Essays in Behavioral Economics

Inauguraldissertation

zur Erlangung des Grades eines Doktors

der Wirtschaftswissenschaften

durch die

Rechts- und Staatswissenschaftliche Fakultät

der Rheinischen Friedrich-Wilhelms-Universität

Bonn

vorgelegt von

Thomas Erik Neuber

aus Kassel

2021

Dekan:	Prof. Dr. Jürgen von Hagen
Erstreferent:	Prof. Dr. Thomas Dohmen
Zweitreferent:	Prof. Felipe Valencia Caicedo, Ph.D.

Tag der mündlichen Prüfung: 19. August 2021

Acknowledgements

This thesis would not exist had I not been supported by many people. Unfortunately, I cannot name everybody here, but I still want to mention some of them.

First of all, I am thankful to my three Ph.D. supervisors. They all must be credited with having had a tremendous impact on the kind of research that I have been doing and for helping me make my first steps in the world of academia. Meeting Thomas Dohmen and finding in him my first supervisor was a decisive turning point during my time as a Ph.D. student. His open mind and curiosity inspired me to think creatively about new topics. Over the years, I have spent many happy hours discussing research with him, and his diligence as a supervisor was of great help. I am also indebted to Felipe Valencia Caicedo, my second supervisor, who was probably the first who gave me extensive academic guidance. I have greatly benefited from his deep knowledge and also his contagious enthusiasm for research and academia. Lastly, Armin Falk has served as my third supervisor and as my repeated coauthor. Working with him has been immensely enriching on all levels. I received ample criticism, but he never made it sting. Instead, I always felt valued and was motivated by the progress I saw I was making. Great thanks also extend to my other coauthors, Jana Hofmeier, Nora Szech, and Philipp Strack. Working with them was a true pleasure. Beyond the immediate research output, I have also learned a lot about approaching projects and research problems.

Over the years, my education and research have been supported by several different institutions—or, to be more precise, the people who are running them and do excellent jobs. I benefited from the work done at the Bonn Graduate School of Economics (BGSE) and the Institute for Applied Microeconomics (IAME). Holger Gerhardt deserves special thanks, as his support for young researchers far exceeds his professional obligations as the manager of the BonnEconLab – not least to the benefit of me. I have also greatly enjoyed being part of the Institute on Behavior and Inequality (briq). Finally, I have been supported by the Collaborative Research Center (CRC) TR 224 and the ECONtribute cluster of excellence during the last years. Jointly, the different institutions have provided an outstanding research environment.

I am also immensely glad about the experience I have had with my fellow Ph.D. students. From the first day, it never felt like a competition but more like a common effort. The latter has particularly manifested itself within the subgroup of my cohort that specialized in applied microeconomics. I have had many fruitful discussions with Mikhail Ananyev, Si Chen, Lukas Kießling, Marta Kozakiewicz, Axel Wogrolly, and Christian Zim-

pelmann. On a personal level, I am thankful for the friendships I have made, among those mentioned above as well as with Gašper Ploj and others. I am also thankful to my parents, whom I always know behind me. The greatest support has been my girlfriend Jana, who held me up when, at times, the situation appeared "adverse." She has always restored my faith that I would succeed—and here we are.

Contents

List of Figures

List of Tables

Introduction

Economics usually conceptualizes individual behavior as the result of external states, such as budgets and prices (or respective beliefs), and preferences. The latter are typically taken as given, and the power of economic analysis has traditionally been seen in explaining observed variation in behavior *without* resorting to variation in preferences.[1] Economists had not much to say about determinants of preferences, preference expression, or behavioral phenomena that are specific consequences of heterogeneity in preferences. It has only been rather recently that this has broadly started to change, and the five chapters of this thesis all contribute to this line of research in behavioral economics.

Underneath this overarching topic, several different threads are weaving together the elements of this thesis. The first such theme is introspection. It can be seen as a next step in taking preferences as not simply given. The idea is that the challenge of understanding preferences does not only apply to researchers. Instead, economic agents (people) themselves regularly engage in reasoning about their own preferences and convictions or those of others. As a result, specific personal preferences can affect decisions where they would otherwise be irrelevant (Chapters 1 and 2), and self-reports can carry varying amounts of information (Chapter 4). The second theme is prosocial and moral decision-making, particularly behavior reminiscent of Kant's categorical imperative. Two experiments suggest that people tend to act like they would want others to act (Chapters 1 and 2). Faced with individual powerlessness in the face of others' immoral behavior, however, many people sacrifice their moral standards. Still, deontological reasoning appears to remain relevant (Chapter 3). The last theme to be mentioned here is a methodological one: the combination of measures based on observed behavior and others that are responses to qualitative survey questions, i.e., self-reports. The relationship is explicitly studied (Chapter 4) and allows for the identification of central underlying mechanisms (Chapters 1 and 2) as well as scalability to a global sample of respondents (Chapter 5). Besides these recurring themes, every chapter is self-contained, and this introduction proceeds with considering the individual chapters.

The first two chapters study individual behavior in the domains of fairness and helping behavior. The central proposition is that emotional introspection is a key mechanism by which people understand normative obligations and needs, meaning that people intuitively tend to act like they would want *others* to act. Both chapters provide evidence from

[1] See George J. Stigler and Gary S. Becker. 1977. "De Gustibus Non Est Disputandum". *American Economic Review* 67 (2): 76–90.

laboratory experiments in which subjects make decisions that affect others. If they were themselves in the positions of those affected others, different subjects would prefer different choices. However, for the decision they actually have to make, standard theory predicts that this preference heterogeneity should be irrelevant. In the first chapter, *Egocentric Norm Adoption*, preferences are exogenously induced by the use of incentives. Each subject makes two decisions over allocations of points worth money within a group of two other participants. The sets of possible allocations entail different normative tradeoffs, and subjects have no personal stakes in their own decisions. However, they are affected by others' decisions: each subject is part of a group, and the members of different groups simultaneously decide over others' allocations along a circle. The main finding is that subjects' decisions are biased towards the normative principles aligned with their own interests, favoring other players whenever they share those interests.

The second chapter, which is joint work with Jana Hofmeier, studies the closely related phenomenon of *imperfect empathy*. Here the experiment leverages preexisting differences in subjects' preferences. In the first step, the design elicits subjects' distaste for eating dried insects by eliciting the willingness to pay (WTP) for avoidance. In the second step, each subject acts an active *sender* and a passive *receiver*. The previous elicitation procedure is repeated, but this time with the following modification: senders report their WTPs to avoid randomly matched receivers have to eat. Crucially, senders have full information about receivers' preferences. Contrary to standard theory, the results show that not only receivers' preferences matter for decisions but also senders' own preferences. Moreover, closer inspection of the data reveals that average helping is higher among pairs of sender and receiver whose preferences are similar rather than dissimilar. Since, typically, helping benefits receivers more than it costs senders (in the experiment), it follows that dissimilarity within pairs reduces welfare. This empirical finding has important implications. For example, systematic differences in consumption preferences between net payers and recipients could undermine public support for public welfare systems.

The third chapter, which is joint work with Armin Falk and Nora Szech, moves the focus from the individual to the group. It studies how the diffusion of being pivotal affects immoral outcomes. In the main experiment, subjects decide about agreeing to kill mice and receiving money versus objecting to the killing and foregoing the monetary amount. In a baseline condition, subjects decide individually about the life of one mouse. In the main treatment, subjects are organized into groups of eight and decide simultaneously. Eight mice are killed if at least one subject opts for killing. The fraction of subjects agreeing to kill is significantly higher in the main condition compared with the baseline condition. The results are replicated in a charity context that also considers sequential decision-making. This second experiment further shows that the observed effects increase with experience, i.e., when the experiment is repeated. For both experiments, we elicit beliefs about being pivotal, which we validate in a treatment with non-involved observers. We show that beliefs are a main driver of our results.

The fourth chapter, which is joint work with Armin Falk and Philipp Strack, relates to the first two in that it is concerned with introspection. Here, the idea is that even

understanding one's own preferences is a challenging process, which we call "limited self-knowledge". Our interest here is to explore this idea in the context of understanding survey responses. First, we develop a choice model of survey response behavior under the assumption that the respondent has imperfect self-knowledge about her individual characteristics. Second, we develop a consistent and unbiased estimator for self-knowledge based on the model. Third, we run an experiment to test the model's main predictions in a context where the researcher knows the true underlying characteristics. The data confirm the model's predictions as well as the estimator's validity. Finally, we turn to a large panel data set, estimate individual levels of self-knowledge, and show that accounting for differences in self-knowledge significantly increases the explanatory power of regression models. Several examples illustrate how using the estimator may improve inference from survey data.

The last chapter, which is joint work with Thomas Dohmen, is itself based on survey data and contributes to the literature on long-term determinants of cross-cultural variation in preferences. It contributes to our understanding of patience by studying the persistent effect that statehood during the last two millennia has had on patience around the globe. It shows that state history and individuals' levels of patient behavior exhibit a hump-shaped relationship, consistent with recent findings for the association between historical statehood and economic development. The relationship is robust to various controls, including contemporary institutions and even economic development. Results for migrants indicate that the portable component of the main effect is negative. This pattern is consistent with a model where state history has a persistently positive effect on patient behavior through the emergence of patience-promoting norms, which are substitutes for intrinsic patience but not portable. This interpretation suggests that the overall effect of state history on present-day patient behavior masks partial crowding-out of intrinsic patience.

Chapter 1

Egocentric Norm Adoption*

Abstract

Social norms pervade human interaction, but their demands are often in conflict. To understand behavior, it is thus crucial to know how individuals resolve normative tradeoffs. This chapter proposes that sincere judgments about the relative importance of conflicting norms are shaped by personal interest. We show that people tend to follow norms from which they benefit themselves, even in contexts where their own decisions only affect others. In a (virtual) laboratory experiment, each subject makes two decisions over allocations of points within a group of two other participants. The sets of possible allocations entail different normative tradeoffs, and subjects have no personal stakes in their own decisions. However, they are affected by others' decisions: each subject is part of a group, and the members of different groups simultaneously decide over others' allocations along a circle. We find that subjects' decisions are biased towards the normative principles aligned with their own interests, thereby favoring other players whenever these share those interests. Subjects' beliefs about the choices made by others suggest a largely unconscious mechanism. Moreover, survey answers indicate that the effects are driven by self-centered reasoning: subjects who report pronounced perspective-taking are less biased.

*I am thankful to Thomas Dohmen and Armin Falk for many discussions, and to Raphael Epperson, Jana Hofmeier, Philipp Strack, Axel Wogrolly, Florian Zimmermann, and Christian Zimpelmann for helpful comments. The study was registered in the AEA RCT Registry under the unique identifying number *AEARCTR-0005774*. Funding by the German Research Foundation (DFG) through CRC TR 224 (Project A01) is gratefully acknowledged.

1.1 Introduction

People care about adhering to social norms, but different norms are often in conflict.[1] Due to opposing prescriptions, it is unclear in many situations what constitutes appropriate and fair behavior. The economic literature has considered this issue from two different angles. One has been to elicit people's *true* attitudes regarding specific tradeoffs (Konow, 2003; Cappelen et al., 2007), often using *impartial spectators* who decide as third parties over allocations between others (Konow, 2000, 2009; Cappelen et al., 2013). The other approach has been to study how people decide about normative tradeoffs when they are affected by their own choices. It has been found that people exploit "moral wiggle room" to excuse selfish behavior (Dana, Weber, and Kuang, 2007). Thus, the two existing approaches either mute self-interest or introduce it directly. However, in many economically relevant situations, an indirect channel might be important: personal interest shapes normative views and is thereby even relevant when, in a particular situation, there are no incentives to behave selfishly.

This chapter proposes that people tend to follow norms aligned with personal interest, even when their own actions do not secure them any advantage. Consider a court case and an unprejudiced judge who neither personally knows any involved party nor has any personal interest in the matter under review. However, the judge shares a certain case-relevant feature with one of the parties, e.g., being male in the context of gender discrimination. It is then easier for the judge to empathize with the male side's interests, possibly leading to a biased decision. Similarly, corporate leaders might think what their staff policies would have meant for themselves at earlier stages of their careers and—perhaps unconsciously—are therefore reluctant towards affirmative action policies. In both cases, people make decisions affecting others that reflect what kind of general behavior is beneficial for themselves, apparently because personal interest has shaped their relative support for different norms. For this phenomenon, we introduce the term *egocentric norm adoption*.

In applied settings, people's interests are correlated with various characteristics, and the potential repercussions of actions are often complex. To provide clean evidence for egocentric norm adoption, we designed a laboratory experiment with three central features: First, subjects are affected by others' choices over normative tradeoffs. Second, subjects' interests are exogenously varied, i.e., they are randomly allocated to roles that profit or lose from certain norms. Third, they also decide in the same decision contexts themselves but over others, such that they are not affected by their own decisions. Specifically, pairs of subjects are randomly assigned to groups. For the two members of each group, subjects from other groups choose allocations of points. The possible allocations involve tradeoffs between two different fairness norms, where each of the principles favors one of the group members. Subjects simultaneously decide over the allocations in other groups along a circle: Group 1 decides over Group 2, Group 2 over Group 3, ..., and Group N over Group 1. Therefore, no subject can influence their own payoff. The experiment consists of

[1] For the general importance of social norms in economics, see, e.g., Elster (1989) and Ostrom (2000).

two decision contexts: the *EF Procedure* trades-off equality against efficiency,[2] while the *EQ Procedure* involves equality and equity, i.e., the principle that divisions of a surplus should reflect individual contributions. Subjects have distinct roles for each procedure that determine from which respective normative principles they profit, and the roles of subjects in adjoining groups are crossed. Before making any decisions, each subject knows that she shares exactly one role with each player over whom she decides. This feature allows us to distinguish the context-specific effect proposed in this chapter, whereby subjects' own *interests* matter, from any person-specific effects, like favoritism towards a specific player.

The experiment's main result is that subjects' decisions over others are biased in favor of their own roles, thereby favoring one of the players in the EF Procedure and the other player in the EQ Procedure.[3] Thus people tend to follow norms from which they would personally benefit if they were adhered to by *others*. Alger and Weibull (2013) have argued that from an evolutionary perspective, such behavior should be expected. They have also drawn a connection to Kant's categorical imperative. However, the behavior of subjects in our experiment seems to follow intuition rather than principled reasoning. After subjects have decided, we elicit their beliefs about the choices of others, not conditioning on roles. Beliefs show very similar biases to those observed for decisions, suggesting that the main effect arises mostly unconsciously. As part of the questionnaires at the end of the experiment, we measure different aspects of empathy. In line with the interpretation of self-centered reasoning driving the results, we find that decisions are less biased among subjects who report pronounced perspective-taking.

Throughout their lives, people gain or lose depending on the prevalence of various normative principles. Hence, egocentric norm adoption suggests that people living under different circumstances develop different normative views. Therefore, it can potentially explain some of the heterogeneity in decisions made by impartial spectators, or what Cappelen et al. (2007) call the "pluralism of fairness ideals." Consider, e.g., the subjects that Cappelen et al. classify as *libertarian*, who believe that even *random* productivity differences should be reflected in payoffs. Perhaps, these individuals have adopted this normative view because they have benefited themselves from random events outside the experimental context. This reasoning is supported by the finding that, among a sample of adolescents in Norway, high-socioeconomic status (SES) spectators exhibited less egalitarianism than their low-SES counterparts (Almås et al., 2017).

How the concept of egocentric norm adoption can potentially explain economically relevant attitudes can be seen in greater detail from three stylized facts about support for public redistribution. (i) Support for national redistribution is decreasing in family income, as Alesina and Giuliano (2011) show with data from the World Value Survey (WVS). This

[2]Throughout the chapter, we will denote the tradeoff between equality and efficiency as a *fairness* tradeoff, although efficiency in itself might not be considered a fairness criterion. However, efficiency is nonetheless relevant for fairness judgments (see Konow, 2001).

[3]The term *bias* here refers to systematic differences in subjects' behavior with no normative justification. A different approach would be to define bias relative to some normative benchmark. That could be the average decision of impartial spectators (see Konow, 2000, 2009; Cappelen et al., 2013) or subjects that are part of the same experiment but uninformed about their own roles.

relationship is found even though most people have virtually no individual power over political decisions, implying that they have no economic motives for self-deception. (ii) Using US data from the General Social Survey, the same article also finds a negative association between support for national redistribution and family income when the respondents were 16 years old, conditional on current family income. The fact that attitudes persist when interests change indicates that they are genuine. Attitudes towards redistribution appear to be influenced by personal interest, but induced shifts can even show in (temporal) contexts where they are unconnected to self-interest. (iii) Support for foreign aid among people in donor countries is *increasing* in income, as Chong and Gradstein (2008) show with data from the WVS. Thus, while the rich and the poor favor their likes concerning national redistribution, the picture is reversed for global redistribution. The above pattern can neither be satisfyingly explained by plain self-interest nor by group cohesion due to socioeconomic status. However, egocentric norm adoption delivers a parsimonious explanation for all three findings: people hold genuine normative views that are more than excuses for selfishness, but their views are nonetheless guided by personal interest. People who are poor within their countries support more national redistribution because they would benefit themselves. They are truly convinced of their normative views and stick with them even if their own situation changes. However, the poor in a rich country support less *global* redistribution, as they suspect an outflow of resources that would otherwise be spent on them.

The experiment's results suggest a certain behavioral mechanism that underlies the phenomenon of egocentric norm adoption: people find it easier to empathize with others' interests if these interests coincide with their own. This mechanism explains why the effects are also present in beliefs and why they decrease in perspective-taking, i.e., people's tendency to "put themselves in others' shoes." The psychological literature has noted that people who are in a given emotional state find it difficult to predict reactions of themselves or others in different emotional states (see Van Boven et al., 2013). The implications of such egocentric empathy gaps have been explored by Van Boven, Dunning, and Loewenstein (2000) in the context of the endowment effect. People who own an object get "attached" to it, and they project their heightened valuations upon potential buyers. Regarding wider economic questions, however, it appears that the topic has received virtually no attention.[4] This chapter is part of a research agenda to explore the economic implications of empathy and its egocentric nature. The related Chapter 2 is concerned with how people's willingness to help depends on how much they would appreciate the same kind of help themselves. In the experiment, *senders* can pay money to avoid that *receivers* have to eat different food items containing dried insects. They know what receivers would be willing to pay for themselves, which mutes the role of beliefs. All subjects act as senders but might be selected to act as receivers at the end of the experiment. The main result is that people pay more for others if they also pay more for themselves. This relationship holds between different subjects and also exists within individual subjects' decisions across different items.

[4]For a general discussion of why emotions should be given a more prominent role in the economic literature, see Elster (1998). For the particular relevance of empathy, see also Singer and Fehr (2005).

Subjects are thus imperfectly empathic in acting not only upon receivers' preferences but also upon their own.

The experiments presented in the current chapter and in Chapter 2 both stress the negative side of egocentrically biased empathy, i.e., its egocentric aspect. As discussed above, the mechanism is a likely explanation for disagreement about fairness standards and distributive policies, even between people who personally are unaffected and could thus claim to be impartial. However, there is also (or, perhaps, primarily) a positive message, which is the presence of empathy: people *do* think about how they would feel about their behavior themselves and act accordingly. In the experiment in Chapter 2, this is indeed quite apparent: Many people are willing to give substantive amounts, just not optimally targeted at the receiver–item-combinations where the benefit for others would be largest. Similarly, egocentric norm adoption can have positive consequences in many social situations and, in particular, promote cooperation between individuals with shared interests. It can, e.g., motivate people to vote in large elections because they would like others who share their political preferences to do the same. More generally, egocentric norm adoption can help to overcome collective action problems and to supply public goods because people in such situations share the same interests.[5] This insight also has practical implications for effective communication in the face of collective action problems. During the current COVID-19 pandemic, e.g., an important policy goal is convincing people to wear face masks, which deliver more protection to people around the wearer than to the wearer herself. In light of this chapter's findings, it is important to stress people's self-interest in *others* wearing face masks. Realizing their own stakes, people should consider the norm of wearing masks important and more readily comply with it themselves.

The remainder of the chapter is organized as follows. Section 1.2 reviews the related literature. Section 1.3 then introduces in detail the experimental design. The derivation of the hypotheses follows in Section 1.4. Section 1.5 presents the main results. Subsequently, Section 1.6 conducts an analysis of heterogeneity in the observed effects. Finally, Section 1.7 summarizes the chapter discusses the results.

1.2 Related Literature

The present chapter is related to multiple strands of literature that previously have been mostly unconnected. First, it is related to the literature on motivated reasoning and beliefs (Kunda, 1987, 1990; Oster, Shoulson, and Dorsey, 2013; Bénabou and Tirole, 2016). In particular, an extensive literature has been concerned with motivated beliefs in the domain of fairness. In an early contribution, Messick and Sentis (1979) find evidence for self-serving fairness views in a hypothetical setting regarding the remuneration for work conducted by oneself and another person who has worked for a longer or shorter time, respectively. In the economic literature, Konow (2000) elicits fairness views as real decisions over allocations between others. Konow shows that subjects who behaved unfairly

[5] This explanation is complementary to other contributing factors such as altruism (Becker, 1974), warm glow (Andreoni, 1990), and reciprocity (Fehr and Gächter, 2000; Falk and Fischbacher, 2006).

due to selfish incentives subsequently adjust their fairness views and interprets this as evidence for cognitive dissonance reduction (Festinger, 1957; Akerlof and Dickens, 1982).[6] Dana, Weber, and Kuang (2007) add "moral wiggle room" to the dictator game by reducing transparency and find decreased giving. Several further contributions have studied how people who are facing monetary incentives to behave unfairly exhibit more selfishness under circumstances which permit sustaining a positive self-image (Gino, Norton, and Weber, 2016). Among the identified kinds of "excuses" are competing (fairness) norms (Rodriguez-Lara and Moreno-Garrido, 2012; Bicchieri and Chavez, 2013; Barron, Stüber, and Veldhuizen, 2019; Kassas and Palma, 2019), sharing the benefits of unethical behavior (Gino, Ayal, and Ariely, 2013), possible misdemeanor of those to be treated unfairly (Di Tella et al., 2015), ambiguity or risk over the efficacy of prosocial behavior (Haisley and Weber, 2010; Exley, 2016), and supposed mistakes in decision-making (Exley and Kessler, 2019). In all of these contributions, biases in fairness views are induced by direct monetary incentives. Self-serving fairness views have also been documented in bargaining contexts, contributing to bargaining impasse between parties who do not sufficiently appreciate the other side's arguments (Thompson and Loewenstein, 1992; Loewenstein et al., 1993; Babcock et al., 1995; Babcock and Loewenstein, 1997; for a successful replication, see Hippel and Hoeppner, 2019). This bias is in line with research showing that people who successfully convince themselves of a particular argument in their favor are better at convincing others (Smith, Trivers, and Hippel, 2017; Schwardmann and Weele, 2019), for which Schwardmann, Tripodi, and Weele (2019) provide additional evidence in the field setting of a debating competition.[7]

Our chapter contributes to the above literature by demonstrating bias in a context without any motives that would conflict with objective fairness. In the experiment, subjects do not need to legitimize any past actions, their decisions do not affect their payoffs, and they do not need to be convincing. Instead, a given subject could do what she objectively believes to be fair and—maybe—hope that others disagree with her view, thereby allocating more points to her than her own decisions would imply. The subject could even think that receiving more points than she would allocate to someone in her own position would happen to be a fair outcome, perhaps because she feels especially deserving as a person or is in particular need of money. The observed bias is evidence that such reasoning is not the whole story. Epley and Caruso (2004) have suggested that people are convinced of self-serving ethical judgments as a result of egocentrically biased affective reactions (see

[6] However, Cerrone and Engel (2019) show that revealing one's fairness view is not sufficient to eliminate subsequent selfish behavior.

[7] Concerning the mechanism behind self-persuasion, Babcock et al. (1995) show that the egocentric bias in fairness views is reduced to statistical insignificance when subjects only learn about their roles only after having read the instructions, i.e., self-persuasion seems to work through differential information encoding. Similarly, in the context of self-interested financial advice, Gneezy et al. (2020) show that self-deception about the truly best options is more pronounced when advisors know about the selfish incentives already before they make their private evaluations. Zimmermann (2020) empirically shows that another mechanism to arrive at motivated beliefs is selective memory. The findings show that creating and sustaining motivated beliefs is an active mental process.

Zajonc, 1980; Haidt, 2001; Slovic et al., 2002) that are automatic and unconscious.[8] This chapter agrees and shows that egocentric perceptions of potential outcomes do not just affect how people feel about narrowly-defined situations that involve themselves. Instead, egocentrism also translates into people's actions and how they treat others, apparently because it alters different norms' perceived importance. The experiment thereby shows that egocentrism can have consequences in situations where people could genuinely claim that they are free from any "conflict of interest" (see the examples in Section 1.1).

The chapter is thus also related to a second strand of literature concerned with in-group–out-group bias. This research area started from the observation that experimental subjects tend to favor other subjects from their own group over subjects from other groups even when the criteria used to form groups are "minimal" (Tajfel, Billig, and Bundy, 1971; Billig and Tajfel, 1973). This finding is now commonly explained with social identity theory (SIT; Turner, Brown, and Tajfel, 1979). The latter starts from the premise that part of individuals' identity is their social identity, which they derive from group memberships. People increase their self-esteem by adopting more favorable beliefs about in-group members than out-group members, as evident in ratings (Mullen, Brown, and Smith, 1992), and treating the former better than the latter. Owing to the observations that individuals usually belong to many social groups and that those groups overlap, there is an interest in effects from crossing group categorizations between individuals (Brown and Turner, 1979), i.e., the relations between in-groups, single out-groups, and double-outgroups. An additive pattern seems to prevail: in evaluations, people behave as if they count the number of dimensions in which another person belongs to their in-group and subtract the number of out-groups to which the given person belongs (Crisp and Hewstone, 1999). Chen and Li (2009) examine the effects of minimal groups within the setting of commonly used paradigms of experimental economics. They find that, relative to out-group members, members of a subject's in-group experience more altruism, increased positive reciprocity, and decreased negative reciprocity. In another economic lab experiment, Cassar and Klein (2019) show that group identity can also be induced by common experiences of success or failure, leading to corresponding favoritism in decisions over redistribution.

Our chapter relates to this literature in that egocentric norm adoption can give rise to a phenomenon akin to in-group–out-group bias. People treat others well if they share the same economic interests. If economic interests in a particular situation coincide among some groups of people and differ for others, discrimination arises between "interest groups." The experiment rules out classical in-group–out-group bias by crossing roles between adjoining groups. Subjects know that both group members for whom they choose an allocation are in one of their in-groups and one of their out-groups, such that SIT would not

[8] Regarding the aspect of unconsciousness, a psychological literature has been concerned with how judgments regarding, e.g., the quality of an applicant, can be "contaminated" by affective reactions (Wilson and Brekke, 1994), finding that people's awareness of their internal processes is insufficient to overcome the resulting biases. Relatedly, Bocian and Wojciszke (2014) show that others' immoral behavior is judged less harshly by observers if the latter themselves profited from the behavior.

make any prediction for differential treatment.[9] Moreover, the crossing of roles implies that egocentric norm adoption favors a different participant for each of the two decisions that a subject makes.

Finally, the present research is related to a mostly theoretical literature on "Kantian" behavior, which proposes that human behavior is following a version of Kant's categorical imperative to "[a]ct only in accordance with that maxim through which you can at the same time will that it become a universal law" (Kant, 1996, p. 73). Loosely speaking, the economic literature says that a subject has Kantian moral concerns if she prefers using strategies that would benefit her also if everyone else also adopted them. Roemer (2010) shows that in the presence of externalities, equilibria arising from Kantian maximization dominate Nash equilibria. Alger and Weibull (2013) show that under assortative matching of individuals who interact, evolution should converge to a mixture of selfish and Kantian preferences.[10] Leeuwen, Alger, and Weibull (2019) empirically investigate the presence of deontological preferences. They do so by letting subjects play both roles in different two-player dilemmas, eliciting their beliefs about others' strategies, and structurally estimating subjects' preferences. Intuitively, Kantian preferences predict strategies that would work especially well if subjects played with themselves in different roles. In the sequential prisoner's dilemma, e.g., those cooperating as the first mover also tend to cooperate with a high probability as the second mover.[11] As has also been shown by Blanco et al. (2014), this correlation can, to a large extent, be explained by beliefs about others' behavior, i.e., by false consensus, but not entirely. Since there is no experimental treatment involved, several different preference-based explanations for this finding are possible (see Blanco et al., 2014). A latent class analysis conducted by Leeuwen, Alger, and Weibull (2019) indicates that deontological preferences do well in explaining the observed patterns. Like the literature on Kantian behavior, this chapter proposes that people mainly care about their own outcomes and exhibit rule-based behavior.

Conceptually, we bridge the above literature to the much larger literature on social norms, an obvious ingredient of rule-based behavior. Moreover, we suggest that the process of selecting behavioral rules is not driven by principled philosophical reasoning, as the reference to Immanuel Kant would suggest, but mainly unconscious, which is confirmed by our finding of biased beliefs. Empirically, we do not rely on interpreting individual-level patterns in behavior but are the first to use the aspect of egocentrism. Identification relies on exogenously induced interests—i.e., on roles—, and egocentric norm adoption is thereby cleanly identified. The results from our experiment show that egocentrism plays a vital role in how people select behavioral rules. This property is clearly opposed to the idea of deontological ethics, but as it turns out, a realistic characterization of people's intuitive behavior.

[9] However, players in the experiment also have *names* X and Y in each group, which are independent of roles (see Section 1.3). SIT predicts bias in favor of players sharing subjects' own names, which we find in Section 1.5.3.

[10] See also Bergstrom (1995) for an early contribution and Alger and Weibull (2019) for a review.

[11] A similar approach is used by Costa-Gomes, Ju, and Li (2019), who find what they call "role-reversal consistency."

1.3 Experiment

Throughout our lives, we lose or benefit from different normative principles. Egocentric norm adoption predicts that this shapes our normative views. In the experiment, we randomly vary which principles align with subjects' personal interests or are opposed to them. These manipulations are small regarding subjects' overall lives, but they are salient during the experiment. Thus, they allow for a causal test of whether personal interest influences adherence to different norms.

People first learn about their own group and their personal interest in the two allocation procedures, i.e., their roles. It is made salient from the beginning of the experiment that they cannot influence their own payoffs. Next, they are informed about the details of the group for which they decide. After everything has been firmly understood, subjects make their two decisions. These are followed by the elicitation of beliefs about other subjects' choices, and the experiment concludes with several questionnaires. The full translated instructions can be found in Appendix 1.C.

1.3.1 Design

A multiple of four participants takes part in each experimental session. Pairs of participants are randomly allocated to groups, numbered consecutively from 1 to N. In each group, one participant is called *Player X*, and the other participant's name is *Player Y*. All participants receive a fixed participation fee of €4 and, during the experiment, points each worth €0.01. Importantly, no player makes any decision regarding their own group. Instead, groups simultaneously decide over players in other groups along a circle, i.e., Group 1 decides over Group 2, Group 2 decides over Group 3, ..., and Group N decides over Group 1. Every player makes two decisions over allocations of points for the players in the respective succeeding group, each according to a different procedure. One decision is about the tradeoff between equality and efficiency (EF Procedure); the other is about the tradeoff between equality and equity, i.e., attribution of responsibility (EQ Procedure). For the EF Procedure, one player in each group takes the role that profits from efficiency, while the other player profits from equality. In the chapter, we denote the former role by A and the latter by B. For the EQ Procedure, we denote roles by a and b, where Role a profits from equity and Role b from equality. The labels of roles do not appear in the instructions, and they are determined independently of subjects' names (X and Y). The instructions do not use the labels for the procedures, either. Instead, these are called "Procedure 1" and "Procedure 2," depending on their randomly determined order on the subject level. Any two players in any two adjoining groups share exactly one role. Figure 1.1 visualizes this structure, where tuples after players' names denote their roles in the EF and the EQ Procedure, respectively.

Estimation Task The EQ Procedure requires that subjects can contribute to the success of their groups. Therefore, all subjects have to engage in an estimation task. The task precedes all other instructions of the experiment, and we tell subjects that a precise

$$\cdots \Rightarrow \begin{matrix} X\colon\ (A,a) \\ Y\colon\ (B,b) \end{matrix} \Rightarrow \begin{matrix} X\colon\ (A,b) \\ Y\colon\ (B,a) \end{matrix} \Rightarrow \begin{matrix} X\colon\ (B,b) \\ Y\colon\ (A,a) \end{matrix} \Rightarrow \begin{matrix} X\colon\ (B,a) \\ Y\colon\ (A,b) \end{matrix} \Rightarrow \cdots$$

Figure 1.1: Example for roles in successive groups

estimate will increase their chances of receiving additional money during the experiment. On their computer screens, subjects see a three-second countdown, after which they see an image for two seconds. The image shows a certain number of blue dots on a yellow background. Immediately after the image has disappeared, subjects have 15 seconds to enter an estimate for the number of dots that they saw. Their task is to minimize the absolute difference between their estimate and the actual number of dots.[12] Before the actual task, subjects complete an identical trial task with a different number of dots. The respective images that subjects see are the same for all participants, showing 40 dots for the trial task and 53 for the actual task. Neither of these numbers is revealed to subjects.

After the estimation task, subjects learn about the experiment's basic setup, i.e., the circular decision structure. The instructions spell out precisely who makes decisions concerning the group to which they belong themselves and for which group they will make decisions. A highlighted box emphasizes that they will in no way be able to influence the allocation of points within their own group. Players first learn about names and roles within their own group and the potential payoff consequences for themselves and their partners. Afterward, they are informed about the structure of the group for which they decide. This order mimics typical real-life situations in which people know about their interests (e.g., being rich or poor) before considering a particular decision problem (voting over a redistributive policy measure).

Efficiency (EF) Procedure The EF procedure concerns the tradeoff between equality in points for both individual players and efficiency regarding the total number of points. The possible allocations of points are shown in Table 1.1.

Table 1.1: Payoffs for the EF Procedure

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	200	300	385	460	525	585	640	690	735	775	811	843	871	896	918	937	953	967	979	990
B	200	190	180	170	160	150	140	130	120	110	100	90	80	70	60	50	40	30	20	10
$\sum$	400	490	565	630	685	735	780	820	855	885	911	933	951	966	978	987	993	997	999	1,000

Columns show the 20 options among which subjects can choose for their respective succeeding groups. The row below the option numbers shows the points that the player in Role A receives as part of each allocation. This number is strictly increasing in the choice

[12] The task of estimating the number of dots follows the one used in Fliessbach et al. (2007). However, the original task asks subjects to make the binary judgment of whether the number of dots was higher or lower than a given integer. Asking for a specific estimate instead allows for a more fine-grained assessment of performance, thereby avoiding ties.

options, but in decreasing increments, i.e., the number of points mimics a strictly concave function. Increases start at 100 points and decrease to a minimum of eleven points. The number of points that the player in Role B receives equals that of the other player only for the first option. Then, it decreases in constant increments from 200 down to 10. The bottom row shows the total number of points, which ranges from 400 to 1,000. Thus, relative to the fully equal outcome, efficiency can be increased by a factor of up to 2.5. However, efficiency gains decrease from 90 points between Option 1 to Option 2 to just a one-point difference between Options 19 and 20. Thus, going from lower to higher options, inequality increases at diminishing returns in terms of efficiency.

Equity (EQ) Procedure At the beginning of the experiment, all players engaged in an estimation task, which they were told would increase their chances of getting additional money (see above). The estimates that subjects gave are used for the EQ Procedure in which the estimate of the player in Role a is compared to the estimate of another player from a non-adjoining group. If the estimate of the player in Role a was better than the other estimate, the group receives 1,000 points, and otherwise, it receives no points. The estimate of the player in Role b does not affect how many points the group receives.[13] Conditional on the player in Role a having secured the points, one allocation needs to be chosen from the 20 options provided in Table 1.2.

Table 1.2: Payoffs for the EQ Procedure

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
a	500	525	550	575	600	625	650	675	700	725	750	775	800	825	850	875	900	925	950	975
b	500	475	450	425	400	375	350	325	300	275	250	225	200	175	150	125	100	75	50	25

As for the EF procedure, Option 1 implements equality of points between roles, i.e., players. For every further option, 25 points are added for the player in Role a (who secured the points), and the same number of points is deducted from the player in Role b (whose performance is irrelevant for the group). Thus higher-numbered options constitute allocations that reflect accountability for the total points that the group received, i.e., a reward for the player who won the points.

The instructions display the potential payoffs like Tables 1.1 and 1.2, except that participants see the names of players (X and Y) instead of roles. The row for Player X is always on top, and that for Player Y below, i.e., the two rows might be reversed.[14] Subjects have to correctly answer three sets of control questions while reading through the instructions, for which they can reread the relevant previous screens. The first set of questions follows the information about their own group. These questions refer to the experiment's structure and roles in the subjects' own groups. For two example options, subjects have to fill in

[13]The fact that subjects cannot learn about their performance and that everybody took part in the same task under the same conditions mutes any self-esteem motives.

[14]The fixed and transparent order facilitates understanding. Subjects find their own payoff in the same row for both procedures. By favoring their own roles, players once give advantage to the subject sharing their own row and once to the subject whose payoff is displayed in the other row.

the amounts of points that both players in their group would receive. A corresponding second set of questions is presented after subjects have learned about the situation within the groups over which they decide. The last set of control questions regards the crossed roles between groups and the below information about the implementation of payoffs. Afterward, subjects make their decisions for the respective succeeding group, one after the other in the subjective-specific order. No option is preselected.

At the end of the experiment, the computer conducts a three-step random procedure to implement a subset of decisions. First, it randomly chooses one of the two procedures. Second, it determines whether decisions come from either all even- or all odd-numbered groups. Third, it determines one subject within each relevant group and implements their respective decisions. Thus, for 50% of subjects, a decision made by another subject is implemented. The 25% of subjects whose own decisions become relevant themselves receive 1,000 points.[15] For the remaining 25% of subjects, their payoff depends on another task independent of their own decisions (see the paragraph on belief elicitation below).

Belief Elicitation After the two decisions, we elicit players' beliefs about choices by others. Specifically, we ask them to guess the average of the choices that subjects from other groups in their session have made for groups that, in terms of the role compositions, are identical to the one for which they have decided themselves. If the decision of a subject's group partner is implemented, i.e., with a probability of 25%, the guess's accuracy determines their payoff. Average choices within the same session are calculated for each procedure, separately for even- and odd-numbered groups, and excluding each subject's own group.[16] Subjects then receive 500 points if their guess is precisely correct and 250 points as long as the correct answer falls into the range of the five options closest to their guess. We elicit the beliefs with tables that look exactly like the ones for the decisions. The tables highlight the range of options for which the currently selected option would still imply 250 points.[17]

Questionnaires The experiment proceeds with a survey asking subjects about fundamental sociodemographic characteristics like age, gender, and income. Moreover, participants complete several questionnaires on personality, preferences, and values. The details

[15] For these subjects, the compensation is thus fixed and thereby independent of their roles. Moreover, the number of points that deciding subjects receive (1,000) is always larger than the payoff for any of the two subjects over whom they decide. These design properties alleviate concerns that subjects' decisions over others might depend on expectations about their own payoffs, e.g., due to aversion towards inequality (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). Also, note that if subjects in Roles B and b should choose more equal options because they wanted to reduce the gap to subjects in Roles A and a (in the succeeding group) *in expectations*, we should observe a negative correlation between choices and beliefs. As Section 1.5.2 will show, the opposite is the case.

[16] This procedure makes sure that the relevant other subjects decided over groups that, abstracting from players' names (who was X and who was Y), are identical to the one for which the respective participant was deciding herself. It also ensures that the roles of comparison subjects are balanced, i.e., that the different roles are present in equal numbers.

[17] For options towards either end of the scale, the interval for which subjects receive 250 points becomes asymmetric around the reported belief. This asymmetry ensures that subjects whose true beliefs are at the extremes have no mechanic incentive to adjust their answers towards the center.

with the corresponding results are presented in Section 1.6. Finally, subjects learn their payoffs and the details of how they came about.

Let us conclude this description of the experimental design by pointing out two noteworthy features that allow for a clean identification of egocentric norm adoption. First, the experiment's structure ensures that subjects' choices do not affect those players on whom their own payoffs depend, avoiding considerations of reciprocity (Fehr and Gächter, 2000; Falk and Fischbacher, 2006). Second, the experimental design comprises two different procedures, such that each player has two roles. Own roles and roles of subjects for whom players decide are crossed, and players thus know that they share exactly one role with each subject over which they decide. Thereby, we distinguish the effect of egocentric norm adoption from in-group–out-group bias in the sense of SIT. According to the latter, preferential treatment is due to elevated attitudes towards in-group members relative to out-group members. Such reasoning is focused on *others*, and it would take into account both of a subject's two roles (hence the interest in how people aggregate crossed categorizations; see Section 1.1). If both procedures were equally relevant for identity, SIT would predict no effect. If one procedure were more important than the other, SIT would predict that a given player favors the same subject in both decisions. In contrast, when people egocentrically adopt norms, they are not focused on others but *themselves*. Preferential treatment is not attached to other people but an individual's roles. Therefore, in the experiment, egocentric norm adoption predicts that a given player favors a different subject for each procedure, i.e., always the one who shares the player's respective own role.

1.3.2 Implementation

The experiment was run from May 13 until May 20, 2020, and implemented as a virtual lab experiment. Seventeen sessions with either 20 or 24 subjects resulted in a total of 372 participants who completed the experiment.[18] Participants were recruited from the subject pool of the *BonnEconLab* using the software *hroot* (Bock, Baetge, and Nicklisch, 2014). The experiment's language was German, and we invited only German-speaking subjects. Participants were mostly university students, and around 60% of subjects were women. For details of the sample composition, see Table 1.B.1 in the appendix. Subjects participated via the Internet. The experiment was programmed using *oTree* (Chen, Schonger, and Wickens, 2016), such that subjects could access it through their web browser using their own devices.[19] They received individual links, such that it was impossible for any subject to participate more than once. Since we ran the experiment during the first phase of the COVID-19 pandemic, subjects presumably participated from home (the university library, e.g., was closed at the time). Contrary to typical online experiments, however, and just as in

[18] We had to exclude four of the 376 participants who initially started the experiment because they either stopped working on the experiment or were unable to answer some of the control questions.

[19] The invitation stated that subjects were required to use a regular desktop or laptop computer. In principle, however, the experiment was also fully functional on smaller devices such as smartphones or tablets.

a usual laboratory experiment, subjects attended specific experimental sessions. They had to participate in the experiment at a pre-specified time and date. Other participants in the same session were taking part simultaneously, and an experimenter was available to answer questions. On the introduction screen, we gave subjects contact details which they could use in case of questions. The experimenter was available via email, telephone, or text.[20] Subjects had already received the contact details before the experiment as part of the automated email communications (invitation, an email with the personal link, reminder). Several subjects asked questions during the experiment, and all contact methods were used.

1.4 Hypotheses

The chapter's main hypothesis is that participants make decisions favoring their own role for the respective procedure. To understand the reasoning behind this conjectured effect in the absence of material incentives or, in fact, *any* instrumental or otherwise self-serving motives, we develop a simple formal framework that attributes biased fairness views not to motivated cognition but the (partial) inability to abstract from one's own role. The framework is inspired by Haidt (2001), who argues that people commonly make ethical judgments based on intuitive reactions and that moral reasoning often takes the form of mere ex-post rationalization. Building on this insight, Epley and Caruso (2004) have conjectured that intuitive moral evaluation in conjunction with automatic egocentrism can explain self-serving ethical judgments. The framework presented here offers a way of formalizing the existing arguments and makes a conceptual contribution by shifting the focus from specific *judgments* to beliefs about generally applicable *norms*. This novel perspective is critical for the resulting behavioral implications: only if people attribute their self-centered affective reactions to the relative importance of norms, the egocentric bias carries over to decisions that do not personally affect them.

Consider, e.g., a metaphor from soccer. A player from a team that a given person supports commits a foul. The intuitive reaction of the supporter is that "this was not a foul." She will perhaps come up with reasons for her judgment, which could take various forms. She could question inferences drawn from video evidence or accuse the opposing player of diving. This kind of reasoning would not affect her judgment of situations between any other teams. However, she could also come to believe that the rules should be changed and that more physical play should be generally permitted. This *would* change her judgment of other situations. In this chapter, we suggest that personal interest changes how people think about the "rules of the game," and not just about situational factors. The formal framework will assume that they rationalize their affective reactions *exclusively* by changing the perceived importance of generally applicable norms, which is the limiting case that makes the argument most transparent. Based on our theoretical conjecture, we derive the testable behavioral implication of egocentric norm adoption.

[20] In contrast to using an online conference platform, these contact methods allowed for one-to-one communication between subjects and the experimenter.

1.4.1 Formal Framework

The starting point of the formal framework is that, while considering the possible choice options, an agent experiences an affective reaction determined by her fairness views but also by the payoff implied for her own relevant role because her perspective is inherently subjective: options implying a high payoff for herself "feel good." Her fairness views and level of subjectivity are, however, imperfectly known to the agent. Instead, she knows which option yields the most positive affective reaction. When being confronted with the choice that she has to make over others, she tries to empathize with those affected. She thus engages in the underlying normative tradeoff and tries to learn about the importance of the involved norms.[21] For this, she uses her affective reactions and asks how they came about. If she is perfectly capable of perspective-taking, she fully realizes the extent of subjectivity underlying her reactions, backs out her true fairness-views, and takes an unbiased decision. However, if she is affected by some degree of egocentrism, i.e., her ability of perspective-taking is imperfect, she underestimates the influence of subjectivity. She arrives at fairness-views that depend on her own roles and at corresponding choices that are egocentrically biased.

Basic Setup

The agent makes one choice for the EF and one for the EQ Procedure, c_{EF} and c_{EQ}, respectively. We assume that the choice set for both procedures is the interval $[1, 20]$, i.e., the agent can choose intermediate options. When considering a given option for one of the procedures, the agent experiences an affective reaction depending on her own respective payoff and the violation of the two norms that are relevant in the respective procedure. We denote by $role_{EF} \in \{A, B\}$ the agent's role in the EF Procedure and by $role_{EQ} \in \{a, b\}$ her role in the EQ Procedure. The agent's affective reaction functions for the two procedures are given by the following equations.

$$React_{EF}(c_{EF}) = \alpha\, Pay(c_{EF}, role_{EF}) - \beta_1\, Ineff(c_{EF}) - Inequal_{EF}(c_{EF}) \tag{1.1}$$

$$React_{EQ}(c_{EQ}) = \alpha\, Pay(c_{EQ}, role_{EQ}) - \beta_2\, Unfair(c_{EQ}) - Inequal_{EQ}(c_{EQ}) \tag{1.2}$$

The influence of the payoff for the own role is determined by the level of subjectivity $\alpha \geq 0$, and the relative weights attached to the efficiency and the fairness norms are $\beta_1 > 0$ and $\beta_2 > 0$, respectively. For Roles A and a, the function *Pay* is strictly increasing in the choice while, for Options B and b, it is strictly decreasing. Thus, it may simply correspond to the number of points. *Ineff* and *Unfair* are both strictly decreasing and strictly convex, as higher options are (decreasingly) more efficient or allocate more points to the responsible player, respectively. On the other hand, $Inequal_{EF}$ and $Inequal_{EQ}$ are both strictly increasing and convex, as higher options imply increasingly unequal payoffs for players. Moreover, we assume that all of the functions are differentiable.

The agent's intuitive reactions are thus best for some options $\tilde{c}_{EF}, \tilde{c}_{EQ} \in (1, 20)$. The

[21] If the agent did not want to exert any effort at all to make her decisions, she would choose randomly.

agent knows how her reactions came about up to the three parameters. In scrutinizing the reasons for her reactions, she forms beliefs $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\gamma}_1$, and $\tilde{\gamma}_2$ about the parameters which, assuming an interior solution, must obey the first order conditions for Equations 1.1 and 1.2. The set of solutions is not atomic, and different combinations of parameters can rationalize the intuitive optimum. For example, a high value of $\tilde{c}_{EF}$ for an agent in Role A could be due to strong subjectivity (large α) or due to strong efficiency concerns (large β_1). The agent starts her inference from prior beliefs about the true parameter values that follow independent Normal distributions with standard deviations of one. For the beliefs about β_1 and β_2, the means are the respective true values, while for belief about α, the mean is multiplied by $\pi \in [0, 1]$. The latter parameter denotes the level of perspective-taking. It captures the ability to recognize how affective reactions depend on roles. The prior belief is given by $\mathcal{N}(\pi\alpha, 1)$.[22] Her decision-relevant beliefs are the values that are most likely given her prior beliefs and the two first-oder conditions.

Lemma 1.1. *Assume positive subjectivity ($\alpha > 0$) and limited perspective-taking ($\pi < 1$). Then:*

1. *The agent underestimates her level of subjectivity, i.e., $\tilde{\alpha} < \alpha$.*
2. *The agent's updated fairness views are egocentrically biased.*
   (a) *If she is in Role A, $\tilde{\beta}_1 > \beta_1$. Otherwise, i.e., if she is in Role B, $\tilde{\beta}_1 < \beta_1$.*
   (b) *If she is in Role a, $\tilde{\beta}_2 > \beta_2$. Otherwise, i.e., if she is in Role b, $\tilde{\beta}_2 < \beta_2$.*

Proof in Appendix 1.A.1. □

Lemma 1.1 formally captures the intuition of egocentric norm adoption: an agent who profits from efficiency-oriented decisions by others will tend to consider this normative principle important. In contrast, an agent who personally loses from efficient allocations will tend to object to the principle. Similarly, an agent who profits from equity-oriented decisions will support the corresponding principle more strongly than an agent who loses from them.

Combining the Procedures

The above framework considers the decisions for the two procedures independently, which suffices for the main predictions. Note, however, that both the EF Procedure and the EQ Procedure involve equality as an overlap in the involved fairness norms, aligned with the interest of roles B and b, respectively. Thus, a participant with Roles B and b always profits from equality, while one with roles Roles B and a or Roles A and b profits from equality according to one procedure and loses in the other. Lastly, the private interest of a participant with roles Roles A and a is always opposed to equality. Using this feature, the setup allows for insights into how egocentrically adopted norms can spill over from their

[22] One could also interpret this assumption in the sense of cognitive dissonance. Subjects would then find it implausible that a wedge exists between their affective reactions and their true fairness judgments. The results on perspective-taking in Section 1.6 support the interpretation in the sense of perspective-taking.

source to other contexts. Formally, let us modify Equations 1.1 and 1.2 in the following way:

$$React_{EF}(c_{EF}) = \alpha\, Pay(c_{EF}, role_{EF}) - \beta_1\, Ineff(c_{EF}) - \gamma\, Inequal_{EF}(c_{EF}) \tag{1.3}$$

$$React_{EQ}(c_{EQ}) = \alpha\, Pay(c_{EQ}, role_{EQ}) - \beta_2\, Unfair(c_{EQ}) - \gamma\, Inequal_{EQ}(c_{EQ}) \tag{1.4}$$

In contrast to the previous assumptions from Equations 1.1 and 1.2, the agent now also forms a belief about the importance of equality, γ. This creates a connection between the two procedures regarding the relative importance of the involved norms, just as it has been intuitively discussed above. As before, the agent knows her true reaction functions up to the now four parameters. All beliefs are the same as before, and the prior belief about γ also follows a Normal distribution with a standard deviation of one, centered around the true value. From the modified assumptions, the below results follow.

Lemma 1.2. *Assume positive subjectivity ($\alpha > 0$) and limited perspective-taking ($\pi < 1$). Then:*

1. *The agent underestimates her level of subjectivity, i.e., $\tilde{\alpha} < \alpha$.*
2. *The agent's updated fairness views are egocentrically biased.*
   - *(a) For roles A and a, it holds that $\tilde{\gamma} < \gamma$. Moreover, $\tilde{\beta}_1 > \beta_1$ and/or $\tilde{\beta}_2 > \beta_2$.*
   - *(b) For roles B and b, it holds that $\tilde{\gamma} > \gamma$. Moreover, $\tilde{\beta}_1 < \beta_1$ and/or $\tilde{\beta}_2 < \beta_2$.*
   - *(c) For roles A and b, it holds that $\tilde{\beta}_1 > \beta_1$ and $\tilde{\beta}_2 < \beta_2$.*
   - *(d) For roles B and a, it holds that $\tilde{\beta}_1 < \beta_1$ and $\tilde{\beta}_2 > \beta_2$.*

Proof in Appendix 1.A.1. □

Lemma 1.2 states that an agent who always loses if others are taking equality-oriented decisions applies a weight to equality that is biased downward. Moreover, there is an upward bias in at least one of the weights that she assigns to the opposing norms, i.e., efficiency and equity. The opposite is true for an agent who always gains from equality. We can make no statement about the weight attached to equality for agents who gain from equality-oriented decisions in one procedure and lose from them in the other. However, for the respective opposing norm involved, the same applies as in Lemma 1.1: agents who benefit from decisions that emphasize efficiency or equity consider the respective norms important, while they otherwise exhibit a downward bias in the attached weight.

1.4.2 Predictions

In making her decisions, the agent tries to be impartial and therefore omits considerations regarding her own role. Using the basic setup of Section 1.4.1, the objective functions that she *wants* to maximize are Equations 1.1 and 1.2, setting the value of α to zero. In the objective functions that the she *actually* maximizes, however, the unknown parameters β_1 and β_2 are substituted by the agent's egocentrically biased beliefs $\tilde{\beta}_1$ and $\tilde{\beta}_2$, respectively.

We again assume interior solutions and use that, by the assumptions from Section 1.4.1, the objective functions are concave. Under these conditions, the agent's choices c^*_{EF} and c^*_{EQ} are uniquely identified by the following first-order conditions.

$$-\tilde{\beta}_1 \, Ineff'(c^*_{EF}) - Inequal'_{EF}(c^*_{EF}) = 0$$
$$-\tilde{\beta}_2 \, Unfair'\big(c^*_{EQ}\big) - Inequal'_{EQ}\big(c^*_{EQ}\big) = 0$$

Both optimas' locations are strictly increasing in the values of $\tilde{\beta}_1$ and $\tilde{\beta}_2$. In conjunction with the egocentric biases shown in Lemma 1.1, this leads to the main hypothesis of the chapter.

Hypothesis 1.1. *For both procedures, subjects make choices favoring their own respective roles.*

To test the hypothesis formally, denote by $r_i^{EF} \in \{A, B\}$ the role of subject i for the EF Procedure and by $r_i^{EQ} \in \{a, b\}$ her role for the EQ Procedure. The subject's choice for the EF Procedure is denoted by $c^*_{i,EF}$ and the one for EQ Procedure by $c^*_{i,EQ}$. Hypothesis 1.1 was preregistered, and in the pre-analysis plan we committed to running the two following regressions:

$$c^*_{i,EF} = \delta_0 + \delta_1 1_A\big(r_i^{EF}\big) + \epsilon_{i,g} \tag{1.5}$$
$$c^*_{i,EQ} = \zeta_0 + \zeta_1 1_a\big(r_i^{EQ}\big) + \eta_{i,g} \tag{1.6}$$

The terms $1_A\big(r_i^{EF}\big)$ and $1_a\big(r_i^{EQ}\big)$ denote indicator functions for roles A and a, respectively. Since subjects in roles A and a would profit from higher-ordered choices by the respective sending group, egocentric norm adoption predicts that both δ_1 and ζ_1 should be positive. Note that the hypothesis requires *both* coefficients to be positive. In Appendix 1.A.2, we show that an upper bound for the joint one-sided p-value is provided by the average of the separate two-sided p-values. This result means that if both coefficients are significantly positive in separate OLS regressions, the null hypothesis of either coefficient being weakly negative can be rejected.

An agent who loses from equality in both procedures (i.e., whose roles are A and a) will initially feel attracted to high choice options. As has been shown in Lemma 1.2, she will view this as strong evidence that she cares little about equality and a lot about at least one of the other norms. The converse is, of course, true for an agent with roles B and b. On the other hand, agents who profit from equality in one procedure and lose from it in the other one will notice that their initially preferred choices are somewhat contradictory as one seems to reflect strong concern about equality while the other does not. These observations lead to the following hypothesis.

Hypothesis 1.2. *Among participants whose private interests are aligned with or opposed to equality for* both *procedures, the effect of their own roles is larger than among other participants.*

In other words, we expect spillovers of roles to the respective other decision contexts, i.e., a positive effect of Role A on the decision for the EQ Procedure and, similarly, a positive effect of Role a on the choice for the EF Procedure.

In the formal framework introduced here, the bias in choices arises unconsciously and is accompanied by distorted beliefs about fairness. Research on the *false-consensus effect* has shown that people typically overestimate the extent to which others share their views, which in the context of this experiment would mean that they project their own bias upon others. We thus have a further hypothesis.

Hypothesis 1.3. *Similarly to decisions, beliefs about others' decisions are biased in favor of subjects' respective roles.*

Since people probably do not fully project their own views upon others but will moderate their predictions to some degree, we expect the effects for beliefs to be a bit smaller than those for the respective decisions.

1.5 Main Results

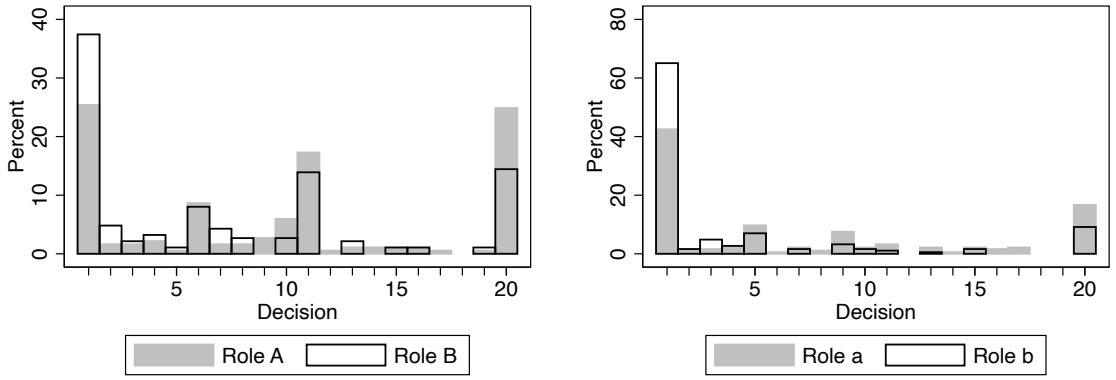
This section presents the main results of the experiment, starting with the decisions in Section 1.5.1 and proceeding with the analysis of beliefs in Section 1.5.2. Section 1.5.3 discusses further observations.

1.5.1 Decisions

Figure 1.2 visualizes the decisions that subjects made in the experiment. The two panels are identically constructed. The left one displays the distribution of decisions for the EF Procedure, and the right panel shows the decisions for the EQ Procedure. In displaying the distributions of decisions, the panels differentiate between the two relevant roles for the respective procedure. For the EF procedure, these are Role A (shaded), profiting from efficiency, and Role B (light), profiting from equality. For the EQ procedure, the relevant roles are a, which is favored by the equity principle, and b, again benefiting from equality. For both procedures and irrespective of roles, the distributions of decisions reveal multiple peaks: one at Option 1, i.e., full equality, one at 20, i.e., least equality, and in the case of the EF procedure, another one at Option 11, which is one of the two options that are closest to the center.

In line with Hypothesis 1.1, differences that depend on subjects' roles are apparent within both procedures. For the EF Procedure, the median of the chosen options by subjects in Role A is 10, while for subjects in Role B, it is only 6. Similarly, the average option chosen by those in Role A is 9.81 and only 7.21 for subjects in Role B, a difference of 0.37 standard deviations. These numbers suggest that, indeed, subjects who would themselves profit from others choosing high options choose higher options themselves than subjects who would personally profit from low options.

Table 1.3 analyses the data in a regression framework, regressing subjects' choices on their roles. Its first two columns show the estimates for the central regressions equations,



(a) EF Procedure (b) EQ Procedure

Notes: The two panels of the figure show subjects' decisions from 1 to 20 split by the respective relevant roles. The left panel shows the data for the EF Procedure. Role A (shaded) profits from higher options while Role B (light) profits from lower options. Similarly, the right panel shows the data for the EQ Procedure. Role a (shaded) profits from higher options while Role b (light) profits from lower options.

Figure 1.2: Decisions by role

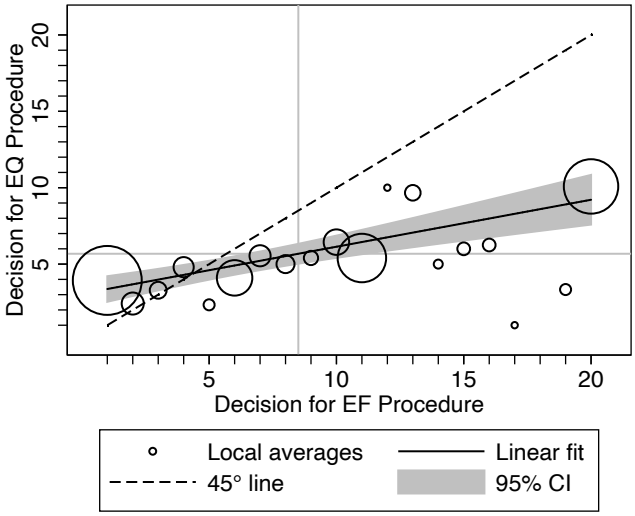
i.e., Equations 1.5 and 1.6. For the EF Procedure, Column 1 shows that the above-mentioned difference in means of 2.6 is statistically significant at any conventional level ($p < 0.001$; two-sided). The same qualitative result of higher choices by subjects in Role A is also confirmed by a non-parametric Mann–Whitney U test ($p < 0.001$; two-sided). The results for the EQ Procedure are qualitatively identical and, in quantitative terms, slightly stronger. Here, the median option chosen by subjects in Role a is 5, while it is 1 for subjects in Role b. The means are 7.25 and 4.11, respectively. The difference between the latter values corresponds to 0.47 standard deviations and is thus even larger than the one observed for the EF Procedure. Column 2 of Table 1.3 shows that this difference is significant ($p < 0.001$; two-sided), and the result is again confirmed by a Mann–Whitney U test ($p < 0.001$; two-sided). Together, the results from both procedures provide clear support for Hypothesis 1.1, namely for egocentric norm adoption: subjects tend to follow fairness evaluations such that if the same standards were adopted by everybody, they would personally profit—and their respective group partners would lose.

While the analyses in Columns 1 and 2 have considered subjects' choices for the two procedures in isolation, it is natural to think that they are related. In particular, the fairness tradeoffs in both procedures involve the criterion of equality, once weighted against efficiency (EF Procedure), and the other time against the equity principle (EQ Procedure). Suppose a subject puts a strong emphasis on equality. In that case, this should manifest itself in low choices for both procedures. On the other hand, one would expect a subject who does not consider equality to be important to make high choices for both procedures. Thus, choices for the two procedures should be positively correlated among subjects. Figure 1.3 displays the empirical relationship between the two decisions that subjects are making. For every option for the EF Procedure on the horizontal axis, the vertical axis shows the respective players' average decisions for the EQ Procedure. The sizes of circles correspond to the relative number of subjects. We observe a clear positive trend. The upward-sloping

Table 1.3: Decisions

Dependent variable	*Decision for succeeding group*			
Procedure	*EF*	*EQ*	*EF*	*EQ*
	(1)	(2)	(3)	(4)
Role A	2.602***		2.609***	1.304*
	(0.727)		(0.725)	(0.677)
Role a		3.140***	1.224*	3.147***
		(0.680)	(0.725)	(0.677)
Constant	7.209***	4.108***	6.593***	3.456***
	(0.498)	(0.428)	(0.585)	(0.475)
Observations	372	372	372	372
R^2	0.033	0.055	0.041	0.064

Notes: Heteroscedasticity-consistent standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.



Notes: The figure groups subjects by their decisions for the EF Procedure. For each option on the horizontal axis, the figure plots the respective subjects' average decisions for the EQ Procedure on the vertical axis. The sizes of circles correspond to the respective numbers of subjects. The dashed line indicates 45 degrees. The gray lines indicate the averages of decisions for the EF Procedure (vertical) and the EQ Procedure (horizontal). The solid black line represents the linear fit from an OLS regression, and the shaded area around it corresponds to the 95% confidence interval based on heteroscedasticity-consistent standard errors.

Figure 1.3: Relationship between the two decisions

regression line confirms the positive relationship. It is based on the disaggregated data and corresponds to a correlation of 0.33 ($p < 0.001$, two-sided). The correlation cannot be due to roles since those are independent.

Given that subjects seem to be consistent in how much weight they attribute to the equality norm in their two decisions, it is useful to consider both procedures jointly. Columns 3 and 4 of Table 1.3 again consider the EF Procedure and the EQ Procedure, respectively, but they include the effects of both roles. Since the roles in the two procedures are independent, the coefficients of Role A for the EF Procedure and Role a for the EQ Procedure remain virtually unchanged compared to Columns 1 and 2.[23] The two other coefficients capture the spillover effects. As predicted by Hypothesis 1.2, both point estimates are positive and consistent with the interpretation that changes in subjects' fairness judgments induced by roles carry over to the respective other procedure similarly to pre-existing differences between different individuals.[24] Individually, both spillover coefficients are weakly statistically significant ($p < 0.1$), and they are jointly significant at the five percent level ($p = 0.02$).[25]

1.5.2 Beliefs

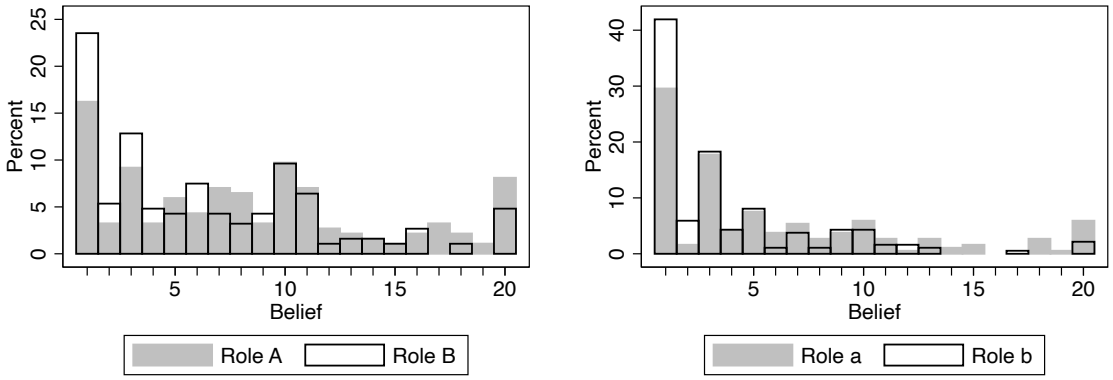
An important question is whether subjects' egocentric behavior is conscious or unconscious, i.e., whether and to which extent subjects realize that their decisions deviate from their true fairness convictions. To address this point, we next analyze subjects' beliefs about average choices made by others.

Figure 1.4 shows the beliefs about others' choices that subjects reported in the experiment. It is constructed in the same way as Figure1.2 for decisions, differentiating between subjects' roles. For both procedures, the distributions of beliefs feature more weight at values around the center than observed for choices. However, concerning role differences, the pattern is the same as for choices: Subjects in roles A and a (shaded) expect higher choices than those in roles B or b (light), respectively. Regression results confirm the visual impression. Role A's effect for the EF Procedure and Role a's for the EQ Procedure are both positive and statistically significant ($p < 0.01$ and $p < 0.001$; see Table 1.B.2 in the appendix). The effects amount to 0.32 standard deviations for the EF Procedure 0.42 standard deviations for the EQ Procedure, which are relative sizes similar to those found

[23] The point estimates slightly differ because, as mentioned earlier in Footnote 18, four subjects did not complete the experiment, and roles are therefore not precisely independent anymore. The slight empirical correlations between roles are random, and the implications for estimates minimal.

[24] Multiplying the effect from Column 1 in Table 1.3 with the slope of the regression line in Figure 1.3 (0.31), one would expect an effect of Role A for the EQ Procedure of 0.80. Similarly, the respective prediction for the effect of Role a for Procedure EF would be 1.08. The observed values in Columns 3 and 4 of Table 1.3 are comparable to these predictions and even slightly larger.

[25] In the appendix, we provide a visual decomposition of the spillover effects by distinguishing between the four possible combinations of roles that subjects might have. In the EF Procedure, the spillover effect is mainly driven by subjects who gain from equality in both procedures and choose very low options (see Figure 1.B.2). Subjects who have to hope for equality attach little relative weight to efficiency. In the EQ Procedure, the spillover effect arises symmetrically: it is driven by subjects who profit from equality in both procedures choosing low options and subjects who lose from equality in both procedures choosing high options (see Figure 1.B.3).



(a) EF Procedure (b) EQ Procedure

Notes: The two panels of the figure show subjects' beliefs about others' average decisions from 1 to 20 split by the respective relevant roles. The left panel shows the data for the EF Procedure. Role A (shaded) would profit from higher options while Role B (light) would profit from lower options. Similarly, the right panel shows the data for the EQ Procedure. Role a (shaded) would profit from higher options while Role b (light) would profit from lower options.

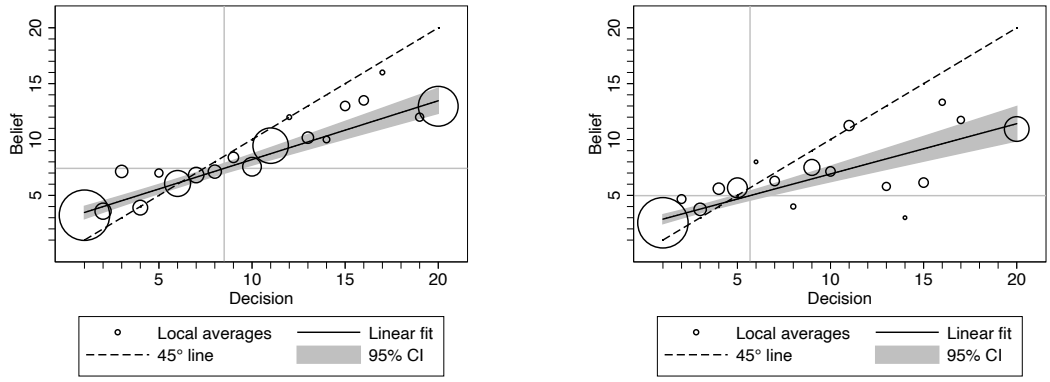
Figure 1.4: Beliefs by roles

for decisions, although slightly smaller.

The results for beliefs resemble those for decisions also in other respects. For decisions, Figure 1.3 has established a strong positive correlation of 0.33 between the two procedures. The corresponding relationship between subjects' beliefs is also clearly positive ($p < 0.001$, two-sided; see Figure 1.B.1 in the appendix), although the correlation coefficient only amounts to 0.18. In light of this finding, it is not surprising that the spillover effects in beliefs are again positive but smaller and not statistically significant (see Table 1.B.2 in the appendix).

The above results might suggest that the effects of roles on decisions arise unconsciously, at least to a large extent. This interpretation presupposes that beliefs and decisions are positively related—which indeed they are. Figure 1.5 plots average beliefs for subjects depending on their decisions and is otherwise constructed just like Figure 1.3. The left panel shows the relationship for the EF Procedure, and the right panel does the same for the EQ Procedure. First, we observe that average beliefs, indicated in both panels by gray lines, are lower than average decisions for both the EF Procedure and the EQ Procedure ($p < 0.001$ and $p = 0.01$, respectively). That means subjects, on average, expect others to assign a higher relative weight to equality than they do themselves. More importantly, we see a clear positive correlation between choices and beliefs ($p < 0.001$), with slopes of 0.53 for the EF Procedure and 0.45 for the EQ Procedure. These associations between decisions and beliefs are instances of the well-established false consensus effect (Ross, Greene, and House, 1977): people have a fundamental disposition to believe that others' convictions are more similar to their own than they really are. The first two columns of Table 1.4 confirm that the effect is strong, regressing decisions on beliefs for the EF and the EQ Procedure, respectively. The estimates for both slope parameters are larger than 0.8.

Thus, we have seen that roles affect beliefs similarly to decisions and that beliefs and



(a) EF Procedure (b) EQ Procedure

Notes: The two panels of the figure group subjects by their decisions for the EF Procedure and the EQ Procedure, respectively. For each option on the horizontal axes, the panels plot the respective subjects' average beliefs about others' decisions for the same procedure on the vertical axes. The sizes of circles correspond to the respective numbers of subjects. The dashed lines indicate 45 degrees. The gray lines indicate the averages of decisions for decisions (vertical) and beliefs (horizontal). The solid black lines represent linear fits from OLS regressions, and the shaded areas around them correspond to the 95% confidence intervals based on heteroscedasticity-consistent standard errors.

Figure 1.5: Decisions and beliefs

Table 1.4: Decisions and beliefs

Dependent variable	*Decision for succeeding group*			
Procedure	*EF*	*EQ*	*EF*	*EQ*
	(1)	(2)	(3)	(4)
Belief (EF Procedure)	0.820***		0.805***	
	(0.0511)		(0.0519)	
Belief (EQ Procedure)		0.809***		0.777***
		(0.0695)		(0.0722)
Role A			1.115**	
			(0.565)	
Role a				1.506**
				(0.606)
Constant	2.417***	1.654***	1.981***	1.058***
	(0.469)	(0.386)	(0.530)	(0.404)
Observations	372	372	372	372
R^2	0.432	0.364	0.438	0.376

Notes: Heteroscedasticity-consistent standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

decisions are closely related. It suggests an unconscious channel: people who are influenced by their own roles might not realize how they are biased. To the extent that this is the case, the effects of roles on decisions should be reduced once we control for subjects' respective beliefs. The results are shown in Columns 3 and 4 of Table 1.4. We first note that the coefficients for beliefs are hardly changed. However, the effects of roles change decisively. Recall that, without controlling for beliefs, the effects were 2.602 for the EF Procedure and 3.140 for the EQ Procedure (see Table 1.3). Now, they are reduced by 57% ($p < 0.01$, two-sided) and 52% ($p < 0.001$), respectively. The remaining effects of roles remain statistically significant ($p = 0.049$ and $p = 0.013$, respectively). However, because our measures of beliefs almost certainly contain some measurement error, the true conditional effects should tend to be even smaller (see Gillen, Snowberg, and Yariv, 2019). In sum, the point estimates suggest that roles' effects on decisions arise to more than one-half unconsciously, and it might even be the case that subjects are entirely unaware of the bias in their decisions.

1.5.3 Further Observations

The experiment allows for some further noteworthy insights. One is that subjects' choices do not seem to be significantly driven by concerns about ex-ante equality. Note that the implications of high choices in terms of procedural fairness are different between affected groups in which one of the players has Roles A and a (*parallel* groups) and other (*crossed*) groups. Within crossed groups, high choices offset each other from an ex-ante perspective because both subjects profit from inequality in one procedure and lose from it in the other. For parallel groups, however, ex-ante inequality from high choices cumulates, since one subject profits from inequality in *both* procedures, while the other loses in both. Thus, one could expect that decisions over players in crossed groups, i.e., those made by subjects in parallel groups, should generally be higher and more strongly positively correlated. Empirically, however, there is no indication of any such differences. The distributions of decisions do not significantly differ between the group types ($p = 0.28$ for the EF Procedure and $p = 1.00$ for the EQ procedure; Kolmogorov–Smirnov test, two-sided). Moreover, the positive correlation between decisions exists for both group types separately ($p < 0.001$ and $p < 0.01$, respectively), and there is no evidence for a difference in the correlations' magnitudes ($p = 0.32$). Despite these similarities in the overall distributions of decisions between the two group types, the effects of players' own roles on decisions are stronger in parallel groups than in crossed groups. In fact, precisely these differences identify the effects of roles across procedures in Columns 3 and 4 of Tables 1.3 and 1.B.2. In light of the above discussion, which excludes concerns about ex-ante fairness as an alternative explanation, the respective coefficients seem interpreted best as evidence for spillover effects.

By the experiment's design, roles do not induce differential proximity between players in adjoining groups due to crossed roles. However, independently of roles, the design deliberately induces *nominal* groups by referring to players in each group as X and Y. Thus, participants decide over allocations between one player with the same name as themselves

and one with a different name. In this dimension, the experiment mimics research on discrimination between *minimal groups* in social psychology (Tajfel, Billig, and Bundy, 1971; Billig and Tajfel, 1973) and economics (Chen and Li, 2009). If subjects favored their nominal in-group, they should choose a high option for EF Procedure if the receiving player sharing their name is in Role A, and a high option for the EQ procedure if the player is in Role a. In line with the literature, subjects exhibit significant nominal in-group bias in both procedures ($p < 0.01$ for the EF and $p < 0.001$ for the EQ Procedure, see Columns 1 and 2 of Table 1.B.3 in the appendix). The effect sizes are smaller than the ones estimated for roles, although the differences are not significant. For beliefs, the corresponding effects are similar but weaker. Both estimated coefficients are positive, and the EQ procedure's effect is significant ($p < 0.001$, see Columns 3 and 4 of Table 1.B.3). Since names were determined independently of roles, the estimated effects of roles and names are virtually unaffected by including the respective regressors jointly (see Table 1.B.4).[26]

A potential concern in many experiments involving human subjects is experimenter demand. It denotes the possibility that subjects try to conform to the experimenters' expectations. This experiment's design mitigates such concerns to the largest possible extent. An important design property is that treatment effects are identified between subjects, as opposed to within-subjects. The between-subject design avoids making subjects aware of the treatment differences or their own counterfactual behavior. In fact, in studying group bias, Chen and Li (2009) rely mainly on a within-subject design and use a between-subject treatment specifically to mitigate experimenter demand effects. As discussed above, this experiment studies (nominal) group bias as well, and this decision was, in part, also made to conceal the purpose of the design. Should subjects have tried to guess the research hypotheses, they might have ended up with the wrong one—or they would have had to balance multiple conflicting motives. Under these conditions, it seems implausible that the observed effects could be as large as observed in our data. For specific "demand treatments," De Quidt, Haushofer, and Roth (2018) find average effects of 0.13 standard deviations. In contrast, the effects observed in this experiment are multiple times as large. The effects are also present for beliefs, which we elicited with an incentivized procedure. Here, subjects would have had to give up their own money to conform to expectations. The data also generally seem well-behaved. For example, Table 1.B.5 in the appendix shows that the randomly determined order of procedures matters for effect sizes in a conceivable way: effects on decisions are stronger for the respective procedure that comes first, although not significantly. Lastly, the next section will show that the treatment effects are not mainly due to a few subjects making extreme decisions but caused by the bulk of subjects exhibiting moderate bias.

[26] In the appendix, we visually inspect the interaction between players' own roles and nominal groups. In the EF Procedure, prominently high options are chosen by subjects whose own role is A and who are in a nominal group with another subject in Role A (see Figure 1.B.4). In the EQ Procedure, the effects of roles and nominal groups appear to be independent (see Figure 1.B.5). Perhaps it matters for these results that payoffs are linear in choice options for the EQ Procedure but not the EF Procedure.

1.6 Heterogeneity

This section aims to relate the observed bias induced by roles to relevant personal attributes of subjects. In terms of understanding the mechanism behind our main results, we are particularly interested in the role of perspective-taking. We also consider the role of different aspects of empathy as well as prosociality. Regarding outcomes, we study the relationship between progressivism and political orientation on the left–right spectrum.

1.6.1 Attributing the Effects to Subjects

A challenge for studying individual heterogeneity in the display of egocentric norm adoption is that the treatment effects of roles are identified not within but only *between* subjects. Therefore, we first convert each subject's two decisions into a single individual-specific *ENA* proxy of egocentric norm adoption and construct a corresponding measure for biased beliefs in the same way. This proxy intuitively measures how a given subject contributes to the observed treatment effects for decisions. We start from the self-evident fact that if there were no treatment effects, it would make no difference for a given subject's decisions to which roles she has been assigned. The average choices conditional on roles would thus coincide with the unconditional average answers. A measure for how much a particular choice contributes to the treatment effect is thus given by how much it deviates from the unconditional average choice in the direction that favors the subject's relevant role.

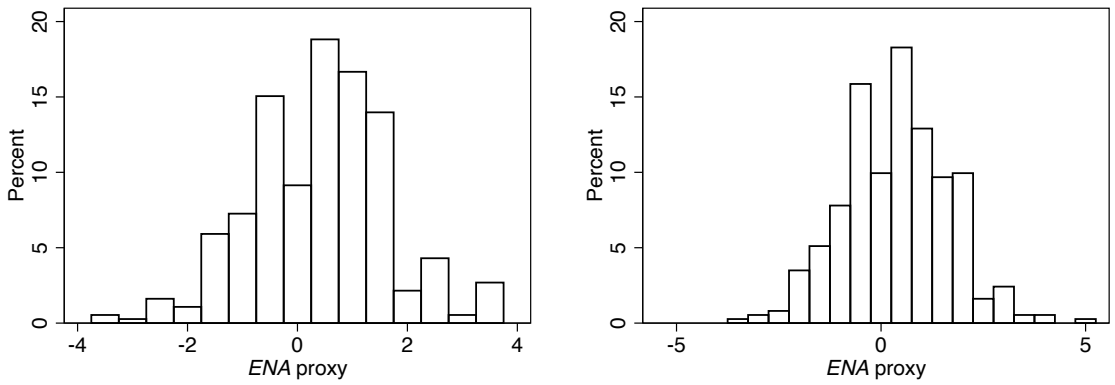
Therefore, we calculate for every decision its deviation from the average of choices for the respective procedure. For better comparability across procedures, we further divide the differences by the respective standard deviation, i.e., we transform subjects' choices into z-scores denoted by z_i^{EF} and z_i^{EQ} for the EF Procedure and the EQ Procedure. The *ENA* proxy is then constructed by adding the respective z-score if a subject has the relevant roles A or a and by subtracting it if the role is B or b. Using the indicator function $1_A\left(r_i^{EF}\right)$ for whether subject i's role for the EF Procedure is A and the analogously defined indicator functions for the three remaining procedure–role combinations, we thus calculate the *ENA* proxy as follows:

$$ENA_i \equiv \left[1_A\left(r_i^{EF}\right) - 1_B\left(r_i^{EF}\right)\right] z_i^{EF} + \left[1_a\left(r_i^{EQ}\right) - 1_b\left(r_i^{EQ}\right)\right] z_i^{EQ} \tag{1.7}$$

Deviations that are aligned with a subject's relevant role contribute to higher values of the *ENA* proxy, while deviations that are opposed to the relevant role lead to a decrease. A subject who makes average decisions for both procedures receives a value of zero, irrespective of her roles. On the other hand, a subject making a high decision for the EF Procedure and a low decision for the EQ Procedure receives a large positive value if her roles are A and b, values closer to zero if her roles are A and a or B and b, and a large negative value of the *ENA* proxy if her roles are B and a.

Of course, the decisions of subjects in part also reflect their true fairness convictions. As can be seen from the above examples, however, the expected effect of any given true fairness preferences on the value of the *ENA* proxy is exactly zero. That is because subjects'

roles determine the signs that Equation 1.7 attaches to the z-scores, and roles are drawn randomly with equal probabilities. The *ENA* proxy thus consists of two components: one is any systematic bias in decisions due to roles, and the other is random noise due to subjects' true fairness convictions. The latter is orthogonal to any subject-specific attributes by construction, while the former might correlate with subjects' personal characteristics.



(a) Decisions (b) Beliefs

Notes: The figure shows the distribution of the *ENA* proxies, one the left for decisions and on the right for beliefs. The respective values have been calculated according to Equation 1.7 for the full sample.

Figure 1.6: Distribution of the *ENA* proxies

Figure 1.6 shows the distributions of the *ENA* proxies for decisions and beliefs, respectively. For decisions, the mean value is 0.42, which is by construction equal to the mean of the two treatment effects in terms of standard deviations (0.37 for the EF Procedure and 0.47 for the EQ Procedure; see Section 1.5.1). This positive average is significantly different from zero ($p < 0.001$, two-sided t-test), in line with the previous findings. The figure shows that the positive average value, i.e., the effects of roles, is not mainly driven by a few subjects at the extremes. Instead, it is also caused by many subjects exhibiting moderate levels of bias towards their roles' interests. When restricting the sample to, e.g., only those 262 out of 372 subjects for whom the value of the *ENA* proxy lies in the interval $[-1.5, 1.5]$, the average value is still significantly positive ($p < 0.001$, two-sided t-test).[27] The picture looks similar for beliefs. Here, the mean value is 0.37 (the average of 0.32 and 0.42; see Section 1.5.2), which is again significantly different from zero ($p < 0.001$, two-sided t-test). As for decisions, it also holds for beliefs that the average value is still significantly positive among moderate values on the interval $[-1.5, 1.5]$ ($p = 0.018$, two-sided t-test).

1.6.2 Survey Measures

After the main experiment, subjects completed several questionnaires that were selected to measure potentially relevant personal characteristics. Below, we introduce the elicited classes of characteristics.

[27] In particular, the restriction excludes all subjects who favor their own roles to the largest possible extent, i.e., those with Roles A and a choosing Option 20 for both procedures, with Roles A and b choosing Option 20 for the EF Procedure and Option 1 for the EQ Procedure, with Roles B and a choosing Options 1 and 20, respectively, and with Roles B and b choosing Option 1 for both procedures.

Empathy To measure empathy, we use the well-established Interpersonal Reactivity Index (IRI) developed by Davis (1980), which consists of four subscales. The first, *perspective-taking*, should be of particular importance for non-egocentric behavior (Davis, 1983). The IRI measures perspective-taking with questions such as "I believe that there are two sides to every question and try to look at them both" (p. 11). Higher scores thus indicate that people typically make an effort to "put themselves in others' shoes," i.e., that they should tend to abstract from their roles in the experiment. Second, *fantasy* measures people's tendency to identify with fictitious characters, e.g., in books or movies. Third, *empathic concern* captures the extent to which people feel for others in need. The above dimensions of empathy are truly directed at others' feelings, and we would expect that they tend to decrease egocentric bias. In contrast, the fourth dimensions of *pdimensionstress* is "self-oriented" (Davis, 1983, p. 114) and addresses whether people feel anxious themselves when they witness others' suffering. Batson, Fultz, and Schoenrade (1987) argue that "empathic distress" is a vicarious feeling that is, in fact, distinct from empathy. In terms of behavior, empathy in its altruistic form facilitates helping (Batson et al., 1981), whereas distress induces an egoistic desire for relief. Therefore, personal distress might be expected to increase egocentric bias.

Prosociality The experiment in this chapter aims to show a bias that speaks of egocentrism. In contrast, the design mutes the role of egoism with the absence of selfish incentives. However, to study prosociality's role empirically, we included the qualitative item for altruism, positive reciprocity, and trust from the Preference Survey Module (Falk et al., 2016; Falk et al., 2018).

Values A leading approach in modern moral philosophy to understand how moral values vary across the political spectrum is Moral Foundations Theory (MFT) (Haidt and Joseph, 2004; Haidt and Graham, 2007; Graham, Haidt, and Nosek, 2009), which traces (cultural) differences in ethical judgments to the respective weights attached to five distinct dimensions of moral intuitions: *harm/care*, i.e., being compassionate with those in need; *fairness/reciprocity*; *ingroup/loyalty*; *authority/respect*; and *purity/sanctity*. We included the 30-item Moral Foundations Questionnaire (MFQ) that was created by a group of researchers around the developers of MFT.[28] As suggested by the developers, we aggregate the five subscales into a single measure of *progressivism*.[29]

$$\begin{aligned} progressivism = &(harm/care + fairness/reciprocity) \div 2 \\ &- (ingroup/loyalty + authority/respect + purity/sanctity) \div 3 \end{aligned}$$

We also include a simple question about people's political attitude on scale from *left* to *right* (European Social Survey, 2014). The variables *progressivism* and *political attitude* turn out

[28] The questionnaire is publicly available on the web (`https://moralfoundations.org/questionnaires/`; retrieved in May 2020).

[29] A very similar measure is used by Enke (2020), who excludes the *purity/sanctity* dimension and focuses on communal vs. universal values in the context of political competition. In our data, the correlation between these two measures based on the same questionnaire is 0.96.

to be highly correlated in the expected direction ($r = -0.51$, $p < 0.001$). Conceptually, we consider the above measures of values as potential *outcomes* of egocentric norm adoption. In contrast, the personality traits of empathy and prosociality are plausible *determinants*.

Personality Controls As control variables, we include the qualitative preference items by Falk et al. (2016) for risk preferences, time preferences, and negative reciprocity. Moreover, the questionnaires included the Big Five personality inventory, which is probably the most widely used framework to study people's personalities. Specifically, we use a translation of the 15-item BFI-S scale developed by Gerlitz and Schupp (2005). The Big Five traits are: *openness*, capturing interest in new experiences; *conscientiousness*, encompassing whether a person is determined and organized; *extraversion*, i.e, how much people like to engage with others; *agreeableness*, measuring altruistic motivation and cooperative behaviors; and *neuroticism*, referring to emotional instability and anxiety.

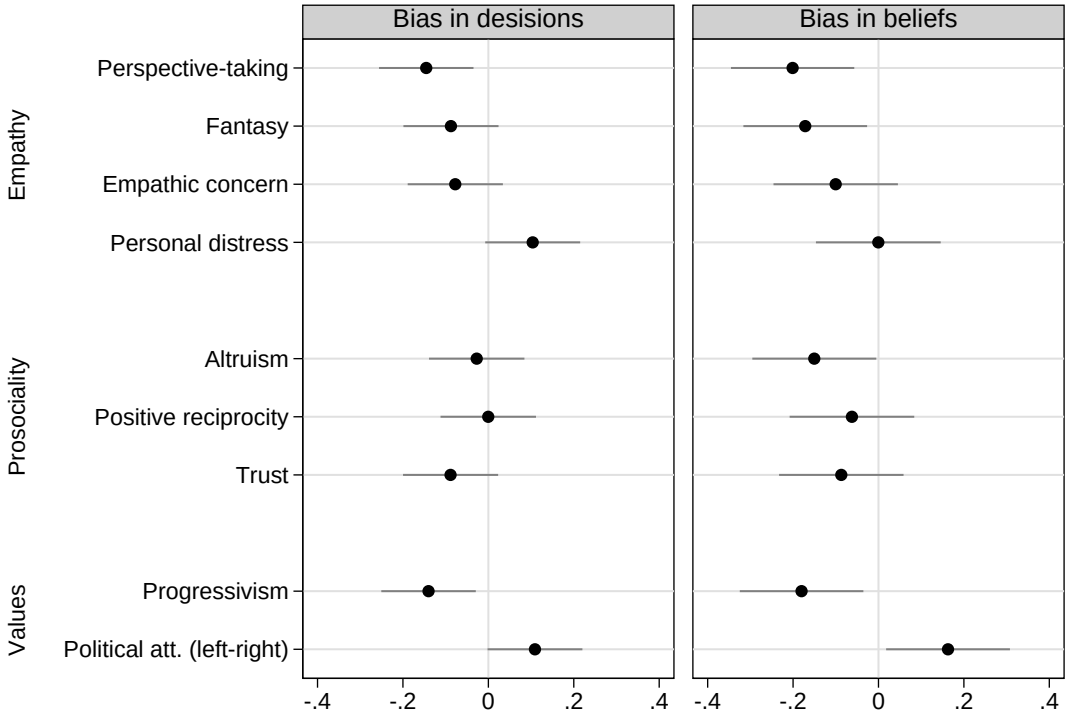
Demographic Controls The elicited sociodemographic controls are subjects' gender (female, male, or diverse) and age, enrollment at a university, and gross monthly income in euros. The latter is converted into log income as $\ln(income + 1)$.

Studying heterogeneity in subjects' behavior is more demanding than the previous analyses in Section 1.5 for two reasons. First, individual subjects' behavior has stronger effects on results since the sample is, intuitively, split, and the resulting subsamples are smaller than the full one. Second, we use answers to unincentivized survey questions, which some subjects might not have taken very seriously or had difficulty answering. Therefore, we restrict the sample to individuals who had no major difficulties understanding the experiment and took answering the survey questions seriously. The experiment included several control questions, and subjects could only progress once they answered them correctly. We automatically recorded the number of incorrectly submitted questions. We exclude those subjects above the 90^{th} percentile of the distribution of mistakes. After the questionnaires, we asked subjects about the reliability of their answers. We exclude subjects who gave answers below the tenth percentile. These restrictions leave us with 312 subjects, for whom all previous results from Section 1.5 replicate. Results for the full sample are provided in Appendix 1.B.[30]

1.6.3 Heterogeneity in Bias

Figure 1.7 considers the correlations between the *ENA* proxies and the different measures of empathy, prosociality, and values. The left panel shows the correlations between the survey measures and egocentric bias in decisions. For the four dimensions of empathy, a pattern arises that is consistent with the theoretical predictions: the three "altruistic"

[30] For the construction of the *ENA* proxy, we still use the z-scores based on the full sample. Alternatively, they can be calculated separately for the restricted sample. The corresponding values of the *ENA* proxy are almost identical ($\rho > 0.99$), and all results that follow remain virtually unchanged.



Notes: The figure shows the Pearson correlation coefficient between the *ENA* proxies and different survey measures. Gray bars indicate 95% confidence intervals. The analysis excludes subjects above the 90^{th} percentile in the distribution of mistakes in the control questions and those whose self-reported reliability regarding the survey answers lies below the 10^{th} percentile, leaving 312 subjects.

Figure 1.7: Correlations with the *ENA* proxies

facets of empathy—perspective-taking, fantasy, and empathic concern—are negatively correlated with egocentric norm adoption, i.e., higher empathy in these regards leads to lower egocentric bias. On the other hand, the "egoistic" side of empathy—personal distress—leads to a stronger egocentric bias. The correlation with perspective-taking, which is the opposite of egocentrism, is significantly negative ($p = 0.01$, two-sided), and the correlation with personal distress is (weakly) significantly positive ($p = 0.07$). The correlations with fantasy and empathic concern are not statistically significant ($p > 0.1$). We do not observe a significant correlation with either of the prosociality measures ($p > 0.1$). In particular, the correlation between the *ENA* proxy and altruism is close to zero, consistent with the irrelevance of selfishness. For moral values, we find a negative correlation with the ENA proxy for progressivism ($p = 0.01$), constructed using the MFQ. People holding liberal values thus seem to exhibit weaker egocentric bias than conservatives. Consistent with this finding, people leaning to the political right show a stronger bias than those leaning to the left ($p = 0.05$).[31]

The panel on the right displays the correlations with the *ENA* proxy for beliefs. Overall, they are remarkably similar to decisions. Again, the correlations with the three other-oriented dimensions of empathy are negative. In this case, they are statistically significant for perspective-taking ($p < 0.01$, two-sided) and also for fantasy ($p = 0.02$). The correlation with empathic concern is insignificant ($p > 0.1$). Thus, people who report little perspective-taking are not only more biased than others, but they also project their bias upon others. This finding suggests that perspective-taking, or the lack thereof, occurs unconsciously. It is in line with the assumptions in the formal framework (see Section 1.4.1) and with the insights provided by Singer and Fehr (2005). Other than observed for decisions, there is no indication of a relationship between bias in beliefs and personal distress. The correlations between the *ENA* proxy for beliefs and the prosociality measures tend to be negative. For one of them—altruism—it is now also statistically significant ($p = 0.04$). The correlations with progressivism and political orientation are very similar to those found for decisions, and they are both statistically significant ($p = 0.01$ and $p = 0.03$, respectively).

In the full sample, correlations of survey measures with the *ENA* proxies for decisions and beliefs are quite similar to those in the restricted sample (see Figure 1.B.6 in the appendix). Importantly, the correlations with perspective-taking remain statistically significant. Other results lose their statistical significance, most likely due to more noise in data. Only the positive correlation between the *ENA* proxy for beliefs and subjects' political attitude remains (weakly) statistically significant.

The analysis of heterogeneity is admittedly descriptive and does not aim at making causal claims. However, because many of the variables considered above are correlated, it would be interesting to see if the observed correlations with the potential determinants, i.e., with the different facts of empathy and prosociality, merely reflect different symptoms of maybe just a single underlying relationship or whether they also hold conditionally on each other. Therefore, we employ a regression framework. All of the reported regressions

[31] Progressivism and political attitude are strongly correlated in my data ($r - 0.51$ in the full and $r = -0.49$ in the restricted sample; both $p < 0.001$, two-sided).

Table 1.5: Heterogeneity

Dependent variable	*ENA proxy*			
Domain	*Decisions*		*Beliefs*	
	(1)	(2)	(3)	(4)
Perspective-taking	-0.176**	-0.185**	-0.187**	-0.180*
	(0.0754)	(0.0754)	(0.0923)	(0.0936)
Fantasy	-0.0999	-0.0886	-0.132	-0.136
	(0.0760)	(0.0786)	(0.0934)	(0.0978)
Empathic concern	-0.00779	-0.0109	0.103	0.119
	(0.0783)	(0.0792)	(0.107)	(0.107)
Personal distress	0.248***	0.261***	0.131	0.123
	(0.0712)	(0.0719)	(0.0986)	(0.0991)
Altruism	-0.00313	-0.00408	-0.161*	-0.166*
	(0.0713)	(0.0718)	(0.0876)	(0.0868)
Positive reciprocity	0.0400	0.0473	-0.0349	-0.0196
	(0.0601)	(0.0611)	(0.0800)	(0.0802)
Trust	-0.111*	-0.104*	-0.0961	-0.0877
	(0.0606)	(0.0618)	(0.0761)	(0.0768)
Personality controls	Yes	Yes	Yes	Yes
Demographic controls	No	Yes	No	Yes
Observations	312	312	312	312
R^2	0.120	0.130	0.093	0.106

Notes: The table reports standardized coefficients. Heteroscedasticity-consistent standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. The analysis excludes subjects above the 90^{th} percentile in the distribution of mistakes in the control questions and those whose self-reported reliability regarding the survey answers lies below the 10^{th} percentile, leaving 312 subjects. The *personality controls* are risk and time preferences along with the Big Five traits. The *demographic controls* are gender (categories: female, male, and diverse), age and squared age, a dummy for being enrolled at university, and the log of monthly gross income in euros, calculated as $\ln(income + 1)$.

include the full set of survey measures previously considered in Figure 1.7. Moreover, all columns control for other personality characteristics (see Section 1.6.2 above), and we report standardized coefficients. The independent variable in the first column is the *ENA* proxy for decisions. The results confirm the results from the correlations. Perspective-taking is associated with less biased decisions ($p = 0.02$) and personal distress with an increase in bias ($p < 0.01$). Otherwise, only the coefficient for trust is (weakly) statistically significant, entering with a negative sign ($p = 0.07$). The results hardly change when also controlling demographic characteristics in Column 2. The coefficients for all control variables can be found in Table 1.B.6 in the appendix. Columns 3 and 4 replicate the previous two for beliefs. The results for perspective-taking are remarkably similar to those for decisions, emphasizing the interpretation of unconsciousness. As already indicated by the respective correlations, personal distress does not appear to be associated with bias in beliefs. The only other coefficients that are (weakly) statistically significant are the negative ones for altruism.

Overall, this section's main result is that perspective-taking seems to play a central role in the emergence of egocentric norm adoption. Among subjects who report high levels of perspective-taking, the bias in decisions is significantly reduced. Subjects reporting little perspective-taking do not only take more biased decisions themselves, but they also project their own bias upon others. Egocentric norm adoption arises unconsciously, and whether individuals overcome it seems to depend on whether they can abstract from their own perspective.

1.7 Conclusion

This chapter has provided experimental evidence for the phenomenon of egocentric norm adoption. If people would benefit from others following certain norms, they adopt these principles themselves. The experiment's central property was that people's own decisions were in no way relevant for their own payoffs but that subjects depended on others' choices in the same decision contexts. Subjects within groups of two players received points according to two procedures. One of these implied a tradeoff between equality and efficiency, and the other involved the norms of equality and equity. Depending on their respective roles, subjects personally gained from one of the norms involved in a procedure and lost from the other. The players of each group decided over the subjects in the respective succeeding groups along a circle. Players' roles (in norms favoring them) were crossed, and players knew that they shared exactly one role with each subject over which they decided. We found an egocentric bias for both procedures and corresponding biases of similar size also in subjects' beliefs about others' behavior. The heterogeneity analysis provides additional support for egocentrism as the critical driving force behind the treatment effects of roles: the bias is largest among subjects who report weak perspective-taking.

Future research on egocentric norm adoption could explore additional potential mechanisms that might underly the phenomenon. This chapter has made the case that an unconscious egocentric bias leads subjects to empathize more with positions that they are

in themselves. Our view is supported by the effect arising largely unconsciously and the role of perspective-taking. However, the presence of one mechanism does not rule out the existence of others. The biases in decisions could, in part, also result from subjects confusing *diagnostic* with *causal* contingencies (Quattrone and Tversky, 1984; Shafir and Tversky, 1992; Acevedo and Krueger, 2005; Krueger and Acevedo, 2007), whereby subjects would try to "induce" a desired behavior by others with their own actions. Similarly, the results for decisions and beliefs could, to some extent, reflect "wishful thinking" (see, e.g., Mijovic-Prelec and Prelec, 2010; Engelmann et al., 2019)—although the evidence seems to suggests that this phenomenon is not present for purely financial stakes (Barron, 2020). Future work could adapt this chapter's experimental design but put some subjects into a position where a non-human random device like a computer determines their own payoffs. If subjects' decisions partly also reflect a direct concern with others' choice behavior, the egocentric biases should become smaller.

For methodology, the chapter's findings and those in Chapter 2 caution against the equivalent use of elicitation procedures for social preferences with or without role uncertainty. Iriberri and Rey-Biel (2011) and Zhan, Eckel, and Grossman (2020) find increased prosociality in (modified) dictator games when it is ex-ante uncertain whether a given subject will be paid as dictator according to one of her own decisions or as a receiver according to a decision by another subject. Egocentric norm adoption can accommodate these findings,[32] and it implies more. Whenever subjects play multiple games within an experiment, researchers who want to avoid bias should be aware that interests can induce norms, potentially creating spillovers between different contexts.

The idea of "acting like one would want others to act" is related to the concept of *rule-utilitarianism* advocated as a normative principle by Harsanyi (1977). Thereby "an individual act should be considered to be morally right if it conforms to the correct moral rule applying to this type of situation – regardless of whether it is the act that will or will not yield the highest possible social utility on this particular occasion" (p. 32). In particular, Harsanyi applies the logic of rule-utilitarianism to voting contexts. He shows that, if people were following rule-utilitarianism, this would, to some extent, resolve the *paradox of voting*. The latter describes the seemingly irrational behavior of people who incur the costs of voting in large elections (e.g., in terms of time) while almost certainly not being pivotal for the outcome (Downs, 1957). Rule-utilitarianism is an abstract normative concept that is probably unfamiliar to most potential voters. In contrast, egocentric norm adoption is grounded in people's intuition. It could explain why people sometimes resemble rule-utilitarians: like the subjects in the experiment in Chapter 2, they incur costs because they would like others to do the same. In the examples discussed by Harsanyi (1977), votes must exceed a certain threshold for the socially optimal option to be implemented, e.g., because a fixed number of votes is cast in favor of the respective alternative option

[32] Grech and Nax (2020) theoretically and empirically analyze the related but more complex difference between standard, non-interactive dictator games with certain roles and interactive dictators games. In the latter, roles are not uncertain, but subjects have two roles, simultaneously serving as recipients and dictators along a "loop." In line with this chapter's predictions and those of Chapter 2, Grech and Nax find less zero-giving in the interactive version of the dictator game than in the non-interactive one.

that is socially suboptimal. Harsanyi does not discuss how these votes come about. Under the label of *ethical voting*, some contributions have made suggestions for positive theories that resolve the paradox of voting. Feddersen and Sandroni (2006a, 2006b) and Coate and Conlin (2004) develop closely related models of voting over two alternative options. Both approaches assume *ethical* voters who follow rules that they would want to be followed by everybody *who favors the same option* as they do themselves, taking as given the behavior of non-ethical voters and ethical voters who favor the opposite option.[33] However, one might still be puzzled why people who behave ethically in terms of incurring (individually useless) voting costs should disagree on the optimal policy. Egocentric norm adoption offers an explanation: people consider options as fair from which they would personally profit, i.e., the selfish option subjectively is perceived as ethically demanded. Thus, a parsimonious behavioral principle explains prosocial behavior in turning out to vote and selfish behavior in terms of supported policies.

The models by Feddersen and Sandroni (2006a, 2006b) and Coate and Conlin (2004) both feature heterogeneous costs of voting. The rules that ethical voters adopt prescribe voting if and only if voting costs do not exceed a certain threshold value. That is because ethical voters aim at maximizing the utility of a group, and winning by an excessive margin would be wasteful. Thus, in the above models, heterogeneity enables coordination between voters who favor the same option. However, this model implication is at odds with experiments on public goods games that feature heterogeneity and find that heterogeneity reduces efficiency (see Fischbacher, Schudy, and Teyssier, 2014 and references therein).[34] Contrary to rule-utilitarianism, egocentric norm adoption is in line with these results since it implies that people will opt for sets of rules in their own favor. Incorporating egocentric rule-following into voting models and testing the resulting predictions would be a further exciting subject of future research.

Beyond voting and the examples in the introduction, many more real-world phenomena can be understood more clearly when considering the egocentric nature of norm adoption. Arguably the most important collective action problem of our time is the fight against global warming, i.e., in particular, the need to reduce global carbon dioxide emissions. It is true for all countries that unilateral action is pointless from a self-interested and strictly (act-) utilitarian perspective since costs are high and private returns (for a given country) are low. This insight applies to China and the United States, which account for 29.7% and 13.9% of global emissions in 2019, respectively (Crippa et al., 2019), but even more so, e.g., to the Marshall Islands, which are a small country in the Pacific Ocean that is part of Micronesia. However, the country is itself endangered by rising sea levels and has announced a plan for reducing carbon dioxide emissions to zero by 2050 (Malo, 2018). That a country with immense stakes takes bold steps against climate change, even when it has

[33]The two models differ in the objectives that individuals pursue: in the model by Feddersen and Sandroni (2006a), ethical voters maximize the utility of all people, while Coate and Conlin (2004) assume that they maximize only the utility of those people who share their own preferences, i.e., of those who are in their group.

[34]Similarly, Kube et al. (2015) find that heterogeneity also makes it more difficult for subjects to agree on efficiency-enhancing institutions, i.e., sets of mandatory rules.

virtually no impact, is what egocentric norm adoption would predict. In this context, the behavioral phenomenon is also closely linked to setting an example (cf. Gächter et al., 2012; Gächter, Nosenzo, and Sefton, 2013). Indeed, Bicchieri et al. (2020) show that observing others breaching a norm erodes people's own propensity to comply with the norm, and others who obey a norm heighten compliance. This finding suggests that acting upon norms that one would want others to follow can be useful in the long term. It thereby provides a potential explanation for why the bias has been evolutionarily successful.

References

Acevedo, Melissa, and Joachim I. Krueger. 2005. "Evidential Reasoning in the Prisoner's Dilemma". *American Journal of Psychology* 118 (3): 431–457.

Akerlof, George A., and William T. Dickens. 1982. "The Economic Consequences of Cognitive Dissonance". *American Economic Review* 72 (3): 307–319.

Alesina, Alberto, and Paola Giuliano. 2011. "Preferences for Redistribution". Chap. 4 in *Handbook of Social Economics*, ed. by Jess Benhabib, Alberto Bisin, and Matthew O. Jackson, vol. 1A, 93–131. North Holland.

Alger, Ingela, and Jörgen W. Weibull. 2019. "Evolutionary Models of Preference Formation". *Annual Review of Economics* 11:329–354.

— . 2013. "Homo Moralis—Preference Evolution Under Incomplete Information and Assortative Matching". *Econometrica* 81 (6): 2269–2302.

Almås, Ingvild, Alexander W. Cappelen, Kjell G. Salvanes, Erik Sørensen, and Bertil Tungodden. 2017. "Fairness and family background". *Politics, Philosophy and Economics* 16 (2): 117–131.

Andreoni, James. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving". *The Economic Journal* 100 (401): 464–477.

Babcock, Linda, and George Loewenstein. 1997. "Explaining Bargaining Impasse: The Role of Self-Serving Biases". *Journal of Economic Perspectives* 11 (1): 109–126.

Babcock, Linda, George Loewenstein, Samuel Issacharoff, and Colin Camerer. 1995. "Biased Judgments of Fairness in Bargaining". *American Economic Review* 85 (5): 1337–1343.

Barron, Kai. 2020. "Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains?" *Experimental Economics*.

Barron, Kai, Robert Stüber, and Roel van Veldhuizen. 2019. *Motivated motive selection in the lying-dictator game*. Discussion Paper SP II 2019–303. Berlin, Germany: Wissenschaftszentrum Berlin für Sozialforschung (WZB).

Batson, C. Daniel, Bruce D. Duncan, Paula Ackerman, Terese Buckley, and Kimberly Birch. 1981. "Is Empathic Emotion a Source of Altruistic Motivation?" *Journal of Personality and Social Psychology* 40 (2): 290–302.

Batson, C. Daniel, Jim Fultz, and Patricia A. Schoenrade. 1987. "Distress and Empathy: Two Qualitatively Distinct Vicarious Emotions with Different Motivational Consequences". *Journal of Personality* 55 (1): 19–39.

Becker, Gary S. 1974. "A Theory of Social Interactions". *Journal of Political Economy* 82 (6): 1063–1093.

Bénabou, Roland, and Jean Tirole. 2016. "Mindful Economics: The Production, Consumption, and Value of Beliefs". *Journal of Economic Perspectives* 30 (3): 141–164.

Bergstrom, Theodore C. 1995. "On the Evolution of Altruistic Ethical Rules for Siblings". *American Economic Review* 85 (1): 58–81.

Bicchieri, Cristina, and Alex K. Chavez. 2013. "Norm Manipulation, Norm Evasion: Experimental Evidence". *Economics and Philosophy* 29 (2): 175–198.

Bicchieri, Cristina, Eugen Dimant, Simon Gächter, and Daniele Nosenzo. 2020. *Observability, Social Proximity, and the Erosion of Norm Compliance*. CESifo Working Paper 8212. Munich, Germany: Munich Society for the Promotion of Economic Research – CESifo.

Billig, Michael, and Henri Tajfel. 1973. "Social categorization and similarity in intergroup behaviour". *European Jounral of Social Psychology* 3 (1): 27–52.

Blanco, Mariana, Dirk Engelmann, Alexander K. Koch, and Hans Theo Normann. 2014. "Preferences and beliefs in a sequential social dilemma: a within-subjects analysis". *Games and Economic Behavior* 87:122–135.

Bocian, Konrad, and Bogdan Wojciszke. 2014. "Self-Interest Bias in Moral Judgments of Others' Actions". *Personality and Social Psychology Bulletin* 40 (7): 898–909.

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch. 2014. "hroot: Hamburg Registration and Organization Online Tool". *European Economic Review* 71:117–120.

Bolton, By Gary E, and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition". *American Economic Review* 90 (1): 166–193.

Brown, Rupert J., and John C. Turner. 1979. "The Criss-cross Categorization Effect in intergroup discrimination". *British Journal of Social and Clinical Psychology* 18 (4): 371–383.

Cappelen, Alexander W., Astri Drange Hole, Erik Ø. Sørensen, and Bertil Tungodden. 2007. "The Pluralism of Fairness Ideals: An Experimental Approach". *American Economic Review* 97 (3): 818–827.

Cappelen, Alexander W., Konow James, Erik Ø. Sørensen, and Bertil Tungodden. 2013. "Just Luck: An Experimental Study of Risk Taking and Fairness". *American Economic Review* 103 (4): 1398–1413.

Cassar, Lea, and Arnd H. Klein. 2019. "A matter of perspective: How failure shapes distributive preferences". *Management Science* 65 (11): 5050–5064.

Cerrone, Claudia, and Christoph Engel. 2019. "Deciding on behalf of others does not mitigate selfishness: An Experiment". *Economics Letters* 183:108616.

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

Chen, Yan, and Sherry Xin Li. 2009. "Group identity and social preferences". *American Economic Review* 99 (1): 431–457.

Chong, Alberto, and Mark Gradstein. 2008. "What determines foreign aid? The donors' perspective". *Journal of Development Economics* 87 (1): 1–13.

Coate, Stephen, and Michael Conlin. 2004. "A Group Rule–Utilitarian Approach to Voter Turnout: Theory and Evidence". *American Economic Review* 94 (5): 1476–1504.

Costa-Gomes, Miguel A., Yuan Ju, and Jiawen Li. 2019. "Role-Reversal Consistency: An Experimental Study of the Golden Rule". *Economic Inquiry* 57 (1): 685–704.

Crippa, M., G. Oreggioni, D. Guizzardi, M. Muntean, E. Schaaf, E. Lo Vullo, E. Solazzo, F. Monforti-Ferrario, J. G. J. Olivier, and E. Vignati. 2019. *Fossil CO2 and GHG emissions of all world countries - 2019 Report*. Tech. rep. Luxemburg: Publications Office of the European Union.

Crisp, Richard J., and Miles Hewstone. 1999. "Differential Evaluation of Crossed Category Groups: Patterns, Processes, and Reducing Intergroup Bias". *Group Processes & Intergroup Relations* 2 (4): 307–333.

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

Davis, Mark H. 1980. "A Multidimensional Approach to Individual Differences in Empathy". *JSAS Catalog of Selected Documents in Psychology* 10:85.

— . 1983. "Measuring Individual Differences in Empathy: Evidence for a Multidimensional Approach". *Journal of Personality and Social Psychology* 44 (1): 113–126.

De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth. 2018. "Measuring and bounding experimenter demand". *American Economic Review* 108 (11): 3266–3302.

Di Tella, Rafael, Ricardo Perez-Truglia, Andres Babino, and Mariano Sigman. 2015. "Conveniently Upset: Avoiding Altruism by Distorting Beliefs about Others' Altruism". *American Economic Review* 105 (11): 3416–3442.

Downs, Anthony. 1957. *An Economic Theory of Democracy*. New York, NY: Harper / Row.

Elster, Jon. 1998. "Emotions and Economic Theory". *Journal of Economic Literature* 36 (1): 47–74.

— . 1989. "Social Norms and Economic Theory". *Journal of Economic Perspectives* 3 (4): 99–117.

Engelmann, Jan, Maël Lebreton, Peter Schwardmann, Joel J. van der Weele, and Li-Ang Chang. 2019. *Anticipatory Anxiety and Wishful Thinking*. Mimeo.

Enke, Benjamin. 2020. "Moral Values and Voting". *Journal of Political Economy*: forthcoming.

Epley, Nicholas, and Eugene M. Caruso. 2004. "Egocentric Ethics". *Social Justice Research* 17 (2): 171–187.

European Social Survey. 2014. *ESS Round 7 Source Questionnaire*. ESS ERIC Headquarters, Centre for Comparative Social Surveys, City University London, London, United Kingdom.

Exley, Christine L. 2016. "Excusing Selfishness in Charitable Giving: The Role of Risk". *Review of Economic Studies* 83 (2): 587–628.

Exley, Christine L., and Judd B. Kessler. 2019. *Motivated Errors*. NBER Working Paper. Cambridge, MA: National Bureau of Economic Research.

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

Falk, Armin, Anke Becker, Thomas Dohmen, David Huffman, and Uwe Sunde. 2016. *The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences*. IZA Discussion Paper 9674. Bonn: Institute for the Study of Labor.

Falk, Armin, and Urs Fischbacher. 2006. "A theory of reciprocity". *Games and Economic Behavior* 54 (2): 293–315.

Feddersen, Timothy, and Alvaro Sandroni. 2006a. "A Theory of Participation in Elections". *American Economic Review* 96 (4): 1271–1282.

— . 2006b. "The calculus of ethical voting". *International Journal of Game Theory* 35 (1): 1–25.

Fehr, Ernst, and Simon Gächter. 2000. "Fairness and Retaliation: The Economics of Reciprocity". *Journal of Economic Perspectives* 14 (3): 159–181.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation". *Quarterly Journal of Economics* 114 (3): 817–868.

Festinger, Leon. 1957. *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Fischbacher, Urs, Simeon Schudy, and Sabrina Teyssier. 2014. "Heterogeneous reactions to heterogeneity in returns from public goods". *Social Choice and Welfare* 43 (1): 195–217.

Fliessbach, Klaus, Bernd Weber, Peter Trautner, Thomas Dohmen, Uwe Sunde, Christian E. Elger, and Armin Falk. 2007. "Social Comparison Affects Reward-Related Brain Activity in the Human Ventral Striatum". *Science* 318 (5854): 1305–1308.

Foot, Philippa. 1967. "The Problem of Abortion and the Doctrine of the Double Effect". *Oxford Review* 5:5–15.

Gächter, Simon, Daniele Nosenzo, Elke Renner, and Martin Sefton. 2012. "Who Makes a Good Leader? Cooperativeness, Optimism and Leading-by-Example". *Economic Inquiry* 50 (4): 953–967.

Gächter, Simon, Daniele Nosenzo, and Martin Sefton. 2013. "Peer Effects in Pro-Social Behavior: Social Norms or Social Preferences?" *Journal of the European Economic Association* 11 (3): 548–573.

Gerlitz, Jean-Yves, and Jürgen Schupp. 2005. *Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. Dokumentation der Instrumententwicklung BFI-S auf Basis des SOEP-Pretests 2005*. Research Notes 4. Berlin, Germany: Deutsches Institut für Wirtschaftsforschung (DIW).

Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2019. "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study". *Journal of Political Economy* 127 (4): 1826–1863.

Gino, Francesca, Shahar Ayal, and Dan Ariely. 2013. "Self-serving altruism? The lure of unethical actions that benefit others". *Journal of Economic Behavior and Organization* 93:285–292.

Gino, Francesca, Michael I. Norton, and Roberto A. Weber. 2016. "Motivated Bayesians: Feeling Moral While Acting Egoistically". *Journal of Economic Perspectives* 30 (3): 189–212.

Gneezy, Uri, Silvia Saccardo, Marta Serra-Garcia, and Roel van Veldhuizen. 2020. "Bribing the Self". *Games and Economic Behavior* 120 (311-324).

Graham, Jesse, Jonathan Haidt, and Brian A. Nosek. 2009. "Liberals and Conservatives Rely on Different Sets of Moral Foundations". *Journal of Personality and Social Psychology* 96 (5): 1029–1046.

Grech, Philip D., and Heinrich H. Nax. 2020. "Rational altruism? On preference estimation and dictator game experiments". *Games and Economic Behavior* 119:309–338.

Haidt, Jonathan. 2001. *The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment.*

Haidt, Jonathan, and Jesse Graham. 2007. "When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize". *Social Justice Research* 20 (1): 98–116.

Haidt, Jonathan, and Craig Joseph. 2004. "Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues". *Daedalus* 133 (4): 55–66.

Haisley, Emily C., and Roberto A. Weber. 2010. "Self-serving interpretations of ambiguity in other-regarding behavior". *Games and Economic Behavior* 68 (2): 614–625.

Harsanyi, John C. 1977. "Rule Utilitarianism and Decision Theory". *Erkenntnis* 11 (1): 25–53.

Hippel, Svenja, and Sven Hoeppner. 2019. "Biased judgements of fairness in bargaining: A replication in the laboratory". *International Review of Law and Economics* 58:63–74.

Iriberri, Nagore, and Pedro Rey-Biel. 2011. "The role of role uncertainty in modified dictator games". *Experimental Economics* 14 (2): 160–180.

Kant, Immanuel. 1996. "Groundwork of the metaphysics of morals". In *The Cambridge edition of the works of Immanuel Kant: Practical philosophy*, ed. by Mary J. Gregor, 37–108. Cambridge, United Kingdom: Cambridge University Press.

Kassas, Bachir, and Marco A. Palma. 2019. "Self-serving biases in social norm compliance". *Journal of Economic Behavior and Organization*.

Konow, James. 2001. "Fair and square: the four sides of distributive justice". *Journal of Economic Behavior and Organization* 46 (2): 137–164.

— . 2000. "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions". *American Economic Review* 90 (4): 1072–1091.

— . 2009. "Is fairness in the eye of the beholder? An impartial spectator analysis of justice". *Social Choice and Welfare* 33 (1): 101–127.

— . 2003. "Which Is the Fairest One of All? A Positive Analysis of Justice Theories". *Journal of Economic Literature* 41 (4): 1188–1239.

Krueger, Joachim I., and Melissa Acevedo. 2007. "Perceptions of self and other in the prisoner's dilemma: Outcome bias and evidential reasoning". *American Journal of Psychology* 120 (4): 593–618.

Kube, Sebastian, Sebastian Schaube, Hannah Schildberg-Hörisch, and Elina Khachatryan. 2015. "Institution formation and cooperation with heterogeneous agents". *European Economic Review* 78:248–268.

Kunda, Ziva. 1987. "Motivated Inference: Self-Serving Generation and Evaluation of Causal Theories". *Journal of Personality and Social Psychology* 53 (4): 636–647.

— . 1990. "The Case for Motivated Reasoning". *Psychological Bulletin* 108 (3): 480–498.

Leeuwen, Boris van, Ingela Alger, and Jörgen W. Weibull. 2019. *Estimating Social Preferences and Kantian Morality in Strategic Interactions*. Mimeo.

Loewenstein, George, Samuel Issacharoff, Colin Camerer, and Linda Babcock. 1993. "Self-Serving Assessments of Fairness and Pretrial Bargaining". *Journal of Legal Studies* 22 (1): 135–159.

Malo, Sebastien. 2018. "Marshall Islands marches toward zero greenhouse emissions by 2050". *Reuters*. Visited on 07/05/2021. `https://www.reuters.com/article/us-global-climatechange-summit-idUSKCN1M42GI`.

Messick, David M., and Keith P. Sentis. 1979. "Fairness and preference". *Journal of Experimental Social Psychology* 15 (4): 418–434.

Mijovic-Prelec, Danica, and Drazen Prelec. 2010. "Self-deception as self-signalling: A model and experimental evidence". *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1538): 227–240.

Mullen, Brian, Rupert Brown, and Colleen Smith. 1992. "Ingroup bias as a function of salience, relevance, and status: An integration". *European Journal of Social Psychology* 22 (2): 103–122.

Oster, Emily, Ira Shoulson, and E. Ray Dorsey. 2013. "Optimal Expectations and Limited Medical Testing: Evidence from Huntington Disease". *American Economic Review* 103 (2): 804–830.

Ostrom, Elinor. 2000. "Collective Action and the Evolution of Social Norms". *Journal of Economic Perspectives* 14 (3): 137–158.

Quattrone, George A., and Amos Tversky. 1984. "Causal Versus Diagnostic Contingencies: On Self-Deception and on the Voter's Illusion." *Journal of Personality and Social Psychology* 46 (2): 237–248.

Rodriguez-Lara, Ismael, and Luis Moreno-Garrido. 2012. "Self-interest and fairness: self-serving choices of justice principles". *Experimental Economics* 15:158–175.

Roemer, John E. 2010. "Kantian Equilibrium". *Scandinavian Journal of Economics* 112 (1): 1–24.

Ross, Lee, David Greene, and Pamela House. 1977. "The "False Consensus Effect": An Egocentric Bias in Social Perception and Attribution Processes". *Journal of Experimental Social Psychology* 13 (3): 279–301.

Schwardmann, Peter, Egon Tripodi, and Joël J. van der Weele. 2019. *Self-Persuasion: Evidence from Field Experiments at Two International Debating Competitions*. CESifo Working Paper 7946. Munich, Germany: CESifo.

Schwardmann, Peter, and Joël van der Weele. 2019. "Deception and self-deception". *Nature Human Behavior* 3:1055–1061.

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

Singer, Tania, and Ernst Fehr. 2005. "The Neuroeconomics of Mind Reading and Empathy". *American Economic Review* 95 (2): 340–345.

Slovic, Paul, Melissa Finucane, Ellen Peters, and Donald G. MacGregor. 2002. "Rational actors or rational fools: Implications of the effects heuristic for behavioral economics". *Journal of Socio-Economics* 31 (4): 329–342.

Smith, Megan K., Robert Trivers, and William von Hippel. 2017. "Self-deception facilitates interpersonal persuasion". *Journal of Economic Psychology* 63:93–101.

Tajfel, Henri, M. G. Billig, and R. P. Bundy. 1971. "Social categorization and intergroup behaviour". *European Journal of Social Psychology* 1 (2): 149–178.

Thompson, Leigh, and George Loewenstein. 1992. "Egocentric Interpretations of Fairness and Interpersonal Conflict". *Organizational Behavior and Human Decision Processes* 51 (2): 176–197.

Turner, J. C., R. J. Brown, and H. Tajfel. 1979. "Social comparison and group interest in ingroup favouritism". *European Journal of Social Psychology* 9 (2): 187–204.

Van Boven, Leaf, David Dunning, and George Loewenstein. 2000. "Egocentric Empathy Gaps Between Owners and Buyers: Misperceptions of the Endowment Effect". *Journal of Personality and Social Psychology* 79 (1): 66–76.

Van Boven, Leaf, George Loewenstein, David Dunning, and Loran F. Nordgren. 2013. "Changing Places: A Dual Judgment Model of Empathy Gaps in Emotional Perspective Taking". Chap. 3 in *Advances in Experimental Social Psychology*, ed. by Mark Zanna and James Olson, 48:117–171. Academic Press.

Wilson, Timothy D., and Nancy Brekke. 1994. "Mental Contamination and Mental Correction: Unwanted Influences on Judgments and Evaluations". *Psychological Bulletin* 116 (2): 117–142.

Zajonc, R. B. 1980. "Feeling and Thinking: Preferences Need No Inferences". *American Psychologist* 35 (2): 151–175.

Zhan, Wei, Catherine C Eckel, and Philip J Grossman. 2020. *Does How We Measure Altruism Matter? Playing Both Roles in Dictator Games*. Mimeo.

Zimmermann, Florian. 2020. "The Dynamics of Motivated Beliefs". *American Economic Review* 110 (2): 337–363.

Appendix 1.A Theoretical Details

1.A.1 Proofs

Proof of Lemma 1.1. The first order conditions for Equations 1.1 and 1.2 are as follows.

$$\tilde{\alpha}\, Pay'(\tilde{c}_{EF}, role_{EF}) - \tilde{\beta}_1\, Ineff'(\tilde{c}_{EF}) - Inequal'_{EF}(\tilde{c}_{EF}) = 0$$
$$\tilde{\alpha}\, Pay'(\tilde{c}_{EQ}, role_{EQ}) - \tilde{\beta}_2\, Unfair'(\tilde{c}_{EQ}) - Inequal'_{EQ}(\tilde{c}_{EQ}) = 0$$

Choose any $\tilde{c}_{EF}, \tilde{c}_{EQ} \in (1, 20)$ and fix α at a positive value such that the two remaining true fairness parameters that follow from the first-order conditions are also strictly positive.

$$\beta_1 = \frac{\alpha\, Pay'(\tilde{c}_{EF}, role_{EF}) - Inequal'_{EF}(\tilde{c}_{EF})}{Ineff'(\tilde{c}_{EF})}\,,$$
$$\beta_2 = \frac{\alpha\, Pay'(\tilde{c}_{EQ}, role_{EQ}) - Inequal'_{EQ}(\tilde{c}_{EQ})}{Unfair'(\tilde{c}_{EQ})}\,.$$

Recall that the agent's prior beliefs about the values of the unknown parameters are independently normally distributed with standard deviations of one. The expected values are the true values for β_1 and β_2, while it is $\pi\alpha$ for α, with $\pi \in [0, 1]$. Thus, the likelihood of any set of values under the prior beliefs is

$$\mathcal{L} = \phi(\tilde{\alpha} - \pi\alpha) \times \phi\Big(\tilde{\beta}_1 - \beta_1\Big) \times \phi\Big(\tilde{\beta}_2 - \beta_2\Big)\,,$$

where ϕ denotes the probability density function of the standard normal distribution. The agent maximizes the corresponding log likelihood subject to the first order conditions.

$$\max_{\tilde{\alpha},\tilde{\beta}_1,\tilde{\beta}_2} Constant - \frac{(\tilde{\alpha} - \pi\alpha)^2 + \Big(\tilde{\beta}_1 - \beta_1\Big)^2 + \Big(\tilde{\beta}_2 - \beta_2\Big)^2}{2}$$
$$\text{s.t.} \quad \tilde{\alpha}\, Pay'(\tilde{c}_{EF}, role_{EF}) - \tilde{\beta}_1\, Ineff'(\tilde{c}_{EF}) - Inequal'_{EF}(\tilde{c}_{EF}) = 0$$
$$\tilde{\alpha}\, Pay'(\tilde{c}_{EQ}, role_{EQ}) - \tilde{\beta}_2\, Unfair'(\tilde{c}_{EQ}) - Inequal'_{EQ}(\tilde{c}_{EQ}) = 0$$

In the below notation, derivatives of functions are indicated by small letters and the affective choice as an argument of the functions is omitted. Moreover, define

$$D = \mathit{ineff}^2 \left(\mathit{unfair}^2 + \mathit{pay}(\mathit{role}_{EQ})^2 \right) + \mathit{unfair}^2 \, \mathit{pay}(\mathit{role}_{EF})^2 \ .$$

Observe that D is always strictly positive. The unique solution of the maximization problem has the following properties.

$$\tilde{\alpha} - \alpha = -\frac{(1-\pi)\,\alpha\,\mathit{ineff}^2\,\mathit{unfair}^2}{D} \tag{1.A.1}$$

$$\tilde{\beta}_1 - \beta_1 = -\frac{(1-\pi)\,\alpha\,\mathit{ineff}\,\mathit{unfair}^2\,\mathit{pay}(\mathit{role}_{EF})}{D} \tag{1.A.2}$$

$$\tilde{\beta}_2 - \beta_2 = -\frac{(1-\pi)\,\alpha\,\mathit{ineff}^2\,\mathit{unfair}\,\mathit{pay}(\mathit{role}_{EQ})}{D} \tag{1.A.3}$$

Part 1 of the lemma follows from Equation 1.A.1. Parts 2a and 2b follow from Equations 1.A.2 and 1.A.3. □

Proof of Lemma 1.2. The first order conditions for Equations 1.3 and 1.4 are as follows.

$$\tilde{\alpha}\,\mathit{Pay}'(\tilde{c}_{EF}, \mathit{role}_{EF}) - \tilde{\beta}_1\,\mathit{Ineff}'(\tilde{c}_{EF}) - \tilde{\gamma}\,\mathit{Inequal}'_{EF}(\tilde{c}_{EF}) = 0$$
$$\tilde{\alpha}\,\mathit{Pay}'(\tilde{c}_{EQ}, \mathit{role}_{EQ}) - \tilde{\beta}_2\,\mathit{Unfair}'(\tilde{c}_{EQ}) - \tilde{\gamma}\,\mathit{Inequal}'_{EQ}(\tilde{c}_{EQ}) = 0$$

Choose any $\tilde{c}_{EF}, \tilde{c}_{EQ} \in (1, 20)$ and fix α at a positive and γ at a strictly positive value such that the two remaining true fairness parameters that follow from the first-order conditions are also strictly positive.

$$\beta_1 = \frac{\alpha\,\mathit{Pay}'(\tilde{c}_{EF}, \mathit{role}_{EF}) - \gamma\,\mathit{Inequal}'_{EF}(\tilde{c}_{EF})}{\mathit{Ineff}'(\tilde{c}_{EF})}\,,$$
$$\beta_2 = \frac{\alpha\,\mathit{Pay}'(\tilde{c}_{EQ}, \mathit{role}_{EQ}) - \gamma\,\mathit{Inequal}'_{EQ}(\tilde{c}_{EQ})}{\mathit{Unfair}'(\tilde{c}_{EQ})}\,.$$

Recall that the agent's prior beliefs about the values of the unknown parameters are independently normally distributed with standard deviations of one. The expected values are the true values for β_1, β_2, and γ, while it is $\pi\alpha$ for α, with $\pi \in [0, 1]$. Thus, the likelihood of any set of values under the prior beliefs is

$$\mathcal{L} = \phi(\tilde{\alpha} - \pi\alpha) \times \phi\Big(\tilde{\beta}_1 - \beta_1\Big) \times \phi\Big(\tilde{\beta}_2 - \beta_2\Big) \times \phi(\tilde{\gamma} - \gamma)\ ,$$

where ϕ denotes the probability density function of the standard normal distribution. The

agent maximizes the corresponding log likelihood subject to the first order conditions.

$$\begin{aligned}\max_{\tilde{\alpha},\tilde{\beta}_1,\tilde{\beta}_2,\tilde{\gamma}} \; & \mathit{Constant} - \frac{(\tilde{\alpha}-\pi\alpha)^2 + \left(\tilde{\beta}_1-\beta_1\right)^2 + \left(\tilde{\beta}_2-\beta_2\right)^2 + (\tilde{\gamma}-\gamma)^2}{2} \\ \text{s.t.} \quad & \tilde{\alpha}\, \mathit{Pay}'(\tilde{c}_{EF}, \mathit{role}_{EF}) - \tilde{\beta}_1\, \mathit{Ineff}'(\tilde{c}_{EF}) - \tilde{\gamma}\, \mathit{Inequal}'_{EF}(\tilde{c}_{EF}) = 0 \\ & \tilde{\alpha}\, \mathit{Pay}'(\tilde{c}_{EQ}, \mathit{role}_{EQ}) - \tilde{\beta}_2\, \mathit{Unfair}'(\tilde{c}_{EQ}) - \tilde{\gamma}\, \mathit{Inequal}'_{EQ}(\tilde{c}_{EQ}) = 0\end{aligned}$$

In the below notation, derivatives of functions are indicated by small letters and the affective choice as an argument of the functions is omitted. Moreover, define

$$\begin{aligned}D = {} & \mathit{ineff}^2 \left(\mathit{unfair}^2 + \mathit{inequal}_{EQ}^2 + \mathit{pay}(\mathit{role}_{EQ})^2\right) \\ & + \mathit{unfair}^2 \left(\mathit{inequal}_{EF}^2 + \mathit{pay}(\mathit{role}_{EF})^2\right) \\ & + \left(\mathit{inequal}_{EF}\, \mathit{pay}(\mathit{role}_{EQ}) - \mathit{inequal}_{EQ}\, \mathit{pay}(\mathit{role}_{EF})\right)^2 .\end{aligned}$$

Observe that D is always strictly positive. The unique solution of the maximization problem has the following properties.

$$\tilde{\alpha} - \alpha = -\frac{(1-\pi)\,\alpha \left[\mathit{ineff}^2 \left(\mathit{unfair}^2 + \mathit{inequal}_{EQ}^2\right) + \mathit{unfair}^2\, \mathit{inequal}_{EF}^2\right]}{D} \tag{1.A.4}$$

$$\tilde{\beta}_1 - \beta_1 = -\frac{\begin{array}{r}(1-\pi)\,\alpha\, \mathit{ineff} \big[\left(\mathit{unfair}^2 + \mathit{inequal}_{EQ}^2\right) \mathit{pay}(\mathit{role}_{EF}) \\ - \mathit{inequal}_{EF}\, \mathit{inequal}_{EQ}\, \mathit{pay}(\mathit{role}_{EQ}) \big]\end{array}}{D} \tag{1.A.5}$$

$$\tilde{\beta}_2 - \beta_2 = -\frac{\begin{array}{r}(1-\pi)\,\alpha\, \mathit{unfair} \big[\left(\mathit{ineff}^2 + \mathit{inequal}_{EF}^2\right) \mathit{pay}(\mathit{role}_{EQ}) \\ - \mathit{inequal}_{EF}\, \mathit{inequal}_{EQ}\, \mathit{pay}(\mathit{role}_{EF}) \big]\end{array}}{D} \tag{1.A.6}$$

$$\tilde{\gamma} - \gamma = -\frac{\begin{array}{r}(1-\pi)\,\alpha \big(\mathit{ineff}^2\, \mathit{inequal}_{EQ}\, \mathit{pay}(\mathit{role}_{EQ}) \\ + \mathit{unfair}^2\, \mathit{inequal}_{EF}\, \mathit{pay}(\mathit{role}_{EF}) \big)\end{array}}{D} \tag{1.A.7}$$

Part 1 of the Lemma directly follows from Equation 1.A.4. The results for $\tilde{\gamma}$ of Parts 2a and 2b directly follow from Equation 1.A.7. To see both statements' results for β_1 and β_2, observe that $\mathit{inequal}_{EQ}^2\, \mathit{pay}(\mathit{role}_{EF}) < \mathit{inequal}_{EF}\, \mathit{inequal}_{EQ}\, \mathit{pay}(\mathit{role}_{EQ})$ implies that $\mathit{inequal}_{EF}^2\, \mathit{pay}(\mathit{role}_{EQ}) > \mathit{inequal}_{EF}\, \mathit{inequal}_{EQ}\, \mathit{pay}(\mathit{role}_{EF})$. Thus, for roles (A, a), it cannot hold that $\tilde{\beta}_1 < \beta_1$ and at the same time $\tilde{\beta}_2 < \beta_2$. Conversely, for roles (B, b), it cannot hold that $\tilde{\beta}_1 > \beta_1$ and at the same time $\tilde{\beta}_2 > \beta_2$. Parts 2c and 2d directly follow from Equations 1.A.5 and 1.A.6. □

1.A.2 Hypothesis Testing

We conduct the following statistical hypothesis test:

$$\begin{aligned}H_0: \quad & \delta_1 \leq 0 \vee \zeta_1 \leq 0 \\ H_1: \quad & \delta_1 > 0 \wedge \zeta_1 > 0\end{aligned}$$

Thus, we want to reject the Null hypothesis of either coefficient being weakly negative, i.e., we want to establish that both coefficients are strictly positive. Note that in Equations 1.5 and 1.6, $1_A\big(r_i^{EF}\big)$ and $1_a\big(r_i^{EQ}\big)$ are statistically independent, since all combinations or roles appear with exactly the same frequencies in the experiment. Moreover, ϵ_i and η_i are each pairwise statistically independent of both $1_A\big(r_i^{EF}\big)$ and $1_a\big(r_i^{EQ}\big)$, since assignment to roles is randomized.

To understand the implications of the above discussion for the hypothesis test, consider the following scenario: we have estimated the two regression equations 1.5 and 1.6 and retrieved the p-values p_δ and p_ζ referring to the two-sided significance tests of δ_1 and ζ_1, respectively. The p-value referring to the above hypothesis test is the probability of either of the two t-values under H_0 (t_δ^0 and t_ζ^0) being as large as they are (t_δ and t_ζ), with at least one of δ_1 and ζ_1 being smaller than zero.

$$
\begin{aligned}
p &= P\big(t_\delta^0 \geq t_\delta \wedge t_\zeta^0 \geq t_\zeta \,\big|\, \delta_1 \leq 0 \vee \zeta_1 \leq 0\big) \\
&= P\big(t_\delta^0 \geq t_\delta \wedge t_\zeta^0 \geq t_\zeta \,\big|\, \delta_1 \leq 0\big) \times P(\delta_1 \leq 0 \,|\, \delta_1 \leq 0 \vee \zeta_1 \leq 0) \\
&\quad + P\big(t_\delta^0 \geq t_\delta \wedge t_\zeta^0 \geq t_\zeta \,\big|\, \zeta_1 \leq 0\big) \times P(\zeta_1 \leq 0 \,|\, \delta_1 \leq 0 \vee \zeta_1 \leq 0) \\
&\quad - P\big(t_\delta^0 \geq t_\delta \wedge t_\zeta^0 \geq t_\zeta \,\big|\, \delta_1 \leq 0 \wedge \zeta_1 \leq 0\big) \times P(\delta_1 \wedge \zeta_1 \leq 0 \,|\, \delta_1 \leq 0 \vee \zeta_1 \leq 0) \\
&\leq P\big(t_\delta^0 \geq t_\delta \wedge t_\zeta^0 \geq t_\zeta \,\big|\, \delta_1 \leq 0\big) + P\big(t_\delta^0 \geq t_\delta \wedge t_\zeta^0 \geq t_\zeta \,\big|\, \zeta_1 \leq 0\big) \\
&\leq P\big(t_\delta^0 \geq t_\delta \wedge t_\zeta^0 \geq t_\zeta \,\big|\, \delta_1 = 0 \wedge \zeta_1 \to \infty\big) + P\big(t_\delta^0 \geq t_\delta \wedge t_\zeta^0 \geq t_\zeta \,\big|\, \delta_1 \to \infty \wedge \zeta_1 = 0\big) \\
&= P\big(t_\delta^0 \geq t_\delta \,\big|\, \delta_1 = 0\big) + P\big(t_\zeta^0 \geq t_\zeta \,\big|\, \zeta_1 = 0\big) \\
&= \frac{p_\delta + p_\zeta}{2}
\end{aligned}
$$

The average of the separate two-sided p-values from the OLS regressions is thus an upper bound for p-value of the joint hypothesis test.

Appendix 1.B Empirical Details

Table 1.B.1: Sample composition

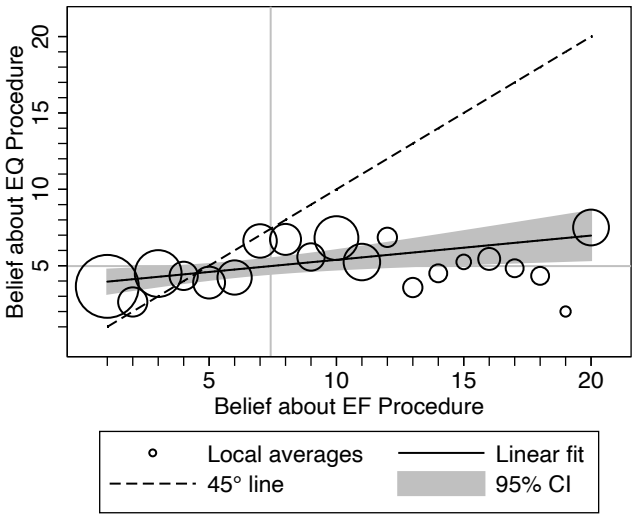
	Obs.	Mean	Median	Min.	Max.
Age	372	25.583	24	18	72
Female	369	0.599	1	0	1
University student	372	0.836	1	0	1
Income	372	741.185	600	0	3500
Log income	372	6.220	6.398595	0	8.160804

Notes: Log income is calculated as $\ln(income + 1)$

Table 1.B.2: Beliefs

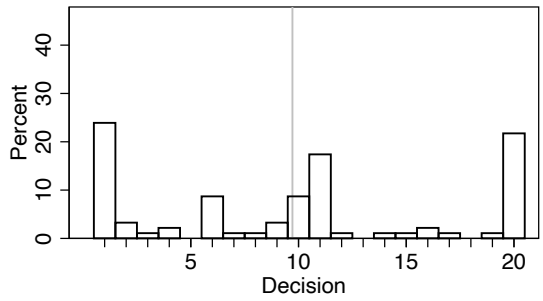
Dependent variable	*Belief about others' average decisions*			
Procedure	*EF*	*EQ*	*EF*	*EQ*
	(1)	(2)	(3)	(4)
Role A	1.849***		1.852***	0.231
	(0.585)		(0.584)	(0.510)
Role a		2.102***	0.639	2.103***
		(0.510)	(0.584)	(0.510)
Constant	6.497***	3.925***	6.176***	3.809***
	(0.390)	(0.299)	(0.457)	(0.396)
Observations	372	372	372	372
R^2	0.026	0.044	0.029	0.045

Notes: Heteroscedasticity-consistent standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

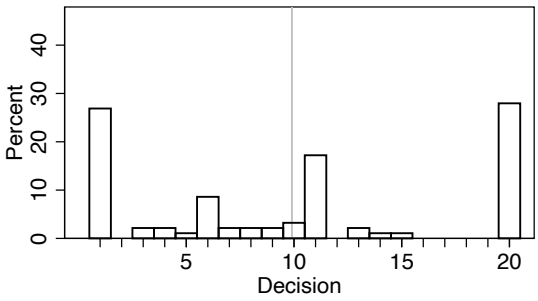


Notes: The figure groups subjects by their beliefs about others' average decisions for the EF Procedure. For each option on the horizontal axis, the figure plots the respective subjects' average belief about others' decisions for the EQ Procedure on the vertical axis. The sizes of circles correspond to the respective numbers of subjects. The dashed line indicates 45 degrees. The gray lines indicate the averages of beliefs about the EF Procedure (vertical) and the EQ Procedure (horizontal). The solid black line represents the linear fit from an OLS regression, and the shaded area around it corresponds to the 95% confidence interval based on heteroscedasticity-consistent standard errors.

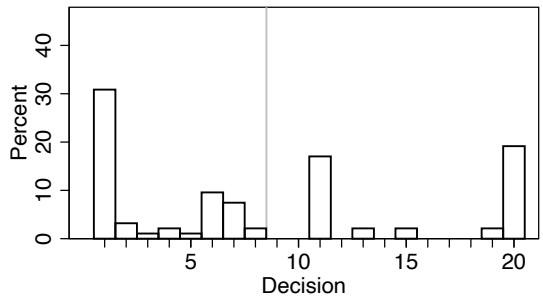
Figure 1.B.1: Relationship between the two predictions



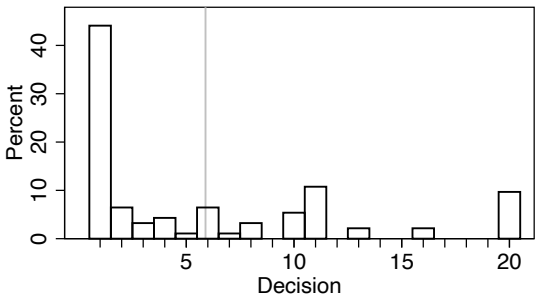
(a) Roles A and a



(b) Roles A and b



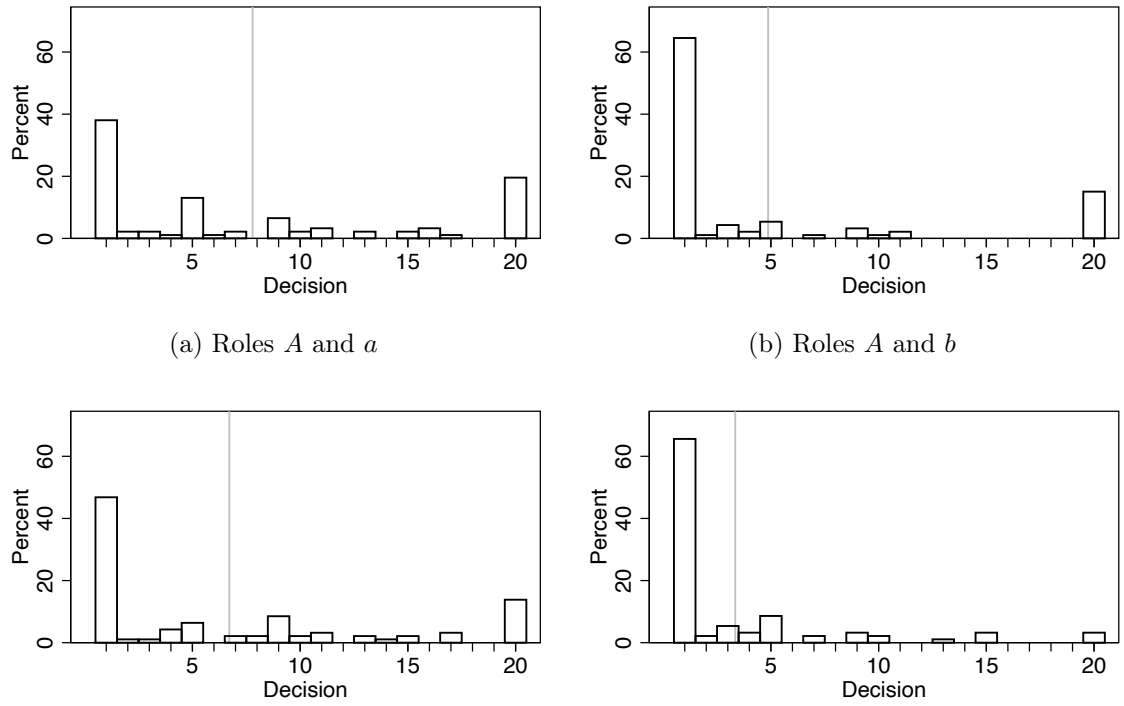
(c) Roles B and a



(d) Roles B and b

Notes: The gray lines indicate the respective average decisions.

Figure 1.B.2: Decisions for the EF Procedure by combinations of roles



(a) Roles A and a

(b) Roles A and b

(c) Roles B and a

(d) Roles B and b

Notes: The gray lines indicate the respective average decisions.

Figure 1.B.3: Decisions for the EQ Procedure by combinations of roles

Table 1.B.3: Nominal group bias in decisions

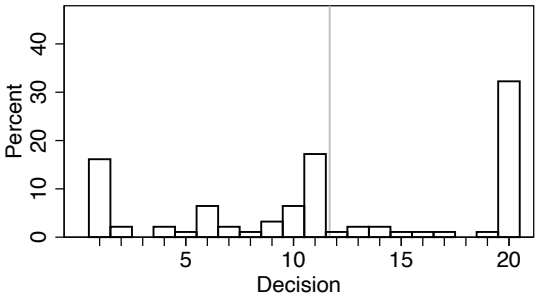
Dependent variable	*Decision*		*Belief*	
Procedure	*EF*	*EQ*	*EF*	*EQ*
	(1)	(2)	(3)	(4)
Same name is A	2.086***		0.958	
	(0.731)		(0.590)	
Same name is a		2.441***		1.715***
		(0.687)		(0.513)
Constant	7.454***	4.457***	6.935***	4.118***
	(0.503)	(0.437)	(0.417)	(0.300)
Observations	372	372	372	372
R^2	0.022	0.033	0.007	0.029

Notes: Heteroscedasticity-consistent standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

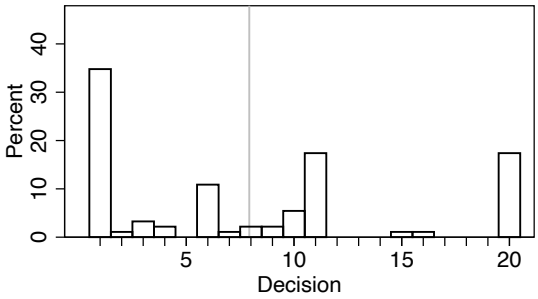
Table 1.B.4: Nominal group bias in decisions (with roles)

Dependent variable	*Decision*		*Belief*	
Procedure	*EF*	*EQ*	*EF*	*EQ*
	(1)	(2)	(3)	(4)
Same name is A	2.086***		0.958	
	(0.720)		(0.583)	
Same name is a		2.441***		1.715***
		(0.669)		(0.502)
Role A	2.602***		1.849***	
	(0.720)		(0.583)	
Role a		3.140***		2.102***
		(0.669)		(0.502)
Constant	6.160***	2.887***	6.016***	3.067***
	(0.617)	(0.439)	(0.500)	(0.367)
Observations	372	372	372	372
R^2	0.055	0.087	0.033	0.073

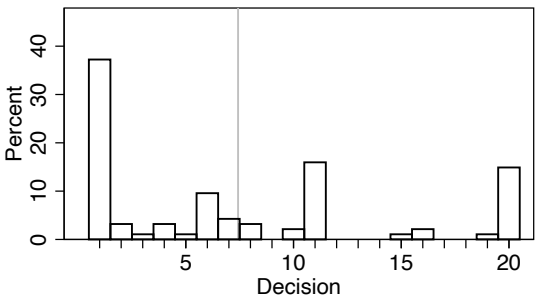
Notes: Heteroscedasticity-consistent standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.



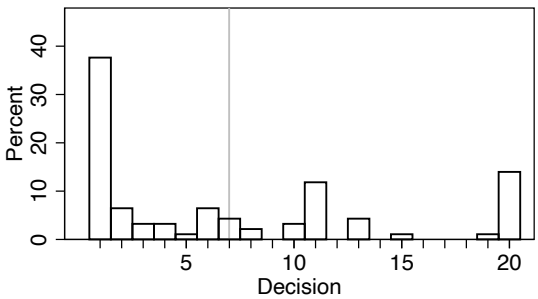
(a) Role A and player with same name is A



(b) Role A and player with same name is B



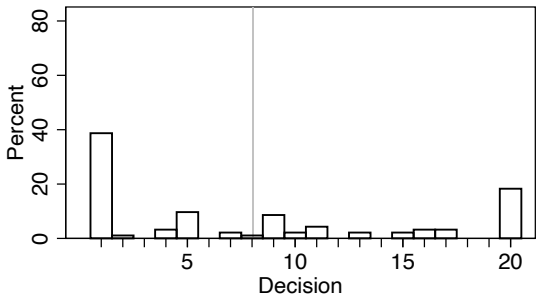
(c) Role B and player with same name is A



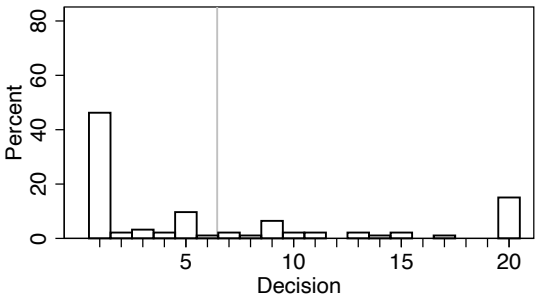
(d) Role B and player with same name is B

Notes: The gray lines indicate the respective average decisions.

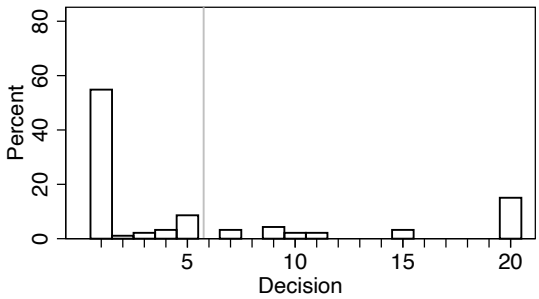
Figure 1.B.4: Nominal group bias in the EF Procedure



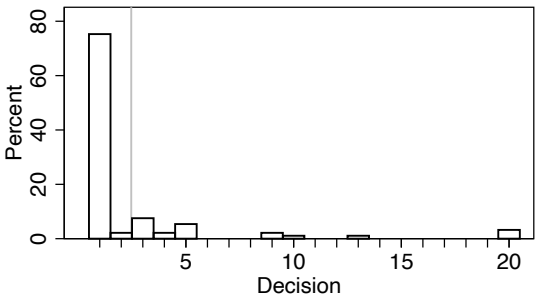
(a) Role a and player with same name is a



(b) Role a and player with same name is b



(c) Role b and player with same name is a



(d) Role b and player with same name is b

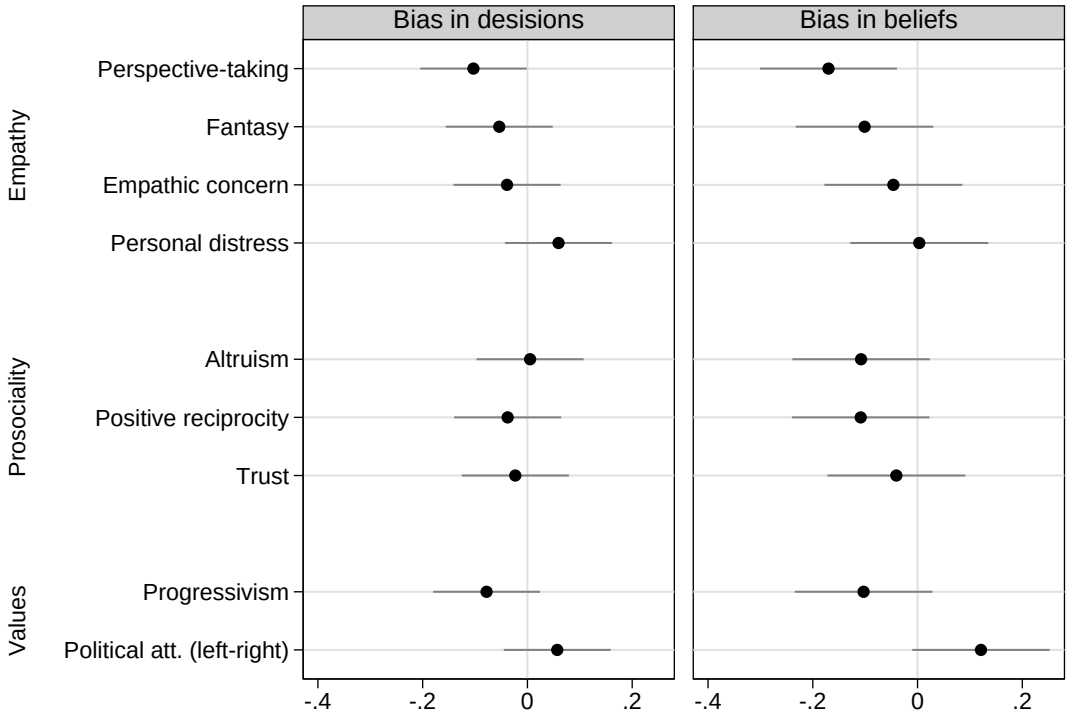
Notes: The gray lines indicate the respective average decisions.

Figure 1.B.5: Nominal group bias in the EQ Procedure

Table 1.B.5: Order effects

Dependent variable	*Decision*		*Belief*	
Procedure	*EF*	*EQ*	*EF*	*EQ*
	(1)	(2)	(3)	(4)
Role A	3.576***		2.502***	
	(1.006)		(0.825)	
Role $A \times EQ$ first	-1.946		-1.292	
	(1.453)		(1.166)	
Role a		2.603***		2.808***
		(1.001)		(0.712)
Role $a \times EQ$ first		1.107		-1.427
		(1.357)		(1.020)
EQ first	1.170	-1.048	0.885	0.727
	(0.992)	(0.848)	(0.778)	(0.598)
Constant	6.596***	4.615***	6.034***	3.573***
	(0.688)	(0.662)	(0.544)	(0.410)
Observations	372	372	372	372
R^2	0.038	0.058	0.030	0.049

Notes: Heteroscedasticity-consistent standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.



Notes: The figure shows the Pearson correlation coefficient for the ENA proxy introduced in Equation 1.7 and the respective survey measure. Gray bars indicate 95% confidence intervals.

Figure 1.B.6: Correlations with the *ENA* proxies (full sample)

Table 1.B.6: Heterogeneity (showing controls)

Dependent variable	*ENA proxy*			
Domain	*Decisions*		*Beliefs*	
	(1)	(2)	(3)	(4)
Perspective-taking	-0.176**	-0.185**	-0.187**	-0.180*
	(0.0754)	(0.0754)	(0.0923)	(0.0936)
Fantasy	-0.0999	-0.0886	-0.132	-0.136
	(0.0760)	(0.0786)	(0.0934)	(0.0978)
Empathic concern	-0.00779	-0.0109	0.103	0.119
	(0.0783)	(0.0792)	(0.107)	(0.107)
Personal distress	0.248***	0.261***	0.131	0.123
	(0.0712)	(0.0719)	(0.0986)	(0.0991)
Altruism	-0.00313	-0.00408	-0.161*	-0.166*
	(0.0713)	(0.0718)	(0.0876)	(0.0868)
Positive reciprocity	0.0400	0.0473	-0.0349	-0.0196
	(0.0601)	(0.0611)	(0.0800)	(0.0802)
Trust	-0.111*	-0.104*	-0.0961	-0.0877
	(0.0606)	(0.0618)	(0.0761)	(0.0768)
Risk taking	0.155**	0.159**	0.244***	0.253***
	(0.0605)	(0.0623)	(0.0868)	(0.0880)
Patience	0.0224	0.0292	0.0300	0.0334
	(0.0627)	(0.0661)	(0.0869)	(0.0914)
Negative reciprocity	-0.126**	-0.138**	-0.0700	-0.0920
	(0.0600)	(0.0612)	(0.0722)	(0.0753)
Openness	0.0302	0.0362	-0.0310	-0.0320
	(0.0593)	(0.0623)	(0.0834)	(0.0855)
Conscientiousness	0.142**	0.145**	0.0994	0.0794
	(0.0619)	(0.0644)	(0.0804)	(0.0842)
Extraversion	-0.0272	-0.0265	0.0721	0.0751
	(0.0609)	(0.0614)	(0.0777)	(0.0795)
Agreeableness	0.0772	0.0839	0.115	0.112
	(0.0733)	(0.0761)	(0.0856)	(0.0849)
Neuroticism	-0.0919	-0.0808	0.00478	0.0157
	(0.0721)	(0.0730)	(0.103)	(0.102)
Female		-0.127		-0.0135
		(0.128)		(0.175)
Other gender		-0.378		0.703
		(0.343)		(0.704)
Age		0.113		-0.0489
		(0.232)		(0.404)
Age^2		-0.000352		-0.000117
		(0.000394)		(0.000811)
University student		-0.0448		-0.130
		(0.0762)		(0.0923)
Log income		0.0340		0.000219
		(0.0601)		(0.0688)
Personality controls	Yes	Yes	Yes	Yes
Demographic controls	No	Yes	No	Yes
Observations	312	312	312	312
R^2	0.120	0.130	0.093	0.106

Notes: The table reports standardized coefficients. Heteroscedasticity-consistent standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The analysis excludes subjects above the 90^{th} percentile in the distribution of mistakes in the control questions and those whose self-reported reliability regarding the survey answers lies below the 10^{th} percentile, leaving 312 subjects. The *personality controls* are risk and time preferences along with the Big Five traits. The *demographic controls* are gender (categories: female, male, and diverse), age and squared age, a dummy for being enrolled at university, and the log of monthly gross income in euros, calculated as $\ln(income + 1)$.

Table 1.B.7: Heterogeneity (full sample)

Dependent variable	*ENA proxy*			
Domain	*Decisions*		*Beliefs*	
	(1)	(2)	(3)	(4)
Perspective-taking	-0.151**	-0.158**	-0.200**	-0.197**
	(0.0706)	(0.0709)	(0.0822)	(0.0827)
Fantasy	-0.0378	-0.0337	-0.0535	-0.0615
	(0.0665)	(0.0681)	(0.0842)	(0.0868)
Empathic concern	0.0276	0.0278	0.140	0.159
	(0.0769)	(0.0776)	(0.0998)	(0.100)
Personal distress	0.166**	0.170**	0.102	0.0926
	(0.0685)	(0.0694)	(0.0871)	(0.0877)
Altruism	0.00246	0.00503	-0.145*	-0.146*
	(0.0642)	(0.0646)	(0.0789)	(0.0779)
Positive reciprocity	-0.0183	-0.0141	-0.104	-0.0935
	(0.0527)	(0.0538)	(0.0786)	(0.0789)
Trust	-0.0570	-0.0541	-0.0570	-0.0517
	(0.0556)	(0.0561)	(0.0727)	(0.0731)
Personality controls	Yes	Yes	Yes	Yes
Demographic controls	No	Yes	No	Yes
Observations	372	372	372	372
R^2	0.078	0.083	0.071	0.079

Notes: The table reports standardized coefficients. Heteroscedasticity-consistent standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. The *personality controls* are risk and time preferences along with the Big Five traits. The *demographic controls* are gender (categories: female, male, and diverse), age and squared age, a dummy for being enrolled at university, and the log of monthly gross income in euros, calculated as $\ln(income + 1)$.

Appendix 1.C Instructions

The instructions have been translated from German. We present the wording for the following example: 24 subjects take part in the respective session, the given example subject has been allocated to Group 6 and has the name Player X, her roles are A and b, and the EF Procedure comes first. In the group over which she decides, i.e., Group 7, Player X's roles are A and a. All subjects saw identical instructions, except for changes regarding the above parameters.

Introduction

Welcome

Welcome, and thank you for participating in today's study! On the following pages, you will first receive information on data protection before entering your bank details. You need to accept the information on data protection to take part in this study. Additionally,

you need a bank account from the euro area. Please note that you are only allowed to participate in the study once.
In case you have questions during the study, you can contact the head of the study at any time via the following channels:

- via telephone: +49 *** ********
- via WhatsApp: +49 *** ********
- via email: ***********@uni-bonn.de

You can find all contact details in the email in which you received the link to this study. Please click "Continue" to proceed.

Information on Data Protection

[omitted]

Your Bank Details

[omitted]

Please Wait

Please wait for a moment. We will continue soon.

Estimation Task

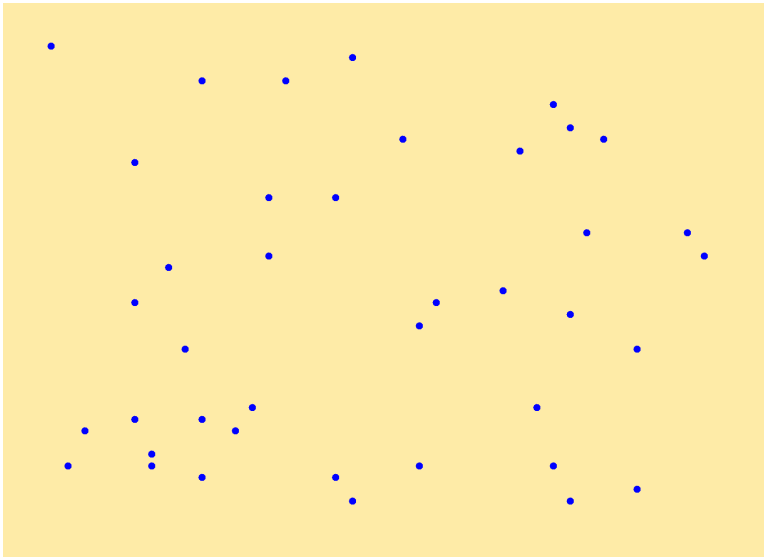
Estimation Task

For your participation in today's study, you receive a basic compensation of €4.00. Additional money can potentially be added during the study. Before the main part of the study begins, we have an estimation task for you.
After a countdown, we will show you a picture for two seconds. The picture has a yellow background and shows a certain number of blue dots. *It is your task to estimate the number of blue dots.*
After the picture has been shown to you, you have 15 seconds to enter your answer in the appropriate field and click "Confirm." The lower the distance between your estimate and the actual number of dots, the better. Every estimate is better than no estimate. The better your estimate, the higher are your chances for additional money.
You will first receive a test task for practicing. In the test task, you will see a different number of dots than in the real task, and your estimate will not have any consequences. Apart from this, the test task is exactly as the actual task. When you are ready to start with the test task, please click "Start."

[Countdown]

3—2—1

[Test Signal]



Test Estimate

Your remaining time for this page: 15—14—. . . —1

The number of dots is:

Are You Ready?

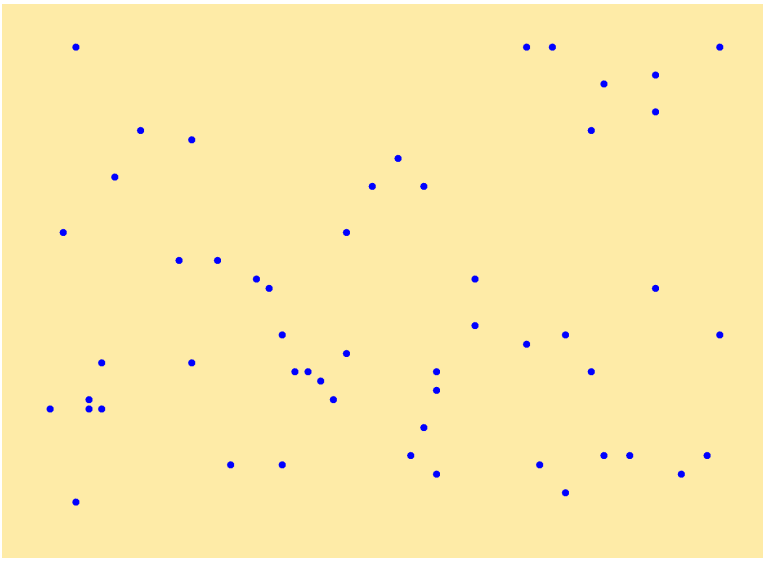
That was the test task. Thank you for your estimate!

You now know how the estimation task works. The real task is next. When you are ready, please click "Start."

[Countdown]

3—2—1

[Test Signal]



Your Estimate

Your remaining time for this page: 15—14—. . . —1
The number of dots is:

Situation

Your Situation: Overview

Thank you! We will later get back to the estimate.
All participants of the currently ongoing study have been grouped into twelve groups of two. The computer has determined the distribution randomly. The groups decide for other groups in a circle: Group 1 decides for Group 2, Group 2 decides for Group 3, and so on. Group 12 decides for Group 1. You yourself are *Player X* in *Group 6*. The other person in the group is *Player Y*.
In every group, points will be distributed between Player X and Player Y, which will later be converted into money. Players can earn up to 1,000 points. 100 points correspond to one euro. Therefore, you will additionally receive between €0 and €10 at the end of the study.

- The distribution of points in your group is decided by the players of Group 5.
- You and Player Y from your group decide over the distribution of points in Group 7.
- You yourself cannot influence the distribution of points in your own group in any way.

The distribution of points takes place twice, once according to *Procedure 1* and once accord-

ing to *Procedure 2*. Which of the two procedures and which decision will be implemented at the end of the study will be decided randomly by the computer. Details follow later. First, we will explain the possible payments for you and Player Y from your group in detail.

Your Situation: Procedure 1

Please read carefully how your payment works according to Procedure 1.
In Procedure 1, there are different possible combinations of points that you and Player Y can receive. In total, you and Player Y from your group can receive between 400 and 1,000 points. You yourself can receive between 200 and 990 points.

- For some options, there are more points for you and Player Y in total, in others less.
- The more points you receive, the fewer Player Y receives.
- The more points you and Player Y receive in total, the more points you receive.

There are 20 options for how the players from Group 5 can distribute the points between you and Player Y from your group.

Option #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Player X	200	300	385	460	525	585	640	690	735	775	811	843	871	896	918	937	953	967	979	990
Player Y	200	190	180	170	160	150	140	130	120	110	100	90	80	70	60	50	40	30	20	10
Total	400	490	565	630	685	735	780	820	855	885	911	933	951	966	978	987	993	997	999	1,000

Every column of the table shows one possible combination of points for you (Player X) and Player Y from your group. The farther right the column, the more you (Player X) receive, and the more you and Player Y receive in total.
Examples

- Option 1 means that you (Player X) receive 200 points, and Player Y from your group also receives 200 points.
- Option 20 means that you (Player X) receive 990 points, and Player Y from your group receives 10 points.

Your Situation: Procedure 2

Please read carefully how your payment works according to Procedure 2.
Points for your group
At the beginning of the study, you answered an estimation task. All other participants answered the same estimate task. At the end of the study, one player from your group will be compared to a player from Group 12—with which your group is otherwise not concerned. The player from your group that has randomly been selected for the comparison is Player Y.

- If Player Y from your group has estimated more accurately than the player from Group 12, your group receives 1,000 points.
- If Player Y from your group has estimated less accurately than the player from Group 12, your group receives no points.
- Your own estimate has no consequences for your group.

"To estimate more accurately" means that the estimate by a player deviates less from the true amount of dots than the estimate by another player. In case the estimates of the two players are the same, the winner is determined randomly. How many points Group 12 receives will be determined by a different comparison of estimates.
How many points your group receives in total, therefore, depends on Player Y from your group but not on you. You will learn the result later on.

Distribution of points

In case Player Y retains the 1,000 points for your group, the distribution of the points has to be decided. There are 20 options for how the players from Group 5 can distribute the points between you and Player Y from your group.

Option #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Player X	500	475	450	425	400	375	350	325	300	275	250	225	200	175	150	125	100	75	50	25
Player Y	500	525	550	575	600	625	650	675	700	725	750	775	800	825	850	875	900	925	950	975

Every column of the table shows one possible distribution of points between you (Player X) and Player Y from your group. The farther right the column, the more receives the player who retained the points for your group with their estimate (Player Y). The other person (you) receives less accordingly. The total number of points is always the same.

Examples

- Option 1 means that Player Y from your group receives 500 points, and you (Player X) also receive 500 points.
- Option 20 means that Player Y from your group receives 975 points, and you (Player X) receive 25 points.

Your Situation: Comprehension Questions

To make sure that you understood everything correctly about your payoff, please answer the following comprehension questions. To re-read the instructions, please click on the respective tab.

Which player are you in your group? (*Player X*, Player Y)
In which group are you? (*6*)
For the players from which group do you decide? (*7*)
The players from which group decide about your payoff? (*5*)

In which group is the player whose estimate is compared to the estimate by Player Y from your group for Procedure 2? (*12*)

How many points do you receive according to Procedure 1? (*minimum: 200, maximum: 990*)

How many points does Player Y from your group receive according to Procedure 1? (*minimum: 10, maximum: 200*)

How are the points distributed according to Procedure 1 if option 4 is chosen for your group? (*you receive: 200, Player Y from your group receives: 170*)

On which of the two estimates from you (Player X) and Player Y from your group does it depend how many points your group receives in total according to Procedure 2? (only on my estimate, *only on the estimate from Player Y from my group*, on my estimate and the estimate from Player Y from my group)

Please also consider the possibility that Player Y might have failed to retain the 1,000 points for your group. In this case, your group receives 0 points in total.

How many points do you receive according to Procedure 2? (*minimum: 0, maximum: 500*)

How many points does Player Y from your group receive according to Procedure 2? (*minimum: 0, maximum: 975*)

Please assume for the following question that Player Y retained the 1,000 points for your group.

How are the points distributed according to Procedure 2 if option 16 is chosen for your group? (*you receive: 125, Player Y from your group receives: 875*)

Your Situation: Comprehension Questions

You have answered all questions correctly! You can now look at the correct answers again. Afterward, please click "Continue."

Task

Your Task: Overview

Thank you for answering the comprehension questions.

You now know the possible payoffs that you and Player Y from your group can receive in this study. Which option is possibly implemented for your group is not decided by you. Instead, you depend on the decisions of the players from Group 5.

Reversely, you and Player Y from your group will make decisions that will possibly be implemented for the players from Group 7. You will make the same types of decisions for Group 7 that the players from Group 5 make for your group.

- In Procedure 1, Player X from Group 7 is in the same position as you are in your group.
- In Procedure 2, Player Y from Group 7 is in the same position as you are in your group.

The other player is in the position of Player Y in your group.
On the following two pages, we will explain exactly which decision you have to make and which options you have.

Your Task: Procedure 1

Please carefully read how the decision for Procedure 1 works that you, as a player from Group 6, make for the players from Group 7.
For Group 7, there are different possible combinations of points that Player X and Player Y from this group can receive in Procedure 1. In total, Player X and Player Y can receive between 400 and 1,000 points. Player X can receive between 200 and 990 points. Player Y can receive between 10 and 200 points.

- For some options, there are more points for Player X and Player Y in total, in others less.
- The more points Player X receives, the fewer receives Player Y.
- The more points Player X and Player Y receive in total, the *more* points Player X receives and the *fewer* points Player Y receives.

There are 20 options for distributing the points between Player X and Player Y from Group 7.

Option #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Player X	200	300	385	460	525	585	640	690	735	775	811	843	871	896	918	937	953	967	979	990
Player Y	200	190	180	170	160	150	140	130	120	110	100	90	80	70	60	50	40	30	20	10
Total	400	490	565	630	685	735	780	820	855	885	911	933	951	966	978	987	993	997	999	1,000

Every column of the table shows one possible combination of points for Player X and Player Y from Group 7. The farther right the column, the more Player X and Player Y receive in total, and the larger the difference between the number of points that Player X and Player Y receive becomes.

Examples

- Option 1 means that Player X receives 200 points, and Player Y also receives 200 points.

- Option 20 means that Player X receives 990 points, and Player Y receives 10 points.

Your Task: Procedure 2

Please carefully read how the decision for Procedure 2 works that you, as a player from Group 6, make for the players from Group 7.

Points for Group 7

At the beginning of the study, the players from Group 7 answered an estimation task. All other participants answered the same estimate task. One player from Group 7 will be compared to a player from Group 1—with which the group is otherwise not concerned—at the end of the study. The player from Group 7 that has randomly been selected for the comparison is Player X.

- If Player X from Group 7 has estimated more accurately than the player from Group 1, the group receives 1,000 points.
- If Player X from Group 7 has estimated less accurately than the player from Group 1, the group receives no points.
- The estimate by Player Y from Group 7 has no consequences for the group.

"To estimate more accurately" means that the estimate by a player deviates less from the true amount of dots than the estimate by another player. In case the estimates of the two players are the same, the winner is determined randomly. How many points Group 1 receives will be determined by a different comparison of estimates.

How many points Group 7 receives in total, therefore, depends on Player X from the group but not on Player Y.

Distribution of points

In case Player X retains the 1,000 points for the group, the distribution of the points has to be decided. There are 20 options for distributing the points between Player X and Player Y from Group 7.

Option #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Player X	500	525	550	575	600	625	650	675	700	725	750	775	800	825	850	875	900	925	950	975
Player Y	500	475	450	425	400	375	350	325	300	275	250	225	200	175	150	125	100	75	50	25

Every column of the table shows one possible distribution of points between Player X and Player Y from Group 7. The farther right the column, the more the player who retained the points for the group with his or her estimate (Player X) receives. The other person (Player Y) receives accordingly less. The total number of points is always the same.

Examples

- Option 1 means that in Group 7, Player X receives 500 points, and Player Y also receives 500 points.

- Option 20 means that in Group 7, Player X receives 975 points, and Player Y receives 25 points.

Your Task: Comprehension Questions

To make sure that you understood everything correctly about your task, please answer the following comprehension questions. To re-read the instructions, please click the respective tab.
How many points does Player X from Group 7 receive according to Procedure 1? (*minimum: 200, maximum: 990*)
How many points does Player Y from Group 7 receive according to Procedure 1? (*minimum: 10, maximum: 200*)
How are the points distributed according to Procedure 1 if you choose option 15 for Group 7? (*Player X from Group 7 receives: 918, Player Y from Group 7 receives: 60*)
On which of the two estimates from the players of Group 7 does it depend how many points the group receives in total according to Procedure 2? (*only on the estimate by player X*, only on the estimate from Player Y, on the estimates from Player X and Player Y)
Please also consider the possibility that Player X from Group 7 might have failed to retain the 1,000 points for the group. In this case, Group 7 receives 0 points in total.
How many points does Player X from Group 7 receive according to Procedure 2? (*minimum: 0, maximum: 975*)
How many points does Player Y from Group 7 receive according to Procedure 2? (*minimum: 0, maximum: 500*)
Please assume for the following question that Player X from Group 7 retained the 1,000 points for the group.
How are the points distributed according to Procedure 2 if you choose Option 5 for Group 7? (*Player X from Group 7 receives: 600, Player Y from Group 7 receives: 400*)

Your Task: Comprehension Questions

You have answered all questions correctly! You can now look at the correct answers again. Afterward, Please click "Continue."

Implementation of the Decisions

Implementation of the Decisions

Thank you for answering the comprehension questions!
At the end of the study, the computer will randomly decide which decisions will actually be implemented.
The exact selection of the decisions happens according to the following three steps:

1. For all groups jointly, one of the two procedures will be selected: Procedure 1 or Procedure 2

2. Among the groups ordered in a circle (1, 2, . . . , 12, 1, 2, . . .), every other group will be selected to make a decision: all groups with even numbers or all groups with uneven numbers.

3. Within each selected group, one player will be randomly determined, whose decision will be implemented for the respective following group.

For the points and consequences of the decisions in your group, therefore exist the following possibilities:

50% – One decision of Group 5 will be implemented for your group.

- If Procedure 1 is selected, you receive between 200 and 990 points and Player Y between 10 and 200 points.
- If Procedure 2 is selected and Player Y. . .
  - did not retain the 1,000 points for your group, you and Player Y from your group both receive 0 points.
  - retained the 1,000 points for your group, you receive between 25 and 500 points, and Player Y receives between 500 and 975 points.

25% – One decision from you will be implemented for Group 7.

- You receive 1,000 points, independent of all decisions made during today's study.

25% – No decision of you or made for you will be implemented.

- Your points will be determined by a different task, independent of your decisions.

Therefore, please keep in mind: All of your decisions can determine other players' payoffs. At the end of today's study, you will learn – apart from your payoff – how exactly your points were determined. All participants stay completely anonymous.

Comprehension Questions

To make sure that you understood everything correctly, please answer the following comprehension questions. To re-read the instructions, please click on the respective tab.
How many points will you receive if one of your decisions is implemented for the players of Group 7? (*1,000*)
Which of the two players can receive more points according to Procedure 1? (*In your group: Player X (you)*, Player Y; *in Group 7 (for which you decide): Player X*, Player Y)
Which player's estimate does it depend on how many points the respective group receives in total according to Procedure 2? (*In your group:* estimate by Player X (you), *on the*

estimate by Player Y; *in Group 7 (for which you decide): on the estimate by Player X*, on the estimate by Player Y)
Which of the two players can receive more points according to Procedure 2? (*In your Group:* Player X (you), *Player Y*; *in Group 7 (for which you decide): Player X*, Player Y)

Comprehension Questions

You have answered all questions correctly! You can now look at the correct answers again. Afterward, please click "Continue."

Decisions

Your Decision: Overview

Thank you for answering the comprehension questions!
You now have all the necessary information to make your decisions four Group 7. We start with Procedure 1. Your decision for Procedure 2 will follow directly afterward.

Your Decision: Procedure 1

Please make your decision for Group 7 for Procedure 1.
The more points Player X and Player Y receive *in total*, the larger the difference between the individual numbers of points becomes.

Option #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Player X	200	300	385	460	525	585	640	690	735	775	811	843	871	896	918	937	953	967	979	990
Player Y	200	190	180	170	160	150	140	130	120	110	100	90	80	70	60	50	40	30	20	10
Total	400	490	565	630	685	735	780	820	855	885	911	933	951	966	978	987	993	997	999	1,000
Decision	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Your Decision: Procedure 2

Please make your decision for Group 7 for Procedure 2.
Reminder: Your decision is only relevant if *Player X* won the 1,000 points for Group 7.

Option #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Player X	500	525	550	575	600	625	650	675	700	725	750	775	800	825	850	875	900	925	950	975
Player Y	500	475	450	425	400	375	350	325	300	275	250	225	200	175	150	125	100	75	50	25
Decision	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Predictions

Your Prediction: Overview

Thank you for making your decisions!
After you have now made your decisions for Group 7, the next part will deal with how players from other groups have made comparable decisions. We are interested in your prediction regarding the following question:

Which options have other players on average chosen for groups such as Group 7?
By enquoteaverage, we mean the arithmetic mean of the chosen options 1 to 20. For example, if half of all participants had chosen Option 5, and the other half had chosen Option 15, the average would be 10 (the example has been selected randomly). We will ask for your prediction twice, once for Procedure 1 and once for Procedure 2.
Please remember the possibilities for your payment:

- With a probability of 50%, a decision from Group 5 for your group will be implemented.
- With a probability of 25% a, decision from you for Group 7 will be implemented.

With the remaining probability of 25%, you receive points for a good prediction. For this, one of your two predictions will be selected randomly and will then be compared to the actual decisions of players *from other groups* (i.e., not with your own decisions or the decisions from Player Y from your group). All considered decisions were made for groups composed like Group 7 regarding the players' situations. The actually chosen options between 1 and 20 will be taken, and the (rounded) mean will be calculated. The closer your prediction is to the actual average, the more points you receive.

If the average of the actually chosen options...

- matches your prediction exactly, you receive 500 points.
- belongs to the four other options that are closest to your prediction, you receive 250 points.
- is further away from your prediction than what was mentioned above, you receive no points.

Your Prediction: Procedure 1

Please enter your prediction for Procedure 1.
Which option have other players chosen on average for groups such as Group 7?

Option #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Player X	200	300	385	460	525	585	640	690	735	775	811	843	871	896	918	937	953	967	979	990
Player Y	200	190	180	170	160	150	140	130	120	110	100	90	80	70	60	50	40	30	20	10
Total	400	490	565	630	685	735	780	820	855	885	911	933	951	966	978	987	993	997	999	1,000
Prediction	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Your Prediction: Procedure 2

Please enter your prediction for Procedure 2.
Which option have other players chosen on average for groups such as Group 7?

Option #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Player X	500	525	550	575	600	625	650	675	700	725	750	775	800	825	850	875	900	925	950	975
Player Y	500	475	450	425	400	375	350	325	300	275	250	225	200	175	150	125	100	75	50	25
Prediction	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Questionnaires

Thank you for your prediction! Please answer a few more questions on the following screens.

How old are you?

What is your gender? (female, male, diverse)

Are you a university student? (yes, no)

If yes: to which area does your major belong?

Have you ever been enrolled in economics? (yes, no)

How much money (in €) do you have available each month after all costs have been subtracted?

We will now ask you for your willingness to behave a certain way. Please use a scale from 0 to 10. 0 means "not at all" and 10 "completely." You can use any number between 0 and 10 to note where you are on the scale.

[Economic Preferences (qualitative questions of the Preference Survey Module)]

From: Armin Falk et al. 2016. *The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences*. IZA Discussion Paper 9674. Bonn: Institute for the Study of Labor

[Big Five (BFI-S)]

From: Jean-Yves Gerlitz and Jürgen Schupp. 2005. *Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. Dokumentation der Instrumententwicklung BFI-S auf Basis des SOEP-Pretests 2005*. Research Notes 4. Berlin, Germany: Deutsches Institut für Wirtschaftsforschung (DIW)

[Interpersonal Reactivity Index]

From: Mark H. Davis. 1980. "A Multidimensional Approach to Individual Differences in Empathy". *JSAS Catalog of Selected Documents in Psychology* 10:85

Other

In politics, it is often talked about "left" and "right." Where would you locate yourself on a scale from 1 for left and 10 for right?

From: European Social Survey. 2014. *ESS Round 7 Source Questionnaire*. ESS ERIC Headquarters, Centre for Comparative Social Surveys, City University London, London, United Kingdom

A tram is out of control and is about to hit five people. By changing a track switch, the tram can be rerouted to another track. Unfortunately, there is another person on that track. Should it be allowed to accept the death of one person (by changing the track) to save the lives of five people? (yes, no)

[`https://de.wikipedia.org/wiki/Trolley-Problem`; accessed in May 2020]

From: Philippa Foot. 1967. "The Problem of Abortion and the Doctrine of the Double Effect". *Oxford Review* 5:5–15
How much can we rely on your answers in the questionnaires? (1: not at all, 10: completely)

Open Questions

How did you make your decision for Procedure 1?
How did you make your decision for Procedure 2?

Chapter 2

Motivated by Others' Preferences? An Experiment on Imperfect Empathy*

Joint work with Jana Hofmeier

Abstract

People care about others. But how do they assess the utility of others when making other-regarding decisions? Do they apply their own preferences, or do they adopt the preferences of the other person? We study this question in a laboratory experiment where subjects in the role of senders can pay money to avoid harm arising to receivers. In a first step, we elicit all subjects' willingness to pay (WTP) for not having to eat food items containing dried insects. We then show senders the WTPs of receivers and repeat the elicitation procedure, but now with receivers having to eat the food items and senders stating their WTPs to spare the receivers from having to eat them. We find that not only receivers' preferences matter for decisions but also senders' own preferences, a phenomenon for which we use the term imperfect empathy. In motivating prosocial transfers, senders' and receivers' WTPs act as complements by reinforcing each other. Conversely, pairs of sender and receiver who are dissimilar generate lower transfers than others. Since transfers usually benefit receivers more than they cost senders, we also find that dissimilarity within pairs reduces welfare. Our results complement the extensive literature on prosocial preferences, which so far abstracts from heterogeneous valuations. The implications might be far-reaching. For example, systematic differences in consumption preferences between net payers and recipients could undermine public support for public welfare systems.

*We thank Thomas Dohmen, Armin Falk, Lorenz Götte, and participants of the CRC TR 224 Conference in Mainz for helpful comments. The study was approved by ethics committee of the University of Bonn (reference number: 174/18). Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 (Project A01) is gratefully acknowledged.

> As we have no immediate experience of what other men feel, we can form no idea of the manner in which they are affected, but by conceiving what we ourselves should feel in the like situation. Though our brother is upon the rack, as long as we ourselves are at our ease, our senses will never inform us of what he suffers. They never did, and never can, carry us beyond our own person, and it is by the imagination only that we can form any conception of what are his sensations.
>
> —ADAM SMITH, *THE THEORY OF MORAL SENTIMENTS*[1]

2.1 Introduction

It is widely documented that people consider others when making decisions: They donate to charities, give blood, or volunteer. These behaviors have often been attributed to social preferences such as altruism (Becker, 1974, 1976), warm glow (Andreoni, 1990), inequity aversion (Fehr and Schmidt, 1999), or reciprocity (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). All of these models have in common that the payoff of others is explicitly incorporated into an agent's utility function. Since we are often interested in the distribution of monetary payoffs, the assumption that people know what is good for others is very plausible. But—given that preferences on goods are not homogeneous—the question arises how these other persons' hedonic benefits, which are not experienced by a given agent herself, are transformed into motives for personal prosocial behavior. It has been claimed that empathy, the ability to feel into others' emotions, is playing a central role. According to the empathy–altruism hypothesis, altruistic motivation arises from empathy felt for a person in need (Batson, 1987), and it has been shown empirically that induced empathy indeed increases prosocial behavior (Coke, Batson, and McDavis, 1978; Klimecki et al., 2016) and cooperation (Batson and Moran, 1999). However, the ability to sympathize with others' emotions is limited. We find evidence that people behave imperfectly empathic: They judge consequences for others not only by the utility that the other person attaches to it but also by their own preferences.

In this chapter, we show that in order to make a monetary transfer to help another person receive a specific good, two requirements have to be fulfilled: First, the receiver of the good needs to show a preference for the transferred good, and second, the sender needs to have a preference for the good as well. This means that people do not only care about the utility that a prosocial action entails for the other person but also which utility they themselves attach to it. We call this type of behavior imperfect empathy (see also Bisin and Verdier, 2001), since people do not only use the other person's preferences to evaluate their actions' consequences on them (perfectly empathic behavior) but also take into account their own preferences.

We run a laboratory experiment in which participants can make prosocial monetary transfers to help other participants. The aim is to find out to which degree people are guided by their own rather than by the receivers' preferences when acting prosocially. Since

[1] Smith, 1859, Part I, Section I, Chapter I.

our interest lies in the emotional accessibility of others' sensations, we use an experimental setup that cleanly isolates such experiences. Following Ambuehl (2017), we let subjects make choices about eating food items that might provoke feelings of disgust, namely dried insects and worms. These "bads" have several important and useful features. First, people have preferences about the consumption of such items. Second, disgust markedly varies between individuals and across items. Third, rational arguments have no power in arguing what is "more disgusting" among the items, e.g., a cricket or a worm. And fourth, we ask people to eat the items and thereby have tight control over consumption.

In Part 1 of the experiment, participants decide how much money between €0 and €20 they themselves would be willing to spend to avoid having to eat several different dried insects. They can decide between receiving a lower payoff and not eating the insect or receiving a high payoff and eating the insect. In Part 2, participants receive information on how much eight other subjects (*receivers*) would each be willing to pay to avoid the insects in the first part. The participants (as *senders*) then decide how much money between €0 and €20 they would be willing to spend on sparing these other subjects from having to eat the dried insects. They can decide between receiving a lower payoff and the receiving subject not having to eat the insect or receiving a high payoff and the receiving subject having to eat the insect. We show that not only the receiver's willingness to pay for avoiding an item has a positive effect on the respective transfer but also the sender's own WTP, and—in particular—the interaction of the two. Calculating the distance between the vectors of subjects' WTPs, we can also show that dissimilarity between senders and receivers decreases expected transfers. Defining welfare as the sum of individual utility from personal consumption, we can further show that dissimilarity reduces welfare. In the last part of the experiment, subjects have the option to alter decisions that others have made for themselves, which gives us a measure of paternalism. We show that imperfect empathy is prevalent among both libertarians and paternalists.

Our chapter makes contributions to the literature on the role of empathy in generating prosocial behavior. It confirms the hypothesis that empathy is a driver of prosocial behavior, where a lack of empathy can decrease the extent of prosocial actions and lead to misallocation of help from the receiving party's perspective. Our results also inform models featuring altruism in conjunction with heterogeneous preferences like they are present in the literature on the intergenerational transmission of preferences (Bisin and Verdier, 2001; Doepke and Zilibotti, 2017). Our finding could furthermore be an additional explanation for in-group–out-group bias that might exist, e.g., along the lines of political affiliation (see Fowler and Kam, 2007). While this bias might even exist between groups defined by attributes that are arbitrary (Tajfel, Billig, and Bundy, 1971), it could also be that people have more similar preferences amongst their in-groups. In this case, imperfect empathy could explain why prosocial behavior is stronger towards members of in-groups than towards members of out-groups. Imperfect empathy is also in line with the literature on the false consensus effect (Ross, Greene, and House, 1977), a bias in which people commonly think that their own preferences and choices are relatively more common than other preferences and other choices. A potential implication of imperfect empathy is that het-

erogeneous preferences reduce the support for redistribution and lower expected welfare. It could therefore be an explanation for the finding that diversity has a negative effect on redistribution and donations (Dahlberg, Edmark, and Lundqvist, 2012; Andreoni et al., 2016) and is thus meaningful from a policy perspective in ever more diverse societies.

The remainder of the chapter is structured as follows. Section 2.2 presents a simple theoretical framework and derives our hypotheses. Section 2.3 describes the laboratory experiment. Section 2.4 presents the results on the aggregate level, on the level of individuals, and distinguishing between libertarians and paternalists. Finally, Section 2.5 summarizes and discusses the results.

2.2 Theory and Hypotheses

We develop a simple theoretical model in which agents derive utility from their own consumption as well as from another person's consumption. When evaluating the other person's consumption, agents use a combination of their own and of the other person's preferences. We use the model to formally derive our hypotheses regarding imperfectly empathic behavior and the consequences arising from dissimilar preferences for the size of transfers and for the overall welfare.

Individual i experiences utility from good x_i and disutility from "bad" y_i; individual j experiences utility from good x_j and disutility from bad y_j. Utilities or disutilities are evaluated by utility functions which are specific to the combinations of individuals and domains. In computing overall utility, consumption value from goods enters additively, while disutility from bad experiences is subtracted. We use money as the numéraire. Therefore, utility from money is simply given by the particular nominal amount of currency.[2] If no consumption takes place, we assume that utility is zero. Individuals receive utility not only from their own consumption but also from the other person's consumption. The total utility of subject i is given by the following expression:

$$U_i(x_i, y_i; x_j, y_j) = u_i(x_i) - v_i(y_i) + \alpha \left(u_j(x_j)^{\beta}\, u_i(x_j)^{1-\beta} - v_j(y_j)^{\beta}\, v_i(y_j)^{1-\beta} \right) \qquad (2.1)$$

The first part of overall utility, $u_i(x_i) - v_i(y_i)$, is utility and disutility derived from i's own consumption. The remaining term is the utility that individual i derives from the other individual j's consumption. The general extent to which i cares about j is determined by her level of altruism α. When evaluating j's utility in a particular domain, i partially relies on both her own relevant utility function and on j's utility function in the respective domain. The degree of reliance on j's preferences is captured by the empathy parameter $\beta \in [0, 1]$. If β is zero, i simply projects her own preferences upon j. If β is one, she fully adopts j's preferences and disregards her own.

The notation can, of course, be extended to further consumption items. We assume in the model above that subjective valuations are complements in generating vicarious

[2] Appendix 2.B provides empirical support for the assumption of linear utility from money in our context.

(dis-)utility by modeling them multiplicatively, while other authors have assumed perfect substitutability (see, e.g., Bisin and Verdier, 2001). Our assumption means that, in order to enjoy someone else's consumption, both the sender and the receiver have to attach utility to the consumed good, or—conversely—they both have to attach disutility to a particular experience to feel that it is bad. The complementarity of assessments gives rise to additional predictions for our experiment, which we develop below and later also test.

2.2.1 Transfer Decisions

We now apply the utility function in Equation 2.1 to decisions about prosocial transfers in our experiment. In the experiment, subjects receive money, which corresponds to good x above, and potentially eat food items, corresponding to bad y. A sender can decide between making a monetary transfer and a receiver not having to eat a food item and not making a monetary transfer and a receiver having to eat a food item. Sender i never has to consume any food item herself, i.e., $v_i(y_i) = 0$, and receiver j always gets a monetary payoff of €20, i.e., $u_j(x_j) = 20$. The sender can now decide to make a monetary transfer $t \in [0, 20]$ so that the receiver does not have to consume item $k \in K$. If the potential transfer of $t \in [0, 20]$ (we abstract from discreteness of choice options) for item k is accepted by the sender, the implied monetary payoff for herself is given by $x_i = 20 - t$ and the receiver does not have to eat, i.e., $v_j(y_j) = 0$ and also $v_i(y_j) = 0$. If she rejects, her payoff is $x_i = 20$ and the receiver has to eat item k, i.e., $y_j = k$. For a transfer to be made, it has to hold that the utility for the sender when making the transfer (the expression on the left-hand side of the equation below) is as least as high as the utility when not making the transfer (the expression on the right-hand side).

$$20 - t + \alpha\, 20 \geq 20 + \alpha \left(20 - v_j(k)^{\beta}\, v_i(k)^{1-\beta}\right)$$

The highest proposed transfer that a sender still accepts, $t^{\star}$ (later simply *transfer*), is therefore given by

$$t^{\star} = \alpha\, v_j(k)^{\beta}\, v_i(k)^{1-\beta} \tag{2.2}$$

Our key hypothesis about decision-making can now be formulated directly in terms of the model parameter β.

Hypothesis 2.1. *People typically exhibit imperfect empathy: transfer decisions depend not only on receivers' preferences but also on senders' own preferences. Formally,* $\beta \in (0, 1)$.

The above hypothesis can directly be tested by estimating the parameter β on the individual level. Moreover, if the hypothesis was true, the partial derivatives of $t^{\star}$ with respect to *both* agents' valuations would be positive, as would be the cross partial derivative. This prediction thus lends itself to reduced-form testing on the level of the subject population using OLS. We expect transfers to depend positively on both the respective sender's and the receiver's valuations and—in particular—on their interaction.

2.2.2 Welfare

In the next step, we theoretically derive predictions about the effect of dissimilarity in preferences between senders and receivers on the size of transfers and on overall welfare. The welfare criterion that we employ is simply the sum of individual utilities from personal consumption.

$$\text{Welfare} \equiv u_i(x_i) - v_i(y_i) + u_j(x_j) - v_j(y_j) \tag{2.3}$$

We predict dissimilarity to decrease welfare through two channels: The size of transfers and the targeting of transfers. The first channel is based on the premise that transfers are, on average, too low from a planner's perspective. This simply follows from the fact that the planner weighs individuals' welfare equally, while people usually care more about themselves than about others, i.e., α is smaller than one. As we show below, dissimilarity in preferences further decreases the size of transfers and thereby amplifies the welfare loss.

To understand the effect of dissimilarity on the size of transfers, consider two subjects, i and j, behaving in accordance with our model and sharing the same parameter values for α and β. We denote their respective individual valuations of some item by $v_i(k) \equiv v_i$ and $v_j(k) \equiv v_j$, and we fix the total level of the two subjects' valuations of items such that $v_i + v_j \equiv \bar{v}$. Both subjects are with equal probability of $1/2$ either sender or receiver. We further assume that $v_i \geq v_j$. This allows us to express the valuations of subjects in terms of the total valuation of both subjects and the distance between the individual valuations: $v_i = \frac{\bar{v}+|v_i-v_j|}{2}$ and $v_j = \frac{\bar{v}-|v_i-v_j|}{2}$. Plugging into Equation 2.2, we can calculate the expected maximum transfer that this pair of subjects generates.

$$\mathbb{E}[t^\star] = \frac{\alpha}{4}\left[(\bar{v}+|v_i-v_j|)^\beta\,(\bar{v}-|v_i-v_j|)^{1-\beta} + (\bar{v}-|v_i-v_j|)^\beta\,(\bar{v}+|v_i-v_j|)^{1-\beta}\right] \tag{2.4}$$

Note that, if we had assumed that $v_i \leq v_j$, Equation 2.4 would be identical. During the derivation, only the order of the two summands would reverse. The assumption about which individual has the higher valuation is thus without loss of generality, as follows from the symmetry of the setup.

The expected maximum transfer given by Equation 2.4 is visualized by Figure 2.1 for $\alpha = 1/2$ and $\bar{v} = 20$. Along the x-axis of the graph, we vary the parameter β, going from a situation where both people fully project their own preferences ($\beta = 0$) to one where they fully adopt others' preferences ($\beta = 1$). On the z-axis, we vary the difference between both subjects' valuations, holding constant the total of the two. The graph starts at the maximum of 20 and ends at a distance of zero, i.e., a situation where both valuations are the same. On the y-axis, the resulting expected maximum transfer $\mathbb{E}[t^\star]$ is depicted. If β is either zero or one, the expected transfer is always at its maximum value of 5. The same is always the case when the two subjects' valuations coincide. Thus, if we were only talking about money, the degree of empathy would not have any effect on expected transfers because there would not exist any heterogeneity in valuations. This is, however, only a special case. If, as we expect, β typically lies in the interior of the interval from zero to one, dissimilarity in preferences strictly decreases expected transfers, which is our

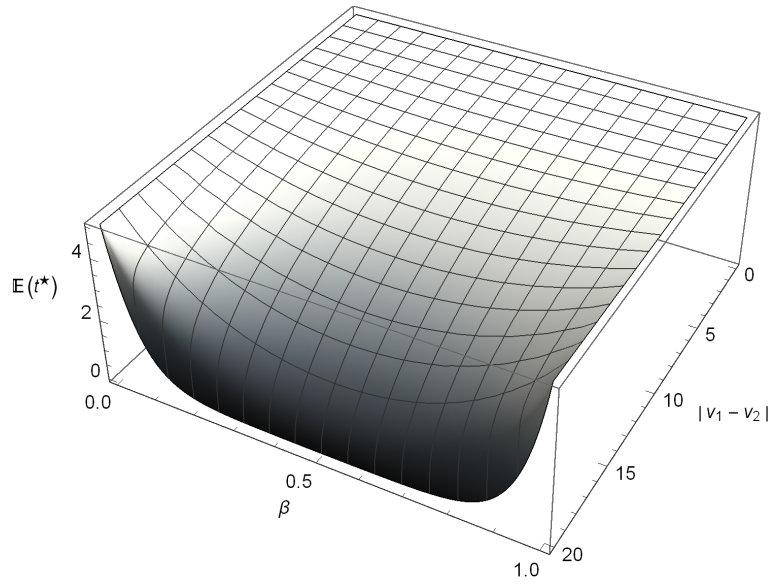


Figure 2.1: Similarity and expected transfers

second hypothesis.

Hypothesis 2.2. *Transfers decrease with the dissimilarity of preferences within pairs of senders and receivers. Formally,* $\partial \mathbb{E}[t^\star] / \partial |v_i(k) - v_j(k)| < 0$.

Proof. See Appendix 2.A □

The second channel through which dissimilarity in preferences decreases welfare is saying that—conditional on a given total amount of transfers that a subject is making—senders give for the wrong items. In the extreme case of $\beta = 0$, a sender might be willing to spend a positive amount when the receiver has no problem with eating the respective food item at all, while she gives nothing in case of an item that the receiver finds repulsive. More generally, values of β which are smaller than one open up a wedge between how the sender evaluates consequences for the receiver and how the receiver himself—and consequently the social planner—evaluates them. This wedge becomes increasingly relevant as valuations of senders and receivers diverge, leading to misallocation of transfers.[3] We thus arrive at the following hypothesis.

Hypothesis 2.3. *The expected net welfare gain from transfer decisions decreases in the dissimilarity of preferences within pairs of senders and receivers.*
Formally, $\partial \mathbb{E}[\textit{net welfare gain}] / \partial |v_i(k) - v_j(k)| < 0$.

To summarize, we expect that senders base their transfer decisions partially on their own valuations, where the latter and the receiver's own valuation are complements in motivating senders to help. This leads transfers to be, on average, lower for pairs of senders

[3] A subtle refinement of the above point would be to consider vicarious experiences—i.e., the sender's feelings when considering consequences for the receiver—as part of welfare. This would reduce the power of the argument about misallocation but not alleviate it. With imperfect empathy and heterogeneous preferences, there always exists a wedge between vicarious valuations and valuations by people themselves, leading to misallocation.

and receivers who are more dissimilar than others. Reduced transfers and misallocation of existing transfers together lead to welfare losses, which again are larger when subjects are dissimilar.

2.3 Experiment

We conducted the experiment at the BonnEconLab in August, September, and December 2018. Subjects were recruited using the software hroot (Bock, Baetge, and Nicklisch, 2014) and a total of 146 participated. In the invitation email, subjects were asked not to sign up for the experiment in case they were vegetarian, followed a special diet due to health, ethical or religious reasons, or had any food allergy. For details on the composition of our sample of subjects, see the summary statistics in Appendix 2.C. Before starting the computerized zTree experiment (Fischbacher, 2007), subjects were informed that they might be asked to eat several types of insects during the experiment. They were then shown a tray with all eight different food items (one buffalo worm, five buffalo worms, one mealworm, ten mealworms, one cricket, one grasshopper, three grasshoppers, and one granola bar containing buffalo worms; see Appendix 2.D for pictures of all food items). Furthermore, they received information about the food items' nutritional innocuousness, and each participant signed a form of consent.

The experiment consisted of four parts. Subjects received a fixed show-up fee, which was set to be either €5 or €7 for everybody participating in the respective session.[4] In addition, subjects were informed that a single decision among all four parts of the experiment would be randomly chosen for implementation and paid at the end of the experiment. All parts were kept as similar to each other as possible. Always, subjects were endowed with €20 and then used multiple price lists (MPLs) in steps of €1 ranging from €1 to €20 to make payments off this amount. Appendix 2.F includes screenshots of the decision screens of all four parts.

In Part 1, we employed separate MPLs to elicit subjects' reservation prices for not having to eat any of the eight food items. Subjects saw one screen per item (see Figure 2.F.1 for an example of a decision screen). On each screen, subjects saw an informative stimuli picture of the respective item on the left and a list of choices in the middle of the screen. The list of choices was made up of 20 rows, each row containing the choice between a payment (going from €1 up to €20) and eating the food item. Subjects had to indicate their choice for one of the two options for each row; a unique switching point was enforced. The order in which the eight items were shown was randomized between subjects. In case Part 1 was selected for implementation at the end of the experiment, one of the 160 rows (20 rows each for eight items) was randomly drawn for implementation. If the subject had chosen to pay the amount indicated in the specific row, she received her show-up fee as well as €20 minus the amount indicated in the row as payment. She did then not have to eat the item. If the subject had chosen to eat the item, she received the show-up fee as

[4] We varied the show-up fee between sessions in order to test our assumption that utility is linear in money.

well as €20 as payment. She furthermore had to eat the item. Subjects who indicated that they would eat the item and refused to do so at the end of the experiment only received their show-up fee.

In Part 2, subjects took the role of a *sender* who had the option to pay for a *receiver* not having to eat a food item. The decision screens were kept very similar to the ones in Part 1 and again contained the same respective stimuli pictures on the left-hand side of the MPLs (see Figure 2.F.2 for an example of a decision screen). On the right side of the screen, subjects additionally saw the WTPs for all eight items that the relevant receiver himself had reported in Part 1. Again, each subject saw eight screens—one for each item. The eight decisions were each made for a different receiver. Receivers were sampled from the pool of subjects taking part in the same session, and each participant appeared as a receiver at least once to allow for potential implementations of a decision in this part. However, receivers were sampled in such a way that the heterogeneity of WTPs between senders and receivers was larger than in the population of subjects.[5] The assignment of receivers to food items was done without any further sophistication. As in Part 1, subjects had to indicate for each row of the choice list if they chose the payment or the insect. In case Part 2 was selected for implementation at the end of the experiment, one of the 160 rows was randomly drawn for implementation. If the sender had chosen to pay the amount indicated in the specific row, she received her show-up fee as well as €20 minus the amount indicated in the row as payment. The receiver of the row did then not have to eat the item and received his show-up fee and €20. If the sender had chosen not to pay, she received the show-up fee as well as €20 as payment. The receiver furthermore had to eat the item and received his show-up fee and €20. Receivers who refused to eat the item even though their senders had indicated that they would not pay only received their show-up fee.

Part 3 elicited subjects' general level of altruism in the domain of money in a way that mimicked the other parts of the experiment as closely as possible. As a default, receivers got an amount which was less than €20, mirroring a situation where they had to eat a food item for which they have a certain willingness to pay, and senders could decide whether they wanted to pay amounts from €1 and €20 to secure the receiver €20 instead of €15, €10, €5, or €0. The order of amounts was again randomized. Since we are not using Part 3 for the analysis, we will not go into more detail here.

Finally, in Part 4, subjects were again shown the same eight receivers as in Part 2. This time, however, they did not decide about engaging in helping behavior but had the option to alter receivers' self-regarding choices from Part 1 without any consequences for themselves. Decision screens looked almost the same as the ones from Part 2 and contained the stimuli picture on the left, the MPL in the middle, and the list of the receiver's WTPs

[5] Receivers were sequentially sampled among subjects in the same experimental session. For each sender, we made eight independent draws pertaining to a specific criterion and found the remaining subject who came closest to the respective point. During four sessions, the criterion was the Euclidean distance towards the potential sender's vector of WTPs. In five sessions, it was a vector of WTPs. Note that identification with senders fixed effects—or on the level of the individual sender—only uses variation in WTPs among receivers of a given sender. The latter variation is the result of simple random matching with fixed, equal probabilities.

from Part 1 on the right (see Figure 2.F.4 for an example of a decision screen). However, the MPL already contained the choices that the respective receiver had marked for himself in Part 1. In case Part 4 was selected for payment at the end of the experiment, one of the 160 rows was randomly drawn for implementation for the receiver. If the sender had chosen the payment indicated in the specific row, the receiver received his show-up fee as well as €20 minus the amount indicated in the row as payment. The receiver did then not have to eat the item. If the sender had chosen the item, the receiver received the show-up fee as well as €20 as payment. He furthermore had to eat the item. Receivers who refused to eat the item even though the other subject had not chosen the payment only received their show-up fee.

After every subject had made their decisions, they were ultimately matched to unilateral pairs of senders and receivers for whom a payoff was implemented. For each subject, a single decision was drawn to be paid out. If Part 1 was implemented for the sender, Part 4 was implemented for the receiver. If Parts 2 or 3 were implemented for the sender, the respective part was also implemented for the receiver. After answering a final survey on the Big Five traits (Gerlitz and Schupp, 2005) and the items of the Interpersonal Reactivity Index, which measures empathy (Davis, 1980), subjects—if necessary—ate their food items and then received their payoffs. If subjects did not comply and refused to eat their food items, they were penalized by only receiving the show-up fee.

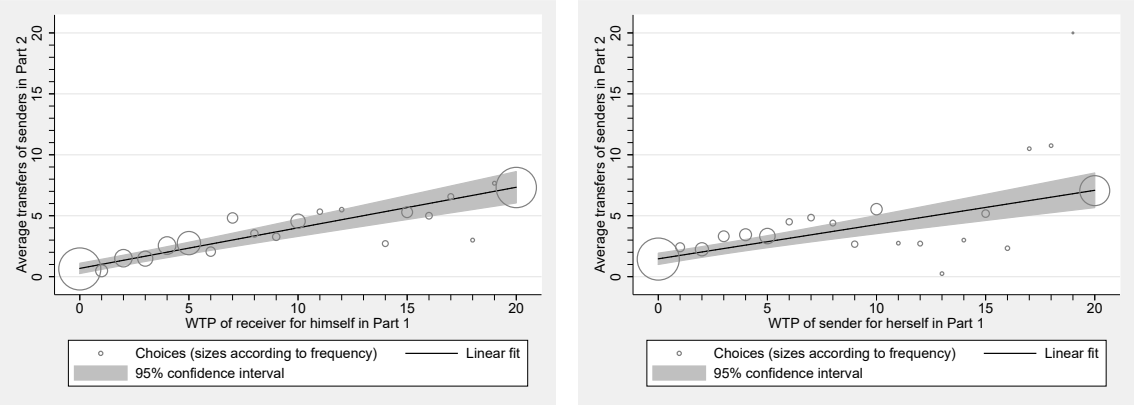
2.4 Results

We start our empirical analysis by estimating the determinants of transfers on the aggregate level by pooling decisions from all individuals. We then proceed by estimating the relationship for each individual separately and recovering individual structural parameters. Next, we turn to the welfare implications by first looking at the effect of dissimilarity on the size of transfers and then directly on net welfare gains. Finally, we show the pervasiveness of imperfect empathy separately among libertarian and paternalist subjects.

2.4.1 Transfer Decisions

In Part 1, participants spend on average €6.57 per item to avoid eating it; 78% of them have a positive WTP for some item. In Part 2, senders spend on average €3.44 per item to spare receivers from eating it; 75% of senders have a positive WTP for some item-receiver combination. Appendix 2.C shows histograms of WTPs for all items separately for Part 1 and Part 2.

Figure 2.2 visualizes how transfers towards receivers in Part 2 depend on the WTPs of receivers and senders from Part 1. Figure 2.2a shows the average size of transfers made for a receiver in Part 2 for every possible WTP of the receiving subject from Part 1. The higher the receiver's WTP, the higher is the average transfer made towards him. The positive relationship is statistically significant at the 1% level. Figure 2.2b shows the average size of transfers made for a receiver in Part 2 for every possible WTP of the sending subject from



(a) Receiver's WTP (b) Sender's WTP

Note: Panel (a) shows the average size of transfers made for a receiver in Part 2 for every possible WTP of the *receiving* subject in Part 1. Panel (b) again shows the average size of transfers made for a receiver in Part 2, this time for every possible WTP of the *sending* subject in Part 1. The lines show fits from OLS estimations, and shaded areas correspond to 95% confidence intervals for standard errors which are clustered at the subject level. Both positive relationships are significant at the 1% level.

Figure 2.2: Individual willingness to pay (WTP) and average transfers

Part 1. The higher the sender's WTP, the higher is the average transfer made towards the receiver. Again, the positive relationship is significant at the 1% level.

To test Hypothesis 2.1, we regress the maximum transfer accepted in Part 2, $t^\star$, on the receiver's willingness to pay $\text{WTP}_{\text{receiver}}$ elicited in Part 1 as well as on the sender's willingness to pay $\text{WTP}_{\text{sender}}$, also elicited in Part 1. Column 1 of Table 2.1 shows the results without any fixed effects. We see that both WTPs—of the receiver and the sender—enter with large and highly significant coefficients, which in fact are not so different in size. An increase of €1 in the WTP of the receiver increases the transfer on average by €0.31, while the same increase in the senders' WTP increases the average transfer by €0.25. A potential concern could be that senders might differ in their general levels of altruism and that this variation is systematically related to their own WTPs. Moreover, particular patterns in receivers' WTPs could trigger responses of senders, irrespective of the particular item in question. To rule out such problems, Column 2 adds sender and receiver fixed effects. Due to sender fixed effects, identification only comes from differences in WTPs between receivers of a given sender and from variation in this sender's WTPs across items. Receiver fixed effects allow accounting, e.g., for some receivers having generally low WTPs and receiving low transfers, irrespective of the particular item and the respective sender. The coefficient for the receiver remains almost unchanged, while the coefficient referring to the WTP of the sender somewhat decreases. The latter points to some degree of "spillovers" in empathy: e.g., a sender who feels strong disgust for worms might also better understand why somebody would strongly dislike eating a grasshopper, even if the grasshopper itself does not seem repulsive for the sender. Despite the proximity of preference domains which we use, variation in preferences within individuals is sufficient to show that there is a strong and significant effect of senders' WTPs on transfers. In Column 3, the square root of the product of the sender's and the receiver's WTP is added to the regression without fixed

Table 2.1: Aggregate analysis of transfers

	Dependent variable: Transfer				
	(1)	(2)	(3)	(4)	(5)
Receiver's WTP	0.308****	0.309****	0.163****	0.160****	0.100***
	(0.0311)	(0.0364)	(0.0300)	(0.0354)	(0.0359)
Sender's WTP	0.252****	0.176****	0.0627	-0.0156	-0.0578
	(0.0345)	(0.0466)	(0.0394)	(0.0582)	(0.0576)
$\sqrt{\text{Sender's} \times \text{receiver's WTP}}$			0.381****	0.369****	0.364****
			(0.0646)	(0.0693)	(0.0702)
Sender fixed effects	No	Yes	No	Yes	Yes
Receiver fixed effects	No	Yes	No	Yes	Yes
Item fixed effects	No	No	No	No	Yes
Observations	1168	1168	1168	1168	1168
Clusters	146	146	146	146	146
(Within-)R^2	0.362	0.197	0.417	0.285	0.171

Note: OLS regression, standard errors are clustered for senders; * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; **** $p < 0.001$; robust standard errors in parentheses.

effects. The coefficient of the receiver's WTP drops by half but is still highly significant, whereas the coefficient of the sender's WTP is not significantly different from zero anymore. However, the interaction term enters with a large and highly significant coefficient. This confirms that WTPs of receivers and senders act as complements in generating transfers. Column 4 again adds sender and receiver fixed effects. In Column 5, we additionally add fixed effects for the eight different food items, accounting for differences in the general levels of transfers. In both Columns 4 and 5, coefficients stay similar, and the qualitative results remain unchanged.

We show in Section 2.4.4 that the above qualitative results also hold within subsamples of our subject populations where senders are restricted to only libertarians or paternalists, respectively. Our empirical results are also insensitive to the size of the show-up fee (see Appendix 2.B), and the assumption of utility from money being linear in the relevant range thus seems innocent. Overall, we find clear support for Hypothesis 2.1. We interpret this as evidence that imperfect empathy is a pervasive phenomenon among our subject population.

2.4.2 Individual-level Analysis

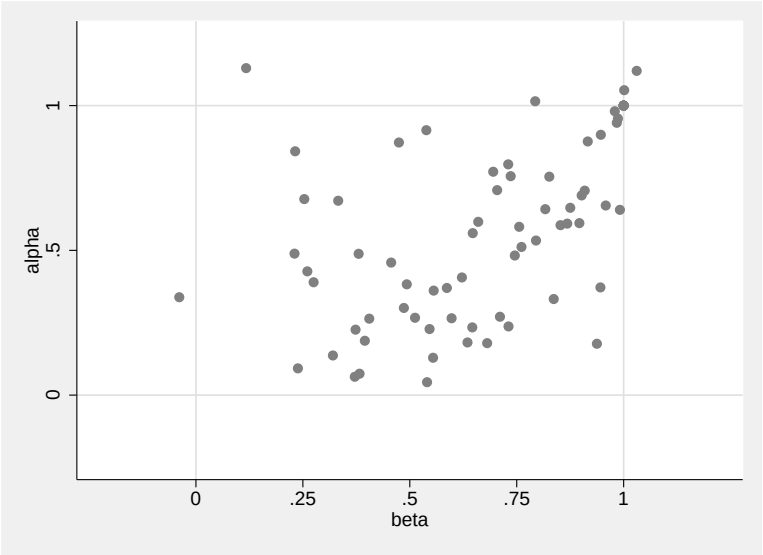
In the next step, we analyze behavior at the level of individuals and recover estimates for the model parameters. To do so, we first linearize Equation 2.2 for the size of transfers by taking the logarithm on both sides.

$$\ln(t^\star) = \ln(\alpha) + \beta \ln[v_j(k)] + (1-\beta)\ln[v_i(k)] \tag{2.5}$$

We estimate Equation (2.5) separately for each individual subject. Note that all quantities except the parameters in the equation are directly observed in our experimental data. $\ln(t^\star)$ is the logarithm of the maximum transfer accepted in Part 2, $\ln[v_j(k)]$ equals the logarithm of the receiver's willingness to pay $\text{WTP}_{\text{receiver}}$ elicited in Part 1, and $\ln[v_i(k)]$ is the logarithm of the sender's willingness to pay $\text{WTP}_{\text{sender}}$ elicited in Part 1.[6] We thus estimate the following linear equation.

$$\ln(t^\star) = \gamma_0 + \gamma_1 \ln(\text{WTP}_{\text{receiver}}) + (1 - \gamma_1) \ln(\text{WTP}_{\text{sender}}) + \epsilon \tag{2.6}$$

The estimates for the general level of altruism are given by $\hat{\alpha} = \exp(\hat{\gamma}_0)$ and those for empathy by $\hat{\beta} = \hat{\gamma}_1$.



Note: The figure shows estimated parameter values for α and β. Only those subjects entered the analysis who made varying choices within Part 1 and Part 2. In addition, six subjects were excluded due to implausible parameter estimates and three further subject were excluded due to large standard deviations of the parameter estimates, leaving 71 observations.

Figure 2.3: Estimates for individual parameters

Figure 2.3 shows parameter estimates for β on the horizontal axis and α on the vertical axis. Variation in senders' WTPs and transfers in principle allows us to identify parameters for 80 subjects, of whom we get reasonable estimates for 71. Among the latter, the vast majority of subjects are assigned estimates which lie inside the ranges of expected values from zero to one. We see large heterogeneity in parameter estimates, and the variation in estimates for β indicates that the effects that we find in the analysis on the aggregate level are not only driven by a small number of subjects. Moreover, the figure shows that, for any given level of general altruism, there exists marked heterogeneity in the empathy parameter. The two thus appear to be distinct characteristics of the individuals.

[6] To avoid missing values at zero, we added 0.1 to all WTPs and transfers.

2.4.3 Welfare

We now turn to study the welfare implications of the decisions that were observed in the experiment. To test Hypothesis 2.2, we regress transfers on two different measures of dissimilarity between sender and receiver. We define partial dissimilarity as the absolute difference between sender i's and receiver j's WTP regarding the relevant item k, divided by its maximum of 20.

$$\text{Partial dissimilarity}_{ijk} = \frac{|\text{WTP}_{ik} - \text{WTP}_{jk}|}{20}$$

Total dissimilarity is the Euclidean distance between the full vectors of sender i's and receiver j's WTPs for all items k, again divided by its potential maximum value.

$$\text{Total dissimilarity}_{ij} = \frac{\sqrt{\sum_{k=1}^{8} \left(\text{WTP}_{ik} - \text{WTP}_{jk}\right)^2}}{20\sqrt{8}}$$

Table 2.2: Similarity and transfers

	Dependent variable: Transfer				
	(1)	(2)	(3)	(4)	(5)
Partial dissimilarity	-4.156****		-4.046****		-3.724****
	(0.718)		(0.728)		(0.718)
Total dissimilarity		-4.244****		-4.393****	-0.652
		(0.970)		(0.973)	(0.789)
Receiver's WTP	0.378****		0.305****		0.301****
	(0.0341)		(0.0428)		(0.0418)
Sender's WTP	0.285****		0.145***		0.143***
	(0.0357)		(0.0442)		(0.0447)
Receiver's average WTP		0.393****			
		(0.0419)			
Sender's average WTP		0.310****			
		(0.0470)			
Sender fixed effects	No	No	Yes	Yes	Yes
Receiver fixed effects	No	No	Yes	Yes	Yes
Item fixed effects	No	No	Yes	Yes	Yes
Observations	1168	1168	1168	1168	1168
Clusters	146	146	146	146	146
(Within-)R^2	0.429	0.305	0.208	0.0773	0.208

Note: OLS regression, standard errors are clustered for senders; * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; **** $p < 0.001$; robust standard errors in parentheses.

Table 2.2 shows the results. Columns 1 and 3 use partial dissimilarity, while Columns 2 and 4 use total dissimilarity. Columns 1 and 2 present the baseline results without any fixed effects. Column 1 uses the receiver's and the sender's WTP to control for level effects. The effect of partial dissimilarity is thus conditional on both parties' own valuations, and it shows that dissimilarity decreases the size of transfers. In Column 2, level effects

are correspondingly controlled for by using the receiver's and the sender's average WTP because total similarity also refers to all items. Total dissimilarity enters negatively and with a similar effect size as partial dissimilarity in Column 1. Columns 3 and 4 replicate the previous two with fixed effects for senders, receivers, and items, making controls for average WTPs redundant. The effects of dissimilarity remain almost unchanged. In Column 5, regressors from the previous two columns are combined. Total dissimilarity has no significant effect beyond the effect through partial dissimilarity, which is in line with Hypothesis 2.2. Interestingly, this means that senders descriptively discriminate against receivers whose preferences are different but only because of imperfect empathy and not because they generally dislike them.

To test Hypothesis 2.3, we first need to derive the welfare consequences of decisions over any proposed transfer level t. The net welfare gain from no transfer is—by definition—zero. If the proposed transfer (i.e., the row on the decision screen) is accepted, the net welfare gain can be calculated according to Equation 2.3. If a transfer of t is made, welfare is given by $20 - t + 20$. If the transfer is not made, welfare is given by $20 + 20 - v_j(k)$, where the latter valuation corresponds to the WTP of the receiver. For any proposed transfer (a row in the MPL), the welfare impact can thus be calculated as follows.

$$\text{Net welfare gain} \equiv \begin{cases} \text{WTP}_{\text{receiver}} - t & \text{if transfer of } t \text{ is made} \\ 0 & \text{if transfer of } t \text{ is not made} \end{cases}$$

By the design of the MPLs used in the experiment, the probability of a transfer being made for a given item and receiver is the maximum size of the transfer, $t^\star$, divided by the number of rows, which is 20. If a transfer is made, the receiver experiences a welfare gain equivalent to his corresponding WTP. The sender loses the transfer amount of the respective row. We can thus calculate the expected net welfare gain of any transfer decision made by senders as follows.

$$\mathbb{E}[\text{net welfare gain}] = \underbrace{\frac{t^\star}{20}}_{\mathbb{P}\,(\text{transfer made})} \left(\text{WTP}_{\text{receiver}} - \underbrace{\frac{t^\star + 1}{2}}_{\mathbb{E}\,[\text{transfer} \,|\, \text{transfer made}]} \right)$$

Table 2.3 shows the results of regressing the expected net welfare gain on partial dissimilarity or total dissimilarity. Columns correspond to the ones in Table 2.2. Columns 1 and 2 present the baseline results without any fixed effects. Column 1 uses the receiver's and the sender's WTPs to control for level effects. We find that partial dissimilarity decreases the expected net welfare gain. In Column 2, level effects are controlled for by using the receiver's and the sender's average WTP. Total dissimilarity also enters negatively, with a magnitude that is comparable to that of partial dissimilarity. Columns 3 and 4 replicate the previous two columns with fixed effects for senders, receivers, and items. The estimated effects remain stable. Columns 5 combines the regressors from Columns 3 and 4, showing that the effect of total dissimilarity is entirely driven by the effect of partial dissimilarity regarding the relevant item. Thus, Hypothesis 2.3 is confirmed. Moreover, we again find

Table 2.3: Similarity and welfare

	Dependent variable: $\mathbb{E}$[net welfare gain]				
	(1)	(2)	(3)	(4)	(5)
Partial dissimilarity	-2.648****		-2.460****		-2.398****
	(0.373)		(0.370)		(0.378)
Total dissimilarity		-2.490****		-2.484****	-0.126
		(0.481)		(0.504)	(0.352)
Receiver's WTP	0.270****		0.260****		0.260****
	(0.0183)		(0.0224)		(0.0224)
Sender's WTP	0.0878****		0.0805****		0.0802****
	(0.0141)		(0.0183)		(0.0186)
Receiver's average WTP		0.267****			
		(0.0214)			
Sender's average WTP		0.0915****			
		(0.0204)			
Sender fixed effects	No	No	Yes	Yes	Yes
Receiver fixed effects	No	No	Yes	Yes	Yes
Item fixed effects	No	No	Yes	Yes	Yes
Observations	1168	1168	1168	1168	1168
Clusters	146	146	146	146	146
(Within-)R^2	0.536	0.331	0.397	0.0665	0.397

Note: OLS regression, standard errors are clustered for senders; * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; **** $p < 0.001$; robust standard errors in parentheses.

no evidence for taste-based discrimination against receivers with different preferences.

2.4.4 Libertarians vs. Paternalists

Since we have shown that people partially rely on their own preferences in choosing the level to which they provide others with help, it is natural to ask whether and how this might be related to paternalistic behavior: if people are not willing to support choices that seem strange to them, they might also want to change them. Nevertheless, imperfect empathy and paternalism are different concepts. First, imperfect empathy pertains to a certain kind of preference, whereas paternalism is a certain kind of behavior. Second, the ranges of relevant applications of both phenomena might overlap (see, e.g., Jacobsson, Johannesson, and Borgquist, 2007) but are not identical: Imperfect empathy is relevant in many situations where restricting others' freedom is not even an option; and paternalism occurs in many contexts where empathy is not relevant but is often driven by, e.g., asymmetric information. Third, it is not clear whether people who make helping behavior depend on their own valuations regard the latter as normatively warranted or would rather—if they were aware of it—object to such behavior and therefore also not want to restrict the freedom of others.

To study the relationship between imperfect empathy and paternalism empirically, we use subjects' choices from Part 4 to classify them as paternalists or libertarians. A subject is only classified as a libertarian if she abstained from altering any other subjects' decisions. All subjects that altered any decision are classified as paternalists. According to this definition, we end up with 74 libertarian subjects and 72 paternalists.

Table 2.4: Libertarians vs. paternalists

	Dependent variable: Transfer							
	Libertarians				Paternalists			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Receiver's WTP	0.412****	0.398****	0.210****	0.219****	0.207****	0.179****	0.134****	0.105**
	(0.0482)	(0.0528)	(0.0554)	(0.0583)	(0.0356)	(0.0386)	(0.0315)	(0.0428)
Sender's WTP	0.238****	0.156***	0.0283	-0.0434	0.251****	0.209***	0.127*	0.0862
	(0.0451)	(0.0574)	(0.0470)	(0.0792)	(0.0550)	(0.0657)	(0.0669)	(0.0808)
$\sqrt{\text{Sender's} \times \text{receiver's WTP}}$			0.442****	0.387****			0.236***	0.224**
			(0.0894)	(0.0979)			(0.0888)	(0.0964)
Sender fixed effects	No	Yes	No	Yes	No	Yes	No	Yes
Receiver fixed effects	No	Yes	No	Yes	No	Yes	No	Yes
Observations	592	569	592	569	576	558	576	558
Clusters	74	74	74	74	72	72	72	72
(Within-)R^2	0.451	0.269	0.516	0.357	0.267	0.142	0.291	0.177

Note: OLS regression, standard errors are clustered for senders; * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; **** $p < 0.001$; robust standard errors in parentheses. Columns with fixed effects include fewer observations than others because some receivers were only matched to a single paternalist or libertarian sender, respectively.

Table 2.4 replicates our main results on transfer decisions from the first four columns of Table 2.1 separately for libertarians in Columns 1–4 and paternalists in Columns 5–8. Comparing Columns 1 and 2 with Columns 5 and 6, respectively, we see that the role of senders' WTPs is slightly weaker among libertarians as compared to paternalists, although the differences in coefficients are not statistically significant. The effect of receivers' WTP

on transfers is stronger amongst libertarians than amongst paternalists. This is plausible since paternalists are subjects that are less willing to accept others' choices, and it is therefore not surprising that they react less to receivers' WTP. More importantly, the effect of senders' WTPs enters with considerable magnitude and high statistical significance within both subsamples. Proceeding towards the comparison of Columns 3 and 4 with Columns 7 and 8, it turns out that the effect of the interactions between senders' and receivers' WTPs is, in fact, stronger among libertarians than among paternalists. The latter finding also alleviates concerns about measurement error driving our results, i.e., about senders trying to "correct" receivers' choices. While the interpretation of WTPs as noisy signals about true valuations would not be able to explain the asymmetry of our results in the first place—i.e., the complementarity of valuations—it would also be incompatible with senders not intervening in other subjects' own decisions: senders care about receivers—they make transfers—and if they thought others were making mistakes, they should save them from doing so. In contrast, in our experiment, even people who put faith in others' personal judgments and who do not show any signs of paternalism exhibit imperfect empathy.

2.5 Conclusion

In this chapter, we show that people behave imperfectly empathic when acting prosocially. They assess consequences arising to others based on a combination of their own and the other persons' preferences. In particular, own and others' valuations act as complements in bringing about helping behavior. We show that this property of imperfect empathy leads to the effect that dissimilar preferences lower the size of transfers as well as overall welfare.

We hereby confirm the hypothesis that empathy plays a role in generating prosocial behavior and show furthermore that a lack of it can lead to lower and —according to a basic libertarian welfare criterion— poorly aimed helping behavior. Our results also inform models featuring altruism in conjunction with heterogeneous preferences. However, the mechanism of imperfect empathy is not only relevant for individual behaviors such as charitable giving or volunteering. It also allows for an alternative perspective on the phenomenon of in-group bias. We observe that transfers are lower if other people have overall different preferences. Within our experiment, however, this effect is entirely driven by imperfect empathy and not by a dislike against subjects who are different. Imperfect empathy might also have implications on the aggregate level for the working of welfare states. If people cannot relate to the consumption choices made by recipients of welfare benefits, this could decrease the willingness to finance such redistributive policies. An implication for policy might be that exposure to individuals with different sets of preferences, e.g., due to cultural or religious backgrounds, could be central to the political sustainability of welfare states in increasingly diverse societies.

References

Ambuehl, Sandro. 2017. *An Offer You Can't Refuse? Incentives Change How We Inform Ourselves and What We Believe*. CESifo Working Paper 6296. Munich: Center for Economic Studies & Ifo Institute.

Andreoni, James. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving". *The Economic Journal* 100 (401): 464–477.

Andreoni, James, A. Abigail Payne, Justin Smith, and David Karp. 2016. "Diversity and donations: The effect of religious and ethnic diversity on charitable giving". *Journal of Economic Behavior and Organization* 128:47–58.

Batson, C. Daniel. 1987. "Prosocial Motivation: Is it ever Truly Altruistic?" *Advances in Experimental Social Psychology* 20 (C): 65–122.

Batson, C. Daniel, and Tecia Moran. 1999. "Empathy-induced altruism in a prisoner's dilemma". *European Journal of Social Psychology* 29 (7): 909–924.

Becker, Gary S. 1974. "A Theory of Social Interactions". *Journal of Political Economy* 82 (6): 1063–1093.

— . 1976. "Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology". *Journal of Economic Literature* 14 (3): 817–826.

Bisin, Alberto, and Thierry Verdier. 2001. "The Economics of Cultural Transmission and the Dynamics of Preferences". *Journal of Economic Theory* 97 (2): 298–319.

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch. 2014. "hroot: Hamburg Registration and Organization Online Tool". *European Economic Review* 71:117–120.

Coke, Jay S., C. Daniel Batson, and Katherine McDavis. 1978. "Empathic Mediation of Helping: A Two-Stage Model". *Journal of Personality and Social Psychology* 36 (7): 752–766.

Dahlberg, Matz, Karin Edmark, and Heléne Lundqvist. 2012. "Ethnic Diversity and Preferences for Redistribution". *Journal of Political Economy* 120 (1): 41–76.

Davis, Mark H. 1980. "A Multidimensional Approach to Individual Differences in Empathy". *JSAS Catalog of Selected Documents in Psychology* 10:85.

Doepke, Matthias, and Fabrizio Zilibotti. 2017. "Parenting with Style: Altruism and Paternalism in Intergenerational Preference Transmission". *Econometrica* 85 (5): 1331–1371.

Dufwenberg, Martin, and Georg Kirchsteiger. 2004. "A theory of sequential reciprocity". *Games and Economic Behavior* 47:268–298.

Falk, Armin, and Urs Fischbacher. 2006. "A theory of reciprocity". *Games and Economic Behavior* 54 (2): 293–315.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation". *Quarterly Journal of Economics* 114 (3): 817–868.

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

Fowler, James H., and Cindy D. Kam. 2007. "Beyond the Self: Social Identity, Altruism, and Political Participation". *Journal of Politics* 69 (3): 813–827.

Gerlitz, Jean-Yves, and Jürgen Schupp. 2005. *Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. Dokumentation der Instrumententwicklung BFI-S auf Basis des SOEP-Pretests 2005*. Research Notes 4. Berlin, Germany: Deutsches Institut für Wirtschaftsforschung (DIW).

Jacobsson, Fredric, Magnus Johannesson, and Lars Borgquist. 2007. "Is Altruism Paternalistic?" *The Economic Journal* 117 (520): 761–781.

Klimecki, Olga M., Sarah V. Mayer, Aiste Jusyte, Jonathan Scheeff, and Michael Schönenberg. 2016. "Empathy promotes altruistic behavior in economic interactions". *Scientific Reports* 6:31961.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics". *American Economic Review* 83 (5): 1281–1302.

Ross, Lee, David Greene, and Pamela House. 1977. "The "False Consensus Effect": An Egocentric Bias in Social Perception and Attribution Processes". *Journal of Experimental Social Psychology* 13 (3): 279–301.

Smith, Adam. 1859. *The Theory of Moral Sentiments.* Glasgow.

Tajfel, Henri, M. G. Billig, and R. P. Bundy. 1971. "Social categorization and intergroup behaviour". *European Journal of Social Psychology* 1 (2): 149–178.

Appendix 2.A Proof of Hypothesis 2.2

The partial derivative of expected transfers given by Equation 2.4 with respect the distance between subjects' valuations is negative whenever β lies in the open interval from zero to one and the distance between individual valuations is larger than one. Valuations are denoted by $v_i, v_j > 0$ and $\bar{v} = v_i + v_j$ denotes the total of both valuations.

Proof.

$$\begin{aligned}
\frac{\partial\,\mathbb{E}[t^\star]}{\partial\,|v_i - v_j|} &= \frac{\alpha}{4}\Bigg[\beta\left(\frac{\bar{v} - |v_i - v_j|}{\bar{v} + |v_i - v_j|}\right)^{1-\beta} - (1-\beta)\left(\frac{\bar{v} + |v_i - v_j|}{\bar{v} - |v_i - v_j|}\right)^{\beta} \\
&\quad - \beta\left(\frac{\bar{v} + |v_i - v_j|}{\bar{v} - |v_i - v_j|}\right)^{1-\beta} + (1-\beta)\left(\frac{\bar{v} - |v_i - v_j|}{\bar{v} + |v_i - v_j|}\right)^{\beta}\Bigg] \\
&= \frac{\alpha}{4}\Bigg\{\beta\Bigg[\Bigg(\underbrace{\frac{\bar{v} - |v_i - v_j|}{\bar{v} + |v_i - v_j|}}_{\in(0,1]}\Bigg)^{1-\beta} - \left(\frac{\bar{v} - |v_i - v_j|}{\bar{v} + |v_i - v_j|}\right)^{-(1-\beta)}\Bigg] \\
&\quad + (1-\beta)\Bigg[\left(\frac{\bar{v} - |v_i - v_j|}{\bar{v} + |v_i - v_j|}\right)^{\beta} - \left(\frac{\bar{v} - |v_i - v_j|}{\bar{v} + |v_i - v_j|}\right)^{-\beta}\Bigg]\Bigg\} \\
&\begin{cases} < 0 & \text{if } \beta \in (0,1) \wedge |v_i - v_j| > 0 \\ = 0 & \text{if } \beta \in \{0,1\} \vee |v_i - v_j| = 0 \end{cases}
\end{aligned}$$

□

Appendix 2.B Robustness Regarding Income Levels

In Section 2.2.1, we have made the assumption that utility from money is linear, which we have used throughout the chapter. We believe that this assumption is innocent since we are concerned with monetary amounts in a range of €0 to €27. However, as a simple robustness exercise, we varied the fixed show-up fee that subjects received between sessions.

In four sessions, subjects received €7 and in five sessions, they received €5. If the level of earnings during the experiment mattered for subjects' decision making, this should voice itself in results that differ between sessions depending on the size of the show-up fee.

Table 2.B.1: High show-up fee vs. low show-up fee

	Dependent variable: Transfer							
	Show-up fee = €7				Show-up fee = €5			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Receiver's WTP	0.292****	0.285****	0.135***	0.120**	0.321****	0.328****	0.179****	0.188****
	(0.0491)	(0.0605)	(0.0428)	(0.0515)	(0.0403)	(0.0450)	(0.0412)	(0.0488)
Sender's WTP	0.259****	0.202**	0.0808	0.0121	0.250****	0.156***	0.0433	-0.0407
	(0.0590)	(0.0878)	(0.0645)	(0.109)	(0.0386)	(0.0506)	(0.0483)	(0.0604)
$\sqrt{\text{Sender's} \times \text{receiver's WTP}}$			0.363****	0.359***			0.415****	0.384****
			(0.0933)	(0.113)			(0.0859)	(0.0824)
Sender fixed effects	No	Yes	No	Yes	No	Yes	No	Yes
Receiver fixed effects	No	Yes	No	Yes	No	Yes	No	Yes
Observations	528	528	528	528	640	640	640	640
Clusters	66	66	66	66	80	80	80	80

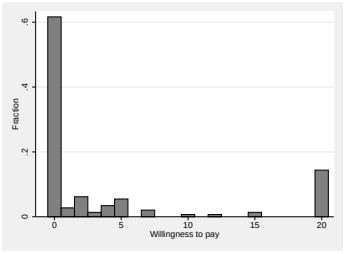
Note: OLS regression, standard errors are clustered for senders; * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; **** $p < 0.001$; robust standard errors in parentheses.

Table 2.B.1 shows the results corresponding to the ones in Table 2.1 split according to the size of the show-up fee. Qualitative results are robust within both subsamples; all differences in coefficients are insignificant. Differences in the income level during the experiment therefore do not seem important for our results.

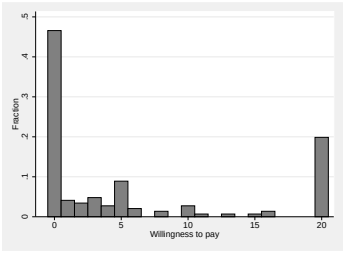
Appendix 2.C Descriptive Statistics

Table 2.C.1: Summary statistics

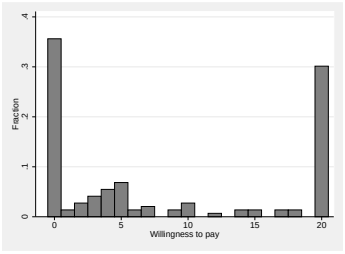
Variable	Mean	Minimum	Maximum	Standard deviation	Observations
Female	0.500	0	1	0.502	146
Age	25.630	18	69	7.741	146
Partial distance	0.416	0	1	0.386	1168
Total distance	0.493	0	1	0.281	1168



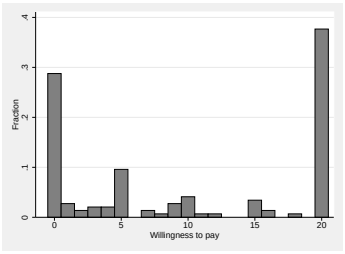
(a) One buffalo worm



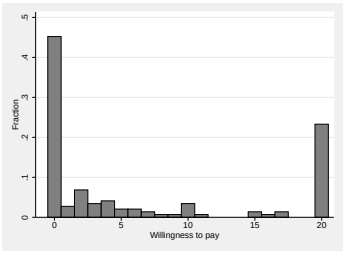
(b) Five buffalo worms



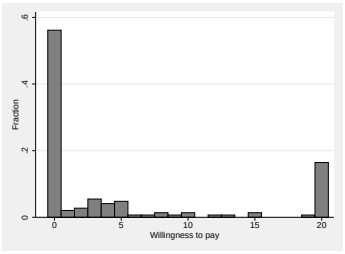
(c) One grasshopper



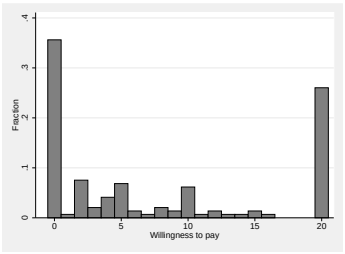
(d) Three grasshoppers



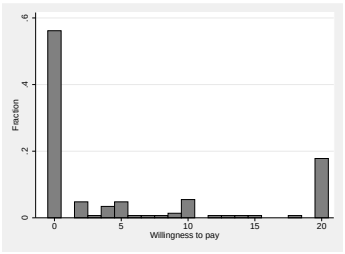
(e) One cricket



(f) One mealworm



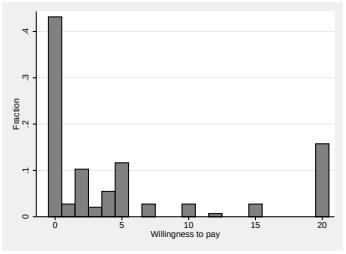
(g) Ten mealworms



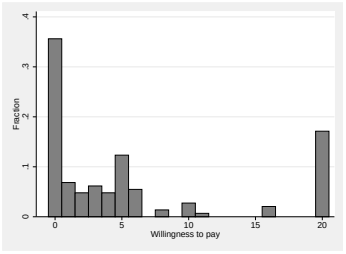
(h) Bar of buffalo worms

Note: The figure shows the distribution of the WTPs for the eight food items of all subjects who acted as senders in Part 2. Shown are the decisions they made for themselves in Part 1.

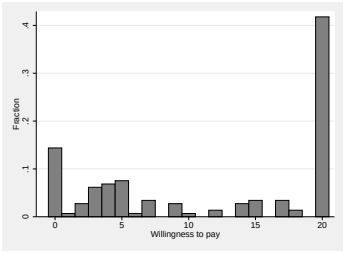
Figure 2.C.1: Senders' willingness to pay (Part 1)



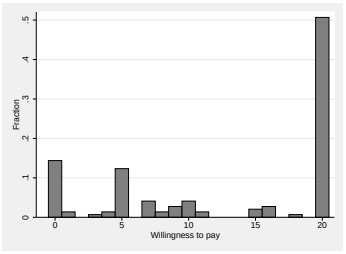
(a) One buffalo worm



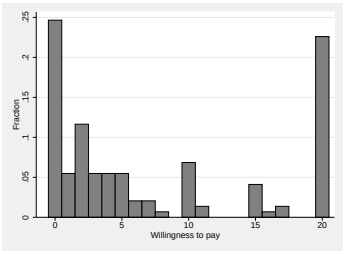
(b) Five buffalo worms



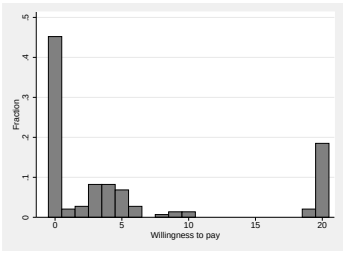
(c) One grasshopper



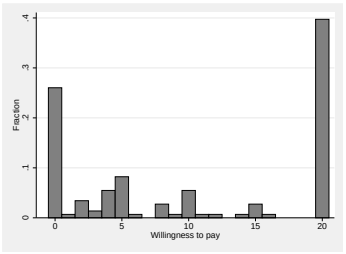
(d) Three grasshoppers



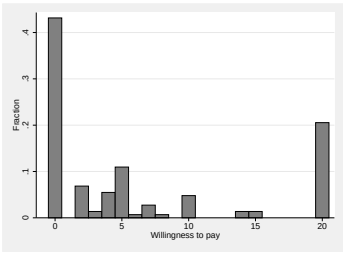
(e) One cricket



(f) One mealworm



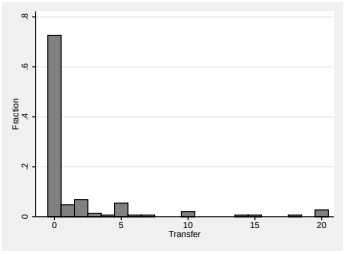
(g) Ten mealworms



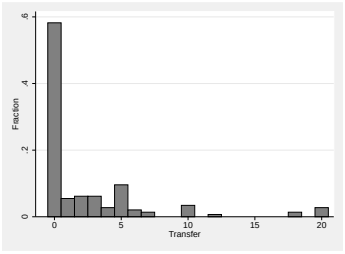
(h) Bar of buffalo worms

Note: The figure shows the distribution of the WTPs for the eight food items of all subjects who acted as receivers in Part 2. Shown are the decisions they made for themselves in Part 1.

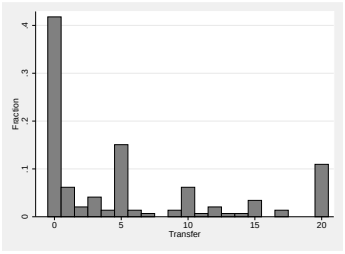
Figure 2.C.2: Receivers' willingness to pay (sampled from Part 1)



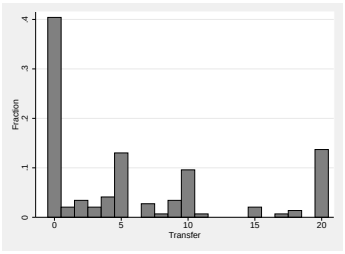
(a) One buffalo worm



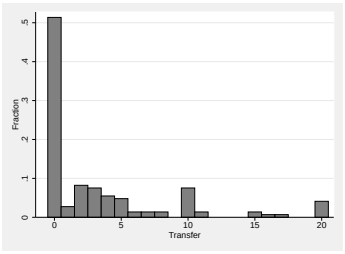
(b) Five buffalo worms



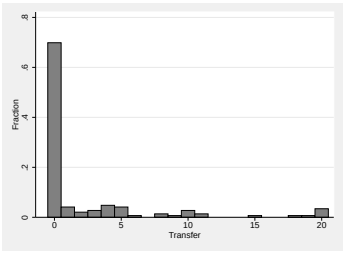
(c) One grasshopper



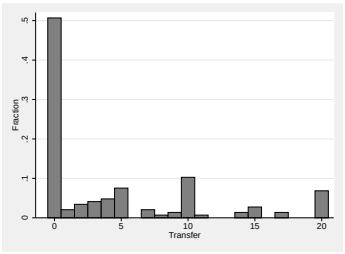
(d) Three grasshoppers



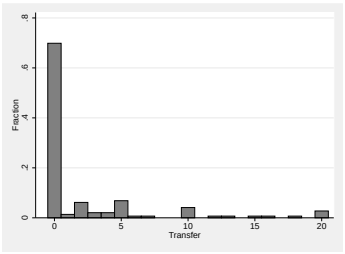
(e) One cricket



(f) One mealworm



(g) Ten mealworms



(h) Bar of buffalo worms

Figure 2.C.3: Transfers (Part 2)

Appendix 2.D Stimuli Pictures

(a) One buffalo worm

(b) Five buffalo worms

(c) One mealworm

(d) Ten mealworms

(e) One grasshopper

(f) Three grasshoppers

(g) One cricket

(h) Bar of buffalo worms

Figure 2.D.1: Stimuli pictures of insects

Appendix 2.E Instructions

Verbal Instructions

"Welcome to today's experiment! Before we start, we would like to provide you with some information.

In this experiment, you will, under specific circumstances, eat dried and non-living insects or food containing dried insects that were farmed in the European Union for the consumption of humans. We will give you information on the specific insects and we will also show them to you. We will now come to your cubicles and show you a tray with the insects and the food containing insects. We will furthermore distribute written information on the nutritional values of the insects and forms of consent. Please read everything carefully and sign the forms of consent; we will afterwards collect the written information and one of the forms. You can keep the other form of consent as a copy. As soon as you have read and signed everything, please hold up your hand to let us know you are ready.

The insects that you are seeing now are crickets, grasshoppers, mealworms, and buffalo worms. You are furthermore seeing a cereal bar containing buffalo worms. You can also touch the insects and the bar carefully. All of these insects were farmed in the European Union for human consumption. They have been certified as food and are completely innocuous for your health; on the contrary, insects are typically very healthy.

Please close your curtains now; we will start the experiment."

2. Welcome and Introduction

Welcome!

For participating in this experiment, you and all other present participants will receive a payment of €5/€7. All further payments, which will depend on your decisions during this experiment, will be added.

Today's study is made up of four parts. You will make decisions that can influence your payment or the payment of another participant in all four parts. At the same time, other participants will make decisions that can in turn influence your payment.

At the end of the experiment, with a probability of 50%, one of your own decisions will be implemented. All single decisions that you make during the course of the experiment will be paid out with equal probability.

With a probability of 50%, a decision another person has made for you will be implemented. All single decisions that a participant took for somebody else will be implemented with equal probability.

Please raise your hand if you have questions at any time. One of the experimenters will then come to you.

Please click "continue" to start the experiment.

3. Instructions Part 1

For Part 1 of this study, you receive a payment of €20. In the following, we would like to know which amount of these €20 you would be willing to pay in order to not eat a specific insect. For every insect or food item containting insects, which were shown to you at the beginning of the study, you will in the following see a list of decisions of the following form:

Option **Payment**: You pay €x and do not have to eat the insect – Option **Insect**: You eat the insect

For each of these decisions, you have to decide between the option **Payment** and the option **Insect**.

In case one of the decisions is implemented at the end of the experiment and you chose option **Payment**, you receive €20 minus the amount that was indicated in the description and you do not have to eat. In case you chose option **Insect** for the chosen decision, you will have to eat the insect before payment and receive €20 without deduction.

Important: In case you choose option Insect and – contrary to your decision - refuse to eat the insect, you will receive a penalty of €20 and receive €0.

If you understood everything, please click "continue" in order to continue with the experiment.

4. Decision Screen Part 1 (see Screenshot in Figure 2.F.1)

5. Instructions Part 2

Thank you for finishing Part 1. You are now starting Part 2.

For Part 2 of this study, you receive a payment of €20. In the following we would like to know which amount of these €20 you would be willing to pay in order for **another participant of this experiment** not having to eat a specific insect. For every insect or food containting insects, which were shown to you at the beginning of the study, you will in the following see a list of decisions of the following form:

Option **Payment**: You pay €x and the other person does not have to eat the insect – Option **Insect**: The other person has to eat the insect

You are making the decision for **eight other participants that are all taking part in this study at this moment**. At the same time, other participants who are also taking part in this study at this moment are making the same decision for you. Before you make your decision for the other participant, you will receive information on how much money this participant was maximally willing to pay for not having to eat the insect in Part 1.

For every decision, you have to decide between option **Payment** and option **Insect**.

In case that at the end of the experiment one of your decisions for somebody else is being implemented, the following will happen: If you chose option **Payment** for the chosen decision, you receive €20 minus the amount that was indicated in the description and the other participant does not have to eat. If you chose option **Insect** for the decision, the other participant has to eat the insect before payment and you receive €20 without deduction.

Important: In case you choose option Insect and – contrary to your decision – the other person refuses to eat the insect, the other person will receive a penalty of €20 and receive €0. You will receive €20.

In case that at the end of the experiment one of the decisions another participant has made for you is being implemented, the following will happen: If the other participant chose option **Payment**, you will receive €20 and you will not have to eat the insect. If the other participant chose option **Insect**, you will have to eat the insect before payment and you will receive €20.

Important: In case the other participant chose option Insect and – contrary to his or her indication - you refuse to eat the insect, you will receive a penalty of €20 and receive €0.

If you understood everything, please click "continue" in order to continue with the experiment.

6. Decision Screen Part 2 (see Screenshot in Figure 2.F.2)

7. Instructions Part 3

Thank you for finishing Part 2. You are now starting Part 3.

For Part 3 of this study, you receive a payment of €20. In the following we would like to know which amount of these €20 you would be willing to pay in order for **another participant of this experiment** to receive a higher payoff. In the following, you will see a list of decisions of the following form:

Option **Payment**: You pay €x and the other person receives €20 – Option **No Payment**: The other person receives €0/€5/€10/€15

For every decision, you have to decide between option **Payment** and option **No Payment**.

In case that at the end of the experiment one of your decisions for somebody else is being implemented, the following will happen: If you chose option **Payment** for the chosen decision, you receive €20 minus the amount that was indicated in the description and the other participant receives €20. If you chose option **No Payment** for the decision, the other participant receives the respective lower payment and you receive €20 without deduction.

In case that at the end of the experiment one of the decisions another participant has

made for you is being implemented, the following will happen: If the other participant chose option **Payment**, you will receive €20. If the other participant chose option **No Payment**, you will receive the respective lower amount.

If you understood everything, please click "continue" in order to continue with the experiment.

8. Decision Screen Part 3 (see Screenshot Figure 2.F.3)

10. Instructions Part 4

Thank you for finishing Part 3. You are now starting Part 4.

In the following we would like to know if you would change the decisions from **Part 1 of another participant**.

As a reminder: In Part 1, every participant decided how much he or she would be willing to pay maximally in order not to eat a specific insect. Every participant saw one list of decisions per insect of the following form and had to decide between option **Payment** and option **Insect**.

Option **Payment**: You pay €x and do not have to eat the insect – Option **Insect**: You eat the insect

You will now see the lists of **eight participants that are all participating in this study at this moment** and their decisions. You can change the decisions of the participants as you want.

In case that at the end of the experiment one of your decisions for somebody else is being selected, this decision will be implemented for the other person. If option **Payment** was chosen, the other participant receives €20 minus the amount that was indicated in the description and does not have to eat. If option **Insect** was chosen, the other participant has to eat the insect before payment and receives €20 without deduction.

Important: In case option Insect was chosen and – contrary to the decision – the other person refuses to eat the insect, the other person will receive a penalty of €20 and will receive €0.

In case that at the end of the experiment one of the decisions another participant has made for you is being implemented, the following will happen: If option **Payment** was chosen, you receive €20 minus the amount that was indicated in the description and you do not have to eat the insect. If option **Insect** was chosen, you have to eat the insect before payment and you receive €20.

Important: In case option Insect was chosen and – contrary to the decision - you refuse to eat the insect, you will receive a penalty of €20 and receive €0.

If you understood everything, please click "continue" in order to continue with the experi-

ment.

9. Decision Screen Part 4 (see Screenshot in Figure 2.F.4)

11. Questionnaire

Thank you for finishing Part 1 to Part 4.

In the following we are asking you to answer some questions.

After you have answered all questions, the experiment is over.

How old are you?

What is your gender?

How high is your monthly income (after taxes and before all expenses)?

In the following, we are interested in how much you are willing to take on risks. Please state your evaluation on a scale from 0 to 10. 0 means „not at all willing to take on risks“ and 10 means „very willing to take on risks“. You can grade your evaluation with the values in between.

0 1 2 3 4 5 6 7 8 9 10

We are now asking you for your willingness to behave a certain way. Please state your evaluation on a scale from 0 to 10. 0 means “not at all willing to do this” and 10 means “very willing to do this”. You can grade your evaluation with the values in between.

How much are you willing to forego something that carries utility for you in order to benefit from it in the future?

0 1 2 3 4 5 6 7 8 9 10

To what extent would you be willing to punish someone who has treated you unfairly even though this has negative consequences for you?

0 1 2 3 4 5 6 7 8 9 10

To what extent would you be willing to punish someone who has treated somebody else unfairly even though this has negative consequences for you?

0 1 2 3 4 5 6 7 8 9 10

To what extent would you be willing to donate to a good cause without expecting something in return?

0 1 2 3 4 5 6 7 8 9 10

Please think about how you would act in the following situation. You are in an unknown area und notice that you got lost. You are asking a stranger for the way. The stranger offers to accompany you to your destination.

Helping you costs the stranger approximately €20 . However, the stranger says that he does not want money from you. You have six gifts with you. The cheapest gift costs €5 , the most expensive gift costs €30 . Would you offer the stranger one of the gifts as a thank you?

Yes/No

Which gift would you give to the stranger?

The gift worth €5
The gift worth €10
The gift worth €15
The gift worth €20
The gift worth €25
The gift worth €30
I don't know

Imagine the following situation: Today, you received an unexpected €1.000.

How much of the money would you donate to a good cause? Donation:

Here are different characteristics a person can have. Probably, some of the characteristics will apply to you personally, whereas others do not. For some, you may be undecided.

Please use the following scale to answer: Value 1 means "does not apply at all" and value 7 means "applies very much". With the values between 1 and 7, you can grade your evaluation.

I see myself as someone, who...

- is a reliable worker 1 2 3 4 5 6 7
- is talkative 1 2 3 4 5 6 7
- is sometimes rude to others 1 2 3 4 5 6 7
- is original, comes up with new ideas 1 2 3 4 5 6 7
- worries a lot 1 2 3 4 5 6 7
- has a forgiving nature 1 2 3 4 5 6 7
- tends to be lazy 1 2 3 4 5 6 7
- is outgoing, sociable 1 2 3 4 5 6 7
- values artistic, aesthetic experiences 1 2 3 4 5 6 7
- gets nervous easily 1 2 3 4 5 6 7
- does things efficiently 1 2 3 4 5 6 7
- is reserved 1 2 3 4 5 6 7

- is considerate and kind to almost everyone 1 2 3 4 5 6 7
- has an active imagination 1 2 3 4 5 6 7
- is relaxed, handles stress well 1 2 3 4 5 6 7

Please indicate for each of the following statements to which extent it applies to you personally. Please state your evaluation on a scale from 1 to 5. A 1 means "describes me very well" and a 5 means "does not describe me well". You can grade your evaluations with the values in between.

- I daydream and fantasize, with some regularity, about things that might happen to me.
- I often have tender, concerned feelings for people less fortunate than me.
- I sometimes find it difficult to see things from the "other guy's" point of view.
- Sometimes I don't feel very sorry for other people when they are having problems.
- I really get involved with the feelings of the characters in a novel.
- In emergency situations, I feel apprehensive and ill-at-ease.
- I am usually objective when I watch a movie or play, and I don't often get completely caught up in it.
- I try to look at everybody's side of disagreement before I make a decision.
- When I see someone being taken advantage of, I feel kind of protective towards them.
- I sometimes feel helpless when I am in the middle of a very emotional situation.
- I sometimes try to understand my friends better by imagining how things look from their perspective.
- Becoming extremely involved in a good book or movie is somewhat rare for me.
- When I see someone get hurt, I tend to remain calm.
- Other people's misfortunes do not usually disturb me a great deal.
- If I'm sure I'm right about something, I don't waste much time listening to other people's arguments.
- After seeing a play or movie, I have felt as though I were one of the characters.
- Being in a tense emotional situation scares me.
- When I see someone being treated unfairly, I sometimes don't feel very much pity for them.

- I am usually pretty effective in dealing with emergencies.
- I am often quite touched by things that I see happen.
- I believe that there are two sides to every question and try to look at them both.
- I would describe myself as a pretty soft-hearted person.
- When I watch a good movie, I can very easily put myself in the place of a leading character.
- I tend to lose control during emergencies.
- When I'm upset at someone, I usually try to "put myself in his shoes" for a while.
- When I am reading an interesting story or novel, I imagine how I would feel if the events in the story were happening to me.
- When I see someone who badly needs help in an emergency, I go to pieces.
- Before criticizing somebody, I try to imagine how I would feel if I were in their place.

How did you decide how much you would be willing to pay for not having to eat an insect?

Based on which criteria did you decide how much money to pay for the other participants?

Do you have any further comments?

Thank you for your participation!

We will begin the payment shortly. Please wait on your seat until your cubicle number is called and then come to the adjoining room to receive your payment.

Appendix 2.F Screenshots

Figure 2.F.1: Screenshot of Part 1

Figure 2.F.2: Screenshot of Part 2

Figure 2.F.3: Screenshot of Part 3

Figure 2.F.4: Screenshot of Part 4

Chapter 3

Diffusion of Being Pivotal and Immoral Outcomes*

Joint work with Armin Falk and Nora Szech

Abstract

We study how the diffusion of being pivotal affects immoral outcomes. In our main experiment, subjects decide about agreeing to kill mice and receiving money versus objecting to the killing and foregoing the monetary amount. In a baseline condition, subjects decide individually about the life of one mouse. In the main treatment, subjects are organized into groups of eight and decide simultaneously. Eight mice are killed if at least one subject opts for killing. The fraction of subjects agreeing to kill is significantly higher in the main condition compared with the baseline condition. In a second experiment, we run the same baseline and main conditions but use a charity context and additionally study sequential decision-making. We replicate our finding from the mouse paradigm. We further show that the observed effects increase with experience, i.e., when we repeat the experiment for a second time. For both experiments, we elicit beliefs about being pivotal, which we validate in a treatment with non-involved observers. We show that beliefs are a main driver of our results.

*We thank K. Albrecht, S. Altmann, R. Bénabou, T. Dohmen, J. Engel, D. Engelmann, S. Gächter, P. Heidhues, D. Huffman, S. Jäger, F. Kosse, B. Kőszegi, F. Krämer, G. Loewenstein, F. Rosar, J. Sobel, N. Schweizer, F. Zimmermann and participants at various seminars for helpful comments. We thank M. Antony, T. Graeber, T. Wölk, and, in particular, S. Walter for excellent support. Falk acknowledges financial support by the German Research Foundation (DFG) through the Leibniz Program and by the European Research Council (ERC Advanced Grant 340950). The study was approved by the ethical committee of the University of Bonn (reference number: 066/12).

3.1 Introduction

This chapter studies how groups favor moral transgression in diffusing responsibility and notions of being pivotal. Intuitively, acting in groups provides an excuse for acting immorally simply because an individual may perceive himself as not or only partly responsible for an outcome. To investigate the consequences of group settings that diffuse being pivotal, we ran two sets of experiments, varying the choice environment and contrast environments where subjects are fully pivotal with contexts where being pivotal is diffused by an exogenous change in organizational design. In the latter, subjects are organized into groups and individual decisions are aggregated such that the individual can easily believe that his decision is unlikely to be pivotal. Organizing people into groups and implementing a decision rule that does not require the support of all members for immoral action enables a simple "replacement logic" (see Sobel, 2010), which denotes the procedural phenomenon whereby people can mutually excuse their immoral behavior with individual powerlessness in the face of others' immoral behavior.

In our main experiment, the paradigm involves the trade-off between life and money. Subjects decide between receiving money and agreeing to kill mice versus not receiving money and objecting to the killing.[1] Importantly, mice used in the experiment are so-called "surplus" mice, all of which would have been killed without our intervention (see Section 3.2). Subjects learn about this default in a post-experimental debriefing. The paradigm is informed by the widely held view that harming others in an unjustified and intentional way is considered immoral.[2] We contrast two treatments: the *Baseline* treatment implements a simple binary choice where subjects either receive €0 for saving a mouse (Option A) or €10 for killing the mouse (Option B). In Baseline, subjects are hence fully pivotal. This condition serves as a comparison benchmark for the main *Simultaneous* treatment, in which eight subjects simultaneously decide between Option A and Option B. As in Baseline, a subject receives no money for choosing Option A and €10 for choosing Option B, irrespective of the other subjects' choices. However, if at least one subject chooses Option B, eight mice are killed. Thus, if a subject believes that at least one other subject is likely to choose Option B, he may no longer consider himself as being pivotal. In line with our argument, we find that the fraction of subjects choosing Option B is significantly higher in Simultaneous than in Baseline, even though—upon being pivotal—killing causes the death of eight mice rather than one. Moreover, the likelihood that a subject chooses to kill mice decreases with his belief of being pivotal. At the aggregate level, all mice are killed in Simultaneous.

Our second choice paradigm involves the binary decision between receiving €10 for oneself or donating €15 to a charity that supports children suffering from cancer. We replicate the two treatments from the first experiment as closely as possible (*BaselineC*

[1] The study was approved by the ethical committee of the University of Bonn.

[2] See, e.g., Gert (2012, Section 1) on "The Definition of Morality," The Stanford Encyclopedia of Philosophy: "In this descriptive sense, although avoiding and preventing harm is common to all, 'morality' can refer to codes of conduct of different societies with widely differing content, and still be used unambiguously."

and *SimultaneousC*) and additionally investigate experience effects, i.e., whether the observed effects become larger if subjects repeat the same experiment one more time. For completeness, we also study a dynamic setting of diffusion of responsibility that mimics SimultaneousC but in which decisions are made sequentially (*SequentialC*). The charity experiment replicates the main effect found for the mouse conditions. The share of subjects choosing the selfish Option B is significantly higher in the simultaneous condition than in the baseline treatment. Moreover, choosing a second time in BaselineC on average does not affect the likelihood of donations but—as expected—induces more selfish choices in both SimultaneousC and SequentialC. In the latter, we additionally find that previous history matters for behavior. In particular, learning that Option B has already been chosen essentially eradicates the choice of Option A further down the line.

Perceptions of being pivotal are central to the mechanism under study and they critically hinge on beliefs about the behavior of others. This is why, in both experiments, we elicit beliefs and confirm that choices are strongly associated with the perceived likelihood of being pivotal. Given the critical role of beliefs, we ran a further treatment with non-involved subjects. In this condition, subjects read the instructions of all three treatments implemented with the charity paradigm and were asked to predict the results from the experiment. These independently elicited beliefs of spectators corroborate our abovementioned findings. In particular, we find that the beliefs of spectators are very similar to those of subjects making a decision.

Organizational contexts that generate replacement arguments are pervasive at various levels of social interaction. They range from state-organized violence and corrupt bureaucracies to cheating in sports, morally dubious market transactions, and malpractice within corporations. We discuss a few examples below. Some examples are more closely related to our simultaneous condition, others to the sequential choice context. Most real-world examples, however, share features of both. In this sense, our experimental group treatments represent limiting cases, where subjects decide either in isolation and complete uncertainty about other individuals' actions (Simultaneous and SimultaneousC) or with perfect information about previous choices and the exact timing and order of moves (SequentialC).

A striking example that closely corresponds to our simultaneous conditions is the practice of firing squads, which comprise of a group of executioners rather than a single person. For all members, shooting entails the personal advantage of avoiding disciplinary measures, and "technologically" one person who shoots his gun is sufficient to bring about the killing. From an individual member's perspective, being pivotal is diffused, as many people shooting at the same time implies that the killing is likely to happen, regardless of whether a particular member fires his gun or not. Moreover, members of firing squads are often randomly issued a gun containing a blank cartridge, which additionally diffuses being pivotal: even if a member of the squad shoots his gun, he remains uncertain whether or not he can effectively cause the killing at all. Apparently, these features reduce feelings of responsibility and facilitate participating in executions.

Corruption is another example that closely resembles the simultaneous decision-making context. Suppose that a citizen wants to gain illegitimate access to a public permit and

therefore intends to bribe an official. He may approach different officials, but he only needs to find one single official who accepts the bribe. Since any official taking the bribe would do so secretly, there is no way to credibly signal honesty. If a given official is sufficiently certain that at least one of his colleagues is corrupt, he may now feel tempted to accept the bribe himself. This logic can give rise to an equilibrium where a large proportion of officials act corruptly. Doping in sports provides a similar example. Most athletes publicly state that they detest doping. However, many are later found guilty, with the road cyclist Lance Armstrong being an infamous example. This places athletes in a dilemma. They might generally object to cheating—at least because it jeopardizes the credibility of their discipline—but believe that others are doped anyway, which makes it seem more acceptable or even necessary to engage in doping themselves.

Reasoning about not being pivotal also helps to explain outcomes in markets that violate traders' own moral or fairness preferences. Here, a replacement argument prevails if traders prefer concluding a trade themselves over letting another trader perform the same transaction, even if trading creates unfair outcomes among the traders or imposes negative externalities on others. In cases where buying decisions create negative externalities, a frequent "excuse" is that "if I don't buy, another buyer will." On the opposite side of the market, suppliers of potentially harmful goods are in a similar situation, arguing that market demand would be met with or without their involvement. British Secretary of State Boris Johnson invoked an argument along these lines in October 2016 after allegations about weapons exported to Saudi Arabia being used for war crimes in Yemen. Faced with a motion in the House of Commons to suspend sales, he retorted that the respective members of parliament should "be in no doubt that we would be vacating a space that would rapidly be filled by other Western countries who would happily supply arms with nothing like the same compunctions or criteria or respect for humanitarian law" (Peck, 2016).[3]

The replacement logic also contributes to corporate crime. For example, Andrew Fastow—chief financial officer (CFO) of Enron from 1998 until 2001, who played a central role in concealing massive losses before the firm's bankruptcy in 2001 and served a six-year sentence in prison—himself drew the following parallel: "But the reality is, if at any point in my career I said 'time out, this is bullshit, I can't do it'... they would have just found another CFO, but that doesn't excuse it. It would be like saying it's OK to murder someone because if I didn't do it someone else would have" (Soltes, 2016, p. 255). The above quote underscores our main hypothesis, while it also highlights that behavior in response

[3]This is a refined version of the discussed argument in pointing at positive "side effects" associated with the United Kingdom taking an active role (see Glover and Scott-Taggart, 1975, pp. 177). Yet, the latter might often represent mere excuses rather than sound justifications.

to uncertainty about being pivotal may nevertheless be perceived as morally repulsive.[4] Note that the replacement logic draws on consequentialist moral thinking. By contrast, deontological moral reasoning would dictate doing the "right" thing regardless of being pivotal or not. The extent to which groups are vulnerable to transgression therefore crucially depends on the share of individuals following consequentialist versus deontological moral reasoning, respectively. We discuss the relative shares in the context of our experiments in Sections 3.3 and 3.4.

Our chapter is related to work on contextual factors affecting fair outcomes in the context of simple dictator, bargaining, or allocation games. While we focus on the role of beliefs about being pivotal, other mechanisms that have been identified to favor "unfair" outcomes are delegation or exploiting moral "wiggle rooms," as discussed, e.g., in Bartling and Fischbacher (2012), Dana, Weber, and Kuang (2007), Hamman, Loewenstein, and Weber (2010), and Serra-Garcia and Szech (2018).[5] Falk and Szech (2013) analyze the malleability of moral outcomes in bilateral and multilateral market situations and Falk (2017) studies the role of status inequality.

The diffusion of being pivotal can be interpreted in terms of higher expected costs of acting morally because, in our group treatments, the probability of reaching a moral outcome when acting morally and foregoing the additional payment is smaller than one. In this sense, our findings are related to work by Andreoni and Miller (2002) and Fisman, Kariv, and Markovits (2007), who show that when exogenously varying the price of giving in simple dictator games, the observed willingness to share varies accordingly. Two important features differentiate our setup from this literature. First, we contrast a monetary benefit for oneself with a *moral* good. It remains to be shown whether people readily engage in trade-offs here as well. Second, we do not set the probability of being pivotal exogenously but it is determined endogenously by the behavior of others, giving rise to equilibrium considerations.[6] Another related strand of literature in social psychology concerns the so-called bystander effect (see, e.g., Latané and Darley [1968] and for a recent overview Fischer et al. [2011]). Typical bystander experiments study helping behavior in response to a staged emergency (e.g., the experimenter becomes injured). What sets our simultaneous treatments apart is that even if a subject opts for the moral outcome, he remains uncertain about whether the moral outcome is implemented or not, similar, e.g., to firing squads. By contrast, in typical bystander experiments, this uncertainty does not

[4] Another example in this vein is the role of the replacement logic in the organization of the Holocaust (Arendt, 1963; Darley, 1992; Lifton, 1986). Lifton (1986) interviewed German doctors stationed in Auschwitz. They were operating in a nightmarish environment, with one of their objectives being to "select" prisoners who would be allowed to live while others would be immediately gassed. Being ordinary doctors, this activity was likely to be morally terrible and self-contradictory to them. Nevertheless, they engaged in the selection procedures. One of the frequently made justifications was that the "horrible machinery would go on," regardless of whether a particular doctor continued to participate. Replacement arguments suggesting the impossibility to stop ongoing moral crime were also used in the Nuremberg Trials as excuses for having participated in various kinds of atrocity under the Nazi Regime (see, e.g., Crawford [2007] and references therein).

[5] On the effects of institutions on values, see also Bowles (1998). On the role of authority, see Milgram [1974] (2009).

[6] For an equilibrium analysis of group decisions in morally relevant contexts, see Rothenhäusler, Schweizer, and Szech (2018).

exist. If a subject opts for helping, the person in need receives help. Furthermore, in a bystander experiment, while deliberating whether to help or not, subjects often observe that others do not help either. In our simultaneous-move setup, this type of social learning is ruled out. When deciding to kill a mouse or not to donate, respectively, subjects do not know whether other subjects also opt for the selfish option. The dynamic properties of observing others, however, are explicitly studied in our sequential treatment. Also, in a bystander experiment, participants need to realize that their help is required (and that it is better to step in than to hope that some other, say, more able helper will step in), while in our setup the consequences of decisions are straightforward. We also note that in our experiment, consequences are real, incentives are exactly specified, and the mechanism (beliefs about being pivotal) is explicitly measured.

The remainder of the chapter is organized as follows. In Section 3.2, we describe the design and implementation of the main experiment and develop our hypotheses. The results are presented in Section 3.3. Section 3.4 covers the charity experiment. We first present a replication of our main results and provide evidence for the validity of elicited beliefs. We then proceed by investigating an additional sequential condition. Finally, Section 3.5 concludes by summarizing the chapter and discussing additional observations.

3.2 Experiment

Avoiding and preventing unjustified harm is central to most notions of morality. It is this notion that informs the "mouse paradigm" used in our main experiment, which involves the trade-off between killing a mouse and receiving money versus saving a mouse life and receiving no money (Falk and Szech, 2013).[7] Subjects are explicitly informed that each mouse is a young and healthy mouse that will live for about two years if saved. For illustrative purposes, we present subjects the picture of a mouse on an instruction screen. We guarantee subjects that mice—if saved—live in an appropriate, enriched environment, jointly with a few other mice. Hence, in case subjects decided to save mice, these mice were kept alive in an enriched environment, with good feed and comfortable nesting material, precisely as stated in the instructions.

3.2.1 Design

Subjects are also informed in detail about the killing process. In the instructions (see Supplementary Appendix 3.C), they read the following passage: "[T]he mouse is gassed. The gas flows slowly into the hermetically sealed cage. The gas leads to breathing arrest. At the point at which the mouse is not visibly breathing anymore, it remains in the cage

[7] Deckers et al. (2016) provide convergent and discriminatory validity of the mouse paradigm as a measure for morality. Killing is negatively related to agreeableness—one of the Big Five facets—which describes a tendency to be compassionate and cooperative rather than suspicious and antagonistic towards others, and positively related to Machiavellianism, measuring a person's tendency to be unemotional and detached from conventional morality. Moreover, killing is not related to disposable income, whether students are professionally involved with animal research or animal experiments, or have a simple preference for animals, as expressed by having a pet at home.

for another 10 minutes. It will then be removed." To further rule out uncertainty about the decision context, subjects are shown a short demonstration video of the killing process. In the video, four mice first move vividly in the cage, then they successively slow down as more and more gas enters the cage. Eventually, they die, with their hearts visibly beating heavily and slowly.

It is important to stress that the mice used in the experiment were so-called "surplus" mice: these mice were bred for animal experiments but proved to be unsuited for scientific research. They were perfectly healthy, but keeping them alive would have been costly. It is common practice in laboratories conducting animal experiments to gas such mice. Thus, as a consequence of our experiment, many mice that would have otherwise all died were saved. Subjects were informed about this default in a post-experimental debriefing.[8]

Treatments

We study the role of diffusion of being pivotal in contrasting two decision environments, one where subjects are fully pivotal (Baseline) and one where being pivotal is diffused by organizing subjects into groups (Simultaneous). The two decision contexts differ in terms of how likely it is that any given subject is pivotal, keeping overall moral and financial consequences identical. In Baseline, each subject decides about the life of one mouse. Subjects face a simple binary choice between Option A and Option B: Option A implies that the mouse will survive and that the subject receives no money, while Option B implies the killing of the mouse and receiving €10. The Baseline treatment informs us about the share of subjects who are willing to kill the mouse for €10 when obviously being pivotal.

In Simultaneous, subjects decide in groups of eight and are endowed with eight mice. As in Baseline, each subject faces an individual binary choice between Option A and Option B: Option A implies that a subject receives no money. If a subject chooses Option B, he receives €10. Individual monetary consequences are independent of other subjects' decisions. All subjects choose simultaneously. They know that if at least one subject chooses Option B, all eight mice are killed. Furthermore, they know that they will not receive feedback on whether the mice are ultimately killed or not (although it is obvious for a subject that the mice die if he chooses Option B). Note that we chose to endow a group with eight mice to keep the number of mice at the aggregate level identical to Baseline. Of course, the valuation of mice lives need not be proportional to the number of saved mice, but keeping numbers identical at the aggregate level allows for a clean comparison of the overall impact of group versus individual decision-making.

In Simultaneous, right after subjects have made their decision, we elicit beliefs about being pivotal. Subjects are asked to indicate the probability that all other seven group members have chosen Option A (*belief_pivotal*). We also ask subjects to estimate how many other subjects in their group have chosen Option B. They can enter any number

[8] While perceptions of the situation may have changed due to this information, the consequences were exactly the same and as stated in the instructions. In future research, it would be interesting to explore whether using an alternative framing would affect decisions in response to institutional changes differently (compare evidence on the so-called omission–commission bias, e.g., in Spranca, Minsk, and Baron [1991]).

from 0 to 7 and are paid €1 for a correct estimate ($belief\_B$).

Procedure

Two hundred and fifty-two subjects—mainly undergraduate university students from all majors—took part in the experiment, 124 subjects in Baseline and 128 in Simultaneous. Each subject participated only in one treatment condition. We used z-Tree as the experimental software (Fischbacher, 2007). Subjects were recruited using the software ORSEE (Greiner, 2004). At the beginning of an experimental session, participants received detailed information about the rules and the structure of the experiment. In all treatments, the experiment started only after all participants had answered several control questions correctly.

To reduce possible communication between subjects across sessions, the experiment was run on two consecutive days in six different rooms at the *Beethovenhalle*, the largest concert hall in Bonn. We set up six parallel, computerized labs in these rooms. Subjects received payments according to the rules of the experiment and an additional show-up fee of €20 to compensate for the remote location. In both treatments, subjects received their payments in a sealed envelope outside the room where the experiment had taken place. This way, neither other subjects nor the experimenters handing over the envelopes knew what a particular subject had earned. This procedure was explained in the instructions.

To ensure credibility, we stated right at the beginning that all statements made in the instructions were true—as is standard in economic experiments—and that all consequences of subjects' decisions would be implemented exactly as described in the instructions. We emphasized orally that the experimenters would personally guarantee the truthfulness of the instructions. Subjects were also invited to send us an email if they wanted to discuss the study.

3.2.2 Hypotheses

Our predictions start from the premise that most subjects follow consequentialist reasoning rather than deontological prescriptions. We expect that subjects in the Simultaneous treatment will engage in strategic considerations, thinking about how other subjects will decide. If they come to the conclusion that the likelihood of being pivotal is sufficiently small, subjects will find it justifiable to opt for the morally problematic Option B. Consequently, we would expect a higher share of subjects opting to kill in the group treatment compared with Baseline, in which subjects know that they are pivotal for certain.

To fix ideas, we normalize the utility from receiving €10 to one and the utility from receiving €0 to zero. There is a subjective moral cost $c_{n,i}$ for subject i associated with the death of $n = 1$ or 8 mice, respectively. Furthermore, we denote by $belief\_pivotal_i \in (0, 1]$ the subjective belief about the probability of being pivotal. If a subject chooses Option A and proves to be pivotal—i.e., killing is averted—utility is given by 0. Otherwise, the resulting level of utility is $-c_{n,i}$. The subjectively expected utility from choosing Option A therefore amounts to $-(1 - belief\_pivotal_i)\, c_{n,i}$. The utility from choosing Option B

is always given by $1 - c_{n,i}$. In making their decisions, deontological subjects disregard cost–benefit considerations and always choose Option A.[9] Any consequentialist subject chooses Option B if and only if the respective utility is at least as large as the subjectively expected utility from Option A or—equivalently—if $c_{n,i} \leq \textit{belief\_pivotal}_i^{-1}$. Obviously, in the individual decision context, it holds for all subjects that $\textit{belief\_pivotal}_i = 1$. By contrast, the belief about the chance of being pivotal in the simultaneous condition depends on beliefs about the behavior of the other subjects in the same group.

This recursive relationship between subjects' decisions in Simultaneous can be understood as a strategic game between eight players whose types are characterized by their subjective moral costs $c_{8,i}$ and their respective moral conceptions, i.e., whether they are deontologists or consequentialists. Types are independently drawn, with $d > 0$ denoting the probability of a subject following deontological ethics and the distribution F of moral costs c_8 being continuous and having full support on the interval (a, b), with $a < 1$ and $b > d^{-7}$. If we additionally impose that subjects hold correct beliefs given by $\textit{belief\_pivotal}_i = p_i$, we can apply the concept of Bayesian equilibrium. According to the above discussion, individual behavior follows a cut-off strategy in which an agent chooses Option B if $c_{8,i} \leq k_i$, with $k_i = p_i^{-1}$, and Option A otherwise. In our setup, a Bayesian equilibrium must feature strategies that are symmetric, i.e, $k_i = k^\star$ for all agents. If any two agents within the same group used cut-off values that were different, the chance of being pivotal would be weakly higher for the agent whose cut-off value was higher. However, a weakly higher probability of being pivotal would imply that the cut-off value should be weakly lower, which is a contradiction.[10]

Consider a candidate k for an equilibrium cut-off value $k^\star$. In conjunction with the distribution of types, it implies a probability of being pivotal, which is given by $p(k) = \{d + (1 - d)\,[1 - F(k)]\}^7$. For an equilibrium cut-off value, a marginal subject for whom $c_{8,i} = k^\star$ needs to be indifferent between the two choice options. An equilibrium cut-off value is thus a fixed point for which $k^\star = p(k^\star)^{-1}$, i.e.,

$$k^\star = \{d + (1 - d)\,[1 - F(k^\star)]\}^{-7}\ . \tag{3.1}$$

The precise number and location of equilibria depends on the distribution of moral types. However, note that $p(k)^{-1}$ is not only strictly increasing in k but also continuous and its values range from 1 to d^{-7}. Thus, equilibrium cut-off values lie in the interval $\left[1, d^{-7}\right]$

[9] Alternatively, one could assume that deontologists take into account moral costs but always act as if they were deciding alone, i.e., they deliberately abstain from equilibrium considerations. Indeed, Kant's categorical imperative requires people to "[a]ct only in accordance with that maxim through which you can at the same time will that it become a universal law" (cited from Kant, 1996, p. 73). Deontologists would then choose Option B if $c_{n,i} \leq 1$ and Option A otherwise (see also Roemer, 2010, 2015). The consequences for our analysis would be minor. For a discussion of the differences between consequentialist versus deontological reasoning, see Bénabou, Falk, and Tirole (2018) and Bénabou et al. (2020).

[10] Formally, assume that strategies are *not* symmetric. Players $1, \ldots, 8$ form a group and—without loss of generality—it holds for their cut-off values that $k_1 < k_8$. For each agent, $p_i = \prod_{j \neq i} \{d + (1 - d)\,[1 - F(k_j)]\}$. It follows that $p_1 \leq p_8$ and thus $k_1 = p_1^{-1} \geq p_8^{-1} = k_8$, which gives the contradiction.

and, by the intermediate value theorem, an equilibrium exists.[11]

As can be seen from equation (3.1), any equilibrium cut-off value $k^\star$ is always weakly larger than one, the latter being the cut-off under individual decision-making. In any equilibrium, the share of subjects choosing to kill is *strictly* larger than under individual decision-making as long as there exist any consequentialists ($d < 1$), for whom we have assumed that some have moral costs smaller or equal than one ($F(1) > 0$). Intuitively, some subjects choosing Option B even when fully pivotal reduce the likelihood of being pivotal for others, causing subjects with moral costs just above one to also choose Option B. Depending on the precise distribution of moral costs and the prevalence of deontologists, this leads other subjects with still higher moral costs to adjust their behavior as well. In practice, the described moral unravelling will most likely reach an equilibrium only after some time and learning, similar to related experimental findings in, e.g., market experiments where reaching an equilibrium typically requires several rounds of repetition. Even if an equilibrium has not been reached, however, the described moral unravelling suggests that the share of subjects choosing Option B should be higher in Simultaneous than in Baseline. This is our first hypothesis.

Hypothesis 3.1. *The share of subjects choosing Option B—thereby taking €10 and agreeing to kill—will be higher in Simultaneous than in Baseline.*

It is worth noting that as long as—for each individual—moral costs $c_{8,i}$ of killing eight mice are higher than moral costs $c_{1,i}$ of killing just one mouse, we tend to underestimate the role of being less pivotal in groups relative to Baseline. We could have endowed groups only with one mouse. In this case, we would expect even larger treatment effects. We opted for eight mice, however, to keep the maximum possible extent of harm fixed at the aggregate level when comparing treatments.

To the extent that an equilibrium has not been reached, subjects will most likely hold heterogeneous beliefs about the likelihood of being pivotal. We elicit these beliefs as part of our experimental procedure. According to the decision rule for consequentialists, the heterogeneity in beliefs should translate into corresponding differences in decisions, which is our second hypothesis.

Hypothesis 3.2. *In the Simultaneous treatment, the likelihood that a given subject opts for taking €10 and killing the mice decreases with the subjective probability assigned to being pivotal.*

In sum, the diffusion of being pivotal in groups leads consequentialist subjects to adjust their behavior. The probability of being pivotal becomes small, making immoral behavior more attractive than when deciding individually. In addition, individual heterogeneity in the belief about the probability of being pivotal should translate into corresponding

[11] Formally, in equilibrium it has to hold that $p(k^\star)^{-1} - k^\star = 0$. Observe that $p(a) - a = 1 - a > 0$ and $p(b) - b = d^{-7} - b < 0$. Since the function $p(k) - k$ is continuous, it follows from the intermediate value theorem that an equilibrium exists.

propensities to choose Option B. Hence, we expect that, on average, Option B is chosen more often in Simultaneous than in Baseline, and that—at the individual level—the likelihood of choosing Option B is inversely related to perceptions of being pivotal.

3.3 Results

In presenting the results of our main experiment, we start with a treatment comparison. We then explicitly study the role of beliefs about being pivotal. According to our model, subjective beliefs along with observed choices imply bounds for each subject's individual moral costs. We use the joint distribution of beliefs and choices to estimate the distribution of subjective moral costs in the population and the prevalence of deontologists. Finally, we explore the implications of our estimates for welfare as well as for the equilibrium to which behavior should ultimately converge.

3.3.1 Choices and Beliefs

Our main result from the mouse experiment is shown in Figure 3.1, where we compare the shares of subjects choosing to kill in Baseline and Simultaneous, respectively. In Baseline, 46.0% of subjects choose Option B. In Simultaneous, the respective share is 58.6%, implying a difference of about 27%. This difference is significant ($p = 0.04$, two-sample test of proportions, two-sided) and confirms Hypothesis 3.1. At the aggregate level, the group impact is striking. While 46% of mice are killed in Baseline, *all* mice are killed in *all* groups in Simultaneous.

We have argued above that individual perceptions of being pivotal are critical in driving the increase in selfish behavior in Simultaneous. Accordingly, we should observe that an individual's willingness to choose Option B decreases with his belief of being pivotal. This is indeed what we find. Recall that we asked subjects about the probability that all other group members had chosen Option A (*belief_pivotal*). Figure 3.2 displays the fraction of subjects choosing Option B depending on this belief. The four categories in Figure 3.2 are based on quartiles of the belief distribution with respective percentage intervals of $[0, 3.5]$, $(3.5, 10]$, $(10, 35]$, and $(35, 100]$. In line with Hypothesis 3.2, the figure shows a clear negative relation between subjective perceptions of being pivotal and the likelihood of choosing Option B (Spearman rank correlation: -0.54, $p < 0.001$).[12]

3.3.2 Implied Moral Costs

In light of our formal framework introduced in Section 3.2.2, the observed heterogeneity in subjective beliefs about being pivotal provides a chance to estimate the distribution of moral costs—within the relevant choice context and subject population. Suppose, e.g., that a consequentialist subject assigns a chance of 50% to the event of being pivotal. If the subject chooses Option B, one can infer that moral costs $c_{8,i}$ are at most $belief\_pivotal_i^{-1} = 2$.

[12] The values of *belief_pivotal*—which we use here—and those of the incentivized *belief_B* are strongly and significantly correlated (Spearman rank correlation: -0.63, $p < 0.001$). The relationship between *belief_B* and choice of Option B is shown in 3.B.1 and confirms the results presented here.

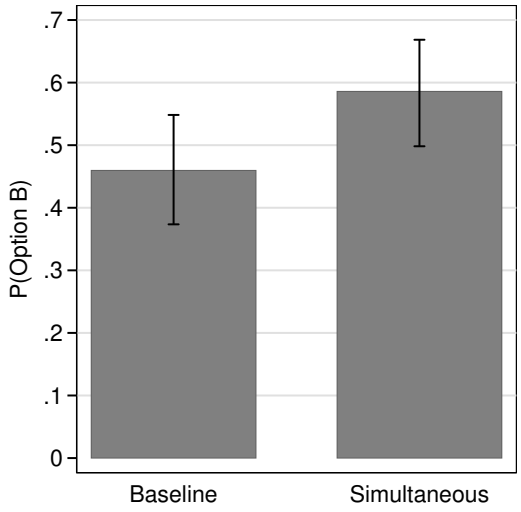


Figure 3.1: Treatment comparison

Notes: Share of subjects choosing Option B in Baseline and Simultaneous. Error bars show 95% confidence intervals (based on logit transformations).

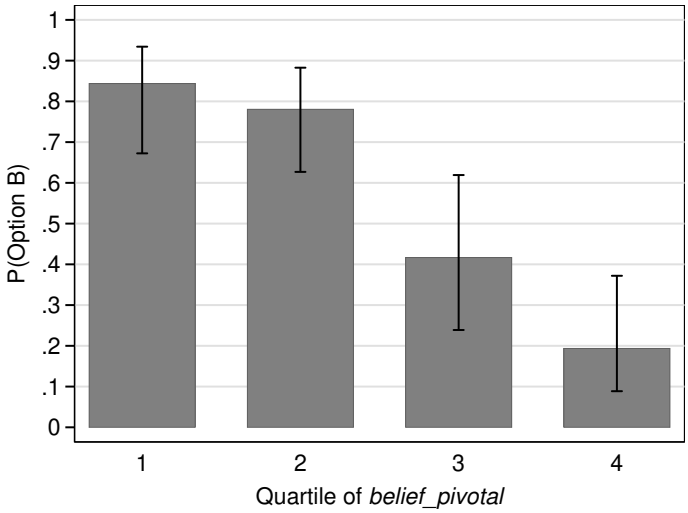


Figure 3.2: Belief quartiles (Simultaneous)

Notes: Share of subjects in Simultaneous choosing Option B depending on their belief of being pivotal. Error bars show 95% confidence intervals (based on logit transformations).

Conversely, if the subject chooses Option A, moral costs must be larger than two. To draw inferences about the distribution in the population, we make the following assumption.

Assumption 3.1. *Moral costs are independent of the perceived likelihood of being pivotal.*

Then the share of consequentialist subjects choosing Option B among all those who believe that they are pivotal with a probability of 50% identifies the value of the distribution function F of moral costs at two. Similarly, the share of subjects choosing Option B among those who believe that they are pivotal with a probability of 25% identifies the value of the distribution function at four, and so on.

To be able to estimate the full distribution of moral costs, we need to impose some additional structure.

Assumption 3.2. *The subjective moral costs of consequentialist subjects follow a log-normal distribution F, with log-costs having mean μ and standard deviation σ.*

We can now write the probability of a given subject choosing Option B in terms of the cumulative distribution function of the standard normal distribution.

$$P(\text{Option B} \mid \textit{belief\_pivotal}_i) = \begin{cases} \Phi\left(\frac{\ln\left(\textit{belief\_pivotal}_i^{-1}\right)-\mu}{\sigma}\right) & \text{for consequentialists} \\ 0 & \text{for deontologists} \end{cases} \tag{3.2}$$

Next, consider a finite mixture model with two latent classes, one capturing consequentialists and the other deontologists. For consequentialists, a probit model is estimated that regresses the likelihood of choosing Option B on the log of the inverse probability of being pivotal and a constant. For deontologists, the probability of choosing Option B is always zero.

$$P(\text{Option B} \mid \textit{belief\_pivotal}_i) = \begin{cases} \Phi\left[\beta_0 + \beta_1 \ln\left(\textit{belief\_pivotal}_i^{-1}\right)\right] & \text{for consequentialists} \\ 0 & \text{for deontologists} \end{cases}$$

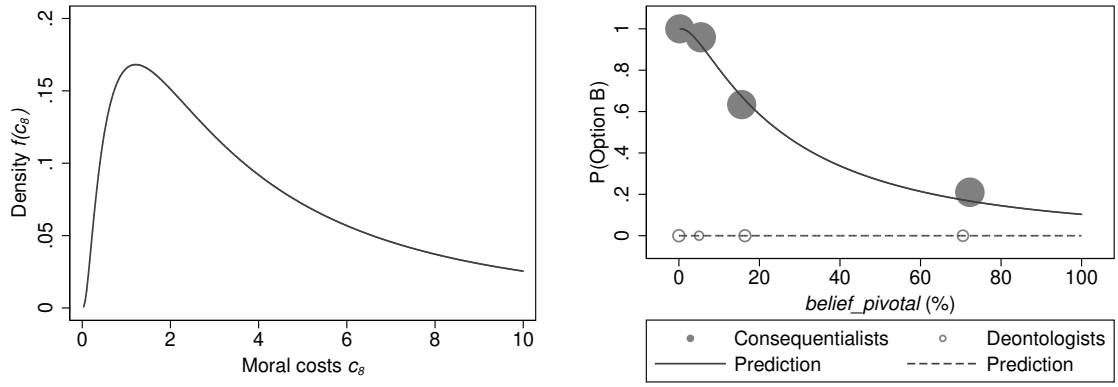
In conjunction with equation (3.2), it follows that

$$\sigma = \beta_1^{-1} \qquad \text{and} \qquad \mu = -\frac{\beta_0}{\beta_1}\,.$$

We estimate this finite mixture model using the expectation–maximization (EM) algorithm, assigning subjects to latent classes in terms of probabilities.[13] The invariance property of maximum likelihood estimates then allows us to convert the point estimates for coefficients into estimates for the parameters of F, as described above.

Figure 3.3 visualizes the results. The left panel shows the density function f of moral costs c_8. The underlying estimates for the distributional parameters are $\hat{\mu} = 1.37$ and $\hat{\sigma} = 1.09$, corresponding to the mean and the standard deviation of log-costs, respectively. The expected value of moral costs is given by 7.098. We further estimate that the share of deontologists within our population of subjects is 13.6%, which is quite close to 17.9%

[13] If *belief_pivotal* is reported as 0%, we treat it as 0.1%.



(A) Estimated moral costs (B) Observed versus predicted decisions

Figure 3.3: Moral costs (Simultaneous)

Notes: The left panel shows the estimated probability density function for moral costs c_8 of consequentialists in the Simultaneous treatment, denoted in multiples of the utility from receiving €10. The right panel plots the implied probabilities of choosing Option B for different beliefs about the chance of being pivotal against observed shares in the experiment. The solid line depicts the predictions for consequentialists, which are given by $F\left(belief\_pivotal^{-1}\right)$. Predictions for deontologists are given by zero and shown as the dashed line. For representing the data, subjects are first partitioned into quartiles of the belief distribution. Then, separately for consequentialists and deontologists within each quartile, probability-weighted average beliefs and shares of Option B are calculated. The sizes of bubbles correspond to estimates for the expected numbers of subjects.

of subjects choosing Option A despite being certain that they will not be pivotal.[14] The right panel uses these estimates to predict subjects' choices depending on their subjective beliefs about being pivotal for consequentialists (solid line) and deontologists (dashed line). Bubbles show observed choice probabilities by quartiles of the belief distribution, again separately for consequentialists and deontologists (solid and hollow, respectively). Deontologists never choose Option B. Consequentialists always choose Option B if they are certainly not pivotal, but this probability decreases to 10.3% if they believe that they are pivotal for sure. If subjects were deciding individually—as in Baseline—but about the lives of eight mice rather than just one, these estimates imply that 8.9% of them would choose Option B, which is much lower than the observed 46.0% opting to kill in Baseline.

3.3.3 Welfare and Equilibria

To conduct a utilitarian welfare analysis based on average utility, we assume that the distribution of moral costs among deontologists is identical to the one for consequentialists, which can reasonably be interpreted as a lower bound. For ease of interpretation, we furthermore assume that utility is linear in money. Then, the average moral costs of killing eight mice across all subjects are equivalent to €70.98. Nonetheless, all mice are killed. All of those subjects who choose Option B secure a monetary payoff of €10, so that the average utility in Simultaneous (for observed behavior) is equivalent to a loss of €65.12. If all subjects had chosen Option B, it would have been equivalent to a loss of

[14] Of course, it may also be the case that some subjects made mistakes. In this sense, deontologists comprise all people whose choice behavior is unresponsive to beliefs about being pivotal.

€60.98. By contrast, if subjects were deciding alone, the utility would be weakly positive for everybody: all deontologists and those consequentialist subjects with moral costs above one (for eight mice) would choose Option A and receive a utility of zero, while consequentialists with moral costs between zero and one would opt for killing and receive utility corresponding to the subjective excess of utility from €10 over their cost of killing. The average level of utility would thus be equivalent to $(1-d)\int_0^1 (1-c)\, f(c)\, \mathrm{d}c \times$ €10, which—according to our estimates—equals €0.31. Interestingly, a utility level of zero could also have been achieved in Simultaneous, had all subjects behaved as deontologists and saved the mice. This increased efficiency captures the intuition regarding why—from an evolutionary perspective—some degree of rule-based moral behavior could indeed be expected (Alger and Weibull, 2013). However, our results point to a dominant role of consequentialist reasoning and question the relatively high fractions of Kantian types in survey data such as the trolley problem (Foot, 1967), where consequences are hypothetical rather than real.

Throughout this section, we have made use of the fact that beliefs about being pivotal are heterogeneous and generally large in comparison to actual outcomes. Both points are evidence that, in Simultaneous, no equilibrium has yet been reached. This is not surprising, given that it typically takes time and experience to arrive at correct beliefs. As a benchmark case, however, we want to conclude the analysis of the mouse experiment by analyzing the predicted equilibrium, i.e., the outcome we would eventually expect given sufficient experience and learning. As has been argued in Section 3.2.2, our experimental setup can generally feature multiple equilibria, which is also true under the additional assumption of log-normally distributed moral costs. Intuitively, this is because choices in favor of Option A by different players act as strategic complements. Any given consequentialist player with moral costs greater than one will refrain from choosing Option B as long others choose the moral option with sufficient likelihood but will behave selfishly if only few others choose Option A. In any case, there must exist at least one (interior) equilibrium, since the existence of deontologists always assures a strictly positive likelihood of being pivotal, which is enough to make some consequentialists with very high moral costs choose Option A. The extent to which their effort to save mice is enough for making yet further consequentialists save the mice as well depends on the prevalence of such high-cost individuals. We can inspect concrete equilibria by plugging our estimates from Section 3.3.2 into the equilibrium condition given by equation (3.1). 3.A provides a visualization, also including an analysis for a hypothetical smaller group size. We find that for the estimated distribution of moral types, the setting in Simultaneous has a unique equilibrium in which the share of consequentialists choosing Option A is virtually zero. Thus, the deterioration of moral behavior in Simultaneous would have been even more pronounced if subjects had held rational beliefs. We would expect convergence to this equilibrium if subjects repeatedly faced decisions like in Simultaneous. In Section 3.4.1, we will find some indication that rational updating of beliefs indeed occurs in a similar setting and that behavior changes accordingly.

3.4 Replication and Extensions

In this section, we employ a different setup. This second choice paradigm involves the binary decision between receiving €10 for oneself or donating €15 to a charity that supports children suffering from cancer. The charity treatments are essentially the same as in the mouse experiment, except that we use a different choice paradigm and study the role of experience as well as an additional sequential condition. As far as possible, we use the same design features, stake sizes (€10 for the selfish option), and wording and framing of choice optionsoptions (the instructions are provided in Supplementary Appendix 3.D). At the beginning of the experiment, subjects are made familiar with the charity, which is devoted to supporting children who suffer from cancer. In particular, the charity is engaged in psychological assistance and organizing leisure activities for children and their families, it helps with follow-up care and school-related issues, and supports parents and siblings as well as clinical research on cancer.

Charity treatments

To check the replicability of our experimental results from the mouse paradigm, we study a baseline (BaselineC, "C" for "charity") and a simultaneous group condition (SimultaneousC), analogous to the mouse conditions. In BaselineC, subjects make the binary decision to either donate €15 (Option A) or keep €10 for themselves (Option B).[15] In SimultaneousC, subjects are in groups of eight and simultaneously choose either Option A or Option B, respectively. Choosing Option B implies receiving €10 and choosing Option A receiving no money, irrespective of the choices of other group members. A donation of €120 (8 × €15) for the charity is only initiated if all group members choose Option A. If one group member or more choose(s) Option B, the donation of €120 is destroyed. To study how a dynamic setting affects the diffusion of responsibility, we further run treatment SequentialC. This treatment is identical to SimultaneousC (including payments, donation, wording, etc.), except that subjects choose sequentially. It is randomly determined at which position a subject is asked to decide, one subject being first, another second, up to position 8. Before making the binary decision (Option A or Option B), subjects are informed about their position (1 to 8) and the previous choice history, i.e., how many subjects have previously chosen A and how many have opted for B. In both SimultaneousC and SequentialC, we also elicit beliefs analogous to Simultaneous in the mouse condition. Subjects are asked to indicate the probability that all other seven group members have chosen Option A. Responses are given in percent using a slider, with higher percentages reflecting a higher perceived likelihood of being pivotal for the respective subject (*belief_pivotal*).[16] We also ask subjects to estimate how many other subjects in their group have chosen Option B, with possible responses from 0 to 7 (*belief_B*). Correct answers are remunerated with €2.

[15] Note that the design choice to donate €15 limits the plausibility of the argument that the €10 kept are spent on an alternative good cause.

[16] Beliefs are elicited in the same way in SimultaneousC and SequentialC, but we note that in the latter, beliefs will depend on position and responses are affected by previous play, e.g., getting to know that Option B has already been chosen.

To measure potential experience effects, all three conditions include a second round, which came to subjects as a surprise.[17] Subjects were told that they will make one more and final decision. In SimultaneousC and SequentialC, subjects learn whether at least one subject in their group has chosen Option B and thereby destroyed the donation and that they will make the same decision in the same group of eight, as in the first round. In SequentialC, they also know that they act in the same order, i.e., each subject chooses at the same position as before. Payoffs and consequences are identical to the first round.

Charity procedures

481 subjects—mainly undergraduate university students from all majors—took part in the experiments, 121 subjects in BaselineC, 120 in SimultaneousC and 240 in SequentialC (30 groups). Each subject participated in only one treatment condition. We used oTree as experimental software (Chen, Schonger, and Wickens, 2016). Subjects were recruited using the software ORSEE (Greiner, 2004). At the beginning of an experimental session, participants received detailed information about the rules and structure of the experiment. In all treatments, the experiment only started after all participants had answered several control questions correctly. The experiments were run at the BonnEconLab in March 2017. Subjects received a show-up fee of €10.

3.4.1 Replication and Experience Effects

We begin by presenting the results for the two treatments that correspond to the ones in our main experiment. The main findings are summarized in Figure 3.4, which displays the share of subjects choosing Option B (not to donate) in conditions BaselineC and SimultaneousC, respectively. The dark bars show results from the first round, the light bars those of the second round (which was unexpected for subjects). Two observations can be made. First, we replicate the main result from the mouse experiment using a different choice paradigm. The share of subjects choosing Option B is significantly higher in SimultaneousC than in BaselineC, with means of 58.3% and 39.7%, respectively ($p = 0.004$, two-sample test of proportions, two-sided). The increase in selfish behavior amounts to 47.0%, which is higher than the respective increase in the mouse condition. At the aggregate level, no single group in SimultaneousC effectively donated. Second, the detrimental effect of group decision-making on prosocial outcomes seems to increase with experience. Comparing the results between periods one and two reveals an increase in the likelihood of immoral choices upon learning the previous outcome of 12.5 percentage points ($p = 0.03$, comparison of means, two-sided and with standard errors clustered at the group level for the second round of SimultaneousC). In sharp contrast, moral behavior is not vulnerable to repetition in BaselineC, with an increase of Option B below one percentage point.

Analogous to the mouse experiment, we find that the association between the belief of

[17] Of our 121 subjects who took part in BaselineC, only 79 took part in an experience condition, i.e., in a second round. For the first two sessions (with 42 subjects) we only ran one round. In the analysis, we therefore either use 121 observations (Round 1) or 79 observations (Round 2), respectively.

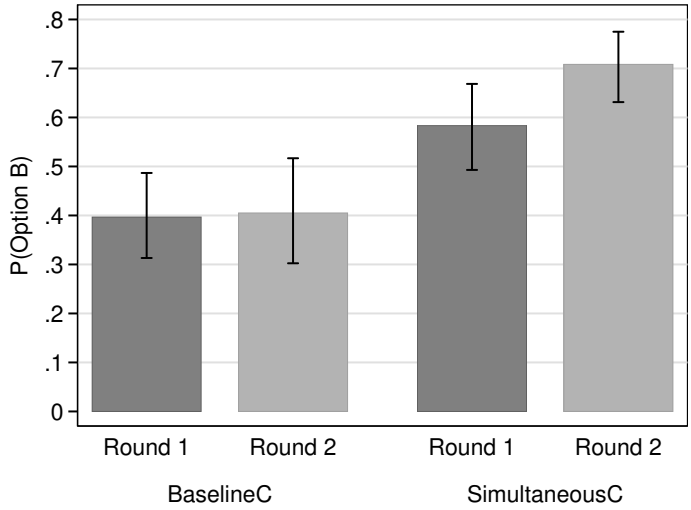


Figure 3.4: Comparison between BaselineC and SimultaneousC

Notes: Share of subjects choosing Option B in BaselineC and SimultaneousC, per round. Error bars show 95% confidence intervals (based on logit transformations), where standard errors are clustered at the group level for the second round of SimultaneousC.

being pivotal and choosing Option B is negative and statistically significant for SimultaneousC.[18] This relationship is shown in Figure 3.5, where we display the share of subjects choosing Option B depending on *belief_pivotal*. In SimultaneousC, among those who believe that they are not pivotal (estimated likelihood of being pivotal of 0%), 17.7% (three out of 17) of subjects choose Option A, presumably reflecting a Kantian kind of moral reasoning.

In both treatments, we observe some subjects switching from one choice option to the other between rounds. In the case of SimultaneousC, this switching is asymmetric, as reflected by the higher share of subjects choosing Option B in the second round. If beliefs about being pivotal are important drivers of behavior, changes in beliefs should have predictive power for switching. In Table 3.1, we regress the choice in Round 2 on the choice in the first period and the change in the belief of being pivotal. There is a significant effect in the expected direction: subjects who consider themselves less pivotal in the second period than in Round 1 indeed become more likely to choose Option B in Round 2.

To summarize, we replicate the main findings from the mouse condition. Subjects are less likely to choose the morally desired action in SimultaneousC than in BaselineC (pertaining to Hypothesis 3.1 from Section 3.2.2) and beliefs about being pivotal seem to be critical (pertaining to Hypothesis 3.2). In addition, we document that selfish outcomes in groups tend to increase with experience in contrast to individual decisions, further supporting the crucial role of beliefs about being pivotal.

[18] Again, both types of beliefs (*belief_B* and *belief_pivotal*) are significantly correlated (Spearman rank correlation: -0.35, $p < 0.001$). For results concerning the relationship between *belief_B* and choice of Option B, see 3.B.1.

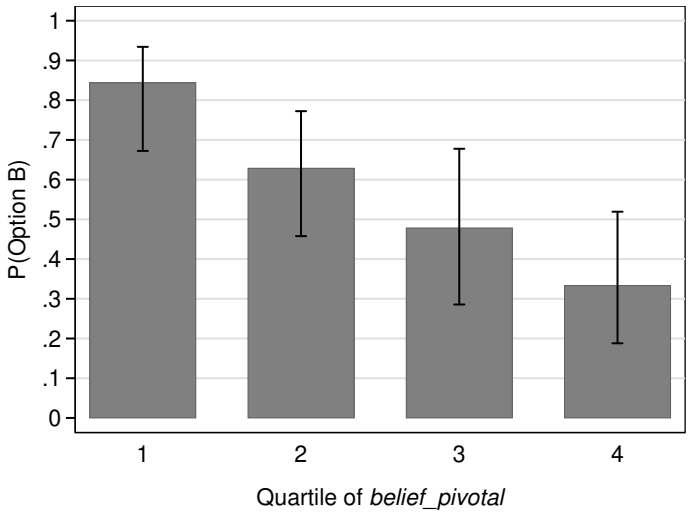


Figure 3.5: Belief quartiles (SimultaneousC)

Notes: Share of subjects choosing Option B in the first round of SimultaneousC depending on the belief of being pivotal. Error bars show 95% confidence intervals (based on logit transformations).

Table 3.1: Switching behavior

	Dependent variable: Option B in Round 2		
	OLS	Probit	Logit
	(1)	(2)	(3)
Option B in Round 1	0.324***	0.333***	0.322***
	(0.0989)	(0.0962)	(0.0973)
Decrease in *belief_pivotal*	0.00398**	0.00385**	0.00389**
	(0.00169)	(0.00162)	(0.00175)
Constant	0.498***		
	(0.0880)		
Observations	120	120	120
Clusters	15	15	15
R^2	0.131		

Notes: Columns 2 and 3 report average marginal effects and average discrete changes due the binary choice in Round 1. Standard errors are clustered at the group level. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

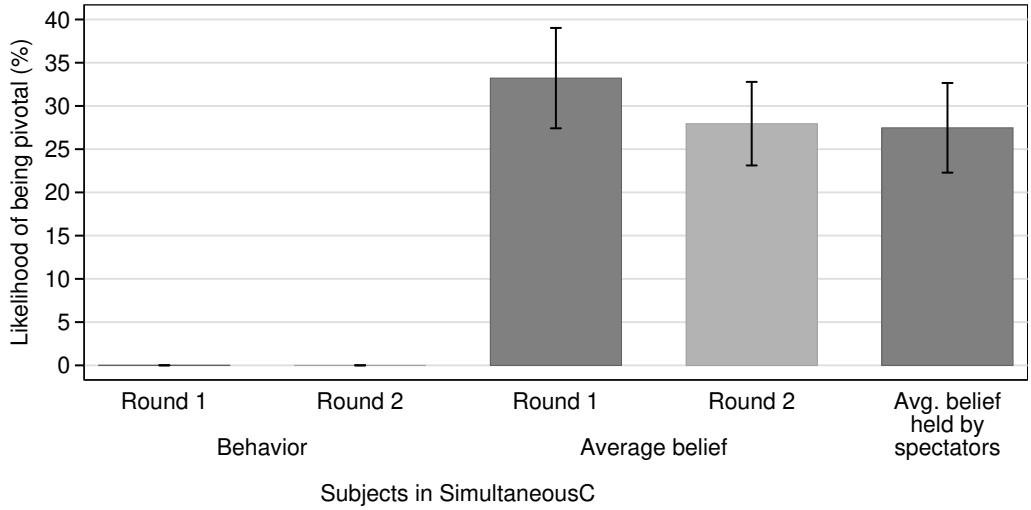


Figure 3.6: Belief comparison (*belief_ pivotal*)

Notes: Likelihood of being pivotal, i.e., the probability that all other seven members of a given subject's group choose Option A (in percent). Error bars show 95% confidence intervals, where standard errors are clustered for the second round.

3.4.2 Belief Experiment

A possible concern in interpreting beliefs is the potential endogeneity of beliefs due to motivated reasoning (Epley and Gilovich, 2016; Gino, Norton, and Weber, 2016). To limit the problem, we incentivized beliefs about the number of other participants choosing Option B in the mouse and charity treatments, such that subjects could earn additional money for accurate estimates. However, we also ran an additional belief experiment with non-involved observers. In the belief experiment, participants read the original instructions of treatments in the charity experiment (avoiding textual redundancies). We then ask them for the probability that a subject is in a group in which all other seven group members choose Option A (*belief_ pivotal*). If the percentage answer (*belief_ pivotal*) is correct within an interval of plus/minus five percentage points, they receive €2. 87 subjects participated in this condition, which was programmed with oTree (Chen, Schonger, and Wickens, 2016) and run at the BonnEconLab in March 2017.[19]

Figure 3.6 shows the results.[20] The actual probability for a subject to be in a group with all other seven group members choosing Option A was 0%, both in Rounds 1 and 2. In no single group, there were more than six subjects choosing Option A. A different way to estimate the actual probability of being pivotal is to use the whole distribution of

[19] Another interesting extension would be to investigate how behavior depends on different sources of being pivotal. There is evidence that endogenously determined probabilities resulting from choices of other group members ("social risk") can give rise to different behavior than probabilities determined by a correspondingly calibrated random device (see, e.g., Bohnet et al., 2008). In particular, if subjects in our experiment cared about fairness in relation to their fellow group members, they would potentially have additional reasons to act selfishly in the respective treatments: they could either wish to equalize their own monetary payoffs with the ones of other group members who are selfish, or they could feel "betrayed" if others did not cooperate in implementing the moral outcome. By contrast, if the probability of being pivotal was exogenously determined, social motives should be less relevant.

[20] For corresponding results regarding *belief_ B*, see 3.B.2.

choices and to calculate the likelihood—given the probability for Option A (41.7%)—of randomly being matched with seven group members who all choose Option A, which is 0.22%. This value is shown in the first bar and the analogous value of 0.02% for Round 2 in the second bar (the probability of Option A in the latter round is 29.2%). Bars 3 and 4 show subjects' average beliefs for Rounds 1 and 2, respectively. It is obvious that subjects heavily overestimate how likely it is that they are pivotal. While the shown average beliefs hide a substantial amount of heterogeneity, almost all subjects perceive themselves as being pivotal with a higher likelihood than what is true. Moving from Round 1 to Round 2, subjects adjust in the correct direction but still heavily overestimate their impact. Importantly, however, average beliefs of the spectators are not significantly different from those of active subjects in the first round of SimultaneousC ($p = 0.59$, Mann–Whitney U test, two-sided). On average, active subjects' beliefs are even slightly higher, suggesting that self-serving belief distortions do not play a dominant role in our main conditions.

3.4.3 Sequential Decision Making

We now turn to the sequential decision-making setup SequentialC. The central findings for this treatment are summarized in Figure 3.7. The overall share of participants choosing Option B in the first round of SequentialC is 72.1%, an increase of 81.7% relative to BaselineC. The difference between the two treatments is statistically significant ($p < 0.001$, comparison of means, two-sided and with standard errors clustered at the group level). This share increases by another 14.2 percentage points towards the second round ($p = 0.06$, comparison of means, two-sided and with standard errors clustered at the group level).[21] At the aggregate level, in both rounds, only two out of the 30 groups in SequentialC do not destroy the donation of €120.

Acting in a chain renders the specific position within the decision process relevant. Subjects deciding first in their group are of particular interest since, in a certain sense, they are in a similar situation as subjects in SimultaneousC. They have no information about others' behavior in the given round and the consequences of the moral choice Option A for them depend on the behavior of seven other subjects. In the first round, 43.3% of first movers choose Option B. Interestingly, this share is not significantly larger than in BaselineC ($p = 0.71$, two-sample test of proportions, two-sided). One can think of several plausible mechanisms contributing to this finding. First, the chance of being pivotal indeed seems to be higher for first movers in SequentialC than for subjects in SimultaneousC. Of the 17 cases in the first round where first movers choose Option A, two result in an actual donation. In light of the simple logic employed in Section 3.2.2, this might even be expected. Conditional on the donation not having been destroyed yet, choosing Option A becomes increasingly attractive the further down the line that a given subject decides, because the donation has to "survive" fewer remaining decisions. Subjects deciding at earlier positions

[21] Again, the two types of beliefs (*belief_B* and *belief_pivotal*) are significantly correlated (Spearman rank correlation: -0.65, $p < 0.001$).

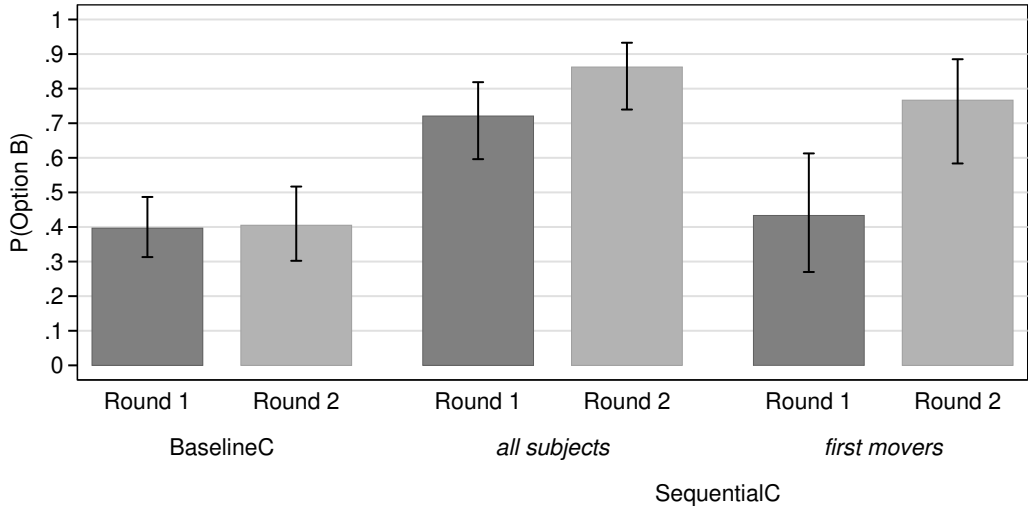


Figure 3.7: Comparison between BaselineC and SequentialC

Notes: Share choosing Option B among subjects in BaselineC, all subjects in SequentialC, and subjects in SequentialC who decide first in their groups, per round. Error bars show 95% confidence intervals (based on logit transformations), where standard errors are clustered at the group level for both rounds of SequentialC.

should anticipate this recovery of moral behavior over positions, incentivizing them to preserve the donation themselves. Second, first movers overestimate their chance of being pivotal. The two surviving donations out of 17 cases where first movers choose Option A correspond to a likelihood of 11.8%, but first movers on average believe that it is 31.3%.[22] This could hint at exaggerated optimism regarding the possibility of acting as a prosocial role model (Gächter et al., 2012; Gächter, Nosenzo, and Sefton, 2013). Third, subjects in SequentialC who choose Option B first in their groups are strongly identified with destroying the donation even in a constellation where, in fact, the decision would not have altered the outcome. This is because subjects at positions 2 to 8 make state-contingent choices, such that their counterfactual behavior remains unknown. In particular, a subject who has chosen Option B will almost certainly observe that all subjects deciding at later positions will do the same but will not know what they *would have done* otherwise. The last two points lose most of their power in the second round. The large majority of first movers who had chosen Option A in the first round will learn that the donation was destroyed, meaning that they have not been pivotal. If they were hoping to be role models, they will feel frustrated. If they did not want to take the blame for choosing Option B first, they will now have a good excuse. Indeed, in the second round, the fraction of subjects who choose Option B increases to 76.7%, which is now significantly different from the second round of BaselineC ($p < 0.001$, two-sample test of proportions, two-sided). It thus seems that with experience, diffusion of being pivotal erodes moral behavior also in the context with sequential decision-making.

Of course, the points discussed above may to some degree also apply to subjects deciding

[22]Note that for all 30 movers in the first round we also find that the belief of being pivotal and choice of Option A are significantly correlated in the expected direction (Spearman rank correlation: -0.68, $p < 0.001$ for *belief_ pivotal* and 0.80, $p < 0.001$ for *belief_ B*).

Table 3.2: Choice dynamics

	Dependent variable: Option B					
	Round 1			Round 2		
	(1)	(2)	(3)	(4)	(5)	(6)
Position (1–8)	0.0552***		-0.0121	0.0210*		0.00108
	(0.0139)		(0.0110)	(0.0115)		(0.00963)
Not destroyed		-0.626***	-0.607***		-0.458***	-0.154
		(0.0626)	(0.115)		(0.122)	(0.126)
Interaction			-0.0170			-0.126***
			(0.0278)			(0.0229)
Constant	0.473***	0.948***	1.013***	0.768***	0.968***	0.962***
	(0.0988)	(0.0228)	(0.0467)	(0.0798)	(0.0161)	(0.0572)
Observations	240	240	240	240	240	240
Clusters	30	30	30	30	30	30
R^2	0.0794	0.450	0.458	0.0196	0.313	0.435
Adj. R^2	0.0755	0.448	0.451	0.0155	0.310	0.428

Notes: OLS regression coefficient estimates, with binary choice option (Option B: destroy donation versus Option A: donate) as the dependent variable. Data come from the SequentialC treatment. *Position* is the position in the move order from 1–8, *Not destroyed* is a dummy that is 1 if all subjects in the respective group have chosen Option A thus far, and *Interaction* is the interaction of the two above variables. Standard errors in parentheses are clustered at the group level (30 groups). * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

on other positions. It is therefore informative to consider the dynamics of choice behavior in this treatment more broadly. In Table 3.2 we explore the role of position and choice history in a simple panel regression framework using both rounds. In Columns 1 and 4, we regress a participant's choice of Option B on his position. Descriptively, subjects are more likely to choose Option B the further down the line that they decide. In Columns 2 and 5, we regress Option B on a dummy indicating that no other group member has chosen Option B yet ("not destroyed"). The respective intercepts are close to one and show that conditional on the donation already having been destroyed, almost all subjects choose Option B. The remaining subjects' decisions could reflect either a lack of attention or understanding (which is unlikely given the control questions and the prominent display of previous play on the decision screen) or a deontological notion of rule-based decision-making. More importantly, in the first as well as in the second round, subjects react strongly to being potentially pivotal, as reflected in the negative and significant coefficients (Columns 2 and 5, respectively). In Columns 3 and 6, we combine position and history and also include the interaction of the two. Turning to Round 1 (Column 3), the coefficients for the position as well as the interaction are insignificant, and the coefficient indicating that Option B has not yet been chosen is essentially identical to the one in Column 2. This suggests that subjects largely ignore their positions. Turning to the second round, this is no longer true. Now, the conditional probability of choosing Option B is generally high but decreases over positions. This can be interpreted as evidence of successful learning

about moral types of subsequent subjects in the same group as well as about the imposed choice mechanism itself.

3.5 Conclusion

This chapter has documented the deterioration of moral outcomes in response to diffusion of being pivotal. Simple organizational changes from an individual decision context to group conditions increase moral transgression at the individual and even more so at the aggregate level.

In our main experiment, subjects decide to either kill mice in return for €10 or to save mice. In Baseline, subjects decide individually about the life of one mouse. In Simultaneous, subjects decide in groups of eight about the lives of eight mice. A single subject is enough to bring about the killing. We observe a statistically significant increase from 46.0% choosing to kill in Baseline to 58.6% in Simultaneous. In the group setting, all mice are killed. Our second paradigm closely resembles that of the first experiment but replaces killing mice with destroying charitable donations of €15 and €120 in the individual and group contexts, respectively. Analogously to the above comparison, we find a significant increase from 39.7% choosing the selfish option in BaselineC to 58.3% in SimultaneousC. To test for experience effects, we repeat the experiment in an unexpected second round. Repetition leaves the share in BaselineC virtually unchanged, while the share increases by another 12.5 percentage points in SimultaneousC. Using the charity paradigm, we also study a sequential context, in which eight subjects decide in a line and know whether the donation has already been destroyed. On average, 72.1% of subjects opt for destroying the donation and the share rises by another 14.2 percentage points towards the second round. Among subjects deciding first in their groups, 43.3% destroy the donation in the first round and 76.7% do so in the second round. Thus, with experience, immoral behavior also deteriorates for first movers in the sequential choice context.

Consequentialism and deontological ethics have been center stage in occidental moral philosophy for the last centuries. Empirical studies using the so-called trolley problem put forward by Philippa Foot (see also, e.g., Greene et al. [2004] and Thomson [1976])[23] have provided support for the relevance of both. However, the evidence highlights the importance of situational and emotional factors. In contrast to the trolley evidence—which uses hypothetical outcomes—subjects in our experiment face real consequences. In all of our group treatments, we elicit beliefs about being pivotal. Subjects consistently respond to notions of being pivotal and only a few subjects appear to follow a Kantian conception. In Simultaneous, 17.9% of subjects who hold the belief that the chance of being pivotal is exactly zero choose Option A. In SimultaneousC, the respective share is almost identical with 17.7%. Finally, in SequentialC, of the 153 individuals for whom the group donation was already destroyed before, eight subjects (5.2%) nevertheless choose

[23]The quandary to be resolved in this problem is to follow either the deontologically warranted option (and not to throw a switch that will divert a trolley and kill one person) or the option preferred from a consequentialist perspective (killing the person to save five others).

Option A. These numbers suggest the existence of deontological reasoning but they are quite low. Our findings thus question the relatively high fractions of Kantian types in survey data.

Using incentivized answers from non-involved observers, we show that there is no indication of subjects forming or reporting self-serving and thus biased beliefs in an attempt to justify selfish behavior in our context. Generally, we find that beliefs about being pivotal are too high. Had they been more realistic, the willingness to engage in selfish behavior may have been even more pronounced. In this sense, it is conceivable that repeated interactions with learning possibilities even further increase the likelihood of immoral outcomes, as we observe in the second round of our experiment using the charity paradigm. Overestimating one's sense of being pivotal could point to a human tendency to overestimate one's impact in general. This may well extend to other (non-moral) contexts and seems worth further investigating, e.g., in voting contexts (Duffy and Tavits, 2008). In this context, Quattrone and Tversky (1984) argue and provide evidence that people use their own actions as prognostic for the behavior of others, therefore trying to "induce" others to behave in a desired way even when no causal impact can exist. Another possible reason for overestimating one's impact could come from a desire for meaning, self-attribution and -determination, as well as for motivating action in general. Such a desire for self-efficacy is already known in the context of the so-called IKEA effect (e.g., Norton, Mochon, and Ariely, 2012).

While the focus of this chapter is to highlight possible negative consequences of organizational design on moral behavior, the reverse inference is, of course, our main interest. Our findings suggest that organizations aiming to promote morality should reduce diffusion of being pivotal and instead attribute individual responsibility to their members.

References

Alger, Ingela, and Jörgen W. Weibull. 2013. "Homo Moralis—Preference Evolution Under Incomplete Information and Assortative Matching". *Econometrica* 81 (6): 2269–2302.

Andreoni, James, and John Miller. 2002. "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism". *Econometrica* 70 (2): 737–753.

Arendt, Hannah. 1963. *Eichmann in Jerusalem*. New York, NY: Viking Press.

Bartling, Björn, and Urs Fischbacher. 2012. "Shifting the Blame: On Delegation and Responsibility". *Review of Economic Studies* 79 (1): 67–87.

Bénabou, Roland, Armin Falk, Luca Henkel, and Jean Tirole. 2020. *Eliciting Moral Preferences*. Working paper. Bonn, Germany: briq – Institute on Behavior & Inequality.

Bénabou, Roland, Armin Falk, and Jean Tirole. 2018. *Narratives, Imperatives, and Moral Reasoning*. NBER Working Paper 24798. Cambridge, MA: National Bureau of Economic Research.

Bohnet, Iris, Fiona Greig, Benedikt Herrmann, and Richard Zeckhauser. 2008. "Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States". *American Economic Review* 98 (1): 294–310.

Bowles, Samuel. 1998. "Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions". *Journal of Economic Literature* 36 (1): 75–111.

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

Crawford, Neta C. 2007. "Individual and Collective Moral Responsibility for Systemic Military Atrocity". *Journal of Political Philosophy* 15 (2): 187–212.

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

Darley, John M. 1992. "Social Organization for the Production of Evil". *Psychological Inquiry* 3 (2): 199–218.

Deckers, Thomas, Armin Falk, Fabian Kosse, and Nora Szech. 2016. *Homo Moralis: Personal Characteristics, Institutions, and Moral Decision-Making*. IZA Discussion Paper 9768. Bonn, Germany: Institute for the Study of Labor (IZA).

Duffy, John, and Margit Tavits. 2008. "Beliefs and Voting Decisions: A Test of the Pivotal Voter Model". *American Journal of Political Science* 52 (3): 603–618.

Epley, Nicholas, and Thomas Gilovich. 2016. "The Mechanics of Motivated Reasoning". *Journal of Economic Perspectives* 30 (3): 133–140.

Falk, Armin. 2017. *Status Inequality, Moral Disengagement and Violence*. CESifo Working Paper 6588. Munich, Germany: CESifo.

Falk, Armin, and Nora Szech. 2013. "Morals and Markets". *Science* 340 (6133): 707–711.

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

Fischer, Peter, Joachim I. Krueger, Tobias Greitemeyer, Claudia Vogrincic, Andreas Kastenmüller, Dieter Frey, Moritz Heene, Magdalena Wicher, and Martina Kainbacher. 2011. "The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies". *Psychological Bulletin* 137 (4): 517–537.

Fisman, Raymond, Shachar Kariv, and Daniel Markovits. 2007. "Individual Preferences for Giving". *American Economic Review* 97 (5): 1858–1876.

Foot, Philippa. 1967. "The Problem of Abortion and the Doctrine of the Double Effect". *Oxford Review* 5:5–15.

Gächter, Simon, Daniele Nosenzo, Elke Renner, and Martin Sefton. 2012. "Who Makes a Good Leader? Cooperativeness, Optimism and Leading-by-Example". *Economic Inquiry* 50 (4): 953–967.

Gächter, Simon, Daniele Nosenzo, and Martin Sefton. 2013. "Peer Effects in Pro-Social Behavior: Social Norms or Social Preferences?" *Journal of the European Economic Association* 11 (3): 548–573.

Gert, Bernard. 2012. "The Definition of Morality". In *Stanford Encyclopedia of Philosophy*, Fall 2012, ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Gino, Francesca, Michael I. Norton, and Roberto A. Weber. 2016. "Motivated Bayesians: Feeling Moral While Acting Egoistically". *Journal of Economic Perspectives* 30 (3): 189–212.

Glover, Jonathan, and M. Scott-Taggart. 1975. "It Makes no Difference Whether or Not I Do It". *Proceedings of the Aristotelian Society, Supplementary Volumes* 49:171–209.

Greene, Joshua D., Leigh E. Nystrom, Andrew D. Engell, John M. Darley, and Jonathan D. Cohen. 2004. "The Neural Bases of Cognitive Conflict and Control in Moral Judgment". *Neuron* 44 (2): 389–400.

Greiner, Ben. 2004. "An Online Recruitment System for Economic Experiments". In *Forschung und wissenschaftliches Rechnen*, ed. by Kurt Kremer and Volker Macho, 63:79–93. GWDG-Bericht.

Hamman, John R., George Loewenstein, and Roberto A. Weber. 2010. "Self-Interest through Delegation: An Additional Rationale for the Principal-Agent Relationship". *American Economic Review* 100 (4): 1826–1846.

Kant, Immanuel. 1996. "Groundwork of the metaphysics of morals". In *The Cambridge edition of the works of Immanuel Kant: Practical philosophy*, ed. by Mary J. Gregor, 37–108. Cambridge, United Kingdom: Cambridge University Press.

Latané, Bibb, and John M. Darley. 1968. "Group inhibition of bystander intervention in emergencies". *Journal of Personality and Social Psychology* 10 (3): 215–221.

Lifton, Robert Jay. 1986. *The Nazi Doctors: Medical Killing and the Psychology of Genocide*. New York, NY: Basic Books.

Milgram, Stanley. 2009. *Obedience to Authority: An Experimental View*. Reprint. New York, NY: Harper Perennial Modern Classics.

Norton, Michael I., Daniel Mochon, and Dan Ariely. 2012. "The IKEA effect: When labor leads to love". *Journal of Consumer Psychology* 22 (3): 453–460.

Peck, Tom. 2016. "If we don't sell arms to Saudi Arabia, someone else will, says Boris Johnson". *The Independent* October 26. Visited on 01/23/2020. `http://www.independent.co.uk/news/uk/politics/if-we-dont-sell-arms-to-saudi-arabia-someone-else-will-says-boris-johnson-a7382126.html`.

Quattrone, George A., and Amos Tversky. 1984. "Causal Versus Diagnostic Contingencies: On Self-Deception and on the Voter's Illusion." *Journal of Personality and Social Psychology* 46 (2): 237–248.

Roemer, John E. 2010. "Kantian Equilibrium". *Scandinavian Journal of Economics* 112 (1): 1–24.

— . 2015. "Kantian optimization: A microfoundation for cooperation". *Journal of Public Economics* 127:45–57.

Rothenhäusler, Dominik, Nikolaus Schweizer, and Nora Szech. 2018. "Guilt in voting and public good games". *European Economic Review* 101:664–681.

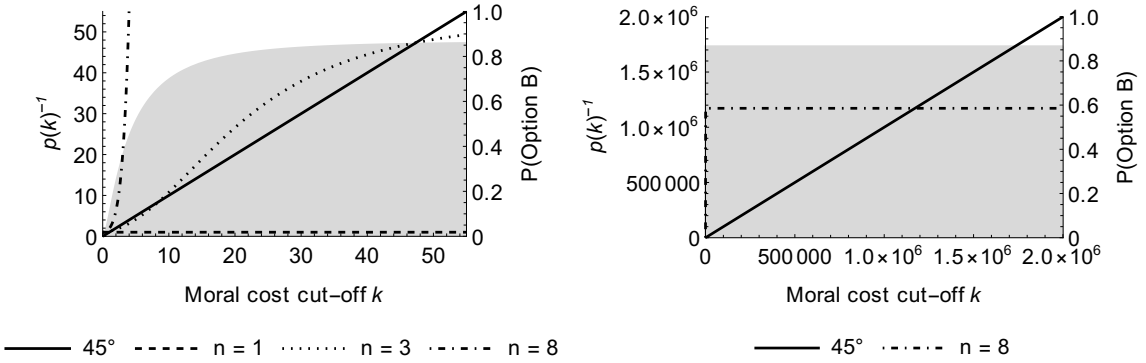
Serra-Garcia, Marta, and Nora Szech. 2018. *The (In)Elasticity of Moral Ignorance*. Working Paper Series in Economics 120. Karlsruhe, Germany: Karlsruhe Institute of Technology (KIT).

Sobel, Joel. 2010. *Do Markets Make People Selfish?* Mimeo.

Soltes, Eugene. 2016. *Why They Do It: Inside the Mind of the White-Collar Criminal*. New York, NY: PublicAffairs.

Spranca, Mark, Elisa Minsk, and Jonathan Baron. 1991. "Omission and commission in judgment and choice". *Journal of Experimental Social Psychology* 27 (1): 76–105.

Thomson, Judith Jarvis. 1976. "Killing, Letting Die, and the Trolley Problem". *The Monist* 59 (2): 204–217.



(A) Different (counterfactual) group sizes (B) Unique equilibrium for $n = 8$

Figure 3.A.1: Equilibria (Simultaneous)

Notes: Numbers on the horizontal axes denote multiples of utility from receiving €10. The 45° line is drawn solidly. Dashed lines visualize the function $p(k)^{-1}$ for different (hypothetical) group sizes n. Values are given on the left axes. The shaded areas represent the cumulative distribution function $(1-d)\,F(k)$ of subjects choosing Option B. Values are given on the right axes.

Appendix 3.A Equilibria

We inspect the equilibrium condition developed in Section 3.2.2 using the parameter values for the share of deontologists d and for the log-normal distribution F of moral costs, μ and σ, estimated in Section 3.3.2 for individuals deciding over the lives of eight mice. To gain a better intuition, we generalize the condition for an equilibrium to exist at a cut-off value of $k^\star$ given by equation (3.1) to the case of a groups size of n.

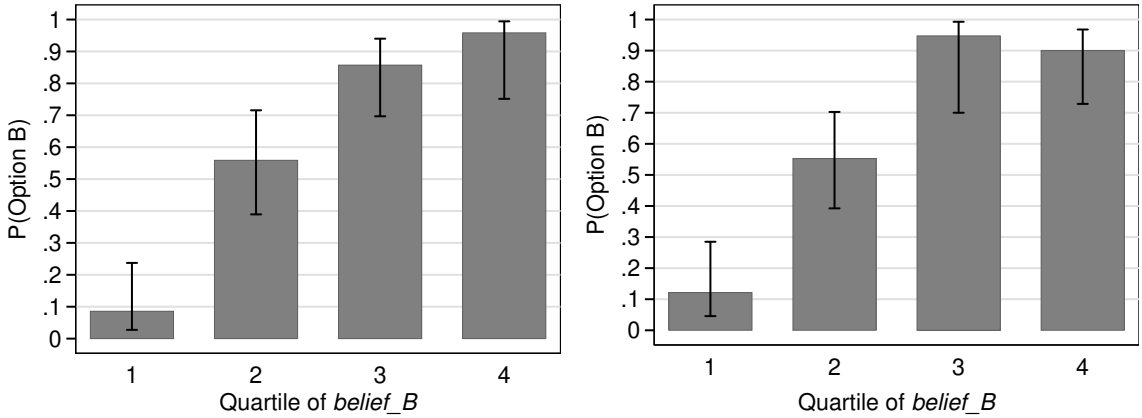
$$k^\star = \{d + (1-d)\,[1 - F(k^\star)]\}^{1-n}$$

The two panels of Figure 3.A.1 provide a visual inspection of this equilibrium condition. Both are identically constructed but vary in their scale. Dashed lines show inverse probabilities of being pivotal as functions of the cut-off value k for moral costs at which subjects switch from choosing Option B to Option A. Equilibria are intersections of dashed lines with the solid 45° line. The left panel of Figure 3.A.1 shows that for $n = 3$ (and still assuming the life of eight mice being at stake), there would exist three equilibria: one at 1.35 in which still only 14.0% of subjects would choose Option B, one at 8.46 in which 65.6% would choose Option B, and one at 47.35 in which 85.5% would do so. For our actual case of $n = 8$, only a single equilibrium exists, which can be seen in the right panel of Figure 3.A.1. In this equilibrium, essentially all consequentialists choose Option B.

Appendix 3.B Results for *belief_B*

3.B.1 Beliefs and Choices

We have established in Figures 3.2 and 3.5 (Sections 3.3.1 and 3.4.1, respectively) that beliefs about being pivotal are strongly associated with the propensity to choose Option B in both Simultaneous and SimultaneousC. We have used *belief_pivotal*, the percentage



(A) Simultaneous (B) SimultaneousC (Round 1)

Figure 3.B.1: Belief quartiles for *belief_B* in Simultaneous and SimultaneousC (Round 1)

Notes: Share of subjects in the respective treatment choosing Option B depending on their belief about the number of other group members choosing Option B (*belief_B*). Error bars show 95% confidence intervals (based on logit transformations).

belief about the likelihood of being pivotal, because it is directly part of the formal analysis in Section 3.2.2. However, we also asked subjects about their belief regarding the number of other subjects in their group who chose Option B (*belief_B*), and the elicitation of this belief was incentivized. We show below that the same kind of relationship as for *belief_pivotal* can also be found for *belief_B*.

Figure 3.B.1a shows fractions of subjects choosing Option B depending on *belief_B* for subjects in Simultaneous. The categories are based on quartiles of the belief distribution and are given by the belief intervals $[0, 2]$, $(2, 4]$, $(4, 6]$, and $(6, 7]$. We see a monotonous increase in the propensity to choose Option B over belief quartiles, which is the expected mirror image of the Figure 3.2.[24] In particular, we see a strong increase between the first two quartiles, and the increases seem to fade out for the higher quartiles. In light of our framework, this is intuitive: subjects who believe that very few others—and thus potentially none—will choose Option B are highly reluctant to do so themselves, while for high expected numbers the precise beliefs do not matter a lot.

Figure 3.B.1b replicates the above relationship in the charity experiment, i.e., for SimultaneousC. Again, we find a general increase of the share of subjects choosing Option B over quartiles, which correspond to intervals of $[0, 2]$, $(2, 4]$, $(4, 5.5]$, and $(5.5, 7]$. Again, the increase between the first two quartiles is pronounced, while the difference between the last two quartiles is insignificant. Thus, the analysis of the relationship between *belief_B*, an indirect measure for the belief of being pivotal, and choice of the immoral option lends additional support to Hypothesis 3.2.

[24] The Spearman rank correlation between *belief_B* and choice of Option B in Simultaneous is 0.65 ($p < 0.001$).

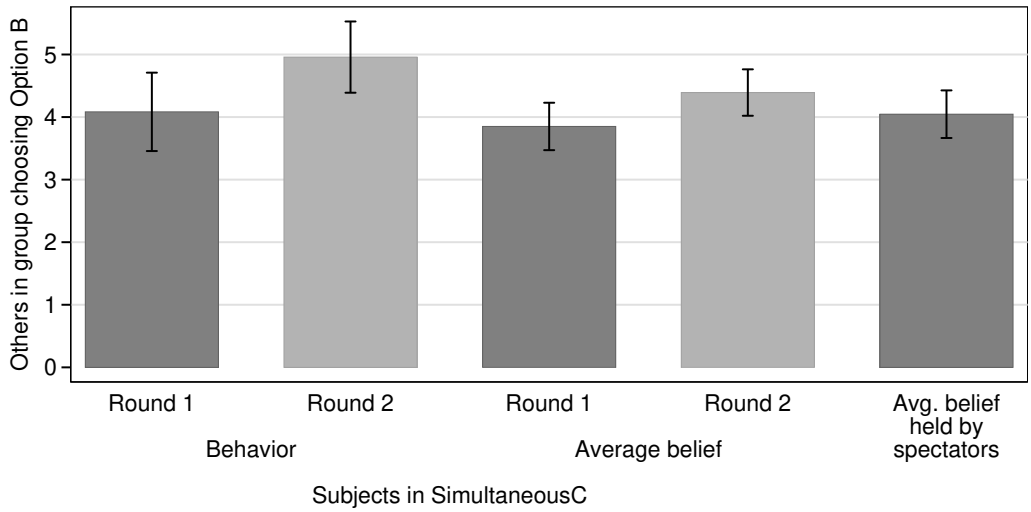


Figure 3.B.2: Belief comparison (*belief_B*)

Notes: Number of other group members choosing Option B (0–7). Error bars show 95% confidence intervals, where standard errors are clustered for the second round.

3.B.2 Belief Experiment

We report results from the belief experiment in which participants read the original instructions of treatments in the charity experiment (avoiding textual redundancies) and reported incentivized estimates corresponding to *belief_pivotal* and *belief_B* in SimultaneousC.

For *belief_B*, results are qualitatively quite similar to those concerning *belief_pivotal* (see Section 3.4.2). Figure 3.B.2 shows the actual behavior of subjects in Rounds 1 and 2 (first two bars), average beliefs in Rounds 1 and 2 (bars 3 and 4), as well as average beliefs of spectators (fifth bar). The number of subjects choosing Option B increases from Round 1 to 2, which is reflected in changes in the beliefs of subjects. In contrast to *belief_pivotal*, however, subjects are overall much more accurate about actual outcomes.[25] Importantly, as for *belief_pivotal*, the average beliefs of active subjects and spectators are not statistically significantly different (comparison of bars 3 and 5 in Figure 3.B.2; $p = 0.59$, Mann–Whitney U test, two sided).

Appendix 3.C Instructions of the Mouse Experiment

Instructions have been translated from German.

Baseline

Thank you very much for your participation!
For your participation, you will, in any case, receive 20 euros. In the following, you can earn an additional amount of money. At the end of the experiment, you will receive your money

[25] A possible explanation is that subjects found estimating absolute numbers easier than estimating a probability.

in an envelope. Neither the other participants of the experiment nor the experimenter will be able to see how much money you have earned.

Please note: Throughout the whole experiment **communication between the participants is not allowed**. On the computer, please only use the functions intended to be used. If you have questions please raise your hand. Your question will then be answered at your cubicle!

Please note: **All statements made in these instructions are true.** This holds for all experiments carried out by the BonnEconLab, and also for this experiment. **In particular, all actions to be taken will be implemented exactly in the way they are described.** If you want to, you will be able to verify the correctness of all statements made in these instructions after the experiment.

In this experiment, there is a **Quiz A** and a **Quiz B**. Both, Quiz A and Quiz B, are simple trivia quizzes with questions from history, geography, sports, and so on. One example question could be: "Capital of Belgium?" There will, respectively, be four possible answers out of which one answer is correct. The posed questions in Quiz A and Quiz B are identical, which means, they are exactly the same regarding their difficulty. You will get three minutes to solve the quiz. The more questions you solve correctly, the more you can earn. **For each question that is answered correctly, you receive 5 cents.**

Details on the mouse

[Picture of a mouse]

In this study, the life of a mouse is entrusted to your care. It is a healthy, young mouse, living with some other mice together in a small group. The expected lifetime of this mouse is approximately two years.

What is the difference between Quiz A and Quiz B?

Quiz A: In Quiz A, at the end of the experiment, you earn no additional money besides the **20 euros** for participation and the mouse stays alive.

Quiz B: In Quiz B, at the end of the experiment, you **get 10 euros in addition**. As another consequence, **the mouse will get killed**.

Details on the killing process

If you opt for the death of the mouse, the mouse is gassed. The gas flows slowly into the hermetically sealed cage. The gas leads to breathing arrest. At the point at which the mouse is not visibly breathing anymore, it remains in the cage for another 10 minutes. It will then be removed.

Summary

In Quiz A you earn no additional money, and the mouse does not get killed. In Quiz B, you earn additionally 10 euros, and the mouse gets killed. The decision is yours. You make

your decision on a decision screen that will be shown as soon as you have answered the control questions on the following screen.

Control questions

In case of Quiz A: How many euros do you receive in addition? ____
Will a mouse be killed? ☐ Yes ☐ No

In case of Quiz B: How many euros do you receive in addition? ____
Will a mouse be killed? ☐ Yes ☐ No

Simultaneous

Introduction as in Baseline

In this study, the life of eight mice is entrusted to your group's care. These are healthy, young mice, living with some other mice together in a small group. The expected lifetime of these mice is approximately two years.

What is the difference between Quiz A and Quiz B?

In the following, we describe the consequences of choosing Quiz A and Quiz B. The choice options and consequences are identical for all eight group members.

Quiz A: In Quiz A, at the end of the experiment, you earn no additional money besides the **20 euros** for participation. This holds for all group members. Each group member who chooses Quiz A receives no additional money.

Quiz B: In Quiz B, at the end of the experiment, you **get 10 euros in addition**. This holds for all group members. Each group member who chooses Quiz B receives 10 euros in addition.
Another consequence is that **eight mice get killed if at least one member of your group of eight chooses Quiz B**. Thus if in total one member of the group, or two, three, four, five, six, seven or eight group members choose Quiz B, eight mice get killed. **Only if no member in your group of eight chooses Quiz B, the mice will not get killed.**

Details on the killing process

If your group opts for the death of the mice, these will be gassed. The gas flows slowly into the hermetically sealed cage. The gas leads to breathing arrest. At the point at which the mice are not visibly breathing anymore, they remain in the cage for another 10 minutes. They will then be removed.

Summary

In Quiz A you earn no additional money. In Quiz B, you earn additionally 10 euros. Whether the mice get killed depends on whether at least one member of your group of eight has chosen Quiz B. You make your decision on a decision screen, which will be shown as soon as you have answered the control questions on the following screen.

Control questions and video

Appendix 3.D Instructions of the Charity Paradigm

Instructions have been translated from German.

BaselineC

Welcome and thank you very much for your interest in today's experiment!

This experiment is part of a research project of the *Bonner Laboratorium für experimentelle Wirtschaftsforschung (BonnEconLab)*.

For your participation, you will, in any case, receive €10.00, which will be handed to you in cash today at the end of the experiment. During the experiment, you will make decisions on the computer. Depending on how you decide, you can earn additional money.

During the experiment, it is not allowed to communicate with other participants. Also, note that the curtain of your cubicle has to be shut throughout the entire experiment. Please now switch off your mobile phone, to make sure that other participants are not being disturbed. On the computer, please only use the functions intended to be used and make all inputs using either the mouse or the keyboard. If you have questions, please contact the conductor of the experiment. To do so, please stick your hand out of the cubicle.

All statements made in this experiment are true. This holds for all experiments carried out by the BonnEconLab, and also for this experiment. **In particular, all actions to be taken will be implemented exactly in the way they are described.** If you want to, you will be able to verify the correctness of all statements made in these instructions after the experiment.

In what follows, we will first ask you to answer a question regarding your mood. Subsequently, the decisions you will have to make will be explained in detail.

How is your current mood?

Please give an answer to this on the following scale from 0 to 10.
0 means that your mood is very bad.
10 means that your mood is very good.
You can choose any integer number on the scale from 0 to 10 to express your current mood.

The donation

This experiment is about a donation to the *Förderkreis für krebskranke Kinder und Jugendliche e.V.*, a regional charity from Bonn.

Every participant, that means also you, will first be entrusted with a donation that will be made to the *Förderkreis für krebskranke Kinder und Jugendliche e.V.* after today's experiment.

During the experiment, you will make decisions that affect this donation. Moreover, the information that follows is also relevant for your personal payoff from this experiment.

Therefore, please carefully read the following instructions. In particular, make sure that you understand all the decisions you can make as well as their potential consequences.

Information about the *Förderkreis*

The *Förderkreis*. The *Förderkreis für krebskranke Kinder und Jugendliche e.V.* **supports young people suffering from cancer and their families comprehensively in dealing with the disease**. The society is committed to psychological support, to organizing free-time activities, as well as to aftercare and to supporting children and adolescents with school. Moreover, indirectly affected individuals like parents and siblings are extensively supported. This takes, for example, the form of a specifically established home for parents and of pedagogic support. Moreover, the *Förderkreis* supports clinical research on cancer.

Projects and tasks of the *Förderkreis*.

- *Klassissimo* school project: offers participation in school lessons using Skype
- *Bärenstark*: support of families at home
- Psychosocial and psychooncological counseling of patients and relatives
- Pedagogic support at the hospital department
- Start-up financing for new positions and financing of specific training of departments' staff.
- Financing of hospital clowns and music therapy
- Aftercare
- Support for clinical research on cancer

Your decision

The donation. You are entrusted with a donation of €15.00, which is supposed to be made to the *Förderkreis für krebskranke Kinder und Jugendliche e.V.* following today's experiment. Whether this amount will, in fact, be transferred

to the *Förderkreis* at the end of the experiment depends on the decisions that you will make.

Anonymity. No other participant in this experiment can see your decisions. The subsequent analysis of all data is done anonymously, such that all your decisions cannot be linked to your identity anymore.

You can choose between two options: **Option A** and **Option B**. **Depending on which of both options you choose, you can earn different amounts of money. Additionally, depending on which option you choose, consequences differ for the donation of €15.00 that was described above.**

In what follows, the consequences associated with choices of **Option A** and **Option B**, respectively, will be described.

Option A. If you choose **Option A**, besides €10.00 for participation you will receive **no** additional money at the end of the experiment.

Option B. If you choose **Option B**, you will **additionally** receive €10.00 at the end of the experiment.

As a further consequence, **the previously described donation of €15.00 will be destroyed**.

Summary. If you choose **Option A**, you do not receive an additional payment and the donation will not be destroyed. If you choose **Option B**, you additionally receive €10.00 and the donation is destroyed. The decision rests with you.

You make your decision on a decision screen, which will be shown as soon as you have answered the control questions on the following screen.

Control questions

In case of Option A. How many euros do you receive in addition? ____
Will the donation be destroyed? ☐ Yes ☐ No

In case of Option B. How many euros do you receive in addition? ____
Will the donation be destroyed? ☐ Yes ☐ No

Your decision

Please now choose between **Option A** and **Option B**.

I choose: ☐ **Option A** ☐ **Option B**

Result

If Option A was chosen: You have decided **not to destroy** the donation.

Therefore, a donation of €15.00 to the *Förderkreis für krebskranke Kinder und Jugendliche e.V.* will be made for you by the BonnEconLab.

If Option B was chosen: You have decided **to destroy** the donation.

Therefore, **no donation will be made**.

Experiment 2

Now follows a second experiment. This experiment is the last experiment. Your final payoff comprises of €10.00 for participation in the experiment, your decision in the first experiment, and, independently, on how you decide in the second experiment.

The decision in the second experiment is the same as in the first experiment. Thus, you can again choose between **Option A** and **Option B**, i.e., you can decide whether a donation will be destroyed or not. The donation is again a donation to the *Förderkreis für krebskranke Kinder und Jugendliche e.V.*

SimultaneousC

Introduction as in BaselineC

Your decision

Your group. You are together with 7 other participants of today's experiment in a group of 8 people. Your group members have been allotted to you at the beginning of the experiment. You will at no point learn which participant is in your group.

Note: **You are making all decisions within this experiment autonomously and independent of the other members of the group.** The consequences of your decisions can depend on the decisions of other group members. On the following screens, all decisions, alternatives, and consequences will be introduced and explained in detail.

The donation. Your group is entrusted with a donation totaling €120.00, which is supposed to be made to the *Förderkreis für krebskranke Kinder und Jugendliche e.V.* following today's experiment. Whether this amount will, in fact, be transferred to the *Förderkreis* at the end of the experiment, depends on the decisions that you and the other members of your group will make.

Anonymity. No other participant in this experiment can see your decisions. This is also true for the other members of your group. The subsequent analysis of all data is done anonymously, such that all your decisions cannot be linked to your identity anymore.

You can choose between two options: **Option A** and **Option B**. **Depending on which of both options you choose, you can earn different amounts of money.**

Additionally, depending on which option you choose and which options the other participants of your group choose independently, consequences differ for the donation of €120.00 that was described above.

In what follows, the consequences associated with choices of **Option A** and **Option B**, respectively, will be described. The choices and the consequences are the same for all 8 participants in your group.

Option A. If you choose **Option A**, besides €10.00 for participation you will receive **no** additional money at the end of the experiment.

This holds for all group members: Each group member who chooses **Option A** receives no additional money.

Option B. If you choose **Option B**, you will **additionally** receive €10.00 at the end of the experiment.

This holds for all group members: Each group member who chooses **Option B** additionally receives €10.00.

As a further consequence, **the previously described donation of €120.00 will be destroyed if at least one of the 8 members of your group chooses Option B**. Thus, if one group member, or if two, three, four, five, six, seven, or eight group members decide for **Option B**, the donation is destroyed. **Only if none of the 8 members of your group chooses Option B, the donation will not be destroyed.**

Summary. If you choose **Option A**, you do not receive an additional payment. If you choose **Option B**, you additionally receive €10.00. Whether the donation to the *Förderkreis* is destroyed depends on whether at least one of the 8 members of your group has chosen **Option B**.

Decisions of participants in your group

Note: The consequences of your choice do not just depend on you but also on the decisions of the other 7 members of your group. This holds in particular for the execution of the donation to the *Förderkreis für krebskranke Kinder und Jugendliche e.V.*: **Only if none of the members of your group has chosen Option B, the donation of €120.00 is made.**

You and the other 7 members of your group decide **simultaneously**. After all group members have made their decision, you learn whether the donation will be made.

At the end of today's experiment, you will also learn how many members of your group have in total chosen **Option A** and how many members of your group have in total chosen **Option B**.

You make your decision on a decision screen, which will be shown as soon as you have answered the control questions on the following screen.

Control questions

Suppose, two/one/no/six group member(s) choose(s) Option B.

You choose Option A: How many euros do you receive in addition? ____
Will the donation be destroyed? ☐ Yes ☐ No

You choose Option B: How many euros do you receive in addition? ____
Will the donation be destroyed? ☐ Yes ☐ No

Your decision

Please now choose between **Option A** and **Option B**.
I choose: ☐ **Option A** ☐ **Option B**

What do you estimate?

How likely is it in your opinion that all other group members have chosen **Option A**? *Please enter a probability (from 0 to 100 percent):* [Slider]

What do you think, how many of the other 7 group members have chosen **Option B**? If you estimate the correct number, you will additionally receive €2.00. *Enter a number between 0 and 7:* ____

Result

If Option A was chosen: You have decided **not to destroy** the donation.

In your group, at least one participant has decided **to destroy** the donation. The donation over €120.00 from you and the other members of your group will therefore **not be made**.

You have not made a correct estimation and therefore do not receive any additional payoff.

Experiment 2

Now follows a second experiment. This experiment is the last experiment. Your final payoff comprises of €10.00 for participation in the experiment, your decision in the first experiment, and, independently, on how you decide in the second experiment.

The decision in the second experiment is the same as in the first experiment. Thus, you can again choose between **Option A** and **Option B**, i.e., you can decide whether a donation will be destroyed or not. The donation is again a donation to the *Förderkreis für krebskranke Kinder und Jugendliche e.V.*

Please note: You are in the same group of 8 participants as in the first experiment.

SequentialC

Introduction as in SimultaneousC

Decisions of participants in your group

Note: The consequences of your choice do not just depend on you but also on the decisions of the other 7 members of your group. This holds in particular for the execution of the donation to the *Förderkreis für krebskranke Kinder und Jugendliche e.V.*: **Only if none of the members of your group has chosen Option B, the donation of €120.00 is made.**

You and the other 7 members of your group decide **one after the other**. Your position is randomly determined by a computer.

When it is your turn, you will learn whether, among the people who have decided before you, someone has already chosen Option B. You will also learn your position within the sequence. Moreover, you will learn how many members of your group have already chosen **Option A** and how many members of your group have already chosen **Option B**. At the end of today's experiment, you will also learn how many members of your group have in total chosen **Option A** and how many members of your group have in total chosen **Option B**.

Please note: If another participant in your group has already decided for **Option B** before it was your task, this means that the donation has already been destroyed. Thus, in this case, your decision has **no** effect any more on whether the donation is made.

Control questions as in SimultaneousC.

Your decision

You are on **position 1** in the order of your group. Consequently, no other member in your group has made a decision yet.

Or:

You are on **position 2** in the order of your group. Consequently, 1 group member has already made a decision.
Of the 1 group members who have decided before you, 1 has decided for **Option A** and 0 for **Option B**.

Or:

You are on **position 3** in the order of your group. Consequently, 2 group members have already made a decision.
Of the 2 group members who have decided before you, 1 has decided for **Option A** and 1 for **Option B**.
Thus, the donation has already been destroyed.

Please now choose between **Option A** and **Option B**.

I choose: ☐ **Option A** ☐ **Option B**

Remaining instructions as in SimultaneousC.

Belief Experiment

In the belief experiment, participants read the original instructions (avoiding redundancies, however), learn how many subjects have taken part in the respective treatment, and are then asked to answer the following questions.

BaselineCB:

- How likely do you think it is that a randomly chosen participant of the just described experiment decides for Option A, i.e., not to destroy the donation?

SimultaneousCB:

- How likely was it for a participant in the experiment to be in a group in which all other 7 group members choose Option A? (answer in percent)
- How likely do you think it is that the donation is not destroyed in such a group in the end, i.e., that all 8 group members choose Option A. (answer in percent)
- Please imagine you are in the new situation at the BonnEconlab that was just described. What do you think: How many of the other 7 members of your group have decided for Option B, i.e., to destroy the donation?

SequentialCB:

- How likely was it for a participant in this experiment to be in a group in which all other 7 group members choose Option A? (answer in percent)
- How likely do you think it is that the donation is not destroyed in such a group, i.e., that all 8 group members choose Option A? (answer in percent)

Please now imagine yourself being in the situation of a participant in the described experiment at the BonnEconLab.

- Imagine, you decide first and choose Option A. How many of the other 7 group members do you think also choose Option A, such that the donation is not destroyed? (answer in percent)
- Imagine, the member at position 1 in your group chooses Option A. You decide second and also choose Option A. How likely do you think it is that all further 6 people in the group also choose Option A, such that the donation is not destroyed? (answer in percent)
- Imagine, the members at positions 1 to 3 in your group all choose Option A. You decide as the fourth and also choose Option A. How likely do you think it is that all further 4 people in the group also choose Option A, such that the donation is not destroyed? (answer in percent)

- You decide last, i.e., as the eighth. How likely do you think it is that all 7 before you have chosen Option A? (answer in percent)
- Please again imagine yourself being in the situation of the described experiment at the BonnEconLab. You decide first. What do you think: How many of the 7 other members of your group decide for Option B, i.e., for destroying the donation?
- Now, please imagine that you decide last in your group, i.e., as the eighth. All 7 group members before you have chosen Option A. Would you then choose Option A or Option B? *(unincentivized)*
- How likely do you think it is that a participant in the just described situation – decided last, all group members before have chosen Option A – also has chosen Option A? (answer in percent)

Chapter 4

Limited Self-knowledge and Survey Response Behavior*

Joint work with Armin Falk and Philipp Strack

Abstract

We study response behavior in surveys and show how the explanatory power of self-reports can be improved. First, we develop a choice model of survey response behavior under the assumption that the respondent has imperfect self-knowledge about her individual characteristics. In panel data, the model predicts that the variance in responses for different characteristics increases in self-knowledge and that the variance for a given characteristic over time is non-monotonic in self-knowledge. Importantly, the ratio of these variances identifies an individual's level of self-knowledge, i.e., the latter can be inferred from observed response patterns. Second, we develop a consistent and unbiased estimator for self-knowledge based on the model. Third, we run an experiment to test the model's main predictions in a context where the researcher knows the true underlying characteristics. The data confirm the model's predictions as well as the estimator's validity. Finally, we turn to a large panel data set, estimate individual levels of self-knowledge, and show that accounting for differences in self-knowledge significantly increases the explanatory power of regression models. Using a median split in self-knowledge and regressing risky behaviors on self-reported risk attitudes, we find that the R^2 can be multiple times larger for above- than below-median subjects. Similarly, gender differences in risk attitudes are considerably larger when restricting samples to subjects with high self-knowledge. These examples illustrate how using the estimator may improve inference from survey data.

*We thank Roland Bénabou, Philipp Eisenhauer, Botond Kőszegi, and participants at various conferences and seminars for helpful comments. We thank Markus Antony for excellent administrative support. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1 – 390838866. Funding by the DFG through CRC TR 224 (Project A01) is gratefully acknowledged. Philipp Strack was supported by a Sloan fellowship.

4.1 Introduction

Survey evidence is a major source of knowledge in the social sciences, including economics. With growing interest in measuring cognitive and non-cognitive skills—such as economic preferences, beliefs, attitudes, and values—survey evidence is gaining increasing relevance in economics (Heckman, Stixrud, and Urzua, 2006; Almlund et al., 2011; Falk et al., 2018). This chapter provides a method to improve the explanatory power of subjective survey data. The method is derived from a simple model of survey response behavior that allows identifying more vs. less informative respondents based only on *patterns* of their response behavior. Hence, this chapter makes two main contributions: it offers a framework for modeling and understanding survey response behavior in general and it derives a method to empirically identify more or less reliable answers, which in turn helps to improve the explanatory power of survey measures.

As a first step, we derive a simple choice model of survey response behavior. In the model, we are serious about the idea that when being asked to report an individual characteristic such as a preference, belief, or some non-cognitive skill, a respondent has to make herself the object of her own self-assessment and makes a *choice*. We assume that there exists a true type (level of each characteristic) but that the respondent is not perfectly aware of her true type. This *limited self-knowledge* is modeled as an imperfect signal that the respondent receives about her true type. Differences in self-knowledge can capture the fact that individuals vary in their capacity to retrieve or memorize relevant information about themselves, engage more or less in reflecting who they are, or that some people simply lack life experience in the domain of interest. We further assume that the respondent wants to minimize the squared distance between her true type and her report, i.e., the interests of the respondent and the researcher are aligned. Conditional on the informativeness of the signal, our agent's Bayesian optimal report is a weighted sum of the population mean of the respective characteristic and her signal. The more informative the signal, the greater the weight placed on the signal relative to the population mean. We analyze the expected variance of respondents' answering behavior conditional on the informativeness of the signal, both over time and between characteristics. We find that the variance *between* characteristics increases in the informativeness of the signal, which mirrors the fact that the more confident a respondent is about her answer, the more she deviates in expectation from the population mean. In contrast, the *within* variance—the variance of responses for a given characteristic over time—is non-monotonic in the signal precision. The intuition is that response behavior is stable over time if a person knows herself either very well or not at all. This result cautions against the use of simple stability to measure the accuracy of signals and reports. Importantly, we show that the ratio of the variance between characteristics and the variance over time (for given characteristics) is equal to the informativeness of the signal. This key result implies that we can use observed variances to estimate individual differences in self-knowledge or the accuracy of respective reports.

We provide several extensions of the model and discuss their implications for expected response behavior. Our first extension relaxes the assumption of exogenous signals and

explores the consequences of endogenous precision. We derive an expression for the choice of signal precision and discuss implications for how the quality of survey responses reacts to incentives. Second, we relax the assumption that respondents are perfectly aware of the signal strength, i.e., how well they know themselves. Instead, we allow for subjective levels of self-knowledge that are higher or lower than actual self-knowledge. While subjective beliefs about self-knowledge affect the distribution of responses, we show that they do not impede the identification of differences in self-knowledge, simply because they cancel out. Third, we allow for individual-specific scale use, i.e., a tendency to report either rather extreme or moderate answers. Again, we show that scale use affects responses but that the identification of self-knowledge remains unchanged. Finally, we relax the assumption that respondents want to report their type truthfully. Instead, we allow for response biases arising from motives such as social desirability or image effects. We study the implications of such motives and show that respondents act similarly as in the case of subjective scale use.

The second step in the chapter is to use the theoretical results in empirical applications—especially the insight that the precision of signals about types can be inferred from the ratio of the between- and the within-variance. We first show that self-knowledge can be estimated using a closed-form estimator before discussing results from a laboratory experiment designed to test the main predictions of the model. Subsequently, we analyze representative panel data to show how accounting for signal precision affects empirical results and explained variance.

To derive an estimator of signal precision—or self-knowledge—from panel data, we essentially consider the ratio between two sample variances, namely the between-variance (the variance of responses between items) and the within-variance (the variance for a given item over time). These are the sample analogs to our theoretically derived variances. We study the asymptotic properties of the estimator and formally show its consistency as well as unbiasedness. Using simulations, we illustrate the performance of the estimator for realistic sample sizes. We study various combinations of the number of respondents, survey items, and waves (periods), respectively. The estimator generally performs well. For example, for 100 respondents, 15 items, and three waves, the rank correlation between the estimated and the true level of self-knowledge is 0.76.

To empirically test the main predictions of the model, it is crucial to observe responses and compare them with respondents' *true* types. However, this is difficult—if not impossible—with typical survey data. Therefore, we ran a laboratory experiment that creates a panel data set with types that are imperfectly known to subjects but perfectly known to the researcher. In particular, subjects in the experiment were paid to accurately report the sizes of 60 male figures shown to them on separate computer screens. This setup allows us to observe subjects' reports and compare them with the respective true types. Results from the experiment confirm the main predictions derived from the model. First, subjects' reports are biased towards the mean, i.e., small sizes are, on average, overestimated, and large sizes are typically underestimated. Second, subjects who are estimated to be more informative actually provide more accurate reports. Based on the estimates, we

split the sample and regress reports on true types. We find that the regression coefficient for the above-median sample is about 2.5 times as large as the respective coefficient for below-median subjects and that the explanatory power in terms of R^2 is about five times as large. Third, we use the experiment to create random variation in signal precision. For this purpose, we randomized subjects into one of two treatments: a Long-treatment in which they saw the figures for 7.5 seconds each, and a Short-treatment in which each figure was presented only for 0.5 seconds. We show that we can use our empirical estimates to predict subjects' treatment status, i.e., we are able to predict whether subjects were assigned to the treatment condition with high or with low signal precision.

Finally, we apply our estimator to a large representative panel data set, the German Socio-economic Panel (SOEP; Goebel et al., 2019). We provide several examples to illustrate how the suggested estimates of self-knowledge can help to increase the explanatory power of regressions based on self-reports. In particular, we use a fifteen-item Big Five personality inventory from multiple waves of the SOEP to estimate self-knowledge. Using these estimates, we form two sub-samples: one with above- and one with below-median values of estimated self-knowledge, respectively. As an illustrative example, we choose the context of risk attitudes, which has received a lot of attention in the literature. We study both determinants and consequences of risk attitudes, measured on an eleven-point Likert scale. To illustrate, we find that the gender effect on the general willingness to take risks is substantially larger for the above-median sample than for the below-median sample. Moreover, the difference in R^2 between the two sub-samples amounts to 36%. Likewise, when we regress the likelihood of receiving performance pay as part of one's compensation on the willingness to take risks, the explained variance (R^2) is 238% higher in the above-median sample than in the below-median sample.

Our chapter is related to multiple strands of the literature. As we take the informational constraints of the agent seriously and study their choice implications, we relate to the work on rational inattention (Sims, 1998, 2003; Caplin and Dean, 2015; Matějka and McKay, 2015; Caplin et al., 2020). This literature focuses on flexible information acquisition and studies what type of information is acquired in a single-agent setting. Our goal is different, and we analyze how to identify agents' levels of information in a situation with many agents who share a common prior. Our framework enables analyzing the provision of incentives in surveys as studied, e.g., in Prelec (2004) and Cvitanić et al. (2017) as well as how contextual factors such as social desirability affect survey responses (see, e.g., Bénabou et al., 2020; Chen et al., 2020). The notion of limited self-knowledge and its economic consequences for the labor market has been studied in Falk, Huffman, and Sunde (2006a, 2006b). The model is also related to work on preferences for consistency, as modeled and tested in Falk and Zimmermann (2017) and applied to survey methodology in Falk and Zimmermann (2013).

Moreover, the chapter contributes to the literature on measurement error in surveys (for an overview, see Bound, Brown, and Mathiowetz, 2001). For the case of classical measurement error—where deviations in answers are independent of the respective true value—, instrumental variables techniques are capable of removing bias. Recently, Gillen, Snow-

berg, and Yariv (2019) have suggested measuring duplicate instances of control variables and using them as mutual instruments. Hyslop and Imbens (2001) consider a model that is related to ours where an agent observes a Normal signal and reports his best estimate of an underlying variable of interest. They analyze the effect of the resulting non-classical measurement error on regression coefficients but do not consider remedies. The focus of our chapter is to estimate the precision of the agent's signal, which allows placing higher weight on subjects with better self-knowledge.

Drerup, Enke, and Gaudecker (2017) estimate a structural model of stock market participation that identifies individuals for whom relevant preferences and beliefs have increased explanatory power. Alternative approaches to deal with measurement error in subjective survey data use structural estimation techniques to recover underlying primitives and choice models, finding that accounting for measurement error yields greater predictive power (Kimball, Sahm, and Shapiro, 2008; Beauchamp, Cesarini, and Johannesson, 2017).[1] Despite not referring to qualitative survey measures, a related contribution comes from Beauchamp et al. (2020), who analyze how accounting for the "compromise effect"—whereby subjects' answers tend towards the center of the provided scale—, can improve estimates for risk preference.

The remainder of the chapter proceeds as follows. Section 4.2 develops the model with its basic framework and extensions. Building upon its insights, Section 4.3 introduces the estimator, presents its theoretical properties, and explores its performance in finite samples. Section 4.4 presents the stylized laboratory experiment. In Section 4.5, we apply the estimator to a large and representative panel and explore its implications for improving estimates. Finally, Section 4.6 concludes.

4.2 Model

In this section, we first introduce a simple framework to model the answering process in surveys, based on limited self-knowledge. Second, we derive how patterns in answering behavior reveal the informational content of responses, providing the intuition for how we later estimate self-knowledge. Finally, we present various extensions of the baseline model to study further important aspects of the answering process and show the robustness of our identification approach.

Introspection and Self-knowledge. The context that we are interested in is a simple survey situation. A researcher asks a respondent (or agent) a question about a specific characteristic, e.g., some preference, personality trait, or belief.[2] The agent's true type is denoted by θ, and we assume that it is normally distributed in the population with mean $\bar{\theta}$

[1] In the psychology literature, processes that underlie response behavior have been studied under the label of *cognitive aspects of survey methodology* (see Sudman, Bradburn, and Schwarz, 1996; Bradburn, Sudman, and Wansink, 2004; Schwarz, 2007). Broadly, our chapter is also related to classical test theory and item response theory (see, e.g., Edwards, 2009; Kyllonen and Zu, 2016; Bolsinova, Boeck, and Tijmstra, 2017).

[2] For example, the researcher may ask the respondent to state her willingness to take risks, her level of agreeableness or conscientiousness, or her belief about her internal or external locus of control.

and variance σ^2. Agents act upon their true types but vary with respect to how well they know their type. Hence, when asked about her type θ, the respondent does not perfectly know herself but instead engages in a process of introspection. The outcome of this process is an informative but noisy signal x about her true type. The signal is normally distributed with a mean equal to the agent's type θ and variance σ^2/τ. The parameter $\tau > 0$ hence indicates the precision of the signal relative to the variance in the population. The higher the value of τ, the more precise the signal that an individual receives about herself. We refer to τ as *self-knowledge*.

Response Behavior. After reflecting on her true type θ, the respondent reports her answer. We assume that she seeks to provide a response r that is as precise as possible, i.e., the interests of the researcher and respondent are perfectly aligned.[3] Formally, the respondent uses her signal x to provide a response r that minimizes the expected quadratic distance to her unknown true type, i.e.,

$$u_\theta(r) = -(r-\theta)^2 \,. \tag{4.1}$$

Hence, she reports her best guess of her type $r = \mathbb{E}[\theta \,|\, x]$. The respondent's prior equals the distribution of types in the population with mean $\bar{\theta}$. Substituting for the expected value of her posterior belief about her type, we obtain by Bayes' Rule that

$$r = \frac{\bar{\theta} + \tau\, x}{1+\tau} \,. \tag{4.2}$$

Intuitively, the higher her self-knowledge τ, the more precise the respondent's signal, and the more weight she puts on her signal relative to the population mean $\bar{\theta}$. In the limit, if she knows nothing about herself, her best estimate is to report the mean of her prior, whereas if she knows herself perfectly, she disregards the prior completely.

This concludes our basic framework. The model defines a mapping from true types to distributions over observable responses, taking into account the notion of limited self-knowledge. In the next subsection, we study how response patterns can be used to identify differences in self-knowledge.

4.2.1 Response Patterns

We now explore the implications of limited self-knowledge for response patterns. We are particularly interested in the variances in reports, both unconditional and conditional on an agent's type. These variances will allow us to identify differences in self-knowledge. In Section 4.3, we will build on these insights when we derive an estimator for an individual's level of self-knowledge in panel data.

[3] For many interview situations, we think that this is a valid assumption. However, there are contexts in which respondents may want to strategically signal a specific type that is actually different from their belief about their true type for reputational or "social desirability" reasons. For a discussion, see Section 4.2.2.

Expected Report. It follows from Equation 4.2 that the expected report conditional on the true type θ equals

$$\mathbb{E}[r \mid \theta] = \frac{\bar{\theta} + \tau\,\theta}{1+\tau}\,. \tag{4.3}$$

For low values of self-knowledge τ, the expected report is close to the population mean $\bar{\theta}$, irrespective of the true type θ. For large values of τ, the expected report converges to the true type θ.

Between-variance. Consider now the variance of conditional expected reports. In the context of panel data, one can think of this theoretical quantity as an approximation of the variance in average reports concerning different characteristics. Following this interpretation (as the variance between different characteristics), we refer to it as the *between-variance*. It is given by

$$\begin{aligned}\sigma^2_{\text{between}} &\coloneqq \operatorname{var}(\mathbb{E}[r \mid \theta]) = \operatorname{var}\left(\frac{\bar{\theta} + \tau\,\theta}{1+\tau}\right) \\ &= \left(\frac{\tau}{1+\tau}\right)^2 \operatorname{var}(\theta) = \left(\frac{\tau}{1+\tau}\right)^2 \sigma^2\,.\end{aligned} \tag{4.4}$$

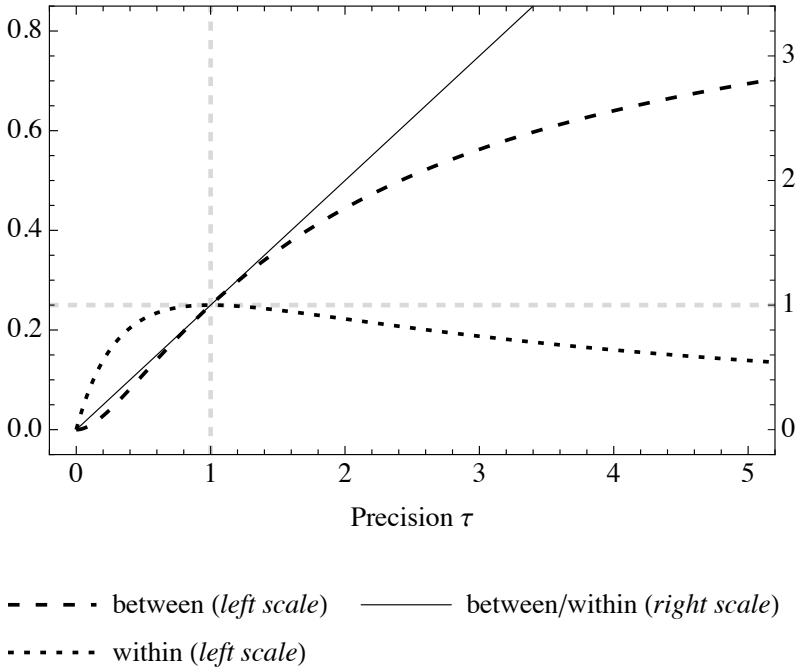
The between-variance is strictly increasing in self-knowledge τ. This reflects the fact that agents with high levels of self-knowledge put relatively little weight on their prior. Instead, they provide reports that tend to deviate from the population mean.

Within-variance. Now consider the variance conditional on an agent's type. This theoretical quantity can be thought of as the variation in responses of an agent responding multiple times to questions about the *same* characteristic. We call this variation the *within-variance* of the agent's reports. It is given by

$$\begin{aligned}\sigma^2_{\text{within}} &\coloneqq \operatorname{var}(r \mid \theta) = \operatorname{var}\left(\frac{\bar{\theta} + \tau\,x}{1+\tau} \,\middle|\, \theta\right) \\ &= \left(\frac{\tau}{1+\tau}\right)^2 \operatorname{var}(x \mid \theta) = \frac{\tau}{(1+\tau)^2}\,\sigma^2\,.\end{aligned} \tag{4.5}$$

The relationship between self-knowledge τ and the within-variance is non-monotonic. For very low levels of τ, the variance is low, simply because the respondent refers to her prior. As τ increases, the variance increases as more weight is placed on the noisy signal. However, as τ further increases, the variance decreases because the signal about the true type becomes increasingly precise. From a researcher's perspective, this pattern implies that stable responses—i.e., similar responses regarding the same characteristics over time—do not necessarily indicate high levels of self-knowledge and precision. The most stable responses come from respondents who know themselves perfectly—or who do not know themselves at all.

Figure 4.1 illustrates the relationship between the two variances and self-knowledge. It plots the between-variance (long dashes) and the within-variance (short dashes) as func-



Note: Variances $\sigma^2_{\text{between}}$ and σ^2_{within} as functions of τ (values on the left axis). The solid line shows the ratio of the two variances, which is equal to τ (values on the right axis).

Figure 4.1: Theoretical variances

tions of self-knowledge τ. As τ goes to zero, both variances converge to zero. This means that the respondent provides the same answer (equal to the prior) to any question. As τ increases, the respondent places higher weight on her signal, which increases both the within- and between-variance. At $\tau = 1$, i.e., when the signal x is exactly as informative as the respondent's prior knowledge about the population, the within-variance reaches its maximum and is equal to the between-variance. Beyond this point, the between-variance further increases and ultimately converges to the variance of true types in the population, σ^2. At the same time, the within-variance strictly decreases and converges to zero, because a respondent with perfect self-knowledge will always provide exactly the same report for a given characteristic.

Both the between- and within-variance contain information about the respondent's level of self-knowledge τ. While a large between-variance is always "good news," indicating high levels of τ, a low within-variance can reflect either high or low levels of τ, respectively. However, considering both variances *jointly* perfectly reveals the level of self-knowledge. In fact, the ratio of the between- and within-variance equals the degree of self-knowledge:

$$\frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{\left(\frac{\tau}{1+\tau}\right)^2 \sigma^2}{\frac{\tau}{(1+\tau)^2}\,\sigma^2} = \tau\,. \tag{4.6}$$

The respective relationship is also shown in Figure 4.1 where, for each level of τ, the thin solid line plots the ratio of the two variances.

Our chapter builds on this insight. We show that the relationship between the variances and self-knowledge is robust to various extensions of the model, construct a finite sample estimator based on this relationship, and show that this estimator indeed predicts the informativeness of subjects' responses both in lab and field data.

4.2.2 Extensions

In this subsection, we study four extensions of the basic framework. The purpose of this exercise is twofold. The first aim is to show that the framework enables integrating additional important aspects of the survey response process in a meaningful way. In particular, we consider the role of costly introspection, deviations of subjective self-knowledge from actual self-knowledge, subjective scale use, and social desirability issues. Second, we show that for the extensions studied, the result that self-knowledge τ can be inferred as the ratio of the between- and within-variance is robust.

Endogenous Precision

Our first extension considers endogenous precision. So far, we have modeled the process of introspection as receiving an exogenous signal with a fixed relative precision τ. However, the cognitive process of introspection requires mental effort, and a respondent has to decide how much effort to invest. For example, the agent chooses how long and intensively she engages in recollecting past behaviors to extract her type and how carefully she evaluates and maps information into a response. We assume that the variance of the signal x is no longer fixed at a given level of σ^2/τ. Instead, τ is chosen by the agent at a cost τ/a for some ability $a \in \mathbb{R}_+$. The utility function (corresponding to Equation 4.1) equals

$$u_\theta(r, \tau) = -m\,(r-\theta)^2 - \frac{\tau}{a}\,.$$

Here, $m \in \mathbb{R}_+$ measures the motivation of the respondent to provide an accurate answer, and it can be thought of as either extrinsic or intrinsic motivation.[4] Assume that $m\,a > \sigma^{-2}$, as, otherwise, incentives are too weak to motivate any effort and a precision of zero is optimal.

Lemma 4.1. *The respondent's optimal precision is given by* $\tau^* = \sqrt{m\,a}\,\sigma - 1$.

The chosen signal precision τ^* is increasing in both incentives m and ability a, i.e., a higher level of incentives or ability generates more precise signals. The proof of Lemma 4.1 is provided in Appendix 4.A.

In the presence of endogenous effort, subjects giving high- vs. low-quality answers can be distinguished by the exact same response patterns as for the case with exogenous self-knowledge. However, the interpretation changes, as differences may now reflect differences in motivation m or ability a. In fact, the model predicts that the higher the incentives, the more reliable and informative the responses. This is exactly the rationale for paying

[4] The former could reflect, e.g., monetary or social approval incentives, while the latter may capture motives such as a desire to respond truthfully and accurately or simply an interest in (promoting) research.

subjects in economic experiments (Smith, 1976; Camerer and Hogarth, 1999) and similar attempts to incentivize survey responses as, e.g., Prelec's Bayesian Truth Serum (Prelec, 2004). In addition, differences may reflect motivational dispositions (e.g., mood, fatigue, boredom) or fundamental differences in "introspection ability" a, such as cognitive skills, memory, and recollection capabilities.

Subjective Self-knowledge

The basic framework assumes that the respondent knows the relative precision τ of her signal x. In other words, she perfectly knows how well she knows herself and weighs her signals accordingly. However, a large body of evidence has shown that individuals often misperceive their own knowledge and skills (Camerer and Lovallo, 1999; Malmendier and Tate, 2005). Applied to our context, respondents may be over-confident and place too much weight on their signal x, or they are under-confident and place too much weight on the prior. In either case, this will result in a wedge between the optimal and the actual response, again potentially complicating inference about respondents' true types.

To model potential biases in perceived self-knowledge, we introduce subjective self-knowledge $\tilde{\tau}$. A respondent has correct beliefs about her self-knowledge if $\tilde{\tau} = \tau$, she is under-confident if $\tilde{\tau} < \tau$, and she is over-confident if $\tilde{\tau} > \tau$. We assume that the agent is naive and that when determining her survey response, she applies relative weights according to her subjective self-knowledge $\tilde{\tau}$. Equation 4.2 changes as follows:

$$r = \frac{\bar{\theta} + \tilde{\tau}\, x}{1 + \tilde{\tau}}$$

Corresponding to Equation 4.4, the between-variance becomes

$$\sigma^2_{\text{between}} = \text{var}(\mathbb{E}[R \,|\, \theta]) = \left(\frac{\tilde{\tau}}{1 + \tilde{\tau}}\right)^2 \sigma^2 \,.$$

Hence, the variability in answers between different items reflects the respondent's subjective self-knowledge but is independent of self-knowledge itself. Intuitively, as the between-variance is based only on the expected response, which is independent of the true precision of the agent's signal τ, the variance is also independent of the true precision of the agent's signal.

This is different for the within-variance, corresponding to Equation 4.5.

$$\sigma^2_{\text{within}} = \text{var}(r \,|\, \theta) = \left(\frac{\tilde{\tau}}{1 + \tilde{\tau}}\right)^2 \frac{\sigma^2}{\tau} \,.$$

The latter depends on both subjective self-knowledge as well as actual self-knowledge. Intuitively, the within-variance of responses is affected by the respondent's subjective self-knowledge $\tilde{\tau}$ through the weight that she places on her signal and by her self-knowledge τ through the variance of the signal.[5]

[5]Observe that only for $\tilde{\tau} \to \infty$, the model predicts classical measurement error.

Importantly, the result from Equation 4.6 about the ratio of the two variances still holds.

$$\frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{\left(\frac{\tilde{\tau}}{1+\tilde{\tau}}\right)^2 \sigma^2}{\left(\frac{\tilde{\tau}}{1+\tilde{\tau}}\right)^2 \frac{\sigma^2}{\tau}} = \tau$$

Hence, while deviations from correct beliefs about the precision of one's signals affect expected response behavior in general, inference about τ remains feasible.

Subjective Scale Use

Empirical research typically assumes that individuals who want to express the same level of agreement or disagreement with respect to a particular survey item will respond in the exact same way. For example, two respondents intending to express the exact same willingness to take risks on a Likert scale would be expected to choose the exact same answer category. However, if response scales are subjectively interpreted, responses may differ. Hence, the mapping from an intended response to some scale may depend on individual-specific notions of how to express a given level of agreement or disagreement. We suggest a simple way how to model this kind of subjective scale use and show that it affects responses in general but not the estimation approach for τ suggested by Equation 4.6.

In particular, assume that an agent has arrived at her *intended* report and now needs to map it to an *actual* report r on an answering scale. This mapping may be individual-specific in the sense that some agents may use more "extreme" answers while others use more "moderate" answers to express the same information. For a given intended response, therefore, two agents may come up with different actual responses. We assume that the agent's response is scaled away from some point $c \in \mathbb{R}$, e.g., the center of the scale, by a factor $\phi \in (0, 1]$. The report and its expected value (corresponding to Equations 4.2.2 and 4.3, respectively) are then given by

$$r = (1-\phi)\, c + \phi \left(\frac{\bar{\theta} + \tau\, x}{1+\tau}\right) \qquad \text{and} \qquad \mathbb{E}[r \mid \theta] = (1-\phi)\, c + \phi \left(\frac{\bar{\theta} + \tau\, \theta}{1+\tau}\right).$$

Depending on ϕ, actual responses may thus be pushed towards the center of the scale, rendering the interpretation of responses more difficult. This holds in particular if ϕ is systematically correlated with underlying types (such as preferences) or group characteristics under study (such as gender or socioeconomic status).

The between-variance (corresponding to Equation 4.4) becomes

$$\begin{aligned}\sigma^2_{\text{between}} &= \text{var}(\mathbb{E}[R \mid \theta]) = \text{var}\left((1-\phi)\, c + \phi\, \frac{\bar{\theta} + \tau\, \theta}{1+\tau}\right) \\ &= \phi^2 \left(\frac{\tau}{1+\tau}\right)^2 \text{var}(\theta) = \phi^2 \left(\frac{\tau}{1+\tau}\right)^2 \sigma^2,\end{aligned}$$

and the within-variance (corresponding to Equation 4.5) becomes

$$\begin{aligned}\sigma^2_{\text{within}} &= \text{var}(r \,|\, \theta) = \text{var}\left((1-\phi)\, c + \phi\, \frac{\bar{\theta} + \tau\, x}{1+\tau} \,\middle|\, \theta \right) \\ &= \phi^2 \left(\frac{\tau}{1+\tau} \right)^2 \text{var}(x \,|\, \theta) = \phi^2\, \frac{\tau}{(1+\tau)^2}\, \sigma^2 \,.\end{aligned}$$

We see that both variances increase quadratically in the scale use parameter ϕ. However, for the ratio of the two, the effect of scale use cancels out, and it still holds that the ratio equals τ.

$$\frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{\phi^2 \left(\frac{\tau}{1+\tau} \right)^2 \sigma^2}{\phi^2\, \frac{\tau}{(1+\tau)^2}\, \sigma^2} = \tau$$

Social Desirability Effects

In some situations, respondents might not want to truthfully report their type but rather provide an answer that is deemed socially desirable. These contexts are likely to arise if the interview situation is not anonymous (audience effects) and/or if items are image relevant. For example, it is plausible that a respondent feels more comfortable reporting that she is an honest rather than a dishonest person. Such concerns can be integrated into our framework by adding a desirable answer $d \in \mathbb{R}$. Respondents' objective now is to minimize the weighted sum of the squared distances to their type and the desirable answer, respectively. The utility function is thus

$$u_{\theta,d}(r) = -\,(1-\psi)\,(r-\theta)^2 - \psi\,(r-d)^2 \;,$$

where $\psi \in [0,1]$ measures the intensity of the preference to report d. The optimal report of a respondent equals the weighted sum of the best guess of her type θ and the desirable answer

$$r = (1-\psi) \left(\frac{\bar{\theta} + \tau x}{1+\tau} \right) + \psi\, d \,.$$

The respondent thus acts as if subject to subjective scale use, as introduced in Section 4.2.2. The main difference between subjective scale use and desirability arises in the context of multiple agents and characteristics: while the scale use parameters (ϕ, c) are naturally agent-specific, the desirability parameters (ψ, d) are naturally specific to the characteristic.

4.3 Estimator

In this section, we derive an estimator for an individual's level of self-knowledge that is based on the insights from Section 4.2. We consider a panel data set comprising $I > 1$ agents and $T > 1$ waves. In each wave t, each agent i answers an identical set of $K > 1$ questions about distinct, time-invariant characteristics, traits, or beliefs. We denote by θ_{ik} the value of the k^{th} characteristic for agent i and assume that characteristics are independently normally distributed in the population with mean $\bar{\theta}$ and variance σ^2. In

contemplating the answer to question k in wave t, agent i generates a signal x_{ikt} that she uses to form her answer r_{ikt}. The signal x_{ikt} is normally distributed with mean θ_i and variance σ^2/τ_i, independent of all other signals, such that the optimal response is given by

$$r_{ikt} = \frac{\bar{\theta} + \tau_i \, x_{ikt}}{1 + \tau_i} \, .$$

Given the $K \times T$ answers observed for each agent i, the objective of a researcher is to estimate agents' levels of self-knowledge τ_i. In Section 4.2, we have shown that τ equals the (theoretical) variance among expected answers to different questions (between-variance) divided by the (theoretical) variance among answers to the same questions (within-variance). To construct an estimator $\hat{\tau}_i$, we use the sample variance between average answers for different characteristics as an approximation of the true between-variance and the average sample variance of answers for a given characteristic as an approximation of the true within-variance. Denote agent i's average answer to question k by $\bar{r}_{ik} = \frac{1}{T} \sum_{t=1}^{T} r_{ikt}$ and her average answer over all questions by $\bar{r}_i = \frac{1}{K} \sum_{k=1}^{K} \bar{r}_{ik}$. Our estimator $\hat{\tau}_i$ for the self-knowledge of agent i is given by

$$\hat{\tau}_i = \frac{\frac{1}{K-1} \sum_{k=1}^{K} \left(\bar{r}_{ik} - \bar{r}_i \right)^2}{\frac{1}{K(T-1)-2} \sum_{k=1}^{K} \sum_{t=1}^{T} \left(r_{ikt} - \bar{r}_{ik} \right)^2} - \frac{1}{T} \, . \tag{4.7}$$

The enumerator in the first summand of the expression captures the variation *between* the average answers of an agent for different characteristics, while the denominator expresses the average variation in answers *within* characteristics. Since the expected value of the ratio of two random variables is not the same as the ratio of their respective individual expected values, the denominator is adjusted by a constant factor relative to the unbiased estimator of the within-variance[6] and a correction term of $1/T$ is subtracted from the ratio. These two adjustments are necessary to ensure that the estimator is unbiased.

The following theorem establishes that $\hat{\tau}_i$ is a consistent, unbiased estimator of self-knowledge τ_i and describes its properties.

Theorem. *For every K, T that satisfy $K(T-1) > 4$.*

1. *The estimator $\hat{\tau}_i$ satisfies*

$$\hat{\tau}_i = \left(\tau_i + \frac{1}{T} \right) \frac{K(T-1) - 2}{K(T-1)} F_i - \frac{1}{T} \tag{4.8}$$

   for some random variable F_i that is F distributed with $K-1, K(T-1)$ degrees of freedom for every fixed vector of parameters $\tau_i, \sigma, \bar{\theta}$.

2. *$\hat{\tau}_i$ is an unbiased estimator for τ_i, i.e., $\mathbb{E}[\hat{\tau}_i \,|\, \tau_i] = \tau_i$.*

[6] An unbiased estimator of the within-variance is given by $\frac{1}{K(T-1)} \sum_{k=1}^{K} \sum_{t=1}^{T} \left(r_{ikt} - \bar{r}_{ik} \right)^2$.

3. The standard error of the estimator $\hat{\tau}_i$ is given by

$$\sqrt{\mathbb{E}[(\hat{\tau}_i - \tau_i)^2 \mid \tau_i]} = \left(\tau_i + \frac{1}{T}\right) \sqrt{\frac{2((K-1) + K(T-1) - 2)}{(K-1)(K(T-1) - 4)}}. \tag{4.9}$$

4. $\hat{\tau}_i$ is a consistent estimator and converges to τ_i at the rate $1/\sqrt{K}$ in the number of attributes, and for all $K > 4$ it satisfies the following upper bound independent of the number of repeated observations T:

$$\sqrt{\mathbb{E}[(\hat{\tau}_i - \tau_i)^2 \mid \tau_i]} \leq \frac{2\tau_i + 1}{\sqrt{K-4}}$$

The proof of the theorem is provided in Appendix 4.A. Part 4 of the theorem shows that for retrieving precise estimates, additional questions are more valuable than additional waves. This is the case because, intuitively, having additional questions adds to the precision of estimating both the between as well as the (average) within-variance, whereas additional waves only improve the precision of the estimated within-variance. Therefore, as K goes to infinity, the estimator converges to the true value even for just two waves, while the precision of the estimator is always limited for a finite number of questions.

Remark. *As we show in the proof of the theorem in Appendix 4.A, the properties of the estimator extend unchanged to the model with endogenous effort, subjective self-knowledge, and subjective scale use. We state the properties here without these extensions for ease of exposition.*

Next, we illustrate our model and the behavior of the estimator using numerical simulations. For all illustrations, agents' levels of self-knowledge τ_i are drawn from a uniform distribution with support $[0.1, 5]$, and we abstract from subjective scale use and subjective self-knowledge. The true average value of characteristics $\bar{\theta}$ is set to 5 and the true population variance σ^2 equals 1.

Figure 4.2 displays the joint distribution of the true level of self-knowledge τ_i and the sample within-variance, the sample between-variance, and estimated self-knowledge $\hat{\tau}_i$, respectively. For the within-variance, we observe the expected non-monotonic, hump-shaped relationship with the true level of self-knowledge (Figure 4.2a). The estimates for the between-variance increase in the true level of self-knowledge, but heavily "fan out" for higher levels of true self-knowledge (Figure 4.2b). Our proposed estimator for self-knowledge is strongly concentrated around the 45-degree line and thus highly informative about agents' true levels of self-knowledge (Figure 4.2c).

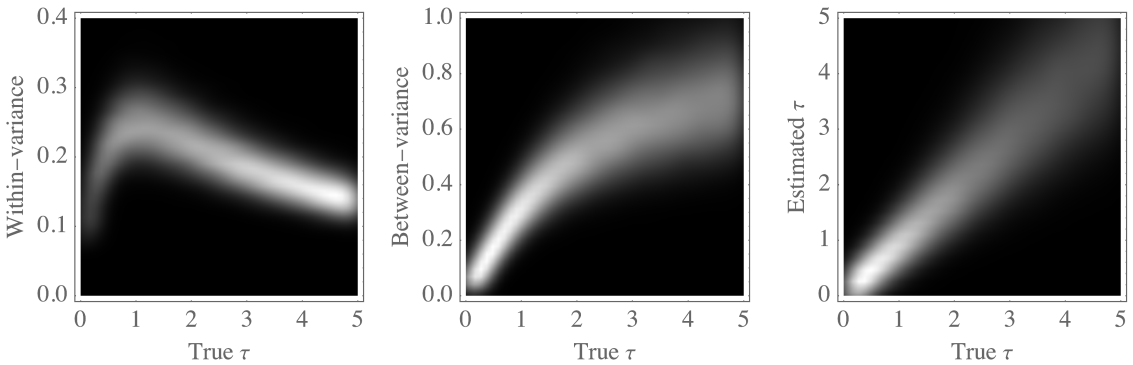
In Table 4.1, we illustrate how the estimator performs for various sample sizes. We consider 100 or 10,000 agents, 15 or 50 characteristics, and 3 or 10 waves, respectively. For each scenario, we run 10,000 simulations and report the average value of three measures for the quality of the estimates: Pearson's correlation and Spearman's rank correlation between estimated and true self-knowledge and the proportion of simulated agents correctly identified as having a level of self-knowledge above or below the value of the median. If

Table 4.1: Accuracy of estimates for different sample sizes

	(1)	(2)	(3)	(4)	(5)
I (respondents)	100	10,000	100	100	100
K (characteristics)	15	15	50	15	50
T (waves)	3	3	3	10	10
Correlation	0.68	0.68	0.87	0.76	0.91
Rank correlation	0.76	0.77	0.90	0.82	0.93
Median split	80%	80%	88%	83%	90%

our estimator had no informational value at all, we would expect a correlation and rank correlation of zero and 50% of correctly-assigned agents in the median split.

The values of the correlation and the rank correlation coefficients of 0.68 and 0.76 shown in Column 1 for $I = 100$, $K = 15$, and $T = 3$ suggest that the estimator is already informative about self-knowledge for modest sample sizes. This is confirmed by 80% of hypothetical agents being assigned to the correct half of the sample in terms of self-knowledge. In Column 2, the number of hypothetical agents is increased to 10,000. The quality of predictions remains almost exactly unchanged, reflecting the fact that our estimator does not use population information. However, as can be seen from Column 3, estimates strongly benefit from a larger number of characteristics (50 instead of 15), in line with Part 4 of the theorem. Relative to these increases, the increase in performance from a higher number of answers per characteristic in Column 4 (ten instead of three) is not quite as large (in line with Part 4 of the theorem, which shows that the standard error does not vanish in T). Column 5 combines the number of characteristics from Column 3 with the number of waves from Column 4, reaching the best performance, with correlation coefficients above 0.9 and a median split result of 90%. In sum, we find that the estimator performs reasonably well with a modest number of fifteen characteristics and three waves, and its performance can be increased, in particular, by a larger number of characteristics.



(a) Within-variance (b) Between-variance (c) Estimated τ

Note: Kernel-density estimates, where lighter shading corresponds to a higher estimated density. Each panel is based on the same 100 simulations, each with I =1,000 hypothetical individuals, for whom reports about $K = 50$ characteristics are observed $T = 3$ times. The panels use Gaussian kernels with bandwidth selection according to Silverman's rule.

Figure 4.2: Simulations

4.4 Experimental Evidence

This section presents experimental evidence to provide an empirical test of the model's main predictions. The idea of the experiment is to create a choice environment where the researcher observes subjects' reports (allowing to estimate τ) while at the same time *knowing the true state* θ. Accordingly, we can study whether our estimator is successful in identifying subjects whose reports are relatively more informative than those of others. In addition, we *exogenously* vary the quality of the signals that subjects receive about true types. In particular, we run two treatments with either high or low signal quality and test whether our estimator of τ is capable of predicting subjects' treatment status, i.e., whether a subject received high- or low-quality signals. Such tests are difficult—if not impossible—with non-experimental data, where true states are unknown to the researcher and the precision of signals cannot be exogenously varied.

4.4.1 Design of the Experiment

To create a choice environment with known types θ and an exogenous variation in knowledge τ, the experiment exposed subjects to a simple, repeated, and incentivized estimation task. The setup mimics a panel data set where respondents are repeatedly asked to respond to a set of different questions.

Types. The requirement that the researcher knows true types implies that we cannot work with individual characteristics such as personality traits, preferences, or IQ, simply because these cannot be known with certainty. To implement types known to the researcher (θ_i), we thus presented subjects a series of abstract figures. In particular, subjects saw a total of 60 screens, each showing a stylized male figure of varying size. On each screen, the figure was randomly located at one of four different parts of the screen, i.e., at either the upper left, the upper right, the lower left, or the lower right part of the screen, respectively. The sizes of the figures were drawn from a normal distribution that closely matches the actual height distribution of men in Germany (based on data from the Socio-economic Panel, SOEP). In particular, sizes were matched into eleven size categories (in meters) with likelihoods as shown in Table 4.2. For example, Category 3 represents male persons of sizes between 1.66 and 1.70 meters, occurring with a likelihood of 11.1%. Subjects received a handout showing this distribution and the corresponding figures underneath (see the instructions in the Online Appendix).

Table 4.2: Choice categories

0	1	2	3	4	5	6	7	8	9	10
<1.56	1.56–1.60	1.61–1.65	1.66–1.70	1.71–1.75	1.76–1.80	1.81–1.85	1.86–1.90	1.91–1.95	1.96–2.00	>2.00
0.1%	0.8%	3.8%	11.1%	21.1%	26.1%	21.1%	11.1%	3.8%	0.8%	0.1%

Note: top row: categories; middle row: sizes (in meters); bottom row: respective likelihoods.

Subjects were informed that a total of 15 distinct sizes were independently drawn from the eleven categories and shown four times. Specifically, subjects saw four blocks, each comprising these 15 distinct sizes. This procedure hence implements a panel structure, i.e., for every subject i, we observe a total of 60 reports for $K = 15$ characteristics in $T = 4$ periods. The location of the male figures was randomly determined for each screen.

To facilitate the estimation task and vary the presentation style of the screens, next to the male figure, subjects also saw a "reference category," i.e., either an elephant or a cat. Subjects were informed that—unlike for the male figures—the size of the two animals was always exactly the same. The height of the elephant was 3.50 meters, and it was 0.40 meters for the cat. Conditional on the randomly determined location of the male figures, the location and type of the reference category (elephant or cat) were also randomly drawn for each screen.

Payoff Function. Subjects had an incentive to estimate the shown size of the male figure as precisely as possible. The payoff function, π, implements a quadratic loss function and corresponds exactly to Equation 4.1 in the model, with

$$\pi(r) = -(r - \theta)^2,$$

where θ indicates the true type (size of the male figure) and r a subject's report. For the payoff, one of the 60 screens was randomly selected. For the selected screen and respective report, subjects received €10 minus the product of €0.10 and the squared difference between the true type and the report. For example, if a subject was shown a male figure of size Category 1 (1.56 meters – 1.60 meters) and estimated a size according to Category 8 (1.91 meters – 1.95 meters), the subject received $€10 - (1 - 8)^2 \times €0.10 = €5.10$. Note that we chose an endowment of €10 to rule out losses even if the difference between the true and the estimated type was maximal.

Signal Precision and Treatments. To exogenously vary the precision τ of the signal, we ran two treatments that only differed in terms of how long subjects saw each of the 60 screens. In the treatment *Long*, subjects saw each screen for 7.5 seconds, in contrast to treatment *Short*, where they saw each screen only for 0.5 seconds. Treatments were randomly assigned within each lab session. Each subject participated in one treatment condition only.

Procedural Details. 199 subjects—mostly undergraduate university students from all majors—took part in the experiment, 101 subjects in the treatment *Long* and 98 in the treatment *Short*. We used z-Tree as the experimental software (Fischbacher, 2007). Subjects were recruited using the software hroot (Bock, Baetge, and Nicklisch, 2014). At the beginning of an experimental session, participants received detailed information about the rules and the structure of the experiment. In all treatments, the experiment only started after all participants had correctly answered several control questions. The experiments

were run at the BonnEconLab in May 2019. For participation, subjects received a show-up fee of €5.

4.4.2 Hypotheses and Results

Our experimental data are well suited for testing several hypotheses derived from our model:

Hypothesis 4.1. *Average reports are linear in true types and biased towards the population average of the true types, i.e., towards five.*

The first hypothesis follows from an optimal report being the weighted sum of the population average $\bar{\theta}$ and the received signal x (see Equation 4.2). It can only be tested because, in our experiment, we know the true type. Graphically, we would expect average reports for different true types to lie on a straight line that is rotated clockwise around the point $(5, 5)$, i.e., we would expect upward bias for small true values, no bias for average true values, and downward bias for large true values.

Hypothesis 4.1 uses that knowledge τ is finite for any subject. Hypotheses 2–4 additionally exploit individual-specific information about τ, either in terms of treatment differences (Short vs. Long) or using the estimator introduced in Section 4.3.

Hypothesis 4.2. *Estimates $\hat{\tau}$ are larger for the subjects in the Long-treatment than for those in the Short-treatment.*

An implication of Hypothesis 4.2 is that the estimates for τ should have reasonable power for predicting subjects' treatment status. Thus, we expect that we can blindfold ourselves regarding the treatment status and be able to tell only from the patterns in answers to which treatment a given subject was assigned.

Regardless of which approach is used to make inferences about τ (the treatment status or the estimator), the following further hypothesis should hold.

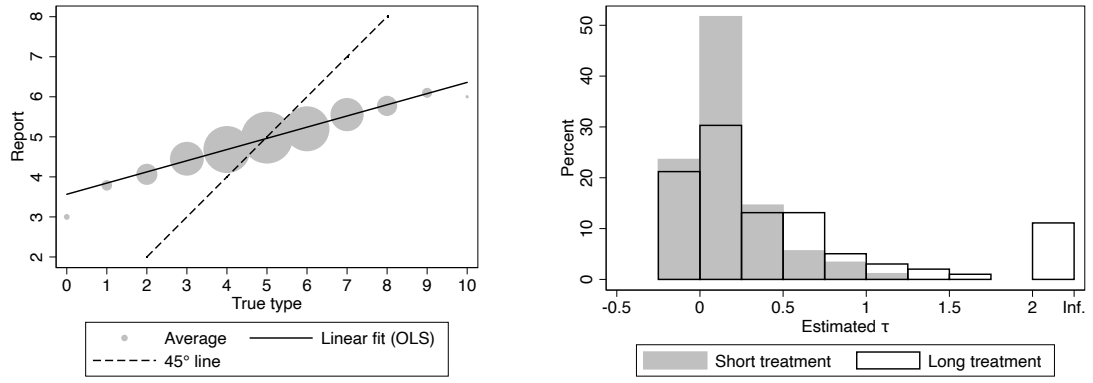
Hypothesis 4.3. *The lower subjects' level of knowledge τ, the stronger the reports' bias towards the average value of the characteristic, i.e., five.*

This hypothesis is a refinement of Hypothesis 4.1. It states that when estimating figure sizes, subjects realize and take into account their *individual-specific* level of τ, which may reflect ability or treatment status.

Hypothesis 4.4. *The higher the level of τ in a given population, the stronger the predictive power of reports for true types.*

Hypothesis 4.4 is our main hypothesis. It states, in particular, that using reports of subjects for whom we have high values of $\hat{\tau}$ yields higher explanatory power of reports in comparison to using either all subjects or subjects with low levels of $\hat{\tau}$.

Figure 4.3a provides a visual test of Hypothesis 4.1. It plots true types against observed reports, pooled for both treatments. Gray bubbles represent average reports for given true



(a) Reports vs. true types

(b) Estimates of τ by treatment

Figure 4.3: Results from the experiment

types, with their sizes reflecting the respective number of observations (which is largely determined by the sampling distribution). Relative to the dotted 45-degree line, the fitted ordinary least squares (OLS) line is rotated clockwise around the point $(5, 5)$. Its slope of 0.279 is significantly smaller than one, i.e., answers are biased towards the population average (see Column 1 of Table 4.3 below for details).

To test the further hypotheses, we apply the estimator from Equation 4.7 to our experimental data. Recall that a given subject saw each of the sizes that were drawn for her exactly four times. Therefore, we treat the respective four answers given by a subject as referring to the same characteristic. We hence observe $K = 15$ characteristics and $T = 4$ waves.[7] Figure 4.3b shows the distribution of $\hat{\tau}$, separately for the Short and the Long-treatment (gray and transparent, respectively).

In support of Hypothesis 4.2, estimates of τ are higher for subjects in the Long-treatment than for those in the Short-treatment ($p < 0.001$, Mann–Whitney U test). Conversely, this implies that our estimator predicts subjects' treatment status. A simple probit regression of an indicator variable for the Long-treatment on our estimates for τ yields a significant positive coefficient value (average marginal increase in the predicted probability $= 0.37$; $p < 0.001$, two-sided).

For the tests of Hypotheses 4.3 and 4.4, we turn to Table 4.3. Column 1 corresponds to the fitted line shown in Figure 4.3a, regressing reports on true types within the full sample. Columns 2 and 3 replicate Column 1 separately for the two treatments, Short and Long. In comparison with the pooled sample, the slope is flatter for the Short and steeper for the Long-treatment. The three possible pairwise differences in slopes (full sample, Short-treatment, and Long-treatment) are all statistically significant ($p < 0.001$, two-sided). This is in line with a successful treatment manipulation of τ and with Hypothesis 4.3. In Columns 4 and 5, we split the sample by $\hat{\tau}$. As predicted, for subjects with above-median values of $\hat{\tau}$, the estimated coefficient for the relationship between reports and true types is larger than for below-median subjects (Column 4) and the whole sample (Column 1).

[7] The estimator uses the information that, e.g., signals 3, 18, 33, and 48 showed the same true type, but it does *not* use the information what that type was.

Table 4.3: Relationship between reports and true types

Subjects	*Dependent variable:* Report				
	all	*by treatment*		*by $\hat{\tau}$*	
		Short	*Long*	*low*	*high*
	(1)	(2)	(3)	(4)	(5)
True type	0.279***	0.190***	0.366***	0.162***	0.404***
	(0.0167)	(0.0169)	(0.0260)	(0.0138)	(0.0250)
Constant	3.565***	3.973***	3.177***	4.100***	3.012***
	(0.0799)	(0.0859)	(0.122)	(0.0741)	(0.116)
Observations	11940	5880	6060	5640	5640
Clusters	199	98	101	94	94
R^2	0.134	0.0793	0.189	0.0491	0.243
ΔR^2		139%, $p < 0.001$		394%, $p < 0.001$	

Note: The table reports OLS estimates. The sample underlying Columns 4 and 5 excludes eleven subjects for whom there exists no variation in answers and, therefore, no estimates for τ are available. The p-values for the respective sizes of ΔR^2 are each based on 10,000 permutations (Heß, 2017). Standard errors clustered at the subject level in parentheses. $^{*}\,p < 0.05$, $^{**}\,p < 0.01$, $^{***}\,p < 0.001$.

Again, all three possible pairwise differences are statistically significant ($p < 0.001$, two-sided).[8]

To test Hypothesis 4.4, which states that the predictive power of reports for true types should increase in a population's level of τ, we again draw on Table 4.3 and compare the R^2-values within the two pairs of sub-samples (Columns 2–5). The data confirm our hypothesis: relative to the Short-treatment, the value of R^2 in the Long-treatment is more than doubled (comparison of Columns 2 and 3; $p < 0.001$, two-sided). For the two sub-samples based on the estimator $\hat{\tau}$, the difference is even larger (comparison of Columns 4 and 5): the R^2 for the above-median sample is about five times as large as the respective R^2 for the below-median subjects ($p < 0.001$, two-sided). In addition to supporting our hypothesis, these comparisons show that the estimates $\hat{\tau}$ are more informative than knowledge about subjects' treatment status. This is remarkable, given that our estimator only uses the pattern of subjects' responses. We conclude this section with a discussion about two further analyses (i) using individual-level data and (ii) using survey items on the quality of answers.

Individual-level Data. Recall that each subject in the experiment made 60 estimation decisions. This means that we can run regressions of these 60 reports on the respective true states *separately for each individual*. The resulting individual-specific value of R^2 is informative about how well a subject is able to discriminate between different true states, and it is therefore informative about τ. Moreover, the individual slope parameter reveals how much weight is assigned to signals, and thus it is informative about the level of subjective

[8]Note that the difference between the sub-samples in Columns 4 and 5 is more pronounced than the one between Columns 2 and 3: the coefficient in the low-τ sub-sample (Column 4) is smaller than the one for the Short-treatment in Column 2 ($p = 0.060$, two-sided), and the high-τ coefficient in Column 5 is larger than the Long-treatment coefficient in Column 3 ($p = 0.045$, two-sided).

knowledge, or confidence, $\tilde{\tau}$. Several observations can be made. First, in individual-level regressions, the values of R^2 and the slope parameters are strongly positively related, with a rank correlation of 0.83 ($p < 0.001$, two-sided, $N = 188$).[9] This positive correlation supports the central assumption of the model that agents with more knowledge (measured in terms of R^2) place more weight on their signals (measured in terms of coefficients). Second, the individual-level values of R^2 allow us to further test the validity of our estimator from Section 4.3, the latter *not* using information about the true types. We find that the individual values of R^2 are strongly correlated with the values of $\hat{\tau}$: the rank correlation is 0.83 ($p < 0.001$, two-sided, $N = 188$). In fact, this relationship can be analyzed even more thoroughly. In light of our model, the R^2-values can be transformed into alternative estimates of τ according to the formula $\hat{\tau}_{\text{alt.}} = R^2 / \left(1 - R^2\right)$.[10] The Pearson correlation between the alternative estimates and our main estimates $\hat{\tau}$ is 0.98 ($p < 0.001$, two-sided, $N = 188$). This finding is not mechanic, since the identification approaches behind the two estimators rest on entirely different information in the data: the R^2-based measure uses the information about true states, while our main estimator only uses information about which of the states are identical across the four waves.

Survey Items on Self-knowledge. We have argued that accounting for differences in (self-)knowledge can help to improve estimates, and we have suggested an estimator based on the pattern of behavior. An alternative to using this estimator could be to simply ask respondents directly how accurate or reliable they think their responses are. The use of such survey items appears to be fairly common. At the end of the experiment, we asked two such items and can compare their discriminatory power to that of our estimates $\hat{\tau}$. In particular, we asked subjects "how difficult" they thought the estimation task had been and "how sure" they were about their answers. The answers to both questions were provided on seven-point Likert scales. Reassuringly, responses to these two items are strongly negatively correlated ($\rho = -0.59$; $p < 0.001$, two-sided). To obtain a single measure, we take the first principal component of these two items. The rank correlation between this measure of *self-reported precision* and our estimate of knowledge $\hat{\tau}$ is only 0.05 and statistically insignificant ($p = 0.46$, two-sided, $N = 188$). However, the rank correlation between self-reported precision and the individual-level values of R^2 is also just 0.08 ($p = 0.29$, two-sided, $N = 188$), i.e., very small and, in particular, much smaller than the respective correlation of 0.83 between R^2 and $\hat{\tau}$. These results suggest that—in contrast to our estimator—, the survey items of self-reported precision contain only very limited information.

[9] As in Table 4.3, the eleven subjects for whom there exists no variation in answers (all of them always chose the answer "5") have to be excluded.

[10] For the derivation, see the last paragraph of Appendix 4.C.1.

4.5 Applications

In this section, we apply our estimator to data from the German Socio-economic Panel (SOEP),[11] a large, representative panel data set. The main objective is to show that by using estimates of self-knowledge, $\hat{\tau}$, we can increase the explanatory power of regressions that involve self-reports. In particular, we estimate τ using answers to the Big Five personality inventory from multiple waves and split the relevant samples by the respective median levels of $\hat{\tau}$, exactly as it was done in Section 4.4 for the data from the experiment (see Table 4.3). We illustrate differences in explanatory power (R^2) between the resulting sub-samples in the context of risk preferences, using self-reported measures of individual willingness to take risks. Following recent work on consequences and determinants of risk preferences, we use the preference measures to explain economic outcomes (with risk measures on the right-hand side) and explore determinants such as gender (with risk measures on the left-hand side).

4.5.1 Data and Measures

Our measure of self-knowledge is constructed using the fifteen-item Big Five inventory that was included in the 2005, 2009, 2013, and 2017 waves of the SOEP (Gerlitz and Schupp, 2005). The respective questions are particularly suitable for our purposes since they are meant and designed to cover independent traits that are stable over time (see, e.g., Cobb-Clark and Schurer, 2012). We use the maximum number of waves available for a given respondent, i.e., two waves for 47.4%, three waves for 22.1%, and four waves for 30.4% of the respondents (N = 21,157). The estimator introduced in Section 4.3 assumes that types are identically distributed for different characteristics. Empirically, however, the means and variances of answers might differ for different characteristics. Therefore, we add the following modification to our estimation procedure:

1. We construct normalized responses n_{ikt} as the difference between agent i's response r_{ikt} and the average response $\bar{r}_k$, divided by the standard deviation s_k of agents' average responses $\bar{r}_{ik}$ for the given characteristic k.

$$n_{ikt} = \frac{r_{ikt} - \bar{r}_k}{s_k}, \qquad \text{with} \quad s_k = \sqrt{\frac{1}{I-1}\sum_{i=1}^{I}(\bar{r}_{ik} - \bar{r}_k)^2}$$

2. Analogous to Equation 4.7, we use the standardized answers to apply the following estimator.

$$\hat{\tau}_i^{POP} = \frac{\frac{1}{K-1}\sum_{k=1}^{K}\left(\bar{n}_{ik} - \bar{n}_i\right)^2}{\frac{1}{K(T-1)-2}\sum_{k=1}^{K}\sum_{t=1}^{T}\left(n_{ikt} - \bar{n}_{ik}\right)^2} - \frac{1}{T} \tag{4.10}$$

As we show in Appendix 4.B.1, for $I \to \infty$, this population-based estimator retains the properties that were stated in the theorem in Section 4.3. In Appendix 4.B.2, we also

[11] Socio-Economic Panel (SOEP), data for years 1984–2017, version 34, SOEP, 2019, doi:10.5684/soep.v34.

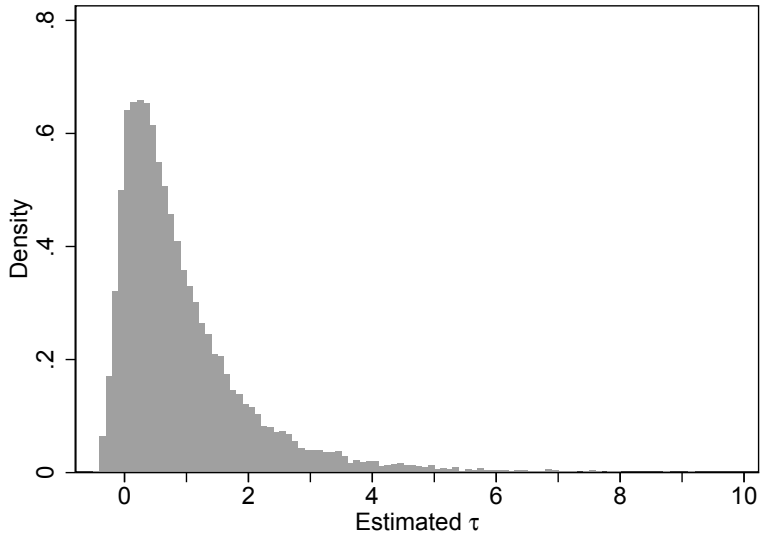
consider the case that characteristics are correlated, with results showing that the estimator remains informative.

Figure 4.4 shows the empirical distribution of $\hat{\tau}$ in the SOEP sample. We see considerable variation in these estimates, suggesting substantial heterogeneity in latent self-knowledge. The median value is 0.64, and for about 66% of respondents, the estimate $\hat{\tau}$ is smaller than one.

The main focus of this section is to show how empirical relationships between non-cognitive skills and economic outcomes are attenuated due to limited self-knowledge. However, the concept of self-knowledge might also be interpreted as an individual trait, i.e., an interesting object in itself: high or low self-knowledge can be thought of as an integral part of one's personality, reflecting individual differences in life experience, cognitive skills, or parental influence. Before turning to the main analyses, we therefore briefly consider potential determinants of τ, treating it as an individual trait.

In Table 4.4, we present results from regressions of estimated self-knowledge $\hat{\tau}$ on a set of plausibly exogenous determinants, in particular gender and age, as well as education. As shown in Column 1, self-knowledge is very weakly correlated with gender, with an R^2 of virtually zero. With respect to age, Column 2 reveals a hump-shaped relationship with self-knowledge. Descriptively, the latter increases until the age of about 43 years and then declines. However, the coefficients and the values of R^2 are fairly small. Given that self-knowledge might reflect differences in cognitive skills, we also consider an association with education (see Column 3). The correlation is significant and indicates that one more year of education is, on average, associated with an increase of about 0.06 in the level of self-knowledge. In Column 4, we regress estimated self-knowledge simultaneously on all of the previously considered variables. Education seems to dominate, as becomes apparent when comparing the values of R^2 between the columns. However, even the combined R^2 of 0.033 is fairly low, suggesting that the estimates of self-knowledge contain much information above and beyond socio-demographic characteristics.

If self-knowledge is a trait, it might be intergenerationally transmitted, similar to, e.g., risk aversion, trust, patience, and social preferences (Dohmen et al., 2012; Kosse and Pfeiffer, 2012; Alan et al., 2017; Kosse et al., 2020). Such transmission could come from various sources, e.g., imitation, exposure to similar social environments, or genetic dispositions. Among the SOEP participants for whom we have estimates of self-knowledge, we can match 3,573 respondents to their mothers and 2,964 respondents to their fathers. Figure 4.5 scrutinizes the relationship between parents' estimated levels of self-knowledge and the respective estimates for their children. For this purpose, each survey respondent is assigned to the respective decile in the distribution of $\hat{\tau}$ in the full sample. The figure depicts the average deciles for children conditional on the respective parental deciles, separately for mothers (left panel) and fathers (right panel). As the corresponding regression lines indicate, parents' and children's estimated levels of self-knowledge are positively related ($p < 0.001$, two-sided, separately for both cases). In terms of the precise underlying values of $\hat{\tau}$ (i.e., not in terms of deciles), the rank correlation between children and their parents is 0.17 in the case of mothers and 0.16 for fathers.



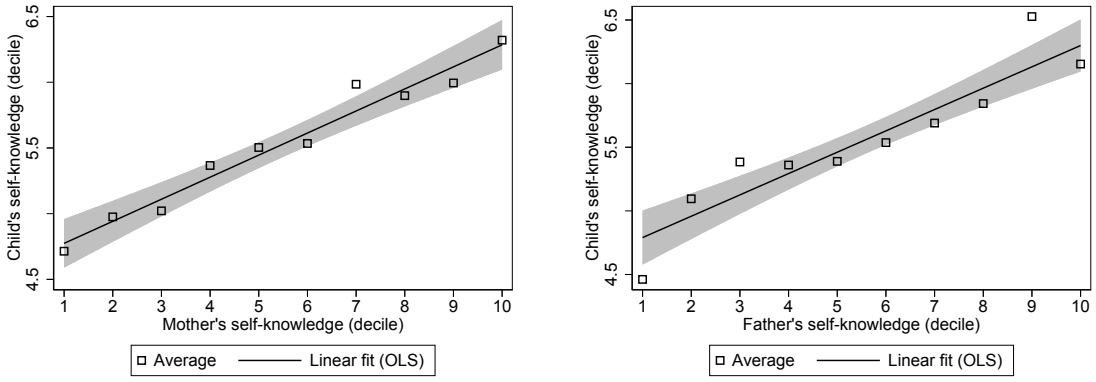
Note: Distribution of $\hat{\tau}$ in the German SOEP. Estimates that are larger than ten (48 out of 21,157) are not displayed.

Figure 4.4: Distribution of $\hat{\tau}$ in the SOEP

Table 4.4: Correlations with τ

	Dependent variable: $\hat{\tau}$			
	(1)	(2)	(3)	(4)
Female	-0.0286*			-0.00105
	(0.0139)			(0.0151)
Age (in '11)		0.0126***		0.00822**
		(0.00215)		(0.00269)
Age^2 (in '11)		-0.000148***		-0.0000971***
		(0.0000207)		(0.0000254)
Edu. years (in '11)			0.0623***	0.0598***
			(0.00292)	(0.00296)
Constant	0.926***	0.696***	0.139***	0.0338
	(0.0103)	(0.0523)	(0.0359)	(0.0747)
Observations	20946	20946	16158	16158
R^2	0.000	0.004	0.031	0.033

Note: The table reports OLS estimates. Individuals for whom $\hat{\tau}$ lies above the 99$^{\text{th}}$ percentile are excluded. Age as well as years of education refer to the year 2011, i.e., the center of the relevant time interval (2005–2017). Heteroskedasticity-robust standard errors in parentheses. $^{*}\,p < 0.05$, $^{**}\,p < 0.01$, $^{***}\,p < 0.001$.



(a) From mother (b) From father

Note: Graphs plot average deciles to which children's $\hat{\tau}$ belong for the deciles of $\hat{\tau}$ for mothers (left) and fathers (right). Deciles refer to the full sample, i.e., thresholds are the same for parents and children. The shaded areas around the regression lines represent 95% confidence intervals. Standard errors are clustered at the level of the respective parents.

Figure 4.5: Intergenerational transmission of self-knowledge

4.5.2 Predicting Outcomes

To illustrate how accounting for individual estimates of $\hat{\tau}$ can increase explanatory power in the context of non-cognitive skills, we study the effect of risk attitudes on various economic outcomes. Similar to the analysis of the experiment in Table 4.3, we split the respective samples of SOEP respondents into two groups: individuals with either low self-knowledge (below the median value of $\hat{\tau}$) or high self-knowledge (above the median level of $\hat{\tau}$). This way, we refrain from imposing any functional form assumptions about how self-knowledge affects the estimates. In light of the model and the experimental results, we would expect to see larger explanatory power for the above-median sample than for the below-median sample, reflected in larger values of R^2.

Table 4.5 presents empirical results for three different economic outcomes related to risk attitudes: holding risky financial securities, receiving performance-related pay, and smoking. These outcomes were selected based on prior research, arguing that they should—and actually are—related to risk attitudes (Dohmen and Falk, 2011; Dohmen et al., 2011). The measures that we use to elicit risk attitudes are survey items that ask about willingness to take risks in specific domains on eleven-point Likert scales. In particular, the items refer to the willingness to take risks concerning one's *financial matters*, *career*, and *health*, respectively. Columns 1–3 show results from OLS regressions without further controls, and Columns 4–6 replicate the analyses controlling for a set of socio-demographic characteristics, namely (squared) age, gender, body height, years of education, parental education, log net household income, log wealth, and log debts. Columns 1 and 4 consider the full sample and confirm a positive and significant relationship between risk attitudes and the respective outcomes.

Our main interest concerns the pairwise comparisons between Columns 2 and 3 and those between Columns 5 and 6, where we show results for individuals with estimated levels of self-knowledge below and above the median. In all instances, the values of R^2 are higher for individuals with high self-knowledge than for individuals with low self-knowledge, and the values for the full sample are between those for the two sub-samples.[12] This holds both with and without controls included in the regressions (in the regressions with controls, we refer to the partial R^2). In all cases, the explanatory power among high self-knowledge respondents is much larger than among the ones with low self-knowledge, ranging from an 87% increase (smoking, without controls) up to an increase of 610% (performance pay, with controls). As the respective p-values show, the differences in explanatory power are statistically significant. We note that these results hold for a non-cognitive skill—risk attitude—that is different and mostly unrelated to the set of traits that we used to estimate $\hat{\tau}$ (the Big Five). This suggests that self-knowledge does, in fact, generalize to different aspects of people's personalities.

[12] In Appendix 4.C, we discuss how to interpret differences in the estimated coefficients.

Table 4.5: Predictive power of domain-specific attitudes towards risk

	Without controls			Including controls		
Sample:	*pooled*	*below*	*above*	*pooled*	*below*	*above*
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable: *Risky financial securities*						
Risk attitude	0.0698***	0.0519***	0.0863***	0.0523***	0.0383***	0.0665***
	(0.00264)	(0.00359)	(0.00369)	(0.00280)	(0.00383)	(0.00398)
(Partial) R^2	0.0827	0.0520	0.114	0.0498	0.0295	0.0737
Observations	9095	4548	4547	7472	3736	3736
ΔR^2		119%,	$p < 0.001$		150%,	$p < 0.001$
Dependent variable: *Performance pay*						
Risk attitude	0.0128***	0.00808***	0.0174***	0.00977***	0.00491	0.0150***
	(0.00174)	(0.00221)	(0.00268)	(0.00199)	(0.00252)	(0.00310)
(Partial) R^2	0.00870	0.00412	0.0139	0.00487	0.00142	0.0101
Observations	5758	2879	2879	4464	2232	2232
ΔR^2		238%,	$p = 0.03$		610%,	$p = 0.02$
Dependent variable: *Smoking*						
Risk attitude	0.0199***	0.0175***	0.0229***	0.0125***	0.00923***	0.0158***
	(0.00154)	(0.00220)	(0.00216)	(0.00175)	(0.00248)	(0.00246)
(Partial) R^2	0.0119	0.00888	0.0166	0.00481	0.00258	0.00782
Observations	15162	7581	7581	11652	5826	5826
ΔR^2		87%,	$p = 0.04$		203%,	$p = 0.06$

Note: The table reports OLS estimates, with binary dependent variables taking the values zero and one. If not stated otherwise, all the data refer to the year 2009. Regressions are based only on respondents who are 18 years or older, and those for performance pay include only respondents up to the age of 66 who work full-time and receive wages or salaries. *Risky financial securities* are, in the SOEP, a residual category of securities without a fixed interest rate, like stocks or options ("other securities"). Since the relevant question was asked on the household level in 2010, the units of observation in the respective regressions are households in that year. *Performance pay* indicates that an employee receives payments from profit-sharing, premiums, or bonuses. Smoking refers to 2010. The variable *risk attitude* in each of the panels refers to the respective domain-specific question asked in the SOEP. The contexts are *financial matters* for holding risky financial securities, *career* for performance pay, and *health* for smoking. The controls used in Columns 4–6 are gender, age, squared age, body height in 2010, years of education, parental education (whether mother and father each have either *Abitur* or *Fachabitur*), log net household income, and log wealth and log debts of the current household in 2007. The last three variables are calculated as ln(euro amount + 1). For the regressions involving *risky financial securities*, all variables are averaged on the household level, and we base our data only on respondents for whom all information is available individually. The p-values for the sizes of ΔR^2 are each based on 10,000 permutations. Heteroskedasticity-robust standard errors in parentheses. $^{*}\, p < 0.05$, $^{**}\, p < 0.01$, $^{***}\, p < 0.001$.

4.5.3 Determinants of Preferences

An active literature seeks to uncover the individual determinants of preferences and personality (e.g., Sutter and Kocher, 2007; Croson and Gneezy, 2009; Falk et al., 2018). Understanding how, e.g., age and gender affect preferences is not only interesting in itself. It is also relevant for gaining a better understanding of group-specific outcomes, such as gender differences with respect to sorting into competitive environments, wage gaps, and occupational choice, to give just one example (see, e.g., Niederle and Vesterlund, 2007; Croson and Gneezy, 2009; Dohmen and Falk, 2011; Buser, Niederle, and Oosterbeek, 2014). Here, we use differences in domain-specific risk attitudes associated with gender and height to illustrate that when accounting for differences in self-knowledge, exogenous determinants of preferences may actually have higher explanatory power and yield larger effect sizes than typically inferred.

Table 4.6 reports the differences associated with gender and body height for two different measures of risk attitudes, both based on the 2009 wave of the SOEP and standardized according to the pooled samples used in Column 1.[13] One measure is the so-called general risk question, asking about the willingness to take risks "in general" and measured on an eleven-point Likert scale, while the other is the first principal component of five domain-specific risk questions, referring to *car driving*, *financial matters*, *sports/leisure*, *career*, as well as *health*. Replicating previous findings,[14] women tend to be less willing to take risks than men, and taller people tend to be more willing to take risks than smaller individuals. Our interest here is to compare samples with high vs. low levels of self-knowledge, as shown in Columns 2 and 3. In all four instances, we consistently find that explanatory power, measured in terms of R^2, is larger among high-τ individuals than among low-τ individuals. These differences are substantial, ranging from 36% to 44%. The p-values for differences in explanatory power imply statistical significance. An inspection of estimated coefficients further shows that the size of coefficients is always (absolutely) larger for the above-median sample than for the below-median sample. Increased effect sizes are at odds with classical measurement error but in line with the predictions of our model. They also mimic the results from our stylized experiment in Section 4.4, where we saw a steeper slope between reports and true states for high-τ relative to low-τ subjects (see Table 4.3, Columns 4 and 5).

A potential concern regarding the interpretation of the above results is selection. The latter would imply that the observed patterns reflect that the true explanatory power, as well as true coefficients, are actually larger among respondents with high self-knowledge. In principle, we cannot rule out such an interpretation with non-experimental data. However, recall from Table 4.4 that the effects of socio-demographic characteristics on $\hat{\tau}$ were rather small. It is therefore unlikely that selection plays a major role in our findings. Still, we address this issue explicitly in Columns 4 and 5 by restoring representativeness with respect to observable characteristics using inverse probability weighting. We estimate

[13] Individuals' height again refers to 2010, due to availability of data.

[14] See in particular Dohmen et al. (2011) but also Croson and Gneezy (2009) and Falk et al. (2018).

Table 4.6: Differences in risk attitudes

Sample:	Unweighted			Weighted	
	pooled	*below*	*above*	*below*	*above*
	(1)	(2)	(3)	(4)	(5)
General risk question					
Female	-0.374***	-0.348***	-0.400***	-0.350***	-0.410***
	(0.0153)	(0.0218)	(0.0214)	(0.0219)	(0.0221)
R^2	0.0348	0.0297	0.0404	0.0305	0.0416
Observations	16654	8327	8327	8327	8327
ΔR^2		36%, $p = 0.05$		36%, $p = 0.05$	
Height (in '10)	0.0219***	0.0202***	0.0236***	0.0196***	0.0249***
	(0.000864)	(0.00121)	(0.00124)	(0.00122)	(0.00127)
R^2	0.0425	0.0349	0.0502	0.0333	0.0556
Observations	15134	7567	7567	7567	7567
ΔR^2		44%, $p = 0.02$		67%, $p < 0.01$	
Domain-specific risk questions: first principal component					
Female	-0.433***	-0.401***	-0.463***	-0.405***	-0.477***
	(0.0164)	(0.0238)	(0.0226)	(0.0238)	(0.0236)
R^2	0.0469	0.0390	0.0558	0.0402	0.0578
Observations	14160	7080	7080	7080	7080
ΔR^2		43%, $p = 0.02$		44%, $p = 0.02$	
Height (in '10)	0.0271***	0.0252***	0.0284***	0.0243***	0.0305***
	(0.000947)	(0.00133)	(0.00135)	(0.00135)	(0.00138)
R^2	0.0648	0.0532	0.0750	0.0498	0.0855
Observations	12858	6429	6429	6429	6429
ΔR^2		41%, $p = 0.01$		72%, $p < 0.001$	

Note: The table reports OLS estimates. All regressions only use respondents who are 18 years or older. The dependent variables are standardized among the respondents who enter the corresponding regression in Column 1. Columns 4 and 5 use inverse probability weights that come from probit regressions of group assignment on gender, a second-order age polynomial, and years of education. Except for height, all data refer to the year 2009. The p-values for the respective sizes of ΔR^2 are each based on 10,000 permutations. Heteroskedasticity-robust standard errors in parentheses. $^{*}\, p < 0.05$, $^{**}\, p < 0.01$, $^{***}\, p < 0.001$.

probit models in which we regress group assignment (below or above median) on gender, a second-order age polynomial, and years of education (all as of 2009). We then invert the predicted probabilities and use them as weights, otherwise replicating the regression from Columns 2 and 3. The results change very little and even tend to become slightly stronger.[15] Thus, the findings suggest that individuals with relatively high levels of self-knowledge do, in fact, contribute more to our understanding of the determinants of non-cognitive skills than their low-τ counterparts.

We conclude this section with a brief discussion of the differences associated with gender and height in the Big Five personality traits. Table 4.D.2 in Appendix 4.D.2 is constructed analogously to Table 4.6 but analyzes the Big Five rather than risk attitudes. It shows that both effects—higher explanatory power and larger effect sizes for high-τ

[15] A corresponding procedure can, of course, also be applied in the context of the differences in predictive power that were analyzed in Table 4.5. Although given the controls that are used, it seems less needed at that point, we still report the corresponding results in Table 4.D.1 in Appendix 4.D.1. The effect sizes decrease a bit, but the results still support the earlier conclusions.

relative to low-τ individuals—are also observed for the Big Five Inventory. As an example, take conscientiousness, which is considered one of the most important personality traits for explaining educational and labor market outcomes (Judge et al., 1999; Hogan and Holland, 2003; Almlund et al., 2011) as well as health and mortality (see, e.g., Bogg and Roberts, 2004; Hill et al., 2011). The comparison of Columns 2 and 3 shows that the gender difference is almost three times as large for the high-τ compared with the low-τ individuals and that the difference in R^2 amounts to more than 500%.[16] This also suggests the possibility that treatment effects of childhood interventions on personality traits (see Heckman and Kautz, 2012) could be even larger than previously assumed.

4.6 Conclusion

In this chapter, we have suggested a theoretical framework of survey response behavior. We assume that respondents try to provide accurate answers but lack perfect self-knowledge. In addition, survey responses may be affected in terms of subjective scale use, inaccurate beliefs about one's self-knowledge, differences in the endogenous precision of reports, as well as image or social desirability effects. The framework is kept deliberately simple but could be extended to allow for a richer and more realistic analysis of survey response behavior. For example, we assume that the outcome of inspecting one's individual characteristics is simply an (exogenous) signal about one's type. It would be interesting to explore cognitive (and emotional) processes involved in this introspection process in more detail, e.g., the role of limited memory and retrieval, how individuals select representative choice contexts to evaluate their characteristics, or how social comparison or life experience affect introspection. The framework also allows for integrating the role and meaning of response times, which could hold strong practical importance. For example, many binary choice experiments in neuroscience and psychology find that accuracy decreases with response time, in the sense that slower decisions are less likely to be correct (Swensson, 1972; Luce, 1986; Ratcliff and McKoon, 2008).[17] An interesting question is how one can integrate response times into our approach to facilitate the identification of precise responses.

We note that while we have interpreted the model in terms of survey response behavior, it can be applied to any elicitation method where subjects make a decision, i.e., in particular to lab and field experiments. For instance, in typical choice experiments to elicit risk or time preferences, the same issues that we discuss in the context of survey responses also arise. In fact, a main difference in experiments is the provision of incentives, which may increase the accuracy of responses (see Section 4.2.2) but do not solve the issues of limited self-knowledge (in the sense of introspective ability), scale use, or social desirability.

A better understanding of the survey response process may also inform the "optimal" design of research. Conditional on survey respondents' behavior, we can ask the question

[16] For the Big Five, we can also show that Cronbach's alpha, a common psychometric measure for scale consistency, is higher among respondents with high (above-median) levels of $\hat{\tau}$. Among low-$\hat{\tau}$ respondents in the SOEP, the average across the five facets is 0.50, while it is 0.67 among high-$\hat{\tau}$ respondents.

[17] Fudenberg, Strack, and Strzalecki (2018) and Alós-Ferrer, Fehr, and Netzer (2021) provide theoretical analyses of the relationship between response times and the accuracy of binary decisions.

of how surveys or other elicitation methods should be designed to extract a maximum amount of information. Such a design perspective would consider research as a principal–agent relationship where agents participate in surveys, experiments, or related research contexts that are designed by researchers who optimize research paradigms conditional on agents' behaviors. Such an approach could be used to investigate how to design survey items and response scales, when and how incentives should be given, or how to design specific modules meant to correct for expected biases.

A key insight of the model is that we can extract individual differences in self-knowledge based on response patterns, in particular by using the ratio of the variance between characteristics and the variance for a given characteristic over time. Building on this finding, we suggest a consistent and unbiased estimator of self-knowledge, discuss its properties, and apply it to experimental data as well as a large panel data set. We show that the estimator reliably identifies individual differences in the informativeness of answers in the laboratory context where we know true states. Splitting the lab sample into individuals giving answers with high vs. low quality, we further show that reports are much closer to true states for the former than for the latter part of the sample. Repeating the same exercise using a representative panel data set and risk attitudes as an example for non-cognitive skills, we show that for subjects with a high level of self-knowledge, the explained variance is significantly higher than for individuals with low levels of self-knowledge. This holds for regressions where risk attitudes are on either the left- or the right-hand side of the regression equation. These applications illustrate the potential of distinguishing between respondents with high vs. low self-knowledge for improving survey evidence. They suggest further econometric implications for the study of measurement error and highlight the potential of integrating self-knowledge into regression frameworks.

References

Alan, Sule, Nazli Baydar, Teodora Boneva, Thomas F. Crossley, and Seda Ertac. 2017. "Transmission of risk preferences from mothers to daughters". *Journal of Economic Behavior and Organization* 134:60–77.

Almlund, Mathilde, Angela Lee Duckworth, James Heckman, and Tim Kautz. 2011. "Personality Psychology and Economics". In *Handbook of the Economics of Education*, ed. by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 4:1–181. 2008. Amsterdam, Netherlands: Elsevier B.V.

Alós-Ferrer, Carlos, Ernst Fehr, and Nick Netzer. 2021. "Time Will Tell: Recovering Preferences When Choices Are Noisy". *Journal of Political Economy* 129 (6): 1828–1877.

Beauchamp, Jonathan, Daniel J. Benjamin, David I. Laibson, and Christopher F. Chabris. 2020. "Measuring and controlling for the compromise effect when estimating risk preference parameters". *Experimental Economics* 23:forthcoming.

Beauchamp, Jonathan P., David Cesarini, and Magnus Johannesson. 2017. "The psychometric and empirical properties of measures of risk preferences". *Journal of Risk and Uncertainty* 54 (3): 203–237.

Bénabou, Roland, Armin Falk, Luca Henkel, and Jean Tirole. 2020. *Eliciting Moral Preferences*. Working paper. Bonn, Germany: briq – Institute on Behavior & Inequality.

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch. 2014. "hroot: Hamburg Registration and Organization Online Tool". *European Economic Review* 71:117–120.

Bogg, Tim, and Brent W. Roberts. 2004. "Conscientiousness and Health-Related Behaviors: A Meta-Analysis of the Leading Behavioral Contributors to Mortality". *Psychological Bulletin* 130 (6): 887–919.

Bolsinova, Maria, Paul de Boeck, and Jesper Tijmstra. 2017. "Modelling Conditional Dependence Between Response Time and Accuracy". *Psychometrika* 82 (4): 1126–1148.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement Error in Survey Data". Chap. 59 in *Handbook of Econometrics*, ed. by James J. Heckman and Edward Leamer, 5:3705–3843. Amsterdam, Netherlands: Elsevier Science B.V.

Bradburn, Norman M., Seymour Sudman, and Brian Wansink. 2004. *Asking Questions: The Definitive Guide to Questionnaire Design*. 448 pp. San Francisco: Jossey-Bass.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek. 2014. "Gender, Competitiveness, and Career Choices". *Quarterly Journal of Economics* 129 (3): 1409–1447.

Camerer, Colin, and Dan Lovallo. 1999. "Overconfidence and Excess Entry: An Experimental Approach". *American Economic Review* 89 (1): 306–318.

Camerer, Colin F., and Robin M. Hogarth. 1999. "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework". *Journal of Risk and Uncertainty* 19 (1-3): 7–42.

Caplin, Andrew, Dániel Csaba, John Leahy, and Oded Nov. 2020. "Rational Inattention, Competitive Supply, and Psychometrics". *Quarterly Journal of Economics* 135 (3): 1681–1724.

Caplin, Andrew, and Mark Dean. 2015. "Revealed Preference, Rational Inattention, and Costly Information Acquisition". *American Economic Review* 105 (7): 2183–2203.

Chen, Yuanyuan, Shuaizhang Feng, James J. Heckman, and Tim Kautz. 2020. "Sensitivity of self-reported noncognitive skills to survey administration conditions". *Proceedings of the National Academy of Sciences* 117 (2): 931–935.

Cobb-Clark, Deborah A., and Stefanie Schurer. 2012. "The stability of big-five personality traits". *Economics Letters* 115 (1): 11–15.

Croson, Rachel, and Uri Gneezy. 2009. "Gender Differences in Preferences". *Journal of Economic Literature* 47 (2): 448–474.

Cvitanić, Jakša, Dražen Prelec, Blake Riley, and Benjamin Tereick. 2017. *Honesty via Choice-Matching*. Mimeo.

Dohmen, Thomas, and Armin Falk. 2011. "Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender". *American Economic Review* 101 (2): 556–590.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde. 2012. "The Intergenerational Transmission of Risk and Trust Attitudes". *Review of Economic Studies* 79 (2): 645–677.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. "Individual risk attitudes: Measurement, determinants, and behavioral consequences". *Journal of the European Economic Association* 9 (3): 522–550.

Drerup, Tilman, Benjamin Enke, and Hans Martin von Gaudecker. 2017. "The precision of subjective data and the explanatory power of economic models". *Journal of Econometrics* 200 (2): 378–389.

Edwards, Michael C. 2009. "An Introduction to Item Response Theory Using the Need for Cognition Scale". *Social and Personality Psychology Compass* 3 (4): 507–529.

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

Falk, Armin, David Huffman, and Uwe Sunde. 2006a. *Do I Have What It Takes? Equilibrium Search with Type Uncertainty and Non-Participation*. IZA Discussion Paper 2531. Bonn, Germany: IZA Institute for the Study of Labor.

— . 2006b. *Self-Confidence and Search*. IZA Discussion Paper 2525. Bonn, Germany: IZA Institite for the Study of Labor.

Falk, Armin, and Florian Zimmermann. 2013. "A Taste for Consistency and Survey Response Behavior". *CESifo Economic Studies* 59 (1): 181–193.

— . 2017. "Consistency as a Signal of Skills". *Management Science* 63 (7): 2049–2395.

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

Fudenberg, Drew, Philipp Strack, and Tomasz Strzalecki. 2018. "Speed, Accuracy, and the Optimal Timing of Choices". *American Economic Review* 108 (12): 3651–3684.

Gerlitz, Jean-Yves, and Jürgen Schupp. 2005. *Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. Dokumentation der Instrumententwicklung BFI-S auf Basis des SOEP-Pretests 2005*. Research Notes 4. Berlin, Germany: Deutsches Institut für Wirtschaftsforschung (DIW).

Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2019. "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study". *Journal of Political Economy* 127 (4): 1826–1863.

Goebel, Jan, Markus M. Grabka, Stefan Liebig, Martin Kroh, David Richter, Carsten Schröder, and Jürgen Schupp. 2019. "The German Socio-Economic Panel Study (SOEP)". *Jahrbücher für Nationalökonomie und Statistik / Journal of Economics and Statistics* 239 (2): 345–360.

Heckman, James J., and Tim D. Kautz. 2012. *Hard Evidence on Soft Skills*. NBER Working Paper 18121. Cambridge, MA: National Bureau of Economic Research.

Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior". *Journal of Labor Economics* 24 (3): 411–482.

Heß, Simon. 2017. "Randomization inference with Stata: A guide and software". *Stata Journal* 17 (3).

Hill, Patrick L., Nicholas A. Turiano, Michael D. Hurd, Daniel K. Mroczek, and Brent W. Roberts. 2011. "Conscientiousness and Longevity: An Examination of Possible Mediators". *Health Psychology* 30 (5): 536–541.

Hogan, Joyce, and Brent Holland. 2003. "Using Theory to Evaluate Personality and Job-Performance Relations: A Socioanalytic Perspective". *Journal of Applied Psychology* 88 (1): 100–112.

Hyslop, Dean R., and Guido W. Imbens. 2001. "Bias from classical and other forms of measurement error". *Journal of Business and Economic Statistics* 19 (4): 475–481.

Judge, Timothy A., Chad A. Higgins, Carl J. Thoresen, and Murray R. Barrick. 1999. "The Big Five Personality Traits, General Mental Ability, and Career Success Across the Life Span". *Personnel Psychology* 52 (3): 621–652.

Kimball, Miles S., Claudia R. Sahm, and Matthew D. Shapiro. 2008. "Imputing risk tolerance from survey responses". *Journal of the American Statistical Association* 103 (483): 1028–1038.

Kosse, Fabian, Thomas Deckers, Pia Pinger, Hannah Schildberg-Hörisch, and Armin Falk. 2020. "The Formation of Prosociality: Causal Evidence on the Role of Social Environment". *Journal of Political Economy* 128 (2): 434–467.

Kosse, Fabian, and Friedhelm Pfeiffer. 2012. "Impatience among preschool children and their mothers". *Economics Letters* 115 (3): 493–495.

Kyllonen, Patrick, and Jiyun Zu. 2016. "Use of Response Time for Measuring Cognitive Ability". *Journal of Intelligence* 4 (4): 14.

Luce, R. Duncan. 1986. *Response Times: Their Role in Inferring Elementary Mental Organization*. New York, NY: Oxford University Press.

Malmendier, Ulrike, and Geoffrey Tate. 2005. "CEO Overconfidence and Corporate Investment". *Journal of Finance* 60 (6): 2661–2700.

Matějka, Filip, and Alisdair McKay. 2015. "Rational inattention to discrete choices: A new foundation for the multinomial logit model". *American Economic Review* 105 (1): 272–98.

Niederle, Muriel, and Lise Vesterlund. 2007. "Do Women Shy Away From Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics* 122 (3): 1067–1101.

Prelec, Dražen. 2004. "A Bayesian Truth Serum for Subjective Data". *Science* 306 (5695): 462–466.

Ratcliff, Roger, and Gail McKoon. 2008. "The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks". *Neural Computation* 20 (4): 873–922.

Schwarz, Norbert. 2007. "Cognitive Aspects of Survey Methodology". *Applied Cognitive Psychology* 21:277–287.

Sