) < \theta(a_j = 0|a_{-j} = 0)$.

Even though descriptive norm and injunctive norm are logically distinct concepts and they do not necessarily align, literature suggests that people sometimes infer injunctive norm from descriptive norm and vice versa (Eriksson et al., 2015; Hoeft and Mill, 2017; Kelley, 1971; Lindström et al., 2018; McGraw, 1985; Trafimow et al., 2001; Welch et al., 2005). If evaluators associate what is common with what is socially appropriate–a "common is moral" phenomenon, they may also update their beliefs about injunctive norms upon observing a profile of actions. Thus, fixing the decision maker's action a_j we have: $N^p(a_j = 1|a_{-j} = 1) > N^p(a_j = 1|a_{-j} = 0)$, and $N^p(a_j = 0|a_{-j} = 1) < N^p(a_j = 0|a_{-j} = 0)$, where the superscript p means "posterior".

Evaluators judge an action to be more moral, if more people are doing it, or if the action is more socially appropriate. The norm-shifting effect gives

$$u(a_j = 1|a_{-j} = 0) < u(a_j = 1|a_{-j} = 1) \tag{3.5}$$

$$u(a_j = 0|a_{-j} = 0) > u(a_j = 0|a_{-j} = 1) \tag{3.6}$$

By combining the contrast effect and the norm-shifting effect, it follows that a reference action has asymmetric effects on judgments for moral and immoral actions. When people judge immoral decision makers, the contrast effect and the norm-shifting effect go hand in hand. Equations (3.3) and (3.6) yield the first hypothesis:

Hypothesis 3.1 (IM). *An immoral decision maker is judged to be nicer if the reference action is immoral than if the reference action is moral:*

$$\phi(a_j = 0|a_{-j} = 0) > \phi(a_j = 0|a_{-j} = 1)$$

However, when it comes to moral decision makers, equations (3.2) and (3.5) show that the norm-shifting effect and the contrast effect work in opposite directions. It is not clear which effect will dominate the other. Therefore, I hypothesize that

Hypothesis 3.2 (M). *Judgments about a moral decision maker does not change whether the reference action is immoral or moral:*

$$\phi(a_j = 1|a_{-j} = 0) = \phi(a_j = 1|a_{-j} = 1)$$

I will test these hypotheses in section 4.2 and explore the contrast effect and the norm-shifting effect in 4.3.

3.3 The Experiment

3.3.1 Experimental Design

In the experiment, subjects are assigned the role of either a decision maker or an evaluator. The decision makers make their choices in a binary dictator game. The evaluators are fully informed of the rules of the binary dictator game. After observing the choices of two decision makers in the game, they make judgments about one of the decision makers.

The binary dictator game offers two payoff schemes: a scheme that gives five euros to both the decision maker and her paired participant, and a scheme that gives seven euros to the decision maker and two euros to her paired participant. Decision makers simultaneously choose one of the two payoff schemes. One of the paired decision makers will be randomly selected and her choice will determine their payoffs. In the subsequent sections, I refer to the payoff scheme (5,5) as the "moral choice" and the payoff scheme (7,2) as the "immoral choice".

After all decision makers have chosen between the two schemes, each evaluator observes the choices of two random decision makers who are not necessarily paired without knowing the realized payoff scheme or their final payments. The evaluator is then asked to judge only one of the two decision makers. The other decision maker's choice is regarded as the reference action. The combination of the choice of the decision maker to be judged and the reference choice assigns evaluators to four conditions: observe two moral choices and judge one of them (T11), observe two different choices and judge the moral one (T10), observe two different choices and judge the immoral one (T01), or observe two immoral choices and judge one of them (T00).

To elicit the moral judgment of the evaluator, I employ both an explicit measure and a behavioral measure. The explicit measure includes a series of questions regarding the decision makers' eight personality traits on a 7-point Likert scale. These personality traits are: risk-aversiveness, selfishness, trustworthiness, patience, intelligence, honesty, kindness and temperament. Among these traits only selfishness is both morality-related and directly related to decision makers' choice in the binary dictator game. Three personality traits (trustworthiness, kindness, honesty) are related to moral judgments but not directly related to the binary dictator game. Risk preference, patience, intelligence, and temperament are neither related to moral judgments nor the binary dictator game.[7] An attention check question is inserted among these questions regarding the personality traits.

The behavioral measure rests on the presumption that an evaluator's willingness to punish/reward a decision maker reflects her moral judgments about the decision maker (McGraw, 1985). Evaluators are asked to allocate a number of real-effort tasks to the decision maker being judged immediately after observing the two choices. The real-effort task is to count the number of "1"s in a sequence such as "101000010111101011101". The evaluator decides how many sequences (minimal 1 and maximal 20) the decision maker needs to solve before she/he gets paid. This task requires no specific skill and it is boring, so it can be seen as some sort of punishment. I use the number of sequences the evaluator assigns to a decision maker as the implicit behavioral measure of moral judgments.

To control for the perceived descriptive norm, evaluators are required to provide their estimate on the proportion of decision makers in previous sessions making the moral choice. Evaluators earn a bonus if their estimates deviate from the true proportion by no more than 5 percentage point. As to the perceived injunctive norm, I apply the elicitation method of Krupka and Weber (2013). Evaluators rate the social appropriateness for choosing each payoff scheme on a 7 point Likert scale. They will earn a bonus if their rating turns out to be the modal rating of all evaluators. Their own ratings are therefore treated as the perceived injunctive norm over the possible actions.

As an attention check, I inserted an irrelevant question in the elicitation of explicit moral judgments and specified the correct answer evaluators should choose

[7] The selection of personality traits except selfishness is to some extent arbitrary. I include them mainly for two reasons. First, I use them as a sanity check. For traits that are not related to morality or the dictator game, I expected the evaluators to rate decision makers similarly along these traits regardless of the decision makers' choices. For traits that are not related to the dictator game but somewhat related to morality, the evaluators may or may not rate the decision makers differently depending on their choices in the dictator game. I expect the biggest difference in the ratings of selfishness. Second, I hide the outcome variable of interest, i.e., morality-related traits, among the unrelated traits to reduce the experimenter demanding effect.

in the question. Those who did not choose the specified answer failed the attention check.

3.3.2 Procedures

The experiment was conducted at CentERlab of Tilburg University.[8] Participants registered in the first session of the experiment were assigned the role of decision maker, which was framed as "Green player" in the experiment.[9] Upon registration, decision makers were informed that the experiment consists of two parts which will take place on the day they showed up and a week later respectively. They would only be paid if they participate in both parts. Participants registered in all the other sessions were assigned the role of evaluator, which was framed as "Red player". Evaluators only visited the lab once. In total, 206 subjects from Tilburg University participated in the experiment, of whom 167 were assigned the role of an evaluator. Evaluators received 6.2 euros on average for half an hour in the lab. Decision makers spent less than half an hour in total in the lab and earned 9.4 euros on average.

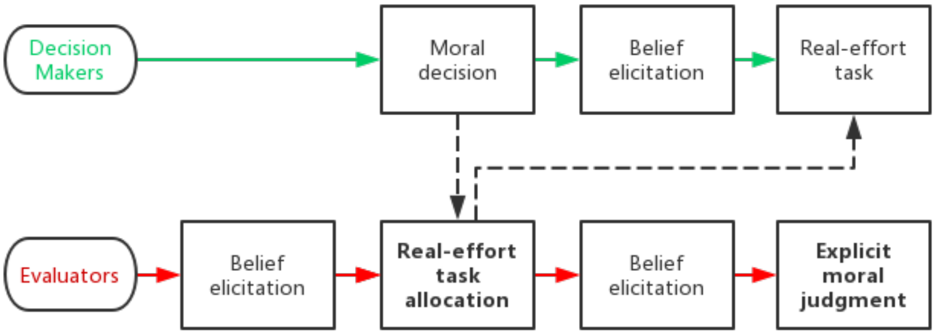


Figure 3.1: Experiment Procedure

Figure 3.1 summarizes the procedure of the experiment. Subjects who were assigned the role of decision maker received instructions on the moral task, the real-effort task, and the schedule upon arrival. In the instructions, it is made clear to the decision makers that their choices in the moral task will be revealed

[8] I pre-registered the experimental design on AsPredicted. For more details of the pre-registration, see: https://aspredicted.org/blind.php?x=q4ti3j.

[9] Data collection of the experiment is split into two waves that took place in June and September respectively because recruitment in June was not satisfying due to a clash with the exam week.

to some evaluators in future sessions in an anonymous way and that the number of real-effort tasks they receive will be dictated by those evaluators. Decision makers were instructed to come back to the lab on the last day of the experiment to finish the real-effort tasks assigned to them before collecting their payments. Decision makers' beliefs about the descriptive norm and the injunctive norm in the moral task were elicited after having made their choices.

Evaluators first read the instructions that decision makers received. This makes sure that evaluators were fully aware of the decision environment the decision makers were confronted with. Next, evaluators learned about their tasks in the experiment, including the belief elicitation and the real-effort task allocation. Given all this information, evaluators' beliefs about the descriptive norm and the injunctive norm in the moral task were elicited. Then each evaluator was presented with the choices of two random decision makers in the moral task and was told to which of the two decision makers they were going to assign real-effort task. After they decided on the number of sequences, they reported perceived social norms again. In the final step, they were asked to judge the decision maker along the eight personality traits.

3.4 Results

I exclude the inputs of 14 evaluators who failed the attention check. [10] The rest of the evaluators fall randomly into 4 groups, depending on the action of decision maker and the reference action. I use Ta_ia_{-i} to identify each group, with a_i being the action of the decision maker and a_{-i} being the reference action. There are 36 evaluators in T11, 42 in T10, 36 in T01, and 39 in T00.

I will first validate the "moral task" by showing evaluators' absolute judgments about the two choices and the decision makers who choose them. Then I move on to testing the hypotheses. Finally, I discuss the contrast effect and the norm-shifting effect in more details.

3.4.1 Absolute moral judgments about decision makers

Do evaluators consider choosing the (5,5) payoff scheme a moral choice and the (7,2) payoff scheme an immoral choice? This section answers the question by

[10] The percentage of subjects who failed the attention check is a bit high, most likely because the attention check appears among the questions about decision makers' personality traits. These questions are not incentivized, which may attract less attention than incentivized questions. Also, there are eight of them, all looking very similar, which may partly explain subjects' inattention.

comparing the *ex ante* perceived injunctive norm over the two available choices and the absolute judgments about moral and immoral decision makers.

The top panel and the bottom panel of Figure 3.2 present the distribution of social appropriateness ratings on choosing the two payoff schemes before subjects observe real choices. The perceived social approval of the two choices differ starkly: 96.1% of the subjects people believe choosing the (5,5) scheme to be socially appropriate ("a bit/somewhat/very socially appropriate"), while 86.3% of the subjects believe choosing the (7,2) scheme to be socially inappropriate ("a bit/somewhat/very socially inappropriate"). Subjects generally agree on the social appropriateness ratings, but there is greater consensus when it comes to choosing the (5,5) scheme than choosing the (7,2) scheme. 64.7% of the evaluators agree on the (5,5) scheme being "very socially appropriate" while only 44.7% of the evaluators agree on the (7,2) scheme to be "somewhat socially inappropriate", which is the modal response.[11]

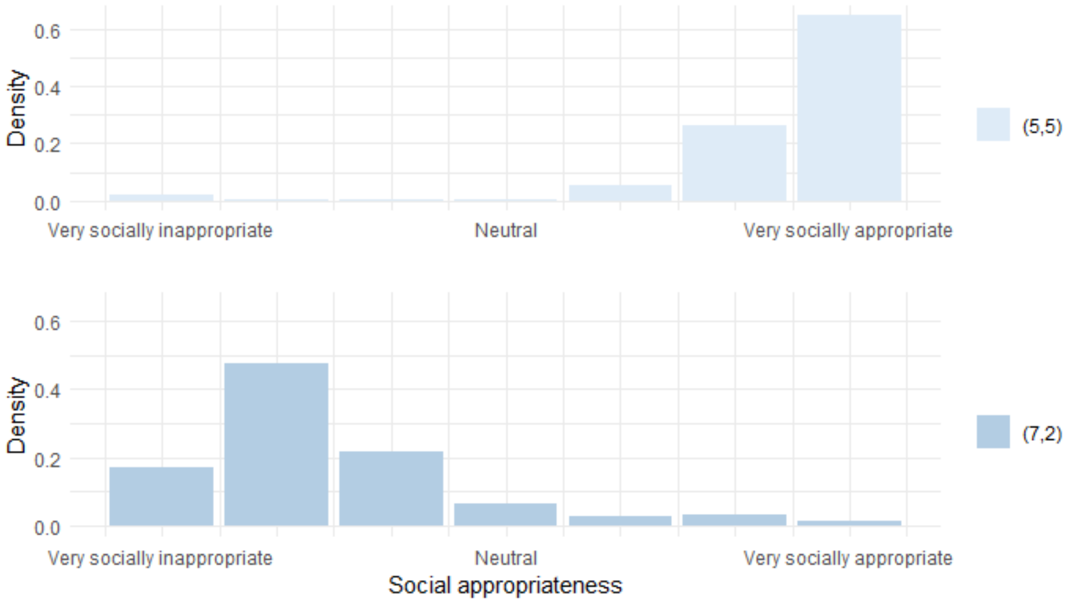


Figure 3.2: Distribution of social appropriateness ratings on choosing the (5,5) payoff scheme and on choosing the (7,2) payoff scheme.

The evaluators judge the decision maker along eight personality traits: risk-attitude, selfishness, trustworthiness, patience, intelligence, honesty, kindness and temperament. Figure 3.3 presents the evaluators' average ratings on these personality traits of the decision makers conditional on their choices. Decision

[11] Note that people's belief of how the society approves of a behavior does not always coincide with how much she/he approves of that behavior (for example, see Bursztyn et al. (2018)).

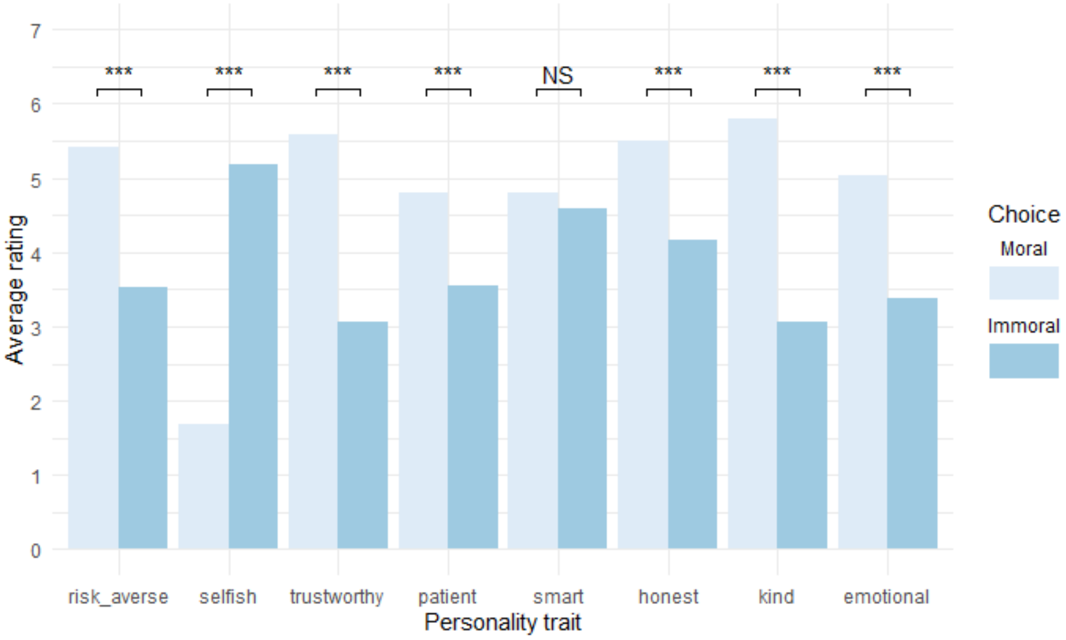


Figure 3.3: Explicit judgments about moral and immoral decision makers, aggregated over the reference actions.

makers who choose the moral option are judged to be less selfish, more trustworthy, more honest, and more kind than those who make the immoral choice. A bit surprisingly, the evaluators also believe that the moral task is informative of the decision makers' risk attitudes, patience, and temperament. They rate moral decision makers as more risk averse, more patient, and more emotional than immoral decision makers. These differences are all significant at 1% level by Wilcoxon-Mann-Whitney test. Only intelligence is thought to be unrelated to the moral task.

The explicit measure of moral judgments about the decision makers is reflected in the behavioral measure of moral judgments as well. Figure 3.4 shows the distribution of the number of real-effort tasks assigned to moral decision makers and immoral ones. 24.4% of the moral decision makers are assigned only 1 task, which is the minimum amount. In sharp contrast, 24% of the immoral decision makers receive the maximal amount of real-effort task. In other words, they are punished to the upper limit. Moral decision makers receive 7.53 tasks on average, significantly less than 13.3, the average number of tasks immoral decision makers receive (Wilcoxon-Mann-Whitney test, $n_1 = 78$, $n_2 = 75$, $p < 0.01$).

In sum, both the explicit measure and the behavioral measure of moral judgments suggest that evaluators judge decision makers distinctively depending on their choices.

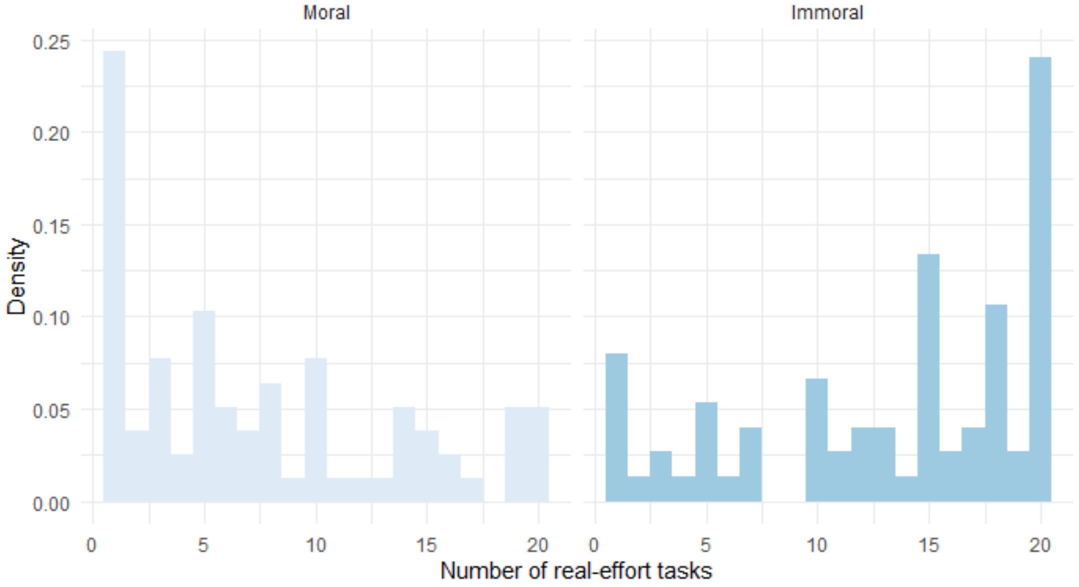


Figure 3.4: Behavioral measure of moral judgments about moral and immoral decision makers, aggregated over the reference actions.

3.4.2 Main results: the effect of reference action on moral judgments

Recall that the reference action is hypothesized to have asymmetric effect on judgments for moral decision makers and immoral decision makers. Specifically, immoral decision makers seem worse when contrasted by a moral reference choice than by an immoral reference action. But it is not so clear in which directions a reference action affects judgments about moral decision makers.

Explicit moral judgments The evaluators judge decision makers along eight personality traits. However, there is no global consensus on which of these traits are in the moral domain. So I construct three different measures of explicit moral judgments that vary in broadness. First, in the narrowest term, the choice between the (5,5) payoff scheme and the (7,2) payoff scheme in the setting is a direct signal of the decision maker's selfishness. Hence, the rating on selfishness is the narrowest measure of explicit moral judgments. The second measure is a composite *morality score* that takes the average of ratings on morally relevant personality traits, i.e. selfishness, trustworthiness, kindness, and honesty. Finally, I construct a composite *image score* by averaging the ratings on all the personality traits that subjects believe to be correlated with the decision makers' choices. For example, if the evaluators think those who choose the moral option are more risk averse than those who choose the immoral option, then the rating on risk attitude is taken into account in the composite image score. As shown in figure 3.3, the evaluators judge moral decision makers and immoral decision makers distinctively along all personality traits except intelligence. Therefore, the rating on intelligence is excluded from the composite image score and the other seven personality traits are counted in.

Explicit moral judgments do not respond to the reference action, whether measured by selfishness, composite morality score, or composite image score. As an example, Figure 3.5 presents the average composite image score across treatments. Moral decision makers score significantly higher (Wilcoxon-Mann-Whitney test, $n_1 = 78$, $n_2 = 75$, $p < 0.01$) than immoral decision makers. However, given the decision maker's choice, the composite image score does not vary with the reference choice. Regressions of the three different variables on the reference choice confirm that explicit moral judgments are independent of the reference choice. The regression results are reported in Table 2 and Table 3 in Appendix C.

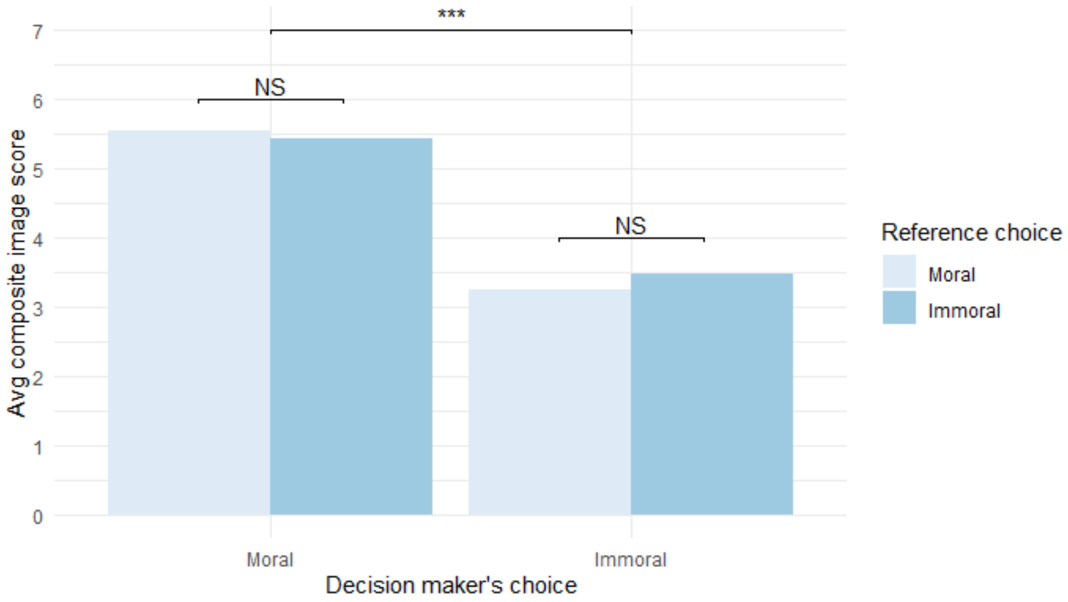


Figure 3.5: Explicit judgments about decision makers sorted by the decision maker's choice and the reference choice.

Behavioral measure of moral judgments Figure 3.6 presents the average number of real-effort task allocated to the decision makers, sorted by the decision maker's choice and the reference choice. Immoral decision makers receive 15.2 real-effort tasks on average when the reference choice is moral but only receive 11.7 real-effort tasks when the reference choice is immoral. The difference is statistically significant (Wilcoxon-Mann-Whitney test, $n_1 = 36$, $n_2 = 39$, $p = 0.04$). That is, evaluators punish immoral decision makers more harshly when they see a moral reference choice than when they observe an immoral reference choice. Meanwhile, the number of tasks assigned to moral decision makers does not change with the reference choice (Wilcoxon-Mann-Whitney test, $n_1 = 36$, $n_2 = 42$, $p = 0.6$).

The results are summarized below.

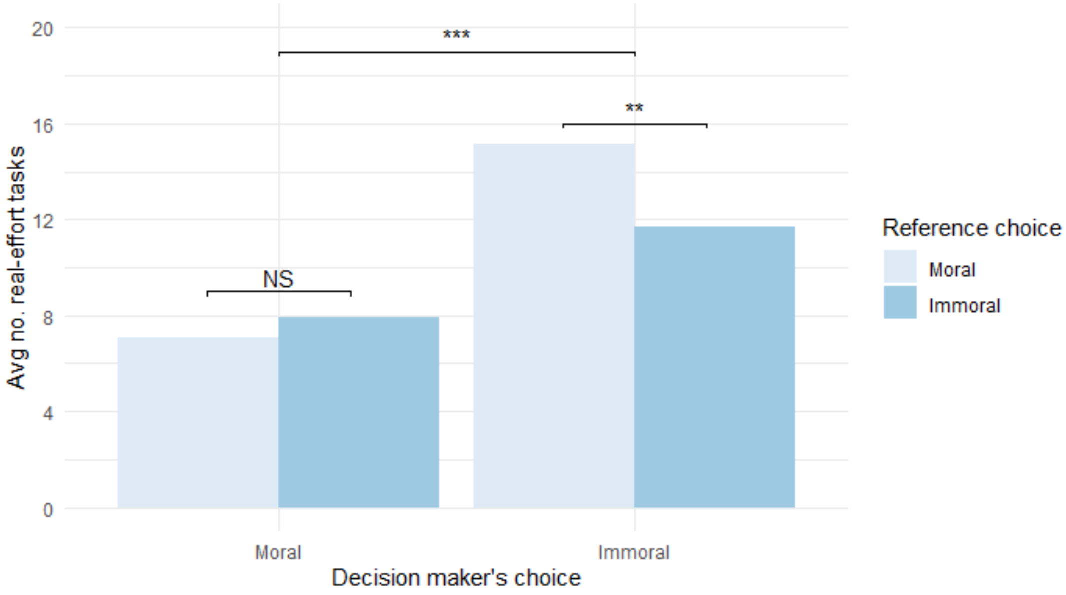


Figure 3.6: Number of real-effort task allocated to decision makers sorted by the decision maker's choice and the reference choice.

Result 3.1. *Immoral decision makers are punished more harshly if the reference choice is moral than if it is immoral. However, explicit moral judgments about immoral decision makers are not affected by the reference point.*

Result 3.2. *The reference choice affects neither the number of real-effort tasks assigned to the moral decision makers nor the explicit judgments about moral decision makers.*

The the relationship between the behavioral measure of moral judgments and the reference choice demonstrates the exact pattern as hypothesized. These results together lend support to both Hypothesis IM and Hypothesis M.

3.4.3 The contrast effect and the norm-shifting effect

In previous section, I find that the behavioral measure of moral judgments reacts to a reference choice in the same way as predicted. In this section, I decompose the overall effect of a reference action on moral judgments into a contrast effect and a norm-shifting effect to see if these two effects are indeed the underlying mechanisms of the previous findings.

The contrast effect The data is split into a sub-sample for moral decision makers and a sub-sample for immoral decision makers. To separate the contrast effect from the norm-shifting effect, I run reduced form individual-level regressions of moral judgments on the reference choice and *ex post* perceived social norms for the moral decision maker sub-sample and for the immoral decision maker sub-sample. The estimated regression is:

$$\phi_i = \beta R_i + \gamma \theta_i^p + \eta N_i^p + \epsilon_i \tag{3.7}$$

where ϕ is the number of real-effort tasks evaluator i assigns to decision maker j.[12] R_i is the reference choice evaluator i observes next to the choice of the decision maker j she judges. θ_i^p is evaluator i's *ex post* belief about the descriptive norm, i.e. the proportion of decision makers choosing action a_j, and N_i^p is her *ex post* belief about the injunctive norm over the decision maker's action a_j.

The dependent variable ϕ–the number of real-effort tasks–takes a greater value if decision maker j is judged to be worse. Therefore, a negative γ or η indicates that moral judgments about a decision maker increases with the perceived descriptive norm or the perceived injunctive norm. When perceive social norms are controlled for, β captures the contrast effect.

The first column of Table 4.1 shows the results from a censored tobit regression for the moral sub-sample. Moral decision makers receive less real-effort tasks if evaluators consider the action as more socially appropriate. However, the reference action has no effect on the amount of real-effort tasks moral decision makers receive, which suggests that the contrast effect does not play a role in this case. Column 2 of Table 4.1 presents the regression results for the im-

[12] I focus on the behavioral measure in this section because it has been shown to respond to a reference action while the explicit measure of moral judgments does not. It makes more sense to explore the mechanisms of an impact when there is an impact. Nonetheless, I run regression results of the explicit measures of moral judgments and report the results in Appendix C. No significant coefficient is found.

moral sub-sample. As the contrast effect predicts, an immoral reference choice decreases the number of real-effort tasks immoral decision makers receive by 4.85.

Result 3.3. *A contrast effect is found only when evaluators are judging immoral decision makers.*

Table 3.1: Determinants of the behavioral measure of moral judgments

	Dependent variable: NO. tasks	
	Moral DMs	Immoral DMs
Immoral reference action	-0.078	-4.851^{**}
	(1.907)	(2.012)
Descriptive norm	-3.583	3.615
	(4.054)	(3.979)
Injunctive norm	-2.632^{***}	0.041
	(0.870)	(0.731)
Constant	25.537^{***}	14.709^{***}
	(6.898)	(3.461)
logSigma	2.034^{***}	2.116^{***}
	(0.103)	(0.108)
Observations	78	75
Akaike Inf. Crit.	434.739	418.716
Bayesian Inf. Crit.	446.522	430.304
Note:	*p<0.1; **p<0.05; ***p<0.01	

In neither sub-sample does the descriptive norm influence the number of real-effort tasks assigned to the decision makers. The perceived injunctive norm over the moral action affects the punishment moral decision makers receive: the more socially appropriate the evaluator beliefs the action to be, the less real-effort tasks she assigns to the decision maker. But the injunctive norm over the immoral action does not affect punishment on immoral decision makers.

Result 3.4. *Perceived descriptive norm is not a significant factor in determining moral judgments.*

Result 3.5. *Perceived injunctive norm is positively correlated with judgments about moral decision makers but not about immoral decision makers.*

Recall that Result 3.1 states that immoral decision makers are punished more harshly when they are compared with a moral decision maker than when they are compared with an immoral decision maker. Result 3.2 states that judgments about moral decision makers are not affected by a reference choice. Result 3.3-3.5 suggest that the contrast effect seems to be the underlying mechanism for

Result 1 and that the norm-shifting effect may be relevant to Result 3.2, which I will examine next.

Shifting the descriptive norm Result 3.4 finds that the descriptive norm is not a significant determinant of moral judgments. On the other hand, the injunctive norm is a major determinant of judgments about moral decision makers, according to Result 3.5. The unanswered question is whether and how individuals update their beliefs about the descriptive norms upon observing two random actions.

The actual percentage of moral choices in the experiment is 46%. Figure 9 in Appendix C presents the distribution of the *ex ante* perceived descriptive norm of the evaluators. Most evaluators are wrong about the descriptive norm; the majority of them are not even close. The distribution of guesses shows that there is no consensus over the descriptive norm.

Evaluators may update their beliefs about the descriptive norm based on the actual choices they observe. To describe the updating process more formally, denote an evaluator i's *ex ante* belief about the proportion of moral choices as θ_i^a and her *ex post* belief about the proportion of moral choices as θ_i^p. In the experiment, evaluator i observes the choices of two decision makers a_j and a_{-j}. The observed descriptive norm is denoted as $\hat{\theta}_i = \frac{a_j + a_{-j}}{2}$. If the observed norm $\hat{\theta}_i$ is greater than she originally believes, she would update her belief upward, and vice versa. I define an adaptive evaluator as someone who updates her belief in the same direction as the observed descriptive norm relative to her prior belief.

Definition 3.1. ***Adaptive evaluator***. *An evaluator is adaptive, if and only if* $(\theta_i^p - \theta_i^a) * (\hat{\theta}_i - \theta_i^a) \geq 0.$[13]

Figure 3.7 demonstrates in details how each evaluator updates her belief about the descriptive norm conditional on her prior belief and the observed descriptive norm. On the x-axis lies the distance between the observed descriptive norm and the evaluators' *ex ante* belief about the descriptive norm: $\hat{\theta}_i - \theta_i^a$. The y-axis shows how much the evaluator actually updates her belief: $\theta_i^p - \theta_i^a$. 90.8% of the evaluators are qualified as adaptive evaluators, identified by dots in the shaded area. Notice that among these adaptive evaluators, 56 of them, which is

[13] Those who do not update their beliefs about social norms are temporarily classified as adaptive evaluators until proven otherwise, because they have a reason not to update: two actual choices are not provide compelling evidence against their initial beliefs. In cases where they observe more choices, they might update in accordance with the observed descriptive norm.

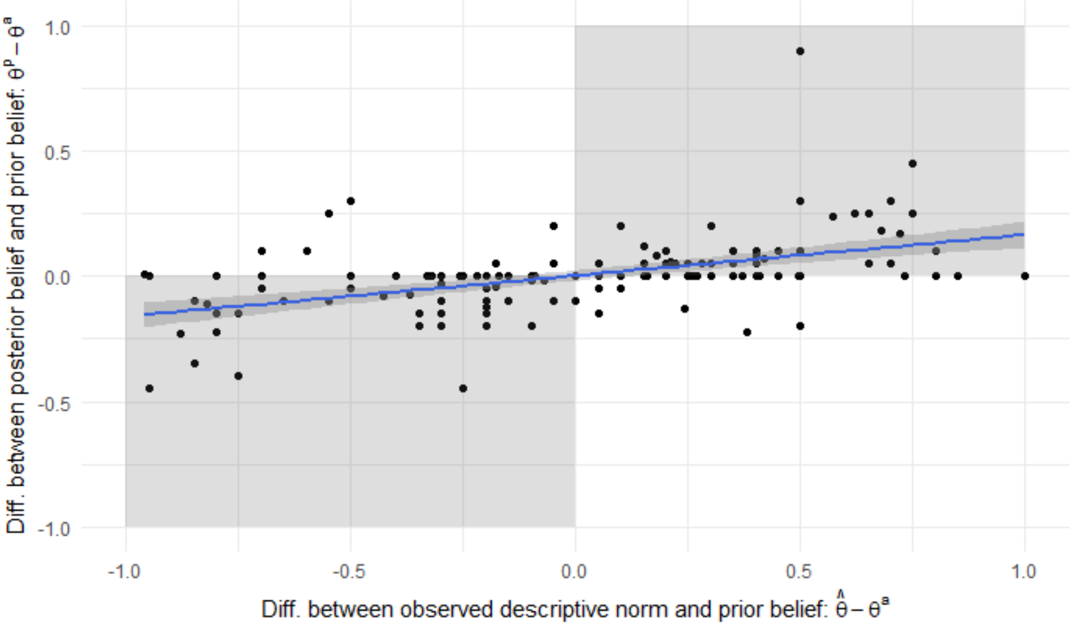


Figure 3.7: Updates of perceived descriptive norm upon observing actual choices

36.6% of all the subjects, do not update their beliefs even when there is a discrepancy between their prior beliefs and the observed descriptive norm.

At the group level, altering the reference action indeed results in different average perceived descriptive norm, but not to a statistically significant extent. Figure 10 in Appendix C presents more details of the average *ex ante* belief and *ex post* belief about the descriptive norm across treatments. The *ex post* beliefs are not distinguishable between T10 and T11 or between T01 and T00. This confirms that the perceived descriptive norm does not likely account for the difference in moral judgments between T01 and T00.

Shifting the injunctive norms If people tend to take what is normal as what is right (Eriksson et al., 2015; Hoeft and Mill, 2017; Kelley, 1971; Lindström et al., 2018; McGraw, 1985; Trafimow et al., 2001; Welch et al., 2005), we should observe a strong correlation between one's *ex ante* perceived descriptive norm and *ex ante* perceived injunctive norm. Yet no such correlation is found in the data, which implies that subjects see the descriptive norm and the injunctive norm as distinct concepts and that a shift in one of them does not automatically lead to a shift in the other.

While more than half of the evaluators update their beliefs about the descriptive norm, most of them do not change their beliefs about the injunctive norms at all. 87.3% of them stick to their *ex ante* belief about the injunctive norm over the moral choice and 82.8 % do not change their belief about the injunctive norm over the immoral choice.

At the treatment level, varying the reference choice does not cause any noticeable difference in *ex post* perceived injunctive norms between T11 and T10 or between T01 and T00, as shown in Figure 11 in Appendix C.

Does the norm-shifting effect play a role? Combining the results in this section, I find little evidence of the norm-shifting effect. First, although observing actual choices shifts the perceived descriptive norm, the perceived descriptive norm is not a significant determinant of moral judgments, as Table 4.1 shows.

Second, observing two choices barely affects perceived injunctive norms in the first place, nor does it lead to any significant difference between treatments. Thus, the norm-shifting effect cannot account for the relativity of moral judgments found in this experiment.

3.5 Conclusion

In this study, I use a lab experiment to show that moral judgments about a decision maker depend on the reference choice people observe. Particularly, subjects punish an immoral decision maker more harshly if they observe a moral reference choice than if they observe an immoral one due to a contrast effect. By eliciting perceived social norms, I show that moral judgments do not depend on the perceived descriptive norm; and observing actual choices do not shift perceived injunctive norms.

This paper shows that moral judgments are malleable and made relative to reference choices. A single reference choice suffices to have a sizable impact on moral judgments and thus has spillover effects on others' judgments and hence behaviors. This result provides a simple alternative explanation–the contrast effect–to the frequent observation that people tend to behave more immorally in groups than in isolation. It takes less than a replacement excuse, a narrative, a competitive market or the support of social norms for people to feel more comfortable choosing the immoral behavior. The result also highlights the possibility that in a minimal social context, seemingly independent decisions can be interdependent. In such cases, individual decision-making should be treated as a game.

The result on the contrast effect suggests that moral judgments do not completely emerge from norm-based reasoning. Instead, moral judgments probably rely heavily on heuristics, which coincides with the findings of moral psychology (Haidt, 2012). The result further implies that the effect of a reference choice on moral judgments is likely transient. As the contrast effect is a fundamental principle of perception, it should disappear once the reference choice is not present. An interesting future line of research would be the direction of the contrast effect as the number of alternatives increases and the strength of the contrast effect when there are more than two decision makers.

While the norm-shifting effect does not account for the finding of this study, this of course does not mean that it never plays a role. As to whether observational experience can change perceived injunctive norms, literature finds mixed evidence in lab settings (Dimant et al., 2020; Hoeft and Mill, 2017). One may reasonably expect the norm-shifting effect to be powerful in some special cases. For instance, when people enter a new cultural environment, their moral standards may adapt quickly to observed social norms. To reconcile these findings and intuitions, factors that determine the tipping point at which injunctive norms evolve remain to be investigated.

Another finding that raises further questions is the asymmetry of the relativity of moral judgments: the reference choice affects moral judgments about immoral decision makers but not about moral decision makers. Does it manifest the moral-immoral asymmetry or the reward-punish asymmetry? Does it have something to do with social norms? These questions provide directions for future research.

Appendix A. Instructions for the evaluators

You will receive a fixed amount of 5 euros by the end of the experiment. In addition, you might earn more depending on your responses in the experiment. Payments will be made to you in cash at the conclusion of the experiment.

There are two roles in this experiment: “Green player” and “Red player”. You are assigned the role of Red player.

Before continuing, please read carefully the instructions that Green players received. Green players have to do two tasks: an allocation task and a counting task. The tasks take place on different days. They have already done the allocation task; they need to do the counting task on a day in November (known to them). All the instructions of the Green players are in italics.

Instructions for Green players

The allocation task You and another Green player are paired. Both of you have to choose between two allocation plans: Plan 1: You receive 5 euros; your paired Green player receives 5 euros. Plan 2: You receive 7 euros; your paired Green player receives 2 euros.

You and your paired Green player will make your choice simultaneously and anonymously. After both of you have submitted the plan you prefer, one of you will be randomly selected, and the choice of this player will be implemented. That is, your earnings and the earnings of your paired Green player will be determined according to the plan submitted by the selected Green player.

For example, Green player A and Green player B are paired. A chooses plan 1 and B chooses plan 2. Then the computer selects B. Therefore, B will receive 7 euros and A will receive 2 euros.

Your choice will be later shown to some Red players in other sessions. Red players can only see your participant ID and which plan you have chosen. But no participant will ever know your identity. You will never know the identities of the Red players or Green players either.

The counting task You will see a sequence with “0”s and “1”s on the screen. Your task is to count the number of “1”s in each sequence. Below you will see an example: 0011100100010011000100101001100110000101 010000110 (answer: 19)

In the instructions of part I, you were already informed that your choice in part I will be shown to some Red players in future sessions. One of these Red players will be able to decide how many sequences you will need to solve correctly on [ ]. The number of sequences a Red player can assign is between 1 and 20. That is, you will need to solve 1 sequence at least and 20 sequences at most.

You will receive your total payments immediately after you have solved all of the sequences correctly. For each sequence, you have multiple chances to correct your answer.

You have finished reading the instructions of the two tasks that Green players received. Now we continue with instructions for you as a Red player.

Instructions, Part I

You will be asked a few questions regarding the allocation task that Green players already did in previous sessions. Please answer the questions as truthfully as you can. These questions give you a chance to earn more money. Your answers in the experiment remain anonymous.

Q1: Please think of the choices of Green players of previous sessions. What is the percentage of Green players who have chosen plan 1? If your guess is close enough (± 5

Q2 and Q3 ask you to evaluate the social appropriateness of choosing each plan.

By "socially appropriate", we mean behavior that most people agree is the "correct" or "ethical" thing to do. You can choose from 7 categories: "very socially inappropriate", "somewhat socially inappropriate", "a bit socially inappropriate", "neutral", "a bit socially appropriate", "somewhat socially appropriate", and "very socially appropriate".

At the conclusion of the experiment, the computer will check for each question which category has been chosen by most participants. All participants who have chosen this category will earn 1 euro per question. For example, there are 10 participants. 8 of them choose "very socially inappropriate". The other 2 choose "very socially appropriate". Then each of the 8 participants who choose "very socially inappropriate" will earn 1 euro.

Q2: Please consider plan 1: You receive 5 euros; your paired Green player receives 5 euros. How socially appropriate do you think it is to choose plan 1 in the allocation task? If your answer corresponds to the category chosen by most participants you will earn 1 euro.

Q3: Please consider plan 2: You receive 7 euros; your paired Green player receives 2 euros. How socially appropriate do you think it is to choose plan 2 in the allocation task? If your answer corresponds to the category chosen by most participants you will earn 1 euro.

Instructions, Part II

The participant IDs and the actual choices in the allocation task of two random Green players will be shown to you. You will be matched with one of them. Your task is to decide how many sequences (as described in the instructions of the counting task) your matched Green player will need to solve correctly before she/he can collect her/his final payoffs. The number of sequences has to be between 1 and 20. If multiple Red players are matched with the same Green player, it will be randomly decided which Red player's decision will be implemented for this Green player.

Appendix B. Illustration of the contrast effect

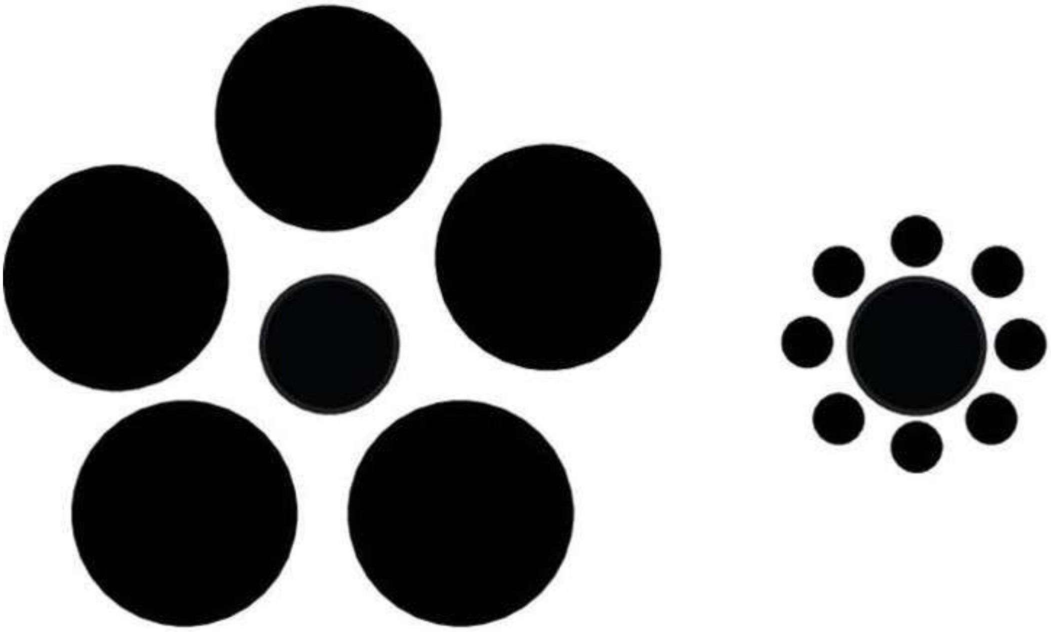


Figure 8: The contrast effect

Appendix C. More results

The impact of a reference action on the explicit judgments about the decision maker

Table 2 and 3present the results of the following estimated regression for the moral decision maker sub-sample and for the immoral decision maker sub-sample respectively:

$$\phi_i = \beta R_i + \epsilon_i \tag{8}$$

where ϕ is the explicit moral judgment. I use three different dependent variables: selfishness, composite morality score and composite image score to measure ϕ. R_i is the reference choice evaluator i observes next to the choice of the decision maker j. The reference point has no effect on explicit moral judgments for either sub-sample.

Distribution of the *ex ante* perceived descriptive norm

Figure 9 shows the distribution of the perceived descriptive norm, the percentage of moral choices specifically, before the evaluators observe any actual choice. The *ex ante* perceived descriptive norm ranges from 0 to 100 with great variation.

Table 2: The impact of reference point on explicit judgments about **moral** DMs

	Moral decision makers:		
	Selfishness	Composite moral score	Composite image score
	(1)	(2)	(3)
Immoral reference Choice	0.887	0.043	−0.125
	(0.673)	(0.221)	(0.172)
Constant	−0.192	5.824***	5.564***
	(0.598)	(0.162)	(0.126)
logSigma	0.897***	−0.034	−0.281***
	(0.153)	(0.086)	(0.081)
Observations	78	78	78
Akaike Inf. Crit.	197.878	218.767	183.933
Bayesian Inf. Crit.	204.948	225.838	191.003
Note: Censored tobit regression.			*p<0.1; **p<0.05; ***p<0.01

Table 3: The impact of reference point on explicit judgments about **immoral** DMs

	Dependent variable:		
	Selfishness	Composite moral score	Composite image score
	(1)	(2)	(3)
Immoral reference Choice	−0.220	0.219	0.241
	(0.682)	(0.283)	(0.201)
Constant	5.739***	3.160***	3.242***
	(0.500)	(0.204)	(0.145)
logSigma	1.033***	0.202**	
	(0.114)	(0.083)	
Observations	75	75	75
R^2			0.019
Adjusted R^2			0.006
Akaike Inf. Crit.	290.996	248.613	
Bayesian Inf. Crit.	297.948	255.565	

Note: Censored tobit regression.Column (3) uses OLS regression because there is no censored data.

*p<0.1; **p<0.05; ***p<0.01

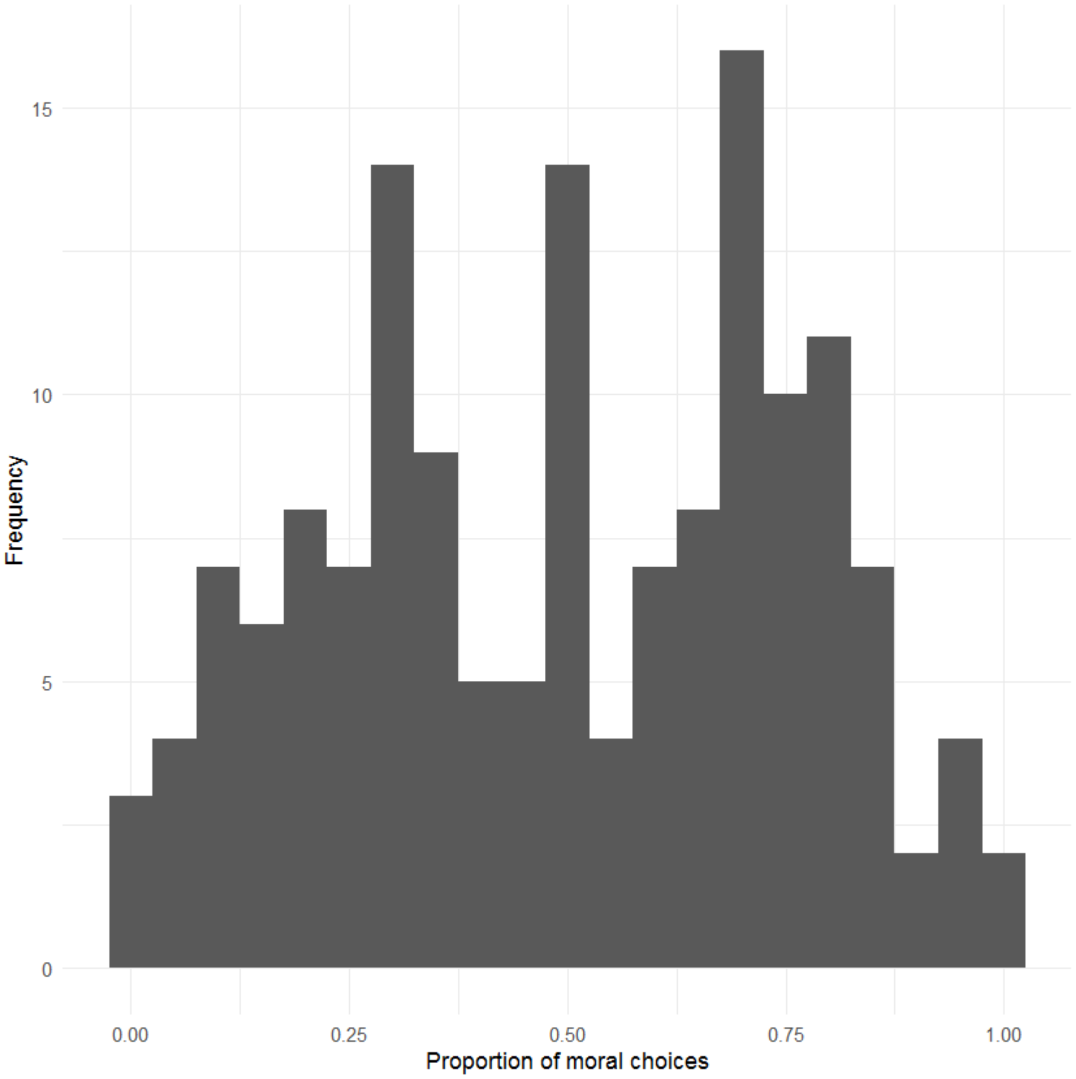


Figure 9: Distribution of the *ex ante* perceived descriptive norm

The *ex ante* and *ex post* perceived descriptive norm

Figure 9 shows the average *ex ante* and *ex post* perceived descriptive norm across treatments. Before the evaluators observe actual choices, all treatments share a similar perceived descriptive norm on average. After the evaluators observe two actual choices, they update their beliefs about the descriptive norm in the direction aligned with Bayesian updating. The differences in perceived descriptive norm between T11 and T10 and between T01 and T11 are of interest, yet neither of them is statistically significant.

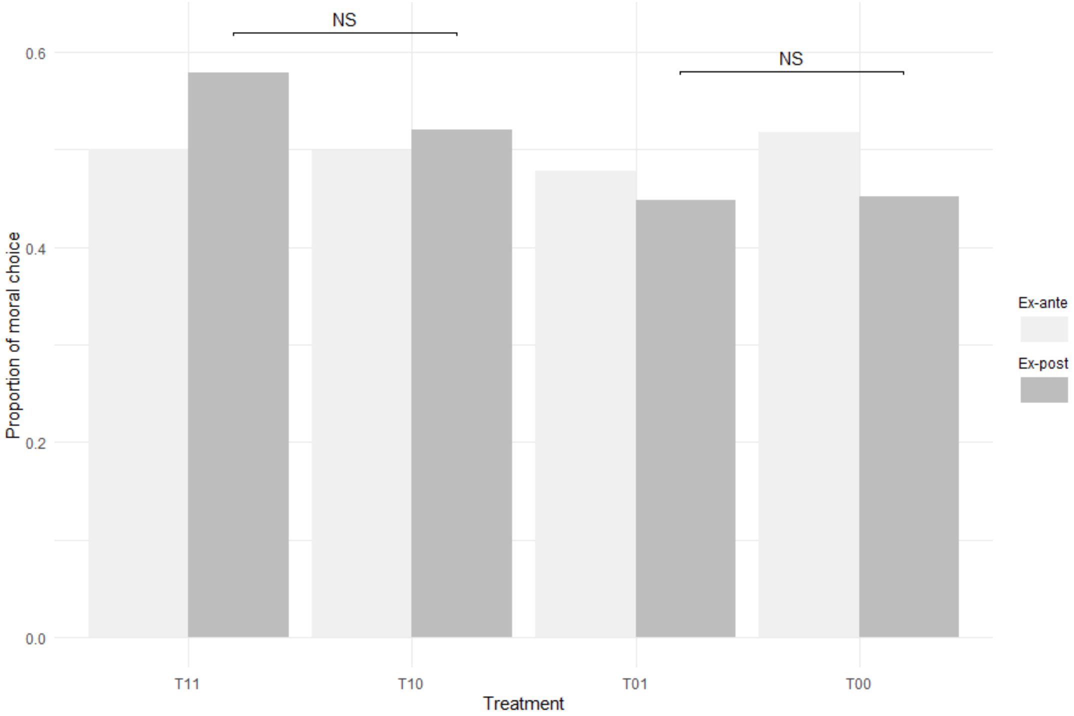


Figure 10: The *ex ante* and *ex post* perceived descriptive norm across treatments

The *ex ante* and *ex post* perceived injunctive norm

Figure 9 shows the average *ex ante* and *ex post* perceived injunctive norms over the two choices across treatments. Perceived injunctive norms are similar across treatments with one exception: the left panel shows that T11 perceives the moral choice as more socially appropriate than T10 due to imperfect randomization.

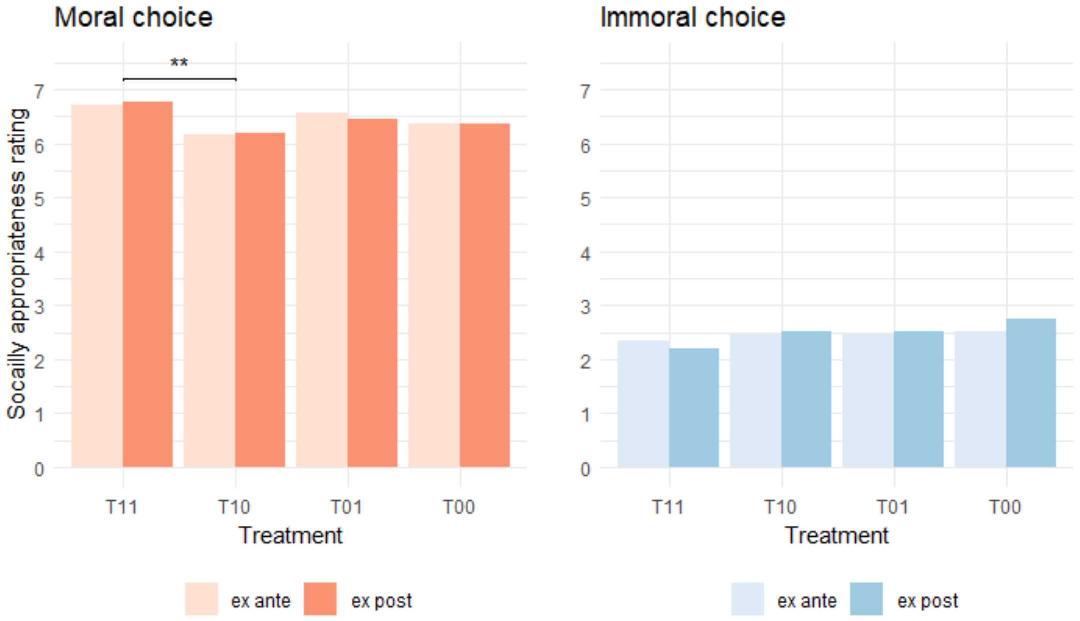


Figure 11: *ex ante* and *ex post* perceived injunctive norms across treatments

References

Abeler, J., Falk, A., Goette, L., and Huffman, D. (2011). Reference points and effort provision. *American Economic Review*, 101(2):470–92.

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, 87(4):1115–1153.

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, 95(9-10):1082–1095.

Allen, V. L. and Wilder, D. A. (1977). Social comparison self-evaluation and conformity to the group.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, 100(401):464–477.

Asch, S. E. (1955). Opinions and social pressure. *Scientific American*, 193(5):31–35.

Bandura, A. and Walters, R. H. (1977). *Social learning theory*, volume 1. Prentice-hall Englewood Cliffs, NJ.

Bartling, B., Weber, R. A., and Yao, L. (2014). Do markets erode social responsibility? *The Quarterly Journal of Economics*, 130(1):219–266.

Bénabou, R., Falk, A., and Tirole, J. (2018). Narratives, imperatives, and moral reasoning. Technical report, National Bureau of Economic Research.

Benabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678.

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Bicchieri, C., Dimant, E., Gaechter, S., et al. (2020). Observability, social proximity, and the erosion of norm compliance.

Bursztyn, L., González, A. L., and Yanagizawa-Drott, D. (2018). Misperceived social norms: Female labor force participation in saudi arabia. Technical report, National Bureau of Economic Research.

Cialdini, R. B. and Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, 55:591–621.

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6):1015.

Conway Dato-on, M. and Dahlstrom, R. (2003). A meta-analytic investigation of contrast effects in decision making. *Psychology & Marketing*, 20(8):707–731.

Crumpler, H. and Grossman, P. J. (2008). An experimental test of warm glow giving. *Journal of Public Economics*, 92(5-6):1011–1021.

Cubitt, R. P., Drouvelis, M., Gächter, S., and Kabalin, R. (2011). Moral judgments in social dilemmas: How bad is free riding? *Journal of Public Economics*, 95(3-4):253–264.

Darley, J. M. and Latané, B. (1968). Bystander intervention in emergencies: diffusion of responsibility. *Journal of Personality and Social Psychology*, 8(4p1):377.

Dimant, E., van Kleef, G. A., and Shalvi, S. (2020). Requiem for a nudge: Framing effects in nudging honesty. *Journal of Economic Behavior & Organization*, 172:247–266.

Ellingsen, T. and Johannesson, M. (2008). Pride and prejudice: The human side of incentive theory. *American Economic Review*, 98(3):990–1008.

Eriksson, K., Strimling, P., and Coultas, J. C. (2015). Bidirectional associations between descriptive and injunctive norms. *Organizational Behavior and Human Decision Processes*, 129:59–69.

Falk, A. and Szech, N. (2013). Morals and markets. *Science*, 340(6133):707–711.

Fehr, E. and Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2):63–87.

Gächter, S. and Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, 531(7595):496–499.

Gerber, A. S. and Rogers, T. (2009). Descriptive social norms and motivation to vote: Everybody's voting and so should you. *The Journal of Politics*, 71(1):178–191.

Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying aversion and the size of the lie. *American Economic Review*, 108(2):419–53.

Grossman, Z. and Van Der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, 15(1):173–217.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Hartzmark, S. M. and Shue, K. (2018). A tough act to follow: Contrast effects in financial markets. *The Journal of Finance*, 73(4):1567–1613.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., et al. (2006). Costly punishment across human societies. *Science*, 312(5781):1767–1770.

Herz, H. and Taubinsky, D. (2017). What makes a price fair? an experimental study of transaction experience and endogenous fairness views. *Journal of the European Economic Association*, 16(2):316–352.

Hoeft, L. and Mill, W. (2017). Abuse of power–an experimental investigation of the effects of power and transparency on centralized punishment. *MPI Collective Goods Preprint*, (2017/15).

Kahneman, D., Knetsch, J. L., and Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *American Economic Review*, pages 728–741.

Keizer, K., Lindenberg, S., and Steg, L. (2008). The spreading of disorder. *Science*, 322(5908):1681–1685.

Kelley, H. H. (1971). Moral evaluation. *American Psychologist*, 26(3):293.

Kenrick, D. T. and Gutierres, S. E. (1980). Contrast effects and judgments of physical attractiveness: When beauty becomes a social problem. *Journal of Personality and Social Psychology*, 38(1):131.

Khalmetski, K. and Sliwka, D. (2019). Disguising lies—image concerns and partial lying in cheating games. *American Economic Journal: Microeconomics*, 11(4):79–110.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524.

Lindström, B., Jangard, S., Selbing, I., and Olsson, A. (2018). The role of a "common is moral" heuristic in the stability and change of moral norms. *Journal of Experimental Psychology: General*, 147(2):228.

Mathes, E. W. and Kahn, A. (1975). Diffusion of responsibility and extreme behavior. *Journal of Personality and Social Psychology*, 31(5):881.

McGraw, K. M. (1985). Subjective probabilities and moral judgments. *Journal of Experimental Social Psychology*, 21(6):501–518.

Mynatt, C. and Sherman, S. J. (1975). Responsibility attribution in groups and individuals: A direct test of the diffusion of responsibility hypothesis. *Journal of Personality and Social Psychology*, 32(6):1111.

Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2008). Normative social influence is underdetected. *Personality and Social Psychology Bulletin*, 34(7):913–923.

Nowak, M. A. and Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437(7063):1291–1298.

Pepitone, A. and DiNubile, M. (1976). Contrast effects in judgments of crime severity and the punishment of criminal violators. *Journal of Personality and Social Psychology*, 33(4):448.

Prentice, D. A. and Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology*, 64(2):243.

Roth, C. and Wohlfart, J. (2018). Experienced inequality and preferences for redistribution. *Journal of Public Economics*, 167:251–262.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, 18(5):429–434.

Sherif, M. (1937). An experimental approach to the study of attitudes. *Sociometry*, 1(1/2):90–98.

Simonsohn, U. (2006). New yorkers commute more everywhere: contrast effects in the field. *Review of Economics and Statistics*, 88(1):1–9.

Simonsohn, U. and Loewenstein, G. (2006). Mistake: The effect of previously encountered prices on current housing demand. *The Economic Journal*, 116(508):175–199.

Trafimow, D., Reeder, G. D., and Bilsing, L. M. (2001). Everybody is doing it: The effects of base rate information on correspondent inferences from violations of perfect and imperfect duties. *The Social Science Journal*, 38(3):421–433.

Welch, M. R., Xu, Y., Bjarnason, T., Petee, T., O'Donnell, P., and Magro, P. (2005). "but everybody does it…": The effects of perceptions, moral pressures, and informal sanctions on tax cheating. *Sociological Spectrum*, 25(1):21–52.

Chapter 4

The persistent impact of information slant

4.1 Introduction

Between 2013 and 2015 a British high-tech company had released about 450,000 male genetically modified (GMO) mosquitoes every week in Brazil, in an effort to eliminate a disease-carrying species of mosquitoes. These GMO mosquitoes are expected to pass on the modified genes that prevent their offspring from reaching the adult stage and reproducing. One article on the result of the practice says:

> The practice was a huge success. The release of GMO mosquitoes worked tremendously well by reducing local mosquito populations by 90% during the trial.

Another article reports:

> The release of GMO mosquitoes seemed to work.The local mosquito population was reduced substantially, yet, almost 10% of them remained active.

Information slant as the above examples conveys a particular viewpoint or inclination by choice of words, tone, and framing without necessarily misrepresenting the underlying facts. Sometimes people are still persuaded by informa-

This chapter is based on a joint project with Sili Zhang.

tion slant despite being aware that it is partial and intended to manipulate their thinking,[1] which may lead to biased or even polarized beliefs.

While it is worrisome that people are conscious and yet susceptible to information slant, it is not clear whether and how people can counteract its impact. In this paper, we examine whether and to what extent a procedure, which gives people an opportunity to acquire more information from both sides, mitigates the impact of information slant. We expect an effect of this procedure because information from the other side increases the awareness and understanding of rationales for the viewpoint of the other side (Mutz, 2002). Moreover, subsequent information may dilute the impact of initial information slant. Intuitively, as people acquire more common information, they may be able to find more common ground on the issue and opinions and attitudes should move toward each other.

In a controlled experiment, we choose a concrete issue, the use of GMO mosquitoes in disease control, and construct two versions of initial information. The initial information is based on the same set of underlying facts but slanted toward either the pro-GMO side or anti-GMO side, as previous examples have shown. A key feature of the design is that subjects are well informed that the information they receive is slanted toward one side of the issue and toward which side exactly. We elicit subjects' attitudes toward the use of GMO mosquitoes after their exposure to the initial information. Attitudes are measured by a self-reported view on the GMO mosquitoes and how people allocate a donation between two organizations that favor GMO mosquitoes or against GMO mosquitoes. Next, we offer subjects an opportunity to acquire more information from both sides and elicit their attitudes for a second time to see whether and in what direction their attitudes change. Additionally, we add a treatment in which subjects are required to read all the subsequent information as a baseline for comparison. This treatment allows us to assess the pure effects of subsequent information on attitudes.

We learn from existing studies that several mechanisms may limit people's capacity to counteract information slant. A first mechanism is that people neglect the fact that the information they receive is one-sided or they are unaware of the potential effect such information may have on them. Enke (2020) finds

[1] A large literature has established that people are sometimes persuaded by information that may be slanted or distorted, for example, political information delivered by partisan-affiliated agents (e.g., DellaVigna and Kaplan (2007)), professional recommendations given by people at stake (e.g., Koehler and Mercer (2009)), or legal arguments written in a deliberately biased way to represent the interests of a specific party (e.g., Brenner et al. (1996)).

that many people do not take into account the unobserved information because it never comes to their mind.[2] A second mechanism is motivated reasoning: people are motivated to favor certain information rather than to counteract it.[3] A third mechanism is that people who already adopt certain positions may suffer from confirmation bias.[4] In particular, they tend to overestimate the strength of information that is aligned with their prior views, and thus are more susceptible to such information.

We help people reach their full capacity of counteracting the impact of information slant by minimizing the influence of the above mentioned mechanisms. To encourage subjects to reflect on and counteract information slant, we first make it clear to them that the information they receive is one-sided and that they are randomly assigned to this side. Furthermore, we ask subjects to estimate the attitudes of those who have seen information from both sides in a incentivized manner. To minimize motivated reasoning and confirmation bias, we target at people who are unfamiliar with and have no prior beliefs about GMO mosquitoes.

Our results show that information slant indeed affects subjects' attitudes: at the aggregate level, subjects' self-reported attitudes toward the use of GMO mosquitoes lean in the direction aligned with the initial slant compared to the attitudes of those who have seen both sides; and subjects allocate a larger fraction of donation to the organization favoring the side aligned with the initial information than those who have seen both sides. The opportunity to acquire information from both sides only slightly mitigates the impact of information slant, as there is mild evidence that attitudes move closer after information acquisition, whereas exogenously imposed information does not narrow the gap between subjects exposed to different sides. We find that the impact of initial information slant may be persistent: it shifts subjects' attitudes, and the shift in attitudes in turn induces confirmation bias which distorts the way people process subsequent information. The induced confirmation bias partly accounts for people's failure of fully counteracting the impact of initial information slant.

A vast literature has established that slanted information affects people's at-

[2] Note that our aim is to study how people deal with information slant embedded in natural language. In an abstract setting as Enke's, information in the form of math language, i.e., numbers, does not leave much room for slant.

[3] Literature shows that people seek, avoid, or misinterpret information in service of self-interest. See, e.g., Bénabou and Tirole (2002); Dana et al. (2007); Ditto and Lopez (1992); Epley and Gilovich (2016); Exley (2016); Grossman and Van Der Weele (2017); Kunda (1990); Taber and Lodge (2006).

[4] A vast psychological literature provides evidence of confirmation bias. See, e.g., Darley and Gross (1983); Klayman and Ha (1987); Lord et al. (1979); Nickerson (1998); Plous (1991). For economic research on confirmation bias, see e.g.,Rabin and Schrag (1999); Fryer Jr et al. (2019); Golman et al. (2016).

titudes and behaviors; for example, see Chiang and Knight (2011); DellaVigna and Kaplan (2007); DellaVigna and Gentzkow (2010); Gentzkow and Shapiro (2006); Grigorieff et al. (2020); Lippmann (1946); Martin and Yurukoglu (2017). However, direct evidence on the persuasion effects of information slant when people are aware thereof is scarce. The closest study to our paper is Brenner et al. (1996). They use a hypothetically legal dispute to show that one-sided evidence leads to biased predictions and judgments even when people know full well that the information available to them represents only one side of the dispute. They find that asking subjects to contemplate the relative strength of the two sides greatly reduces the bias induced by one-sided evidence.

Our study also adds to the growing literature on the demand for information about multi-side issues. Most studies investigate the demand for information in a naturally occurring environment, for example, news consumption. These studies offer a mixed and nuanced view of how people select information sources; see Bakshy et al. (2015); Flaxman et al. (2016); Halberstam and Knight (2016); Garrett (2009); Gentzkow and Shapiro (2006); Iyengar and Hahn (2009). The overall take-away message is that, despite the ability to select information sources, a majority of people do not actively avoid information that challenges their own views. Other studies use more controlled environments to examine how people select information sources. A closest related study to this paper is Chen and Yang (2019). They use a field experiment in China to investigate people's information acquisition behavior when they are given access to uncensored Internet. It is no secret that China develops a sophisticated Internet censorship apparatus to restrict access to regime-threatening information. In this specific context, they find that temporarily offering free access alone does not lead to more acquisition of politically sensitive information, unless people are incentivized to do so. They also find that acquiring politically sensitive information causes substantial and persistent changes in citizens' knowledge, beliefs, attitudes, and intended behaviors. Hjort et al. (2019) investigates how policy-makers demand research findings that may inform them about the effectiveness of policies. They find that policy-makers are willing to pay for policy-related research results, and update their beliefs when informed of those findings. They do not find evidence for confirmation bias or information acquisition driven by motivated reasoning. In an abstract setting, Charness et al. (2018) studies how people choose between two information sources that are biased toward or against their prior beliefs. They find that subjects are systematically prone to confirmatory information. In the current paper, we combine a naturally-occurring issue with a controlled information environment in which people are neutral to both the involved issue and the information environment.

The remainder of this paper is organized as follows. Section 2 presents the experimental design and the main hypotheses. Section 3 provides the main results. Section 4 explores the mechanisms, and Section 5 concludes.

4.2 Experimental design

This section first outlines a pilot experiment which aims at selecting an issue that reduces the influence of motivated reasoning and confirmation bias as much as possible. Based on the results of the pilot, we choose an issue that best fits our need for the main experiment. Then we describe the main experiment.[5] Finally, we describe the outcome variables of interest and formulate our hypotheses.

The reason why we do not use an abstract setting, for example, a setting in which subjects form beliefs about the value of a unknown variable based on the signals they receive, is that we are interested in how people deal with information slant embedded in natural language. While an abstract setting as in Enke (2020) can easily get rid of motivated reasoning, it would also take away the subjectivity that makes information slant.

4.2.1 Pilot: issue selection

An ideal issue for the experiment does not invoke ego driven or self-interest driven reasoning, and people have not formed strong opinions about it so that they would not suffer from confirmation bias. So to maximize the chance of finding an ideal issue, we restricted our attention to a few topics that are either new (recent scientific and technological developments) or exotic to our target subjects. In the pilot, we tested a list of candidate topics:

- Genetically modified (GMO) mosquitoes to prevent disease spread by limiting their reproduction
- Traditional Chinese medicine
- Complementary and alternative medicine
- The General Data Protection Regulation (GDPR)

[5] The experiment and the analysis plan were pre-registered on the AEA RCT Registry (AEARCTR-0005595). For more details, see Appendix D.

For each of the above topics, we asked respondents four questions: how well they know the topic, whether they hold strong prior attitudes toward the topic, whether they are comfortable with evidence against their opinions on the topic, and whether they think these issues are important and relevant. The first two questions are our criteria for selecting the topic.

We recruited 163 Mturk workers for the pilot. Among all the topics that we tested, genetically modified (GMO) mosquito is the topic that respondents have the least knowledge about and the weakest prior attitudes toward. We then constructed a more concrete and specific issue within this topic, which will be explained in the next section.

4.2.2 Main experiment

Overview

The experiment consists of two parts: a pre-screening survey and the main experiment. The pre-screening survey is designed to screen out respondents who already hold strong prior attitudes toward GMO mosquitoes and those who are very knowledgeable about GMO mosquitoes. Subjects who pass the pre-screening are immediately invited to continue with the main experiment.

In the pre-screening survey, participants are first briefly introduced to the topic, GMO mosquitoes. We then ask them: (1) how much do you know about genetically modified mosquitoes without searching online? and (2) what is your general attitude toward GMO mosquitoes? For the first question, they state their answers on a five-point scale from “none at all” to “a great deal”. For the second question, they choose from “extremely negative” to “extremely positive” on a five-point scale. Only those who do not know “a great deal” about GMO mosquitoes and are neither “extremely negative” nor “extremely positive” about GMO mosquitoes are invited to participate in the second part of the study, which is the main experiment.

The issue we are using in the main experiment is “releasing genetically modified mosquitoes in the wild to eliminate disease-carrying mosquitoes”. There are two mutually-exclusive sides of the issue, whether one supports or opposes such practices. For simplicity, we refer to the two sides as pro-GMO and anti-GMO in the remaining of this paper.

Figure 4.1 overviews the experimental setup. We implement a 3x2 (Both/Pro/Anti × Exo/Endo) between-subject design. Two treatment variations occur in two phases of the experiment. In phase I, we vary what information subjects are initially exposed to. In phase II, we vary whether subsequent arguments from both sides are exogenously imposed or can be endogenously selected. In total, we have six treatment groups.

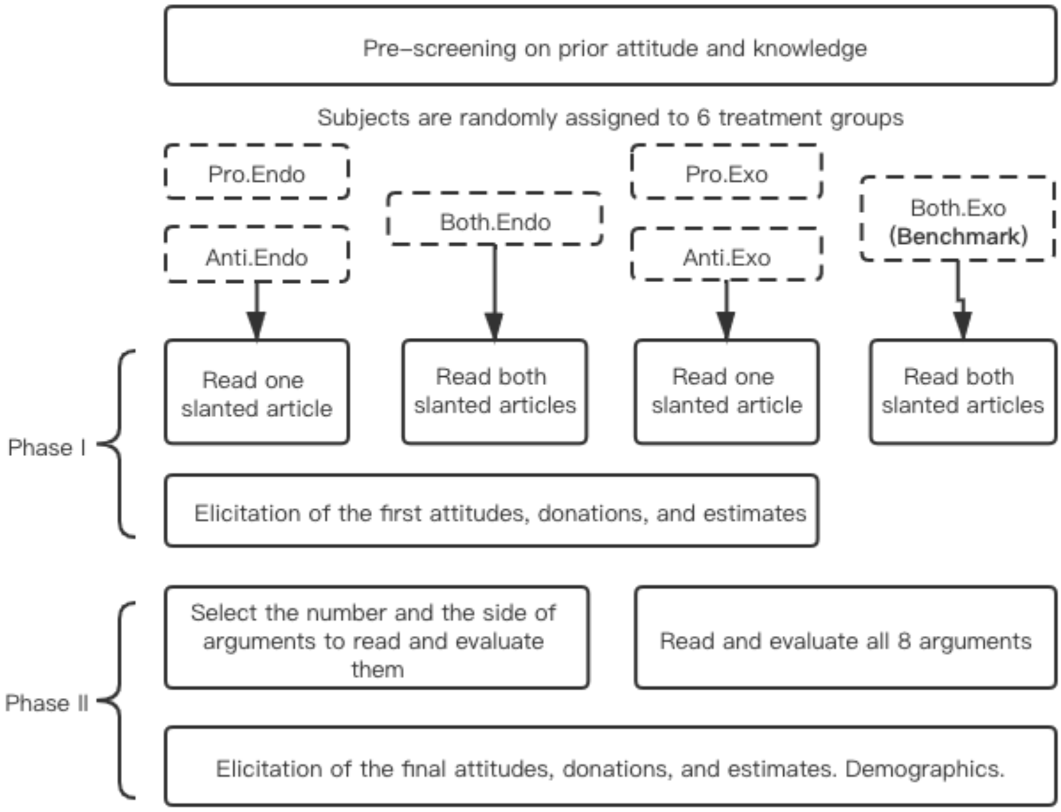

Pre–screening on prior attitude and knowledge
Subjects are randomly assigned to 6 treatment groups
Pro.Endo
Anti.Endo
Both.Endo
Pro.Exo
Anti.Exo
Both.Exo
(Benchmark)
Read one slanted article
Read both slanted articles
Read one slanted article
Read both slanted articles
Phase I
Elicitation of the first attitudes, donations, and estimates
Select the number and the side of arguments to read and evaluate them
Read and evaluate all 8 arguments
Phase II
Elicitation of the final attitudes, donations, and estimates. Demographics.

Figure 4.1: Overview of Treatments

In phase I, subjects are exposed to one of the two slanted versions of information or both. The initial information takes the form of an article on the issue. We construct two versions of the article: a pro-GMO version and an anti-GMO version. These two articles are based on the same underlying facts published in academic journals or news reports but are slanted toward one side of the issue or the other side (see Appendix A). Subjects assigned to the *Both* condition read both versions while those assigned to *Pro*/*Anti* conditions read either the pro-GMO article or the anti-GMO article.

In phase II, subjects have access to subsequent information from both sides. The subsequent information we provide is eight arguments, each of which favors either the pro-GMO side or the anti-GMO side. Half of the subjects are randomly assigned to the *Exo* condition and are required to read and evaluate all arguments from both sides. The other half of the subjects are assigned to the *Endo* condition. They are able to select how many arguments to read from either side.

Both.Exo is the benchmark group. For data analyses, sometimes we pool together the data of several treatment groups. In such case, we name the pooled data by the treatment variation shared by these pooled treatment groups. For example, *Pro* groups are *Pro.Endo* group and *Pro.Exo* group pooled together.

The two phases each consists of several stages as summarized below. The details of each stage will be given in the next section.

- Phase I:
  - Stage 1.1: exposure to one of the slanted articles or both;
  - Stage 1.2: elicitation of attitudes about the use of GMO mosquitoes;
  - Stage 1.3: elicitation of estimates of the benchmark's attitude about the use of GMO mosquitoes;
- Phase II:
  - Stage 2.1: information acquisition/evaluation;
  - Stage 2.2: elicitation of attitudes about the use of GMO mosquitoes;
  - Stage 2.3: elicitation of estimates of the benchmark's attitude about the use of GMO mosquitoes;
  - Stage 2.4: demographics.

Stages

This section explains the experimental setup stage by stage in more detail. Sample Instructions can be found in Appendix C.

Stage 1.1: exposure to one of the slanted articles or both We begin by bringing up the issue, genetically modified (GMO) mosquito. In this stage, subjects read an article or two articles on a genetically modified mosquitoes field practice in Brazil. We write two versions of the article that are based on the same underlying facts published in academic journals or news reports but are slanted towards either the pro-GMO side or the anti-GMO side (see Appendix A).

Subjects in *Both* treatments read both versions which are presented in a random order. Subjects in *Pro* or *Anti* groups receive only the pro-GMO version or only the anti-GMO version respectively. All subjects are explicitly told that there exist two versions. They are also informed which version(s) they will read and that it is randomly determined.

When they finish reading, we ask a few comprehension questions about the article to make sure that the subjects have actually read it and more importantly that they have grasped the critical facts in the article. Subjects should answer all comprehension questions correctly in order to proceed. They can try as many times as they need and if they are unsure about the answers, they can go back to the article(s).

Stage 1.2: elicitation of attitudes toward the use of GMO mosquitoes We use two measures of attitude toward the issue. First, subjects report on a nine point scale whether they think the practice of releasing GMO mosquitoes to prevent the spread of disease is appropriate use of technology or is taking technology too far.[6] Second, we ask subjects to allocate a donation worth of 100 dollars between a pro-GMO mosquito organization[7] and an anti-GMO mosquito organization[8]. We inform subjects that we make the donation according to the average of all responses. Subjects' responses are recorded as their first attitudes.

Stage 1.3: elicitation of estimates of the benchmark's attitude Subjects in the benchmark group skip this stage. Subjects in the other five treatment groups are asked to estimate the average attitude of the benchmark group, more specifically, the average allocation out of 100 dollars to the pro-GMO organization by subjects in the benchmark group. Subjects choose an interval for their estimates. For example, they can state that the average percentage of donation allocated to the pro-GMO organization is between 45% and 55% by dragging the ends of a slider. They will earn a bonus worth up to one dollar if the actual average allocation of the benchmark group falls into the interval they choose. The amount of the bonus decreases linearly in the width of the interval.[9] Their responses are recorded as the first estimates.

[6] This question was used by PEW in its survey on Americans' views on genetic engineering of Animals

[7] The pro-GMO mosquito organization, Liverpool School of Tropical Medicine, is a research institute that has a team dedicated to developing toolkits for the genetic manipulation of insects to fight against diseases.

[8] The anti-GMO mosquito organization, GeneWatch, is a not-for-profit group that monitors developments in genetic technologies. This organization believes that GMO mosquitoes are a mistake and should be avoided in the future.

[9] The amount of bonus equals to 100 cents minus the difference between the upper bound percentage point and the lower bound percentage point of the interval subjects choose. In the above example, the bonus amount equals to $100 - (55 - 45) = 90$ cents.

Stage 2.1: information acquisition/evaluation In this stage, subjects have free access to in total eight arguments from both pro-GMO side and the anti-GMO side. We construct these arguments on the basis of answers posted on Quora. We picked 4 arguments in favor of the pro-GMO side and 4 in favor of the anti-GMO side, and rewrote them such that they are comparable in terms of length and clarity. Each argument takes a native English speaker around 20 seconds to read.

Subjects in the three *Exo* treatment groups are shown all eight arguments while subjects in the three *Endo* treatment groups can choose how many arguments to read from either side. Additionally, subjects are asked to rate how convincing each argument they read is on a seven point scale.

Stage 2.2: elicitation of attitudes toward GMO mosquitoes This stage is a repetition of stage 1.2. Subjects' responses are recorded as their final attitudes.

Stage 2.3: elicitation of estimates of the benchmark's attitude For all subjects except those in the benchmark group, this stage is a repetition of stage 1.3. Their responses are recorded as the final estimates. Additionally, we ask them to estimate the first attitude of their opposing side. That is, subjects in *Pro* groups estimate the average first attitude of those in *Anti* groups and vice versa. Subjects earn an extra bonus of 50 cents if they are correct.[10]

Subjects in the benchmark group skip stage 1.3. Now they are asked to estimate the average attitude of the benchmark group for the first time. We record their responses as the final estimates. So subjects in the benchmark group do not have "first" estimates.

Stage 2.4: demographics Finally, we collect basic demographics (gender, age, education, income, state of residence, religion), political preferences, and news consumption habits.

4.2.3 Outcome variables and Hypotheses

Outcome variables of interest

Self-reported attitude toward GMO mosquitoes The first measure of attitude is self-reported attitude. Specifically, subjects state on a scale of [-4,4]

[10] We ask subjects to submit a whole number as their point estimate. If the point estimate coincides with the true percentage (rounded to a whole number), they earn 50 cents.

whether it is taking technology too far or is appropriate use of technology to release GMO mosquitoes to eliminate disease-carrying mosquitoes.

Donation to the pro-GMO organization The second measure of attitude is the proportion ([0,100]) of donation they would like to allocate to the pro-GMO organization.

Remember that we elicit each measure of attitude, that is, self-reported attitude and donation, in both phase I and phase II. The phase I attitude will be referred to as the first attitude, and the phase II attitude will be referred to as the final attitude. For subjects in the benchmark group, we elicit their attitudes only once at the end of phase II as their final attitude.

Estimate of the benchmark's final donation This outcome variable is self-explanatory, i.e., subjects' estimates of the average final percentage of donation allocated to the pro-GMO organization by the benchmark group. Subjects choose an interval for the estimate. We use the middle point of the interval as the point estimate.

We are most concerned about subjects' attitudes toward GMO mosquitoes. Self-reported attitude is straightforward but cheap talk. We choose a second measure–the donation decision–because it has real consequences and reflects one's attitude toward GMO mosquitoes. The donation decision is incentivized in some sense, but it may be driven by other motives, for example, concerns for the quality of the organization. We also elicit subjects' estimates of the benchmark group's attitude because the estimates may be related to one's own attitude due to social projection and we use this elicitation to incentivize information acquisition. These three outcome variables complement each other and we expect them to correlate with each other.

Information acquisition We are interested in the total number of arguments subjects in *Endo* groups acquire and the relative numbers of pro-GMO and anti-GMO arguments they acquire.

Information evaluation This outcome variable is subjects' evaluation of the convincingness of each acquired argument on a seven point Likert scale: [-3,3].

Hypotheses

In this section, we formulate our hypotheses based on existing literature.

Phase I: the impact of information slant on first self-reported attitudes, donations, and estimates By design, we restrict our attention to subjects who have little knowledge and mild attitudes about GMO mosquitoes prior to our experimental manipulation. It is unlikely that they are attached to either side to begin with. And *Pro* and *Anti* groups are informed that they will receive an article slanted toward one side of the issue and toward which side exactly. In principle, there should be no difference in the attitudes across treatments, since the pro-GMO article and the anti-GMO articles are randomly assigned to subjects and the two articles share the same underlying facts.

However, as literature shows, people can still be affected by information known to be partial.[11] Our first hypothesis is that subjects will still be affected by information slant. Specifically, those exposed to only one of the slanted articles will be "biased"[12] toward the side aligned with initial article relative to those exposed to both slanted articles. That is,

Hypothesis 4.1 (Att)**.** *Pro groups are more positive about the use of GMO mosquitoes than Both groups. Anti groups are more negative about the use of GMO mosquitoes than Both groups.*

Social projection (Katz et al., 1931; Mullen et al., 1985; Mullen and Hu, 1988) suggests that, when people are asked to predict others' preferences based on limited information, "it is reasonable to assume that other people prefer the same things that we do" (Tarantola et al., 2017). In the context of our study, if subjects' attitudes are affected by information slant, that is, we find evidence to support Hypothesis Att, then social projection leads to the hypothesis that subjects exposed to information slant will be biased in the direction of their own attitudes when estimating the benchmark's attitude.

Hypothesis 4.2 (Est)**.** *Pro groups' estimates of the benchmark's attitude are higher than those of Both groups. Anti groups' estimates are lower than those of Both groups.*

Note that phase I of the experiment, in which subjects are exposed to information slant, sets the stage for phase II. We test Hypothesis Att to check if our

[11] For a review on the persuasion effect, see DellaVigna and Gentzkow (2010).

[12] Note that the word "biased" here is not normative as we can never define what is the unbiased attitude toward GMO mosquitoes by any objective standard. We use *Both* groups as the benchmark without claiming that *Both* groups are unbiased or that everyone should possess the same attitude as *Both* groups do. We refer to "bias" in attitudes as the difference between *Pro*/*Anti* treatment groups and *Both* groups.

"manipulation" works. If yes, we move on to the main hypotheses regarding subsequent information acquisition and its effect on attitudes and estimates.

Phase II: subsequent information acquisition Subjects have a monetary incentive to accurately estimate the benchmark's final attitude. We inform the subjects that the benchmark group has seen all of the eight arguments. Therefore, we hypothesize that subjects acquire all the available information in order for an accurate estimate.

Hypothesis 4.3 (InfoAcq)**.** *Subjects in Endo groups acquire all of the eight arguments.*

Since there is a tradeoff between the expected amount of bonus and the time spent on reading the arguments, not acquiring all the arguments does not necessarily indicate irrationality. In cases where Hypothesis InfoAcq is rejected, we then form an alternative hypothesis concerning how subjects choose between arguments from both sides, in particular, whether subjects are confirmation-seeking, contradiction-seeking, or neither. A large literature suggests that people tend to look for information that confirms their prior beliefs (Charness et al., 2018; Frey, 1986; Iyengar and Hahn, 2009; Klayman and Ha, 1987; Pariser, 2011; Prior, 2007; Sears and Freedman, 1967). While contraction-seeking behavior exists in theory, it is extremely rare in empirical studies. Thus, we hypothesize that people display confirmation-seeking behavior at the aggregate level, that is, they acquire more arguments from the side aligned with the initial article than arguments from the other side.

Hypothesis 4.4 (InfoSeek)**.** *Subjects in Pro.Endo group acquire more pro-GMO arguments than anti-GMO arguments. Subjects in Anti.Endo group acquire more anti-GMO arguments than pro-GMO arguments.*

Phase II: attitudes, donations, and estimates after endogenous information acquisition Conditional on finding support for Hypothesis Att, we expect subjects' attitudes and estimates to move toward convergence after information acquisition.

However, we cannot predict how close attitudes and estimates eventually are, since it depends on the degree of divergence and the relative strength of the arguments.

Hypothesis 4.5 (Endo)**.** *In Endo groups, attitudes and estimates of different sides move toward convergence after information acquisition.*

Phase II: the impact of exogenous information There is a chance that subjects in *Endo* groups acquire very few or no argument. In such cases, we are not able to test the effect of subsequent information on attitudes. So we add the *Exo* treatment and ask subjects in *Exo* groups to read and evaluate all the arguments. If more common information helps narrow the gap in attitudes, subjects in *Exo* groups would move closer in attitudes to a greater extent than those in *Endo* groups after information evaluation.

Hypothesis 4.6 (Exo)**.** *In Exo groups, attitudes and estimates of different sides move toward convergence after information acquisition to a greater extent than in Endo groups.*

4.3 Results

4.3.1 Descriptives

We recruited subjects through Mturkdata[13] in May 2020. A total of 969 respondents completed the pre-screening survey. 802 (82.7%) of them are not very knowledgeable and do not have extreme attitudes about GMO mosquitoes.[14] These respondents were immediately invited to participate in the main experiment. After excluding incomplete responses, we collected 716 observations in total, with more than 100 observations in each treatment group.

The pre-screening survey takes less than 1 minute. Respondents earn 5 cents by finishing it. The main experiment that immediately follows takes 15-20 minutes. Respondents who agreed to participate earned at least 2 dollars by completing the main experiment. Additionally, they could earn a bonus depending on their responses to the incentivized questions. The average bonus is 30 cents.

Upon recruitment, we made it clear to Mturk workers that after completing the 1 minute survey, they might be invited to participate in a follow-up study and we mentioned the length and payment of the follow-up study. Participation in the follow-up study was voluntary. Attrition is not a concern, because only 36 invited respondents (less than 5%) did not continue with the follow-up study and they do not differ from those who agreed to continue in terms of prior attitude and knowledge.

[13] Mturkdata is a platform that helps publish studies on the Mturk platform.

[14] The pre-screening question on attitude is: what is your general attitude toward genetically modified (GMO) mosquitoes? Subjects choose their attitude from 5 categories that range from "extremely negative" to "extremely positive".

Table 2 in Appendix E presents summary statistics about the subjects who completed the main experiment.

Randomization check Randomization was done by a randomizer embedded in the survey. It was mostly successful, except that subjects' prior attitudes about GMO mosquitoes are statistically different across treatments at the 10% level; see Appendix E. Therefore, we control for priors in the subsequent analyses.

4.3.2 Main results

In this section, we first examine how much subjects' attitudes and estimates are affected by information slant when they are aware thereof. We then explore to what extent they use the opportunity to acquire additional information from both sides to counteract the impact of information slant. Finally, we discuss the effect of exogenously imposed information.

Does information slant affect attitudes and estimates?

As a manipulation check, we start by looking at the effects of one-sided evidence on subjects' attitudes and estimates. Remember that we use two measures of attitudes: the self-reported attitude towards the practice of using GMO mosquitoes to eliminate mosquito-carrying diseases and the allocation of donation between a pro-GMO organization and an anti-GMO organization. The third outcome of interest is subjects' estimates of the benchmark's allocation decision, which are hypothesized to be correlated with the attitudes.

Figure 4.2 shows the average self-reported attitude toward GMO mosquitoes, average donation to the pro-GMO organization, and the average estimate of the benchmark's attitude among subjects in *Pro* groups, *Both* groups, and *Anti* groups.[15]

The impact of information slant on attitudes is evident. In terms of the self-reported attitude, subjects who read a pro-GMO article are significantly more positive toward the use of GMO mosquitoes than those who read both articles (Mann-Whitney test, $p = 0.014$). And subjects who read an anti-GMO article are significantly more negative toward the usage of GMO mosquitoes than those who read both articles (Mann-Whitney test, $p = 0.012$). information slant affects the allocation of donation in a similar manner but to a smaller extent. Subjects

[15] We had not mentioned anything about the possibility of reading more arguments until phase II. Therefore, we do not distinguish subjects in the *Exogenous* and *Endogenous groups* in phase I.

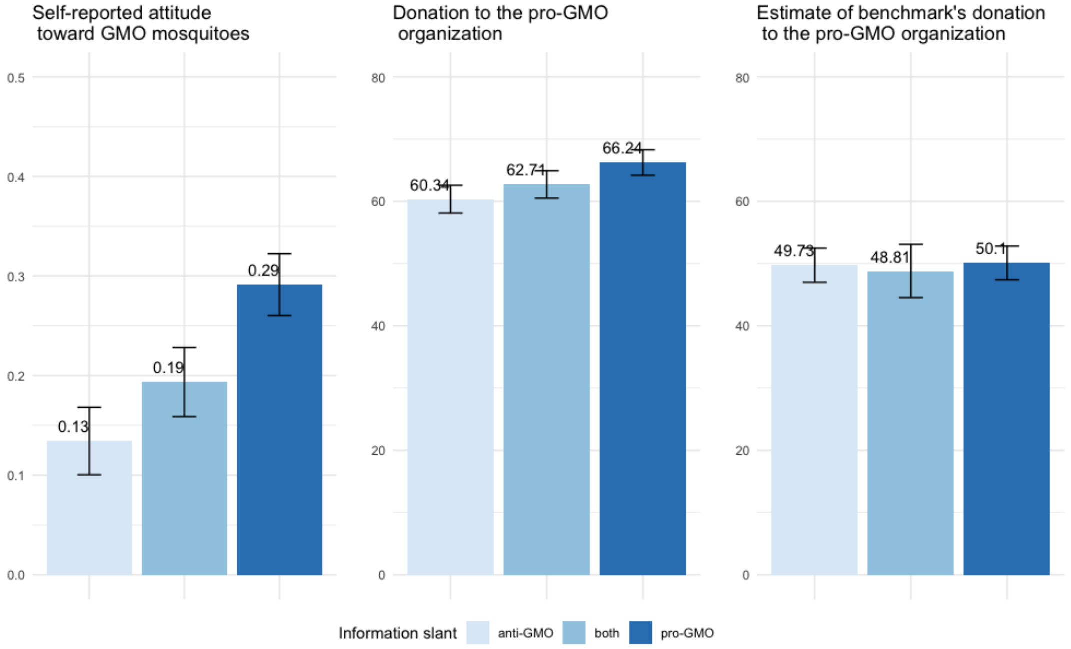


Figure 4.2: Phase I: average self-reported attitude, average donation to the pro-GMO organization, and average estimate of the benchmark's donation to the pro-GMO organization among all subjects. Self-reported attitude is normalized to [-1,1].

in *Pro* groups allocate a significantly larger fraction of donation to the pro-GMO organization than those in the *Anti* groups (Mann-Whitney test, $p = 0.001$), yet neither *Pro* groups nor *Anti* groups are statistically different from *Both* groups (Mann-Whitney test, $p = 0.204$ and $p = 0.235$ respectively).

However, we do not find any difference in the estimates of the benchmark's attitude. Subjects in all treatment groups predominantly locate their guesses around 50%, whereas the actual percentage of final donation allocated to pro-GMO organization made by the benchmark group is 58% (see Figure 11 in the appendix). This result suggests that social projection does not seem to play much of a role when subjects are incentivized to estimate others' attitude.

In sum, the evidence supports Hypothesis Att but not Hypothesis Est. The results are summarized below:

Result 4.1. *(Att) After being exposed to information slant, Pro groups hold a more positive attitude toward GMO mosquitoes than Both groups and Anti groups have a more negative attitude toward GMO mosquitoes than Both groups.*

Result 4.2. *(Est) information slant does not cause divergence among subjects in their estimates of the benchmark group's attitude.*

Next, we look at what people do when they are offered access to more information from both sides and if endogenous information acquisition can narrow the gap in attitudes.

Does information acquisition mitigate the impact of information slant?

Information acquisition pattern We hypothesize that subjects in *Endo* groups acquire all the arguments in order to estimate the benchmark's attitude more accurately.

Figure 4.3 below presents the cumulative distribution of the number of acquired arguments. In fact, only 25.1% of the subjects choose to read all arguments. 68% of them read four or even less arguments. The information acquisition patterns are very similar across the three treatment groups. Therefore, we reject Hypothesis InfoAcq and turn to Hypothesis InfoSeek.

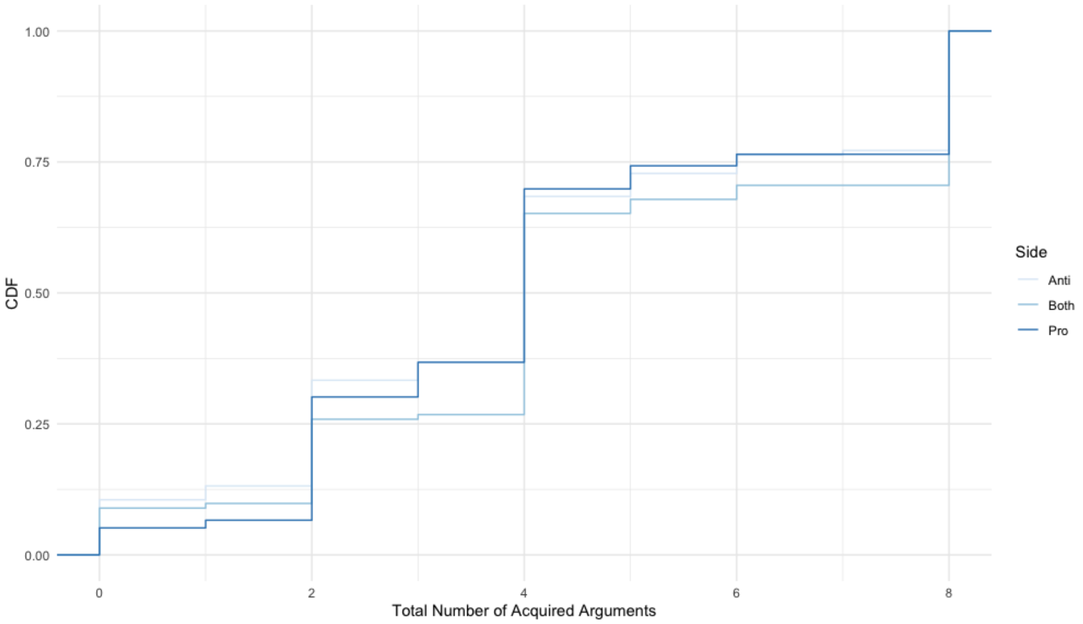


Figure 4.3: CDF of the total number of arguments acquired among subjects in *Pro.Endo* group, in *Both.Endo* group, and in *Anti.Endo* group.

We hypothesize that subjects may acquire more confirming arguments than contradicting ones if they do not acquire all the available arguments. However, subjects do not seem to discriminate between information from different sides: 85.6% of them choose equal numbers of pro-GMO arguments and anti-GMO arguments. We find no difference in the relative numbers of acquired arguments from the two sides across treatment groups, and hence we reject Hypothesis In-

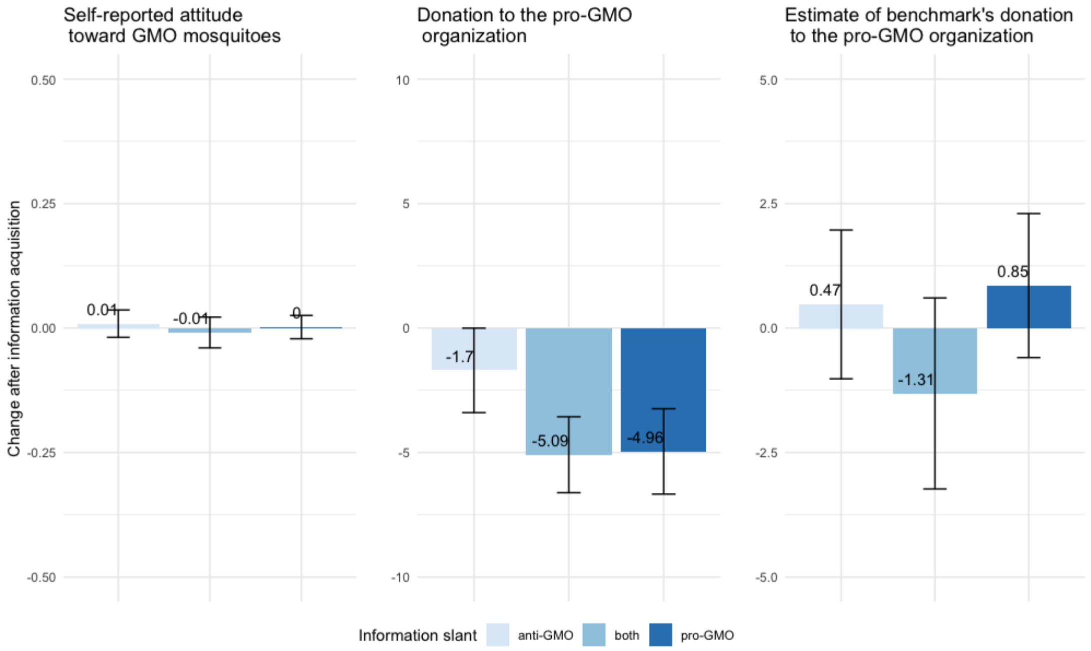


Figure 4.4: Change in self-reported attitude, donation to the pro-GMO organization, and estimate among subjects in *Pro.Endo* group, *Both.Endo* group and *Anti.Endo* group.

foSeek. Taken together, we have the following results:

Result 4.3. *(InfoAcq) The majority of the subjects do not acquire all the available information even when they are incentivized to do so.*

Result 4.4. *(InfoSeek) Subjects do not display a confirmation-seeking pattern in information acquisition; they predominantly choose equal numbers of arguments from both sides.*

Do attitudes and estimates converge after information acquisition? Again, we focus on subjects in *Endo* groups. To see if attitudes are converging, we compare subjects' attitudes, donations, and estimates before and after information acquisition. Figure 4.4 below demonstrates the change in the self-reported attitude, the change in the donation allocation, and change in the estimate of the benchmark's attitude after information acquisition among subjects in *Endo* groups.

While self-reported attitudes and estimates of the benchmark's attitude barely change at the treatment level, we find some evidence that donation decisions move closer; see Figure 8 and Figure 9 in the appendix for more details. Specifically, subjects in *Pro.Endo* and in *Both.Endo* lower their donations to the pro-GMO organization by 5 percentage points, more than subjects in *Anti.Endo* do

(Mann-Whitney test, $p = 0.168$). This renders the difference in the final donation allocation between *Pro.Endo* and *Anti.Endo* non-significant (Mann-Whitney test, $p = 0.279$), which was statistically significant before information acquisition (Mann-Whitney test, $p = 0.042$). In other words, the three treatment groups are converging in their donation decisions after information acquisition. The result provides some support for Hypothesis Endo.

Result 4.5. *(Endo) There is mild evidence that attitudes are converging among people who are initially exposed to diverse information after information acquisition.*

Summing up the results so far, we find that information slant indeed affects subjects' attitudes and causes an evident gap between people exposed to different sides. The gap narrows slightly when subjects are given the opportunity to acquire more information from both sides, although subjects do not exploit this opportunity in the sense that the majority of them acquire no more than half of all the available information. Next, we investigate if imposing all the information on subjects would narrow the gap in attitudes to a greater extent.

Does exogenously imposed information mitigate the impact of information slant?

Contrary to what we hypothesize, there is no evidence of convergence in attitudes among subjects in *Exo* groups. Figure 4.5 presents the change in self-reported attitude, the change in donation allocation, and the change in estimate of the benchmark's attitude among subjects in *Exo* groups by the side of the initial information. The self-reported attitude of *Both.Exo* and *Anti.Exo* are significantly different from *Pro.Exo* and they remain so after subjects read all the arguments; see Figure 10 and Figure 11 in the appendix. The gap in donation allocation is even wider, as *Anti.Exo* lowers their donation after reading the arguments. As a result, the difference in donation allocation between *Pro.Exo* and *Anti.Exo* becomes statistically significant at the 10% level (Mann-Whitney test, $p = 0.062$).

As to the estimates, the benchmark group is the most accurate among the three treatment groups in estimating its own final donation: the average guess of the benchmark group is 55.8% while the truth percentage is 58%. Meanwhile, both *Pro.Exo* and *Anti.Exo* underestimate the donation made by the benchmark group, but their guesses are not statistically different from each other (Mann-Whitney test, $p = 0.72$).

In sum, we reject Hypothesis Exo:

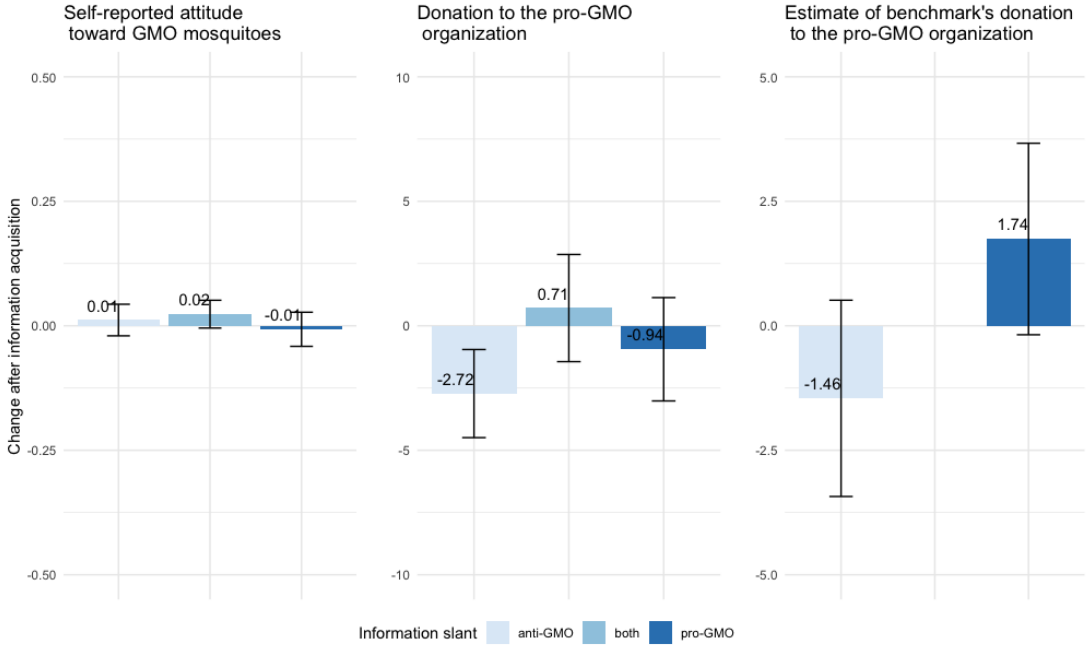


Figure 4.5: Change in average self-reported attitude, average donation to the pro-GMO organization, and average estimate among subjects in *Pro.Exo* group, *Both.Exo* group and *Anti.Exo* group.

Result 4.6. *(Exo) We find no evidence that exogenously imposed information leads to greater convergence of attitudes among subjects who are initially exposed to diverse information than among subjects who can endogenously acquire information.*

While endogenous information acquisition slightly narrows the gap in attitude induced by initial slanted articles, forcing subsequent information from both sides on people does not; it may even backfire. This raises a question as to why imposed information widens the gap in attitudes, which remains to be further investigated.

4.4 Mechanisms

Previous results show that information slant has an impact on people's attitudes even when they are aware thereof. In particular, it causes a gap in attitudes among people who are exposed to diverse information. The opportunity to acquire more information from both sides only slightly mitigates the impact of information slant, while imposing information from both sides seems to do the opposite. In this section, we discuss two mechanisms that may hinder people from counteracting the impact of information slant.

4.4.1 Awareness of the impact of information slant

People do not make efforts to counteract the impact of information slant possibly because they are not aware that information slant may have an impact. In this section, we show evidence that can rule out this possibility.

In stage 2.3, we ask subjects who receive only one article to estimate the first average attitude of their opposing side, i.e. the first allocation decision made by those who read an article of the opposing side. By comparing their estimates of the benchmark group's attitude and the opposing side's attitude, we are able to see if subjects expect a discrepancy in attitudes between those who receive only one slanted article and those who receive both articles. Put it differently, suppose a subject A receives article H_A, a subject B receives article H_B, and a subject C receives both articles: $H_A + H_B$. We ask subject A to estimate the attitude of B and the attitude of C. The difference between these two estimates is subject A's belief about the effects of article H_A on attitudes. If the difference is non-zero, then we may conclude that subject A expects information slant to affect people's attitude.

Figure 4.6 organizes the estimate of the benchmark group's attitude and the estimate of the opposing side's attitude given by *Pro* groups and *Anti* groups. The left panel compares *Pro* groups' average estimate of the benchmark's attitude and their average estimate of *Anti* groups' attitude. *Pro* groups believe that *Anti* groups allocate a smaller fraction (34.4% v.s. 51.3%) of the donation to the pro-GMO organization than *Both* groups would allocate. The right panel of Figure 4.6 shows the *Anti* groups' estimate of the benchmark's attitude and their estimate of *Pro* groups' attitude. They believe that *Pro* groups would allocate a larger fraction (57.7% v.s. 49.2%) of the donation to the pro-GMO organization than *Both* groups. The pairwise differences between the two estimates are significant at 1% level (Mann-Whitney test). In other words, subjects believe that those who have read two articles are more positive than those who only read an anti-GMO article and are more negative than those who only read a pro-GMO article.

The results demonstrate that subjects expect an impact of information slant on attitudes, at least on others' attitudes. In fact, while people correctly anticipate the direction of the impact, they overestimate its magnitude: the true difference in donation allocation between *Pro*/*Anti* groups and *Both* groups is smaller than what subjects expect and the true difference is not statistically significant. Thus, the result that subjects do not fully counteract the impact of information slant cannot be explained by a lack of awareness of either the one-sided nature

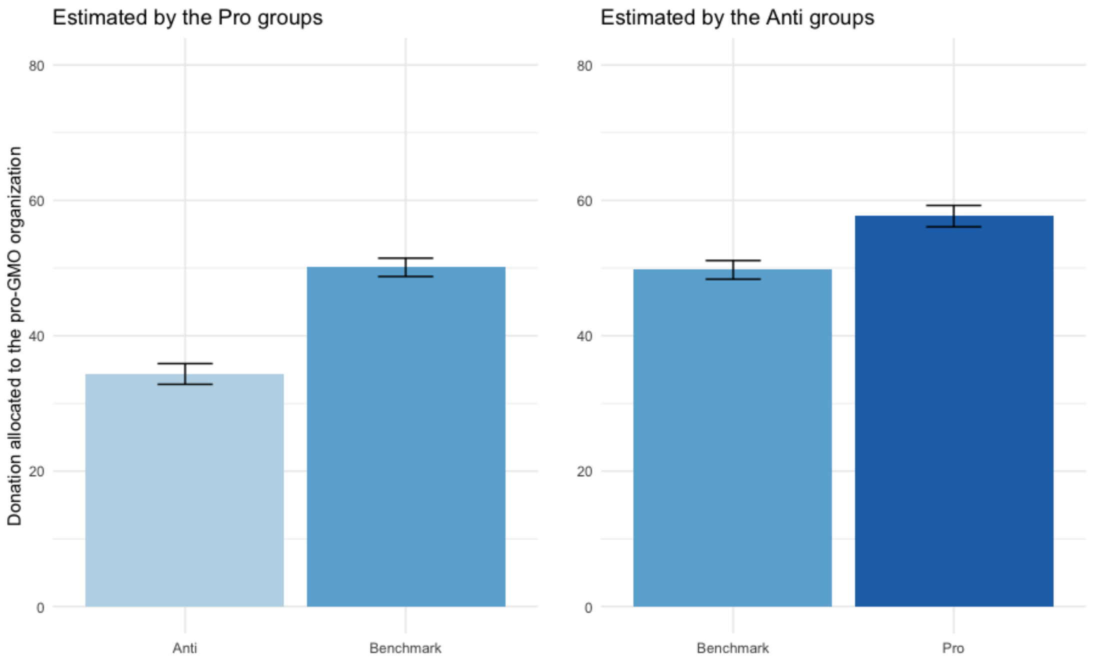


Figure 4.6: Estimates of the attitude of *Benchmark* and the attitude of the opposite side by the Pro groups and the Anti groups.

of the initial information or its impact on attitudes.

4.4.2 Induced confirmation bias

We argue that we minimize the role of confirmation bias by selecting an issue toward which subjects did not have strong priors. However, it could happen that subjects quickly adopt certain attitudes or views over the course of reading the initial slanted article. As information slant causes a shift in attitudes, it may in turn induces a confirmation bias that prevents people from fully counteracting its impact through subsequent information acquisition.

To investigate the induced confirmation bias that comes after initial information, we compare how subjects across treatments rate the convincingness of the arguments they read.[16] The eight arguments are sorted into pro-GMO arguments and anti-GMO arguments. For subjects in *Endo* groups, we randomly select arguments from the pro-GMO argument pool and the anti-GMO argument pool depending on how many arguments they want to read from either side, and we present the selected arguments to them in a random order. Subjects in *Exo* groups read eight arguments in a random order. We use the mean rating on

[16] We demonstrate in more detail subjects' evaluation of each argument and the effects of reading each argument on the attitudes in Appendix H.

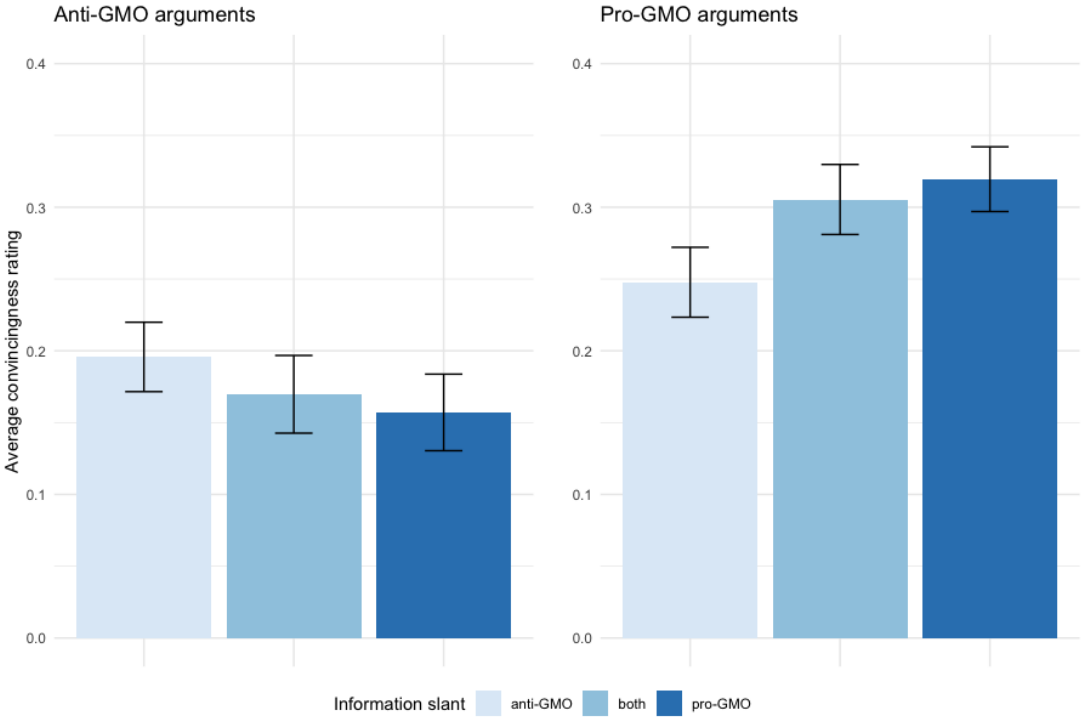


Figure 4.7: Evaluation of pro-GMO and anti-GMO arguments by subjects in *Anti* groups, *Both* groups, and *Pro* groups. Ratings are normalized to [-1,1].

pro-GMO arguments of each subject to measure how she evaluates pro-GMO information and her mean rating on anti-GMO arguments as her evaluation of anti-GMO information, if she chooses to read any. In the absence of confirmation bias, we should find subjects in all treatment groups rate pro-GMO arguments similarly, and likewise anti-GMO arguments.

The right panel of Figure 4.7 shows the average ratings on pro-GMO arguments given by *Pro* groups, *Both* groups, and *Anti* groups. The left panel presents the average ratings on anti-GMO arguments given by *Pro* groups, *Both* groups, and *Anti* groups. Pro-GMO arguments are significantly more convincing in the eyes of *Pro* groups than they are in the eyes of *Anti* groups (Mann-Whitney test, $p = 0.021$). Anti-GMO arguments are more convincing judged by *Anti* groups than judged by *Pro* groups, although the difference is not statistically significant (Mann-Whitney test, $p = 0.16$). The results constitute evidence of the confirmation bias induced by information slant.

How does confirmation bias affect attitudes? Denote the difference between one's evaluation of pro-GMO arguments and one's evaluation of anti-GMO arguments as d_i. The larger d_i is, the more this subject is "biased" in favor of GMO mosquitoes in evaluating the arguments. We then estimate

$$\Delta att_i = \gamma_0 + \theta d_i + \sigma Side_i + X_i\beta + \epsilon_i \tag{4.1}$$

where $\Delta att_i = att_i^{post} - att_i^{pre}$ is the change of individual i's self-reported attitude or donation before and after information acquisition. Self-reported attitude is normalized to [-1,1]. Donation ranges from 0 to 100. $Side_i$ indicates the side of the initial information individual i receives: it equals to 1 if the subject reads a pro-GMO article, -1 if the subject reads an anti-GMO article and 0 if the subject reads both articles. X_i is a vector that controls for the individual's prior attitude toward GMO mosquitoes and her knowledge about GMO mosquitoes in the prescreening survey.

The results are presented in Table 4.1. Subjects become more positive about the use of GMO mosquitoes if they find pro-GMO arguments more convincing relative to anti-GMO arguments. The degree of bias in evaluation is a significant predictor of the change in attitudes. When the bias in evaluation is controlled for, the initial information does not affect the change in attitudes.

Table 4.1: Change in attitude and bias in evaluation

	Dependent variable:	
	Self-reported attitude	Donation
	(1)	(2)
Bias in evaluation: d_i	0.175^{***}	5.242^{***}
	(0.022)	(1.420)
Initial information: $Side_i$	-0.020	-0.759
	(0.014)	(0.944)
Prior attitude	-0.053^{***}	-2.887^{**}
	(0.018)	(1.193)
Prior knowledge	-0.017	2.731^{**}
	(0.016)	(1.060)
Constant	0.005	-3.719^{***}
	(0.014)	(0.949)
Observations	677	677
R^2	0.089	0.033
Adjusted R^2	0.084	0.027

Note: *p<0.1; **p<0.05; ***p<0.01

The results in this section suggest that information slant has a persistent effect: it affects not only people's attitudes, but also the way people process subsequent information. People are "persuaded" by the initial information slant and subsequently suffer from confirmation bias when evaluating further information from both sides, which in turn reinforces the impact of information slant on attitudes. This behavioral mechanism provides a plausible explanation for why offering more common information fails to close the gap between different sides and why forcing more information on people may even backfire.

4.5 Conclusions

In this study, we show that people do not fully counteract the impact of information slant even when they are aware thereof. People could quickly pick up certain attitude that is randomly assigned to them over a course as short as an article. As a result, information slant creates a gap in attitudes among those who are exposed to different sets of information. Offering free access to more information from both sides only slightly narrows the gap in attitudes, while exogenously imposed information does the opposite. Moreover, we find that information slant shifts people's attitudes and in turn induces a confirmation bias when people evaluate subsequent information.

The key feature of our study is that we try to create an environment such that people have the best chance to counteract the impact of information slant. Yet, our results demonstrate how easily information slant can sway people's thinking and how challenging it is to adjust for its impact. These results call for caution on both the supply and the demand side of the information market. On the supply side, it should be noted that warning consumers of potentially biased information is not a substitute of providing balanced information. On the demand side, consumers should perhaps pay more attention to the selection of information sources, as the initial information they are exposed to seems to play a role in determining their attitudes that deliberation struggles to underplay.

This study leaves a couple of directions of future research. For instance, it remains to be investigated whether information slant affects beliefs about verifiable facts related to GMO mosquitoes or other issues in the broader topic of gene-modification and how persistent its impact is. This question also concerns another result that we find intriguing: while subjects exposed to one slanted article believe that those who receive both slanted articles think differently than they do, they stick to their own attitudes. Exploring the impact of information slant on factual beliefs may shed more light on the open question as to whether those who have seen information from both sides are epistemically more superior to those exposed to only one side. In a different direction, it would be interesting to study why exogenously imposed information from both sides seems to increase rather than decrease polarization in attitudes. More specifically, does it matter if information is imposed or endogenously chosen? Are people overloaded with all the available information? The answers may be relevant to developing better debiasing strategies.

Appendix A. Information slant

Pro-GMO article

Between 2013 and 2015 a British high-tech company had released about 450,000 male genetically modified (GMO) mosquitoes every week in Brazil, in an effort to eliminate a disease-carrying species of mosquitoes. These GMO mosquitoes are expected to pass on the modified genes that prevent their offspring from reaching the adult stage and reproducing. The practice was a huge success. The release of GMO mosquitoes worked tremendously well by reducing local mosquito populations by 90% during the trial.

Recently, a team of independent researchers claimed in a published study that some offspring of the GMO mosquitoes survived and reproduced. As a result, the gene of GMO mosquitoes may have been transferred to the local population. “The important thing is something unanticipated happened,” says one of the authors.

In fact, it was not “unanticipated”, nor is it a problem. Some of the authors have shown in their previous work that the transgene is lost from the population over time. Yet, intentionally or unintentionally, that important piece of work is omitted in the current study. “Unfortunately, the study’s authors used dramatized statements to create unfounded concern, ignoring scientific evidence within their own study,” says the company which feels like being unfairly criticized.

Last month, Nature Research, which published the study, added an editor’s note to the paper saying its conclusions are subject to valid criticisms that should be addressed. Nature Research offered the authors the opportunity to submit a correction. Several authors acknowledge the criticisms and they have been struggling to fully address them.

After all, both the study and the company agree that there’s no evidence that these hybrids of GMO mosquito and local mosquito endanger humans more than the wild mosquitoes; nor is there any evidence that the company’s strategy is rendered ineffective.

Anti-GMO article

Between 2013 and 2015 a British high-tech company had released about 450,000 male genetically modified (GMO) mosquitoes every week in Brazil, in an effort to eliminate a disease-carrying species of mosquitoes. These GMO mosquitoes are expected to pass on the modified genes that prevent their offspring from reaching the adult stage and reproducing. The release of GMO mosquitoes seemed to work. The local mosquito population was reduced substantially, yet, almost 10% of them remained active.

Recently, a team of independent researchers warned in a published study that some offspring of the GMO mosquitoes not only survived unexpectedly but were able to reproduce. As a result, the gene of GMO mosquitoes has been transferred to the local population. In other words, the experiment may have unintended consequences and even backfired by creating a hybrid of GMO mosquito and local mosquito.

Although the study and the company agree that there's no evidence yet that these hybrids endanger humans more than the wild mosquitoes or that they'll render the company's strategy ineffective, "the important thing is something unanticipated happened," says one of the authors. One may easily wonder whether manual intervention makes these hybrids become even stronger.

The study triggered an aggressive push-back from the company, which obviously has a lot at stake. "Unfortunately, the study's authors used dramatized statements to create unfounded concern, ignoring scientific evidence within their own study," says the company. It claims that some of the authors did not mention other work of theirs on how the transgene is decreasing in the population over time.

Last month, Nature Research, which published the study, added an editor's note to the paper saying its conclusions are subject to scientific concerns and offered the authors the opportunity to submit a response. While some of the authors disagree with the push-back, they have been working to fully address the point.

Appendix B. Arguments

Pro-GMO 1 Mosquitoes and the diseases they spread (e.g. malaria) kill more people than wars. While the threat is real, traditional methods are proven ineffective in controlling mosquitoes. For instance, because mosquitoes reproduce and evolve within days, chemical pesticides only help them grow resistance as long as they are not entirely eradicated. We definitely need new technology.

Pro-GMO 2 The effect of GMO mosquitoes on eco-system is really minimal. GMO mosquitoes target at a specific disease-carrying mosquito species while leaving other harmless species or beneficial insects alive. There are no birds, fish, or other insects that depend solely on the species to be eliminated. It does not pollinate flowers or regulate the growth of plants. With GMO mosquitoes, we do not add toxic chemicals to the environment, which actually protects the eco-system.

Pro-GMO 3 So far there has been no proven risk of the practice in any aspect. Researchers show that the GMO mosquitoes are not more likely to carry infectious diseases than the mosquitoes were before the experiment. Therefore, the worst thing that could happen is that these experiments fail (they did not), and mosquitoes continue to spread diseases. Even if these GMO mosquitoes do not solve the problem, they do not cause new problems either.

Pro-GMO 4 Experts and scientists must have thought through all the relevant factors and exerted tremendous effort for precaution. After all, it is their job to use science properly to improve our life, so it probably makes sense for most laypeople to trust experts' judgement. When a GMO insect-free cotton was introduced, the USDA, FDA and EPA were all involved in its review and the scientists at each agency scrutinized everything imaginable.

Anti-GMO 1 GMO mosquitoes are not the only available solution since targeted spraying is available. Most pesticides have been proven safe by established researchers. Besides, only areas where humans are present need to be sprayed, which dramatically minimizes the impact of pesticides. Why risk a release of GMO mosquitoes when there is already a solution?

Anti-GMO 2 Mosquitoes are part of a complex ecosystem and food chain. A tiny change in the ecosystem can have huge ripple effects higher in the food chain. By preventing their reproduction, we risk disrupting the entire ecosystem. If the mosquitoes are eliminated, those fish, frogs, birds, other insects, and arthropods that feed on larval or adult mosquitoes might be reduced because of changes in their diet's population. With that, predators of these animals are affected as well.

Anti-GMO 3 Genetic modification might lead to unintended effects out of control. What if their interactions with other organisms in the environment change? There is also the question of what will fill the gap or occupy the niche should the target mosquitoes have been eliminated. Will other pests increase in number? Will targeted diseases be able to switch vectors? Will these vectors be easier or more difficult to control? We should be extremely cautious about something we do not understand well.

Anti-GMO 4 The idea of GMO mosquitoes is against basic laws of nature. It is widely believed that nature has made sure that a certain gene should function in a certain way. Manipulation of it could cause some unwanted gene expressions. Human intervention may interrupt the natural selection and may even make survived mosquitoes become stronger than ever. A huge uncertainty as such is simply beyond the scope of scientists' promise. Science should not be used like that.

Appendix C. Sample instructions (for *Pro.Endo* treatment group)

The stage numbering is not shown to subjects.

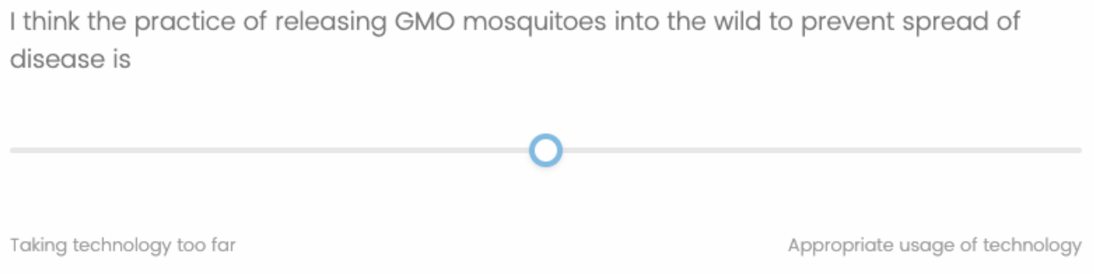
Stage 1.1: exposure to one version of slanted article On the following page, you will read one of the two short articles on a GMO mosquito experiment in Brazil.

Both articles are based on the same underlying facts published in academic articles or news reports. However, they are slanted towards either pro-GMO mosquitoes or anti-GMO mosquitoes. The one that you will read is determined at random.

Please read the article carefully. There will be a few comprehension questions afterward to ensure that the article has been carefully read. You can only proceed if you answer these questions correctly.

This version is slanted towards pro-GMO mosquitoes. (See Appendix A.1 for the article.)

Stage 1.2: elicitation of attitudes What is your opinion on the practice of releasing GMO mosquitoes into the wild to prevent the spread of disease now? Is it taking technology too far or is it appropriate usage of technology?



I think the practice of releasing GMO mosquitoes into the wild to prevent spread of disease is

We will donate 100 dollars to the two organizations below after the study.

Organization 1: Liverpool School of Tropical Medicine (henceforth a pro-GMO organization), a research institute that has a team dedicated to developing toolkits for the genetic manipulation of insects to fight against diseases.

Organization 2: GeneWatch (henceforth an anti-GMO organization), a not-for-profit group that monitors developments in genetic technologies. This organization believes that GMO mosquitoes are a mistake and should be avoided in the future.

Please tell us how you would like to allocate the donation of 100 dollars between the two organizations. We will make the final donation according to the average allocation of all respondents.

I would like to allocate to the pro-GMO organization: _____%;

The rest will go to the anti-GMO organization:_____%.

Stage 1.3: elicitation of estimates Recall that there exist two articles. While you read a pro-GMO mosquitoes article, a previous group of respondents have read both articles.

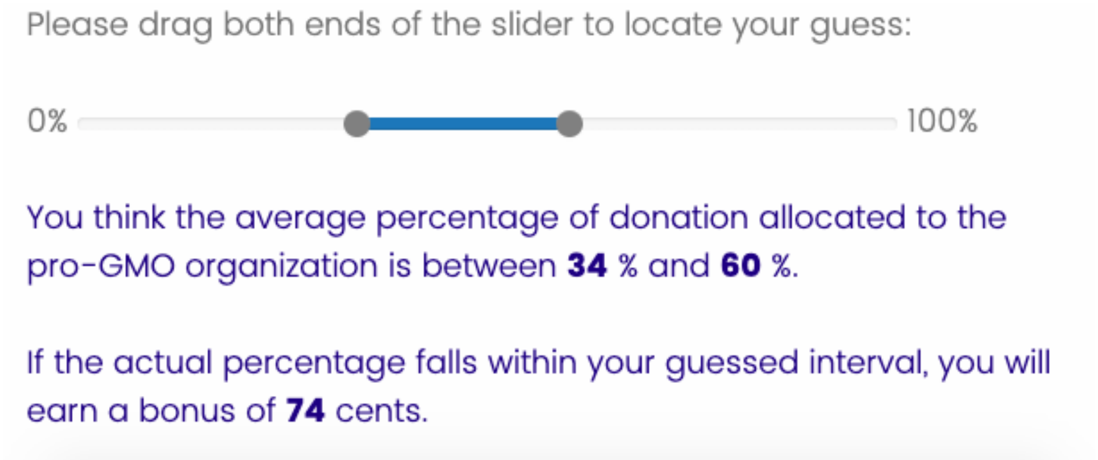
In addition to the articles, they also read eight pieces of arguments made by a separate group of Internet users: 4 arguments from the pro-GMO side and 4 from the anti-GMO side.

After reading two articles and eight arguments, they indicated their preference for the donation allocated between the two organizations.

Your task is to guess the average donation made by this group of respondents, that is, their average percentage of donation allocated to the pro-GMO organization.

This is a bonus task. You can choose an interval for your guess. If the actual average percentage falls into your interval, you will earn a bonus of up to 1 dollar.

Notice that the wider the interval you select for your guess, the smaller the size of the bonus will be. The exact amount of bonus in cents equals to 100 - the width of the interval.



Please drag both ends of the slider to locate your guess:

You think the average percentage of donation allocated to the pro-GMO organization is between **34** % and **60** %.

If the actual percentage falls within your guessed interval, you will earn a bonus of **74** cents.

Stage 2.1: information acquisition In the remaining part of the study, you will have access to the arguments that those respondents have read.

Afterwards, you have a second chance to guess the average allocation of the donation made by those respondents for the bonus task.

You can choose the number and the side of arguments you would like to read before you guess again. (See Appendix B for the arguments.)

Stage 2.2 elicitation of attitudes What is your opinion on the practice of releasing GMO mosquitoes into the wild to prevent the spread of disease now? Is it taking technology too far or is it appropriate usage of technology?

You now have a second chance to revisit your donation allocation decision made in Part I.

Please tell us how you would like to allocate our donation of 100 dollars between the two organizations at the moment. This decision will count for our final donation.

Organization 1: Liverpool School of Tropical Medicine (pro-GMO organization).

Organization 2: GeneWatch (anti-GMO organization).

I would like to allocate to the pro-GMO organization: _____%;

The rest will go to the anti-GMO organization: _____%.

Stage 2.2 elicitation of estimates You now have a second chance to revise your guess. Your answer here will thus replace the first one you made at the beginning and count for the bonus task.

What percentage of the donation do the previous group of respondents that read both articles and all the arguments allocate to the pro-GMO organization, Liverpool School of Tropical Medicine, on average?

Recall that the wider the interval you select for your guess, the smaller the size of the bonus will be. The exact amount of bonus in cents equals to 100 - the width of the interval.

While you read a pro-GMO mosquitoes article, another group of respondents read an anti-GMO mosquitoes article.

Please guess: what percentage of donation does this group from the opposite side allocate to the pro-GMO organization on average when deciding for the first time (after they read the article)?

You will receive a bonus of 50 cents if you guess the percentage correctly.

Stage 2.3: demographics

Appendix D. Pre-registered analysis plan

Research Design

The issue we are using is about a field practice on releasing genetically modified mosquitoes to eliminate disease-carrying mosquitoes. There are two mutually-exclusive sides of the issue, whether one supports or is against the practice. We select this issue based on our pretesting results among participants on Mturk, the flatform from which we plan to collect data. Among all the issues that we pretested, participants have the least strong opinions on GMO mosquitoes and have little knowledge about GMO mosquitoes, which fit the best according to the objective of the study.

Figure 4.1 overviews the experimental setup. We implement a 2x2 (Complete/ Onesided × Exo./Endo.) between-subject design that varies whether subjects are initially exposed to complete or one-sided information about GMO mosquitoes and whether subsequent arguments from both sides are exogenously provided to participants or can be endogenously selected.

The initial exposure of information takes the form of one or two short articles on the practice of using GMO mosquitoes to eliminate disease. It is commonly known to subjects in all treatments that both articles are based on the same underlying facts published in academic journals or news reports but are slanted towards one side or the other. Subjects in Complete treatment read both articles with randomized order. Subjects in OneSided treatment only read one of the two articles and it is made clear that which one they read is determined at random. The side of the evidence, positive or negative, is counterbalanced across participants in OneSided treatments.

After the initial exposure, we provide four arguments from each side of the issue. These arguments were collected from online forums such as Quora. We edit these arguments without changing their substances to make sure that they are similar in clarity and length. While subjects in Exogenous treatments are required to read and evaluate all arguments from both sides, subjects in Endogenous treatments are able to endogenously select how many arguments to read from either side. The presentation order of arguments is fully randomized.

We refer Complete-Exo treatment as the benchmark and the incentive of participants in all treatment groups is to accurately estimate the *final* attitude of the benchmark. The attitude to be estimated is an incentivized allocation of donation between two organizations in favor of different sides. We collect participants' self-reported attitudes, donations and estimates twice within subjects, once after the exposure of evidence but before the introduction of arguments, once after the introduction of arguments. The second elicitation is introduced as a second chance to revisit their donation allocations and estimates. Participants are aware from the beginning that the study consists of two parts, although they are not aware of the existence of a second chance when deciding for the first time.

In total, we have six treatment groups and each subject participates in only one treatment group. As depicted in Figure 4.1, each session may include some of the following blocks depending on the treatment:

- Introduction, pre-screening on background knowledge and strength of the prior of the issue;
- Treatment manipulation with one-sided evidence or complete evidence;
- First elicitation of self-report attitudes;
- First money allocation among organizations in favor of different sides;
- First estimation of the allocation made by the benchmark group;
- Selection or/and evaluation of arguments in favor of different sides to measure patterns of information acquisition;
- Final elicitation of self-report attitudes;
- Final money allocation among organizations in favor of different sides;
- Final estimation of the allocation made by the benchmark group;
- Additional questions to measure covariates of interest.

Data Collection and Exclusion Criteria

We plan to administer the experiment with a total of 900 U.S. respondents, that is, 150 per treatment group in 6 treatment groups, via mTurkdata. The experiment takes about 15-20 minutes to complete.

The experiment includes two screening questions to target participants of interest. In particular, participants will be shortly introduced to the topic we are using, genetically modified mosquitoes. We will then collect their knowledge and opinions on a five point scale. If the participant is extremely knowledgeable (the highest classification in the scale), or has extreme views on genetically modified mosquitoes (the highest or lowest classification in the scale), he or she immediately gets screened out of the experiment.

We also include a few comprehension questions related to the content in the articles as an attention check on whether participants actually read the evidence presented in the experiment. Participants can only proceed if they answer these correctly.

Variables of Interest

1. Primary outcome variables: estimates of the benchmark's final donation *before AND after* the arguments:
   - We incentivize the estimation of the benchmark's final donation in all groups based on accuracy.
   - We incentivize the level of confidence in their estimates by associating bigger incentives with narrower confidence intervals of their estimates.
2. Secondary outcome variables: attitudes *before AND after* the arguments:
   - A self-reported attitude on whether the field practice takes technology too far or is appropriate usage of technology on a Likert scale: [-4,4]
   - An allocation of a donation of 100 dollars between two organizations that favor two different sides: [0,1]
3. Secondary outcome variables: pattern of argument acquisition in treatments with endogenous acquisition:
   - Total number of arguments acquired: [0,8]
   - Ratio of the number of conforming arguments (or confronting arguments) among all acquired arguments: [0,1]
   - Evaluation of convincingness of each acquired argument on a Likert scale: [-4,4]
4. Secondary outcome variables: evaluation of arguments in groups with exogenous information:
   - Evaluation of convincingness of each of all the arguments on a Likert scale: [-4,4]

5. Additional variables of interest:
    - A measure of sophistication/awareness: we collect estimates of the other side's donation among subjects in OneSided-Exo and OneSided-Endo treatments with bonus incentives at the end of the study.
    - Cognitive uncertainty without incentives as in Enke and Graeber (2020).
6. Political preference and daily news consumption:
    - Party affiliation
    - A left/right political scale
    - Daily news source
7. Demographics:
    - Gender
    - Age
    - Educational attainment
    - Income
    - City of residence (to control for participants from Florida, where a similar experiment has taken place)
    - Religiosity

Hypotheses and Planned Analysis

For most of the analysis specified in this section, we mainly rely on across-treatment comparisons using two-sided tests. The reason is that the results will be interesting and informative either way, even though we have ex ante hypotheses in favor of one direction over the other. The exception is 4.1, the first stage of the whole experiment, where we will apply one-sided tests to maximize power.

In addition, we will also conduct complementary regression analysis to include more control variables, such as the number of acquired arguments. Yet, we expect such an exercise to be very much likely underpowered since the total number of different possible combinations of arguments exceeds our planned sample size per treatment, the calculation of which is mainly based on budget consideration rather than a power calculation.

Validating complete exposure

- Final estimates by Complete vs. donations of the benchmark: For those in the treatments Complete, their final estimates of the benchmark's final donations are close to the actual donations of the benchmark.
- *Remark:* In treatment Complete-Endo, subjects' first estimates may differ from the benchmark's actual donations because they have not read all the arguments as subjects in the benchmark do. Yet, this holds constant in all treatments.

First Stage: the Impact of One-sided Evidence

- First attitudes of OneSided vs. first attitudes of Complete: Before information acquisition, the attitudes of subjects in treatments with one-sided exposure will be biased towards their initial one-sided exposure as opposed to those with a complete exposure. Such an impact will also reflect on their first donation decisions.
- First estimates by OneSided vs. first estimates by Complete-Endo: The first estimates by subjects with one-sided exposure will also be biased towards their initial one-sided exposure as opposed to those with a complete exposure.
- *Remark:* Benchmark group only provide estimates once after reading all the arguments since their donation serves as a target.

Insufficient De-biasing in Estimates

- First estimates by Onesided-Endo vs. final estimates by Onesided-Endo vs. actual donations of the benchmark: In treatments with one-sided exposure and endogenous acquisition, the opportunity to acquire or to read arguments will mitigate the impact of initial one-sided exposure but will not fully de-bias subjects in treatments with one-sided exposure.
- First estimates by Onesided-Exo vs. final estimates by OneSided-Exo vs. actual donations of the benchmark: In treatments with one-sided exposure, making it compulsory to read all arguments will mitigate the impact of initial one-sided exposure but will not fully de-bias subjects either. That is, the impact of initial exposure to one-sided evidence is persistent.

- Final estimates by OneSided-Endo vs. final estimates by OneSided-Exo: In treatments with one-sided exposure, estimates by subjects who choose the arguments they read will be more biased towards their initial one-sided exposure as opposed to those who are required to read all arguments.

- First estimates vs. final estimates in Complete-Endo vs. actual donations of the benchmark: Subjects' final estimates are more accurate than their first estimates with the opportunity to acquire the arguments.

- *Remark 1:* We predict a similar across-treatment pattern for our secondary outcome variables, the self-reported attitudes and donation allocations. Yet, we do not use the term ”de-biasing” when referring to these results. The reason is that it is not possible to make normative judgement among these secondary outcomes without making additional assumption on whether the benchmark is closer to “truth”.

- *Remark 2:* It is however possible that the incentive to estimate the benchmark attitude accurately is large enough for some participants to fully compensate for their initial one-sided exposure. To dis-aggregate the difference across treatments, we thus analyze the connection between the effectiveness of de-biasing and the pattern of argument acquisition at the individual level.

Pattern of Argument Acquisition

- Number of arguments acquired by Onesided-Endo vs. number of arguments acquired by Complete-Endo: In Endogenous treatments, those with initial one-sided exposure may acquire relatively more conforming arguments compared to those with initial complete exposure.

- *Remark:* When arguments are not costly, given the incentive to be accurate, a Bayesian participant should acquire all arguments that are read by the benchmark group. The acquisition behaviors of those who do not acquire all arguments are thus not consistent with Bayesian rationality. We have no *ex ante* hypothesis on whether the exposure of one-sided evidence makes people more or less Bayesian.

Pattern of Argument Evaluation

- Evaluation of each argument by OneSided vs. evaluation of corresponding argument by the benchmark: Those with initial one-sided exposure may evaluate conforming arguments more highly compared to those with a complete initial exposure.

- Evaluation of each argument by OneSided-Endo vs. evaluation of corresponding argument by OneSided-Exo: Those with initial one-sided exposure may evaluate conforming arguments even more highly if these arguments are selected endogenously.

- Evaluation of each argument by Complete-Endo vs. evaluation of corresponding argument by the benchmark (a placebo test): Subjects with complete exposure evaluate each argument similarly compared to the benchmark group even though the acquisition is endogenous.

- *Remark:* While our experiment is not designed to test Bayesian learning, the direct comparison of evaluation of arguments across treatments permits us to verify certain deviation from Bayesian updating, e.g. confirming bias, by investigating whether one-sided participants interpret self-reinforcing arguments more favorably compared to the benchmark group.

Link between De-biasing and the Pattern of Argument Acquisition & Evaluation at the Individual Level

- Onesided-Endo: Those who do not read all the arguments are more biased by their initial one-sided evidence in their final estimates of the benchmark's donations.

- Onesided-Endo: Those who acquire relatively more conforming arguments are more biased by their initial one-sided evidence in their final estimates of the benchmark's donations.

- Onesided-Endo: Those who evaluate conforming arguments more highly are more biased by their initial one-sided evidence in their final estimates of the benchmark's donations.

- *Remark: Ex post*, we will estimate the empirical number of arguments one should at least acquire to come up with an estimate that is closer to the benchmark's donations using observations from Endogenous treatments.

Connection with Other Covariates

- We will analyze the connection between our main results and several covariates to shed light on possible mechanisms. The variables that we are particularly interested in include: confidence level in estimates, sophistication about the impact of one-sided evidence and cognitive uncertainty (Enke and Graeber, 2020)

- We will also explore the connection between our main results and political preferences, daily news consumption and other demographics.

Appendix E. Summary statistics and randomization check

Table 2: Summary statistics and randomization check

	B.Exo N=118	B.Endo. N=112	Anti.Endo N=114	Pro.Endo N=136	Anti.Exo N=130	Pro.Exo N=106	Total N=716	p value
prior								0.060
Mean	0.127	0.348	0.289	0.265	0.285	0.104	0.239	
(SD)	(0.790)	(0.654)	(0.713)	(0.680)	(0.729)	(0.703)	(0.716)	
Range	-1-1	-1-1	-1-1	-1-1	-1-1	-1-1	-1-1	
knowledge								0.259
Mean	0.636	0.500	0.474	0.419	0.438	0.509	0.493	
(SD)	(0.802)	(0.816)	(0.743)	(0.736)	(0.671)	(0.680)	(0.743)	
Range	0-3	0-3	0-3	0-3	0-3	0-3	0-3	
gender								0.438
Mean	1.559	1.464	1.596	1.541	1.569	1.585	1.552	
(SD)	(0.499)	(0.501)	(0.510)	(0.500)	(0.542)	(0.514)	(0.511)	
Range	1-2	1-2	1-3	1-2	1-3	1-3	1-3	
age								0.183
Mean	3.864	4.062	3.719	4.044	3.854	4.028	3.929	
(SD)	(1.124)	(1.218)	(1.194)	(1.263)	(1.227)	(1.222)	(1.212)	
Range	2-7	2-7	2-7	2-7	2-7	2-7	2-7	
education								0.745
Mean	2.712	2.812	2.702	2.704	2.646	2.642	2.702	
(SD)	(0.858)	(0.844)	(0.911)	(0.890)	(0.922)	(0.948)	(0.895)	
Range	1-4	1-4	1-4	1-4	1-4	1-4	1-4	
income								0.534
Mean	6.195	6.491	5.746	6.578	6.392	6.142	6.270	
(SD)	(3.859)	(3.770)	(3.231)	(3.757)	(3.569)	(3.479)	(3.620)	
Range	1-12	1-12	1-12	1-12	1-12	1-12	1-12	
party								0.687
Mean	2.110	2.241	2.202	2.200	2.192	2.075	2.172	
(SD)	(0.875)	(0.913)	(0.904)	(0.879)	(0.779)	(0.739)	(0.850)	
Range	1-5	1-5	1-5	1-5	1-5	1-4	1-5	

Appendix F. First and final attitudes and estimates of *Endo* groups

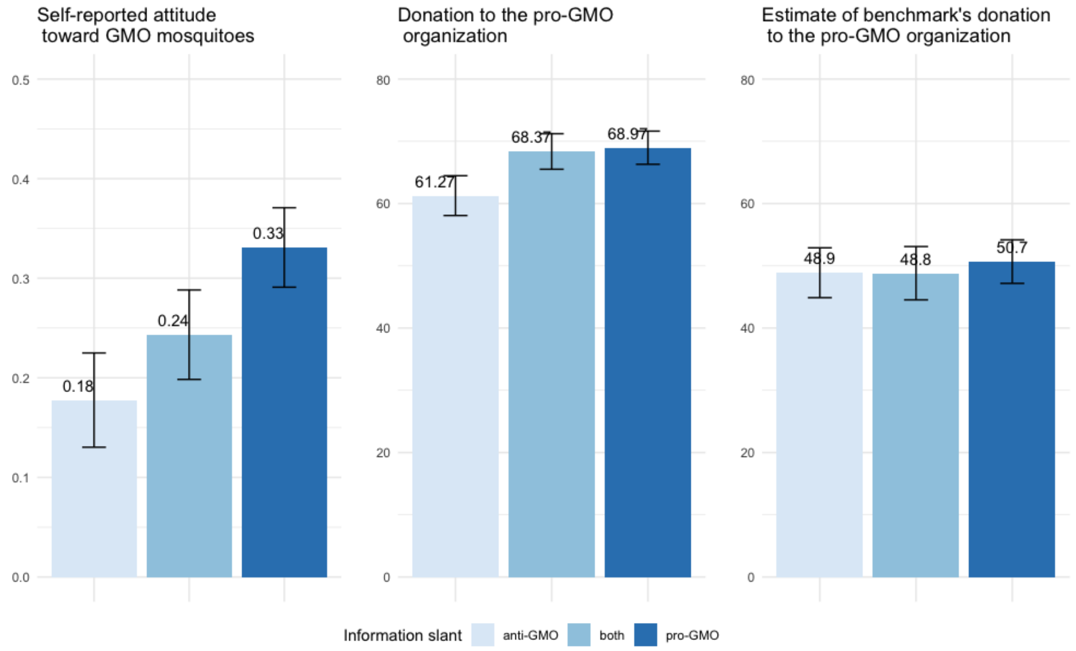


Figure 8: Phase I: first average self-reported attitude, average donation to the pro-GMO organization, and average estimate among subjects in *Pro.Endo* group, *Both.Endo* group and *Anti.Endo* group.

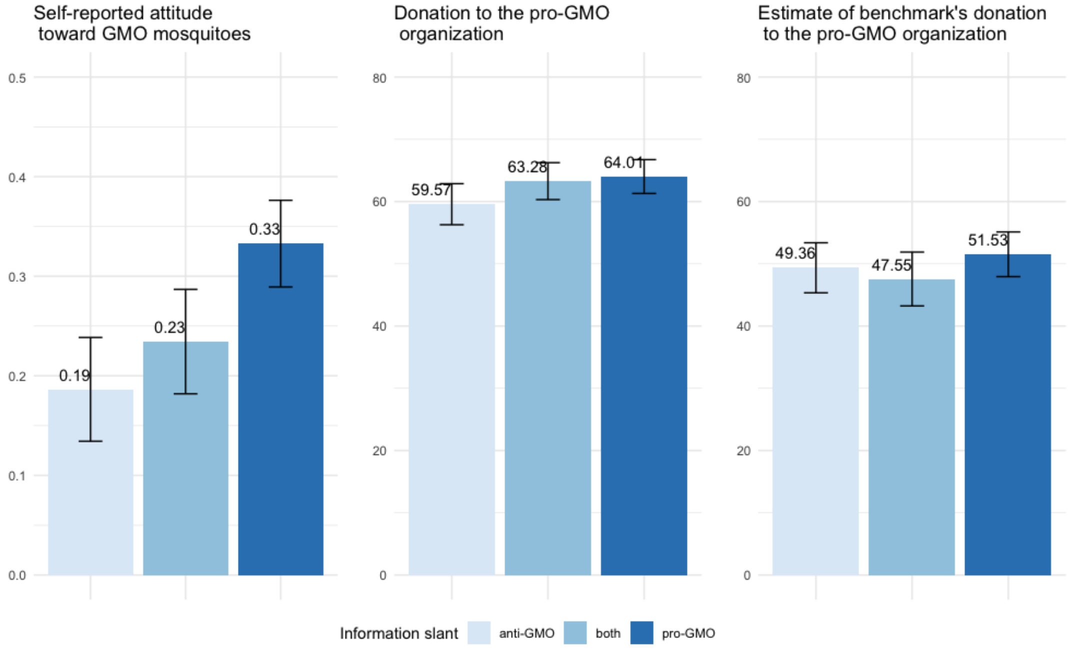


Figure 9: Phase II: final average self-reported attitude, average donation to the pro-GMO organization, and average estimate among subjects in *Pro.Endo* group, *Both.Endo* group and *Anti.Endo* group.

Appendix G. First and final attitudes and estimates of *Exo* groups

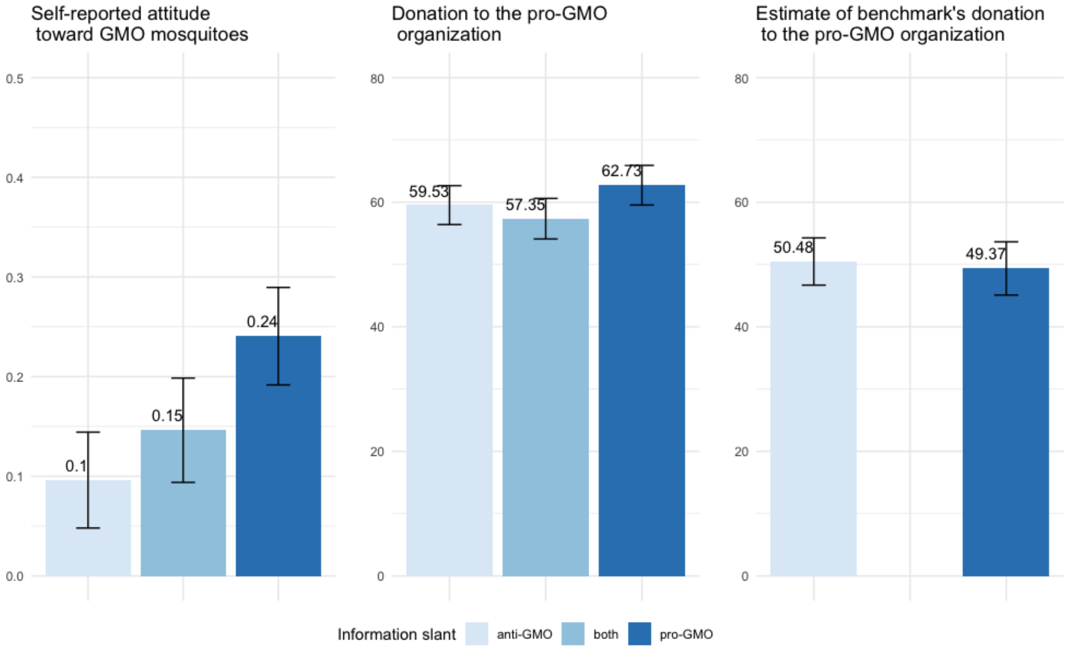


Figure 10: Phase I: first average self-reported attitude, average donation to the pro-GMO organization, and average estimate among subjects in *Pro.Exo* group, *Both.Exo* group and *Anti.Exo* group.

Appendix H. The effects of arguments on attitude

What remains unclear of this experiment is whether the arguments we provide are informative and persuasive at all so that they are able to dilute the impact of one-sided information? Should we expect these arguments to influence one's attitude in the first place? To answer these questions, we first show how subjects evaluate the arguments in general. We then explore to what extent these arguments can influence subjects' attitudes.

Subjects rate each argument from "extremely unconvincing" to "extremely convincing" on a 7-point Likert scale. Figure 12 presents the average rating on the convincingness of each argument by all subjects who have read it. The arguments vary a lot on their convincingness. Five out of eight arguments are rated as convincing. Two arguments are neither unconvincing nor convincing, thus they supposedly have no effect on attitudes. One argument may even backfire since subjects find it slightly unconvincing.

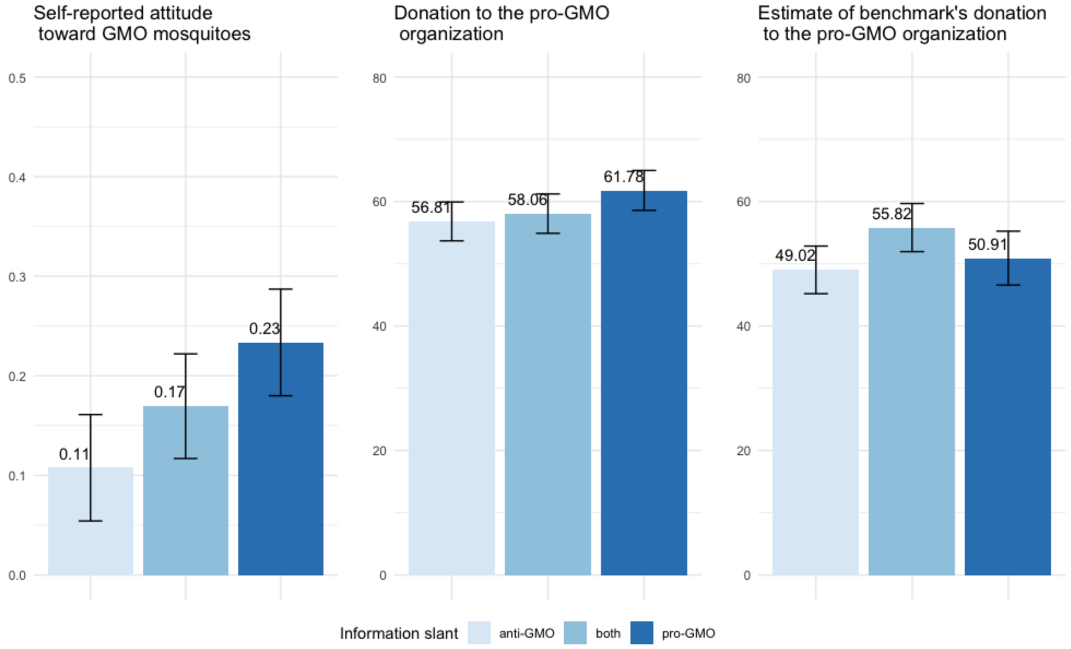


Figure 11: Phase II: final average self-reported attitude, average donation to the pro-GMO organization, and average estimate among subjects in *Pro.Exo* group, *Both.Exo* group and *Anti.Exo* group.

Do these arguments actually affect subjects' attitudes? We estimate the effects of arguments on attitudes, pooling the evaluations by subjects in all treatment groups:

$$\Delta att_i = \gamma_0 + A_i\alpha + X_i\beta + \epsilon_i \tag{2}$$

where Δatt_i is the change of individual i's attitude before and after information acquisition. A_i is the vector of eight dummy variables for the eight arguments: a component equals to 1 if individual i had read the corresponding argument and 0 otherwise. X_i includes the individual's prior attitude toward GMO mosquitoes and knowledge about GMO mosquitoes in the pre-screening survey.

The results are reported in Table 3. One of the four pro-GMO arguments leads to a more positive attitude toward the issue by 0.11 ($p < 0.01$) and increases the donation allocated to the pro-GMO organization by 5.67 percentage point ($p = 0.033$). One of the four anti-GMO arguments causes a more negative attitude toward the issue by 0.16 ($p < 0.01$) and decreases the donation by 5.35 percentage point ($p = 0.045$). Thus, some arguments, although not all of them, are informative enough to be able to influence attitude.

Taken together, the evidence in this appendix shows that the arguments we

Table 3: The effects of arguments on attitude

	Dependent variable:	
	Self-reported attitude	Donation
	(1)	(2)
Prior attitude	0.002	−1.220
	(0.017)	(1.049)
Prior knowledge	−0.002	3.042***
	(0.016)	(1.009)
Pro argument 1	0.063	2.319
	(0.042)	(2.643)
Pro argument 2	0.111***	5.666**
	(0.041)	(2.589)
Pro argument 3	−0.001	0.895
	(0.040)	(2.531)
Pro argument 4	−0.005	0.676
	(0.043)	(2.706)
Anti argument 1	−0.004	0.811
	(0.042)	(2.630)
Anti argument 2	−0.160***	−5.354**
	(0.042)	(2.661)
Anti argument 3	−0.041	−1.671
	(0.041)	(2.609)
Anti argument 4	0.065	2.054
	(0.042)	(2.631)
Constant	−0.016	−7.938***
	(0.032)	(1.991)
Observations	716	716
R^2	0.038	0.033
Adjusted R^2	0.024	0.019
Note:		*p<0.1; **p<0.05; ***p<0.01

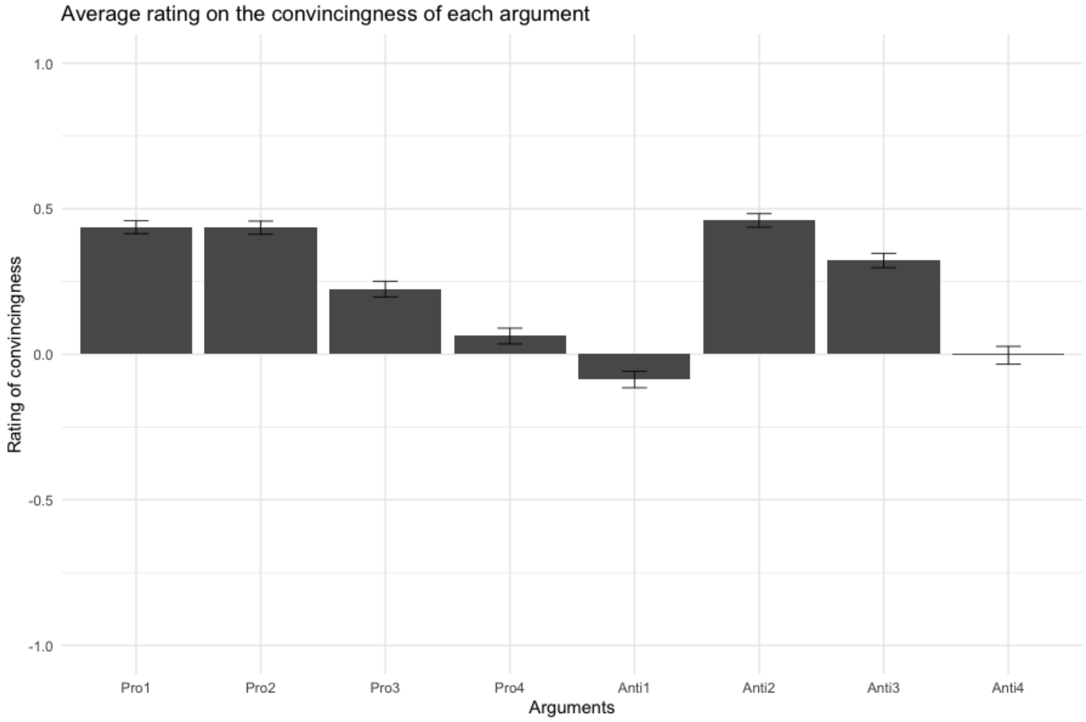


Figure 12: Subjective convincingness rating of arguments. Ratings are normalized to [-1,1].

provide are non-negligible information and that they can influence subjects' attitudes.

References

Bakshy, E., Messing, S., and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

Bénabou, R. and Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, 117(3):871–915.

Brenner, L. A., Koehler, D. J., and Tversky, A. (1996). On the evaluation of one-sided evidence. *Journal of Behavioral Decision Making*, 9(1):59–70.

Charness, G., Oprea, R., and Yuksel, S. (2018). How do people choose between biased information sources? evidence from a laboratory experiment.

Chen, Y. and Yang, D. Y. (2019). The impact of media censorship: 1984 or brave new world? *American Economic Review*, 109(6):2294–2332.

Chiang, C.-F. and Knight, B. (2011). Media bias and influence: Evidence from newspaper endorsements. *The Review of Economic Studies*, 78(3):795–820.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Darley, J. M. and Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, 44(1):20.

DellaVigna, S. and Gentzkow, M. (2010). Persuasion: empirical evidence. *Annual Review of Economics*, 2(1):643–669.

DellaVigna, S. and Kaplan, E. (2007). The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234.

Ditto, P. H. and Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, 63(4):568.

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

Epley, N. and Gilovich, T. (2016). The mechanics of motivated reasoning. *Journal of Economic Perspectives*, 30(3):133–40.

Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies*, 83(2):587–628.

Flaxman, S., Goel, S., and Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1):298–320.

Frey, D. (1986). Recent research on selective exposure to information. *Advances in Experimental Social Psychology*, 19:41–80.

Fryer Jr, R. G., Harms, P., and Jackson, M. O. (2019). Updating beliefs when evidence is open to interpretation: Implications for bias and polarization. *Journal of the European Economic Association*, 17(5):1470–1501.

Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of Computer-Mediated Communication*, 14(2):265–285.

Gentzkow, M. and Shapiro, J. M. (2006). Media bias and reputation. *Journal of Political Economy*, 114(2):280–316.

Golman, R., Loewenstein, G., Moene, K. O., and Zarri, L. (2016). The preference for belief consonance. *Journal of Economic Perspectives*, 30(3):165–88.

Grigorieff, A., Roth, C., and Ubfal, D. (2020). Does information change attitudes toward immigrants? *Demography*, 57(3):1117–1143.

Grossman, Z. and Van Der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, 15(1):173–217.

Halberstam, Y. and Knight, B. (2016). Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. *Journal of Public Economics*, 143:73–88.

Hjort, J., Moreira, D., Rao, G., and Santini, J. F. (2019). How research affects policy: Experimental evidence from 2,150 brazilian municipalities. Technical report, National Bureau of Economic Research.

Iyengar, S. and Hahn, K. S. (2009). Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication*, 59(1):19–39.

Katz, D., Allport, F. H., and Jenness, M. B. (1931). Students' attitudes; a report of the syracuse university reaction study.

Klayman, J. and Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94(2):211.

Koehler, J. J. and Mercer, M. (2009). Selection neglect in mutual fund advertisements. *Management Science*, 55(7):1107–1121.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3):480.

Lippmann, W. (1946). *Public opinion*, volume 1. Transaction Publishers.

Lord, C. G., Ross, L., and Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098.

Martin, G. J. and Yurukoglu, A. (2017). Bias in cable news: Persuasion and polarization. *American Economic Review*, 107(9):2565–99.

Mullen, B., Atkins, J. L., Champion, D. S., Edwards, C., Hardy, D., Story, J. E., and Vanderklok, M. (1985). The false consensus effect: A meta-analysis of 115 hypothesis tests. *Journal of Experimental Social Psychology*, 21(3):262–283.

Mullen, B. and Hu, L.-t. (1988). Social projection as a function of cognitive mechanisms: Two meta-analytic integrations. *British Journal of Social Psychology*, 27(4):333–356.

Mutz, D. C. (2002). Cross-cutting social networks: Testing democratic theory in practice. *American Political Science Review*, pages 111–126.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220.

Pariser, E. (2011). *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Plous, S. (1991). Biases in the assimilation of technological breakdowns: Do accidents make us safer? *Journal of Applied Social Psychology*, 21(13):1058–1082.

Prior, M. (2007). *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge University Press.

Rabin, M. and Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *The Quarterly Journal of Economics*, 114(1):37–82.

Sears, D. O. and Freedman, J. L. (1967). Selective exposure to information: A critical review. *Public Opinion Quarterly*, 31(2):194–213.

Taber, C. S. and Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3):755–769.

Tarantola, T., Kumaran, D., Dayan, P., and De Martino, B. (2017). Prior preferences beneficially influence social and non-social learning. *Nature communications*, 8(1):1–14.

CENTER DISSERTATION SERIES

CentER for Economic Research, Tilburg University, the Netherlands

No.	Author	Title	ISBN	Published
579	Julius Rüschenpöhler	Behavioural Perspectives on Subsistence Entrepreneurship in Emerging Markets	978 90 5668 580 5	January 2019
580	Khulan Altangerel	Essays on Immigration Policy	978 90 5668 581 2	January 2019
581	Kun Zheng	Essays on Duration Analysis and Labour Economics	978 90 5668 582 9	January 2019
582	Tatiana Zabara	Evolution of Entrepreneurial Teams in Technology-Based New Ventures	978 90 5668 583 6	February 2019
583	Yifan Yu	Essays on Mixed Hitting-Time Models	978 90 5668 584 3	April 2019
584	Daniel Martinez Martin	Unpacking Product Modularity, Innovation in Distributed Innovation Teams	978 90 5668 585 0	April 2019
585	Katalin Katona	Managed Competition in Practice Lessons for Healthcare Policy	978 90 5668 586 7	April 2019
586	Serhan Sadikoglu	Essays in Econometric Theory	978 90 5668 587 4	May 2019
587	Hoang Yen Nguyen	Emotions and Strategic Interactions	978 90 5668 588 1	May 2019
588	Ties de Kok	Essays on reporting and information processing	978 90 5668 589 8	May 2019
589	Yusiyu Wang	Regulation, Protest, and Spatial Economics	978 90 5668 590 4	June 2019
590	Ekaterina Neretina	Essays in Corporate Finance, Political Economy, and Competition	978 90 5668 591 1	June 2019
591	Ruth Wandhöfer	Technology innovation in Financial Markets: Implications for Money, Payments and Settlement Finality	978 90 5668 592 8	June 2019

No.	Author	Title	ISBN	Published
592	Andinet Worku Gebreselassie	On communicating about taboo social issues in least developed countries: The case of Ethiopia	978 90 5668 593 5	June 2019
593	Filip Bekjarovski	Active Investing	978 90 5668 594 2	June 2019
594	Miguel Sarmiento	Essays on Banking, Financial Intermediation and Financial Markets	978 90 5668 595 9	June 2019
595	Xiaoyin Ma	Essays on Alternative Investements	978 90 5668 596 6	June 2019
596	Victor van Pelt	A Dynamic View of Management Accounting Systems	978 90 5668 597 3	June 2019
597	Shuai Chen	Marriage, Minorities, and Mass Movements	978 90 5668 598 0	July 2019
598	Ben Gans	Stabilisation operations as complex systems: order and chaos in the interoperability continuum	978 90 5668 599 7	July 2019
599	Mulu Hundera	Role Conflict, Coping Strategies and Female Entrepreneurial Success in Sub-Saharan Africa	978 90 5668 600 0	August 2019
600	Hao Hu	The Quadratic Shortest Path Problem – Theory and Computations	978 90 5668 601 7	September 2019
601	Emerson Erik Schmitz	Essays on Banking and International Trade	978 90 5668 602 4	September 2019
602	Olga Kuryatnikova	The many faces of positivity to approximate structured optimization problems	978 90 5668 603 1	September 2019
603	Sander Gribling	Applications of optimization to factorization ranks and quantum information theory	978 90 5668 604 8	September 2019
604	Camille Hebert	Essays on Corporate Ownership and Human Capital	978 90 5668 605 5	October 2019
605	Gabor Neszveda	Essays on Behavioral Finance	978 90 5668 606 2	October 2019

No.	Author	Title	ISBN	Published
606	Ad van Geesbergen	Duurzame schaarste - Een kritische analyse van twee economische duurzaamheids-paradigma’s geïnspireerd door de filosofie van Dooyeweerd	978 90 5668 607 9	October 2019
607	Richard T. Mason	Digital Enrollment Architecture and Retirement Savings Decisions: Evidence from the Field	978 90 5668 608 6	November 2019
608	Ron Triepels	Anomaly Detection in the Shipping and Banking Industry	978 90 5668 609 3	November 2019
609	Feng Fang	When performance shortfall arises, contract or trust? A multi-method study of the impact of contractual and relation governances on performance in Public-Private Partnerships	978 90 5668 610 9	November 2019
610	Yasir Dewan	Corporate Crime and Punishment: The Role of Status and Ideology	978 90 5668 611 6	November 2019
611	Mart van Hulten	Aiming for Well-Being through Taxation: A Framework of Caution and Restraint for States	978 90 5668 612 3	December 2019
612	Carlos Sandoval Moreno	Three essays on poverty measurement and risk protection	978 90 5668 613 0	December 2019
613	Harmke de Groot	Core strength or Achilles’ heel: Organizational competencies and the performance of R&D collaborations	978 90 5668 614 7	December 2019
614	Peter Brok	Essays in Corporate Finance and Corporate Taxation	978 90 5668 615 4	December 2019
615	Pascal Böni	On the Pricing, Wealth Effects and Return of Private Market Debt	978 90 5668 616 1	December 2019
616	Ana Martinovici	Revealing Attention: How Eye Movements Predict Brand Choice and Moment of Choice	978 90 5668 617 8	December 2019
617	Matjaz Maletic	Essays on international finance and empirical asset pricing	978 90 5668 618 5	January 2020
618	Zilong Niu	Essays on Asset Pricing and International Finance	978 90 5668 619 2	January 2020

No.	Author	Title	ISBN	Published
619	Bjorn Lous	On free markets, income inequality, happiness and trust	978 90 5668 620 8	January 2020
620	Clemens Fiedler	Innovation in the Digital Age: Competition, Cooperation, and Standardization	978 90 5668 621 5	October 2020
621	Andreea Popescu	Essays in Asset Pricing and Auctions	978 90 5668 622 2	June 2020
622	Miranda Stienstra	The Determinants and Performance Implications of Alliance Partner Acquisition	978 90 5668 623 9	June 2020
623	Lei Lei	Essays on Labor and Family Economics in China	978 90 5668 624 6	May 2020
624	Farah Arshad	Performance Management Systems in Modern Organizations	978 90 5668 625 3	June 2020
625	Yi Zhang	Topics in Economics of Labor, Health, and Education	978 90 5668 626 0	June 2020
626	Emiel Jerphanion	Essays in Economic and Financial decisions of Households	978 90 5668 627 7	July 2020
627	Richard Heuver	Applications of liquidity risk discovery using financial market infrastructures transaction archives	978 90 5668 628 4	September 2020
628	Mohammad Nasir Nasiri	Essays on the Impact of Different Forms of Collaborative R&D on Innovation and Technological Change	978 90 5668 629 1	August 2020
629	Dorothee Hillrichs	On inequality and international trade	978 90 5668 630 7	September 2020
630	Roland van de Kerkhof	It's about time: Managing implementation dynamics of condition-based maintenance	978 90 5668 631 4	October 2020
631	Constant Pieters	Process Analysis for Marketing Research	978 90 5668 632 1	December 2020
632	Richard Jaimes	Essays in Macroeconomic Theory and Natural Resources	978 90 5668 633 8	November 2020

No.	Author	Title	ISBN	Published
633	Olivier David Armand Zerbib	Asset pricing and impact investing with pro-environmental preferences	978 90 5668 634 5	November 2020
634	Laura Capera Romero	Essays on Competition, Regulation and Innovation in the Banking Industry	978 90 5668 635 2	December 2020
635	Elisabeth Beusch	Essays on the Self-Employed in the Netherlands and Europe	978 90 5668 636 9	December 2020
636	Sophie Zhou	Essays on the Self-Employed in the Netherlands and Europe	978 90 5668 637 6	November 2020
637	Vincent Peters	Turning modularity upside down: Patient-centered Down syndrome care from a service modularity perspective	978 90 5668 638 3	December 2020
638	Pranav Desai	Essays in Corporate Finance and Innovation	978 90 5668 639 0	January 2021
639	Kristy Jansen	Essays on Institutional Investors, Asset Allocation Decisions, and Asset Prices	978 90 5668 640 6	January 2021
640	Riley Badenbroek	Interior Point Methods and Simulated Annealing for Nonsymmetric Conic Optimization	978 90 5668 641 3	February 2021
641	Stephanie Koornneef	It’s about time: Essays on temporal anchoring devices	978 90 5668 642 0	February 2021
642	Vilma Chila	Knowledge Dynamics in Employee Entrepreneurship: Implications for parents and offspring	978 90 5668 643 7	March 2021
643	Minke Remmerswaal	Essays on Financial Incentives in the Dutch Healthcare System	978 90 5668 644 4	March 2021
644	Tse-Min Wang	Voluntary Contributions to Public Goods: A multi-disciplinary examination of prosocial behavior and its antecedents	978 90 5668 645 1	March 2021
645	Manwei Liu	Interdependent individuals: how aggregation, observation, and persuasion affect economic behavior and judgment	978 90 5668 646 8	March 2021

Manwei Liu (Hubei, China, 1990) received her Bachelor degree in Economics at Renmin University of China in 2012. She obtained her Master degree in Economics at Renmin University of China in 2014. After that, she started in August, 2014 as a PhD candidate at the department of economics at TiSEM, Tilburg University.

Individuals are influenced by the social environment and social interactions they engage in. This dissertation contains an introductory chapter and three chapters all concerning the interdependence of individuals. It explores how observation, aggregation, and persuasion affect judgment and economic decision making. All of the three chapters combine behavioral insights with experimental methods.

Through a collective-choice rule, groups can aggregate individual preferences into a collective decision. Chapter 2 explores the role of collective-choice rules in self-governance in a public goods game, and shows that collective-choice rules affect cooperation directly and indirectly. By observing what others do, people extract information from the environment. Chapter 3 demonstrates that observing a reference action suffices to affect moral judgments. Chapter 4 studies how people deal with information slant and studies the persistent effect of persuasion on judgments and decisions.

ISBN: 978 90 5668 646 8
DOI: 10.26116/center-lis-2104

TILBURG SCHOOL OF ECONOMICS
AND MANAGEMENT

CentER Dissertation Series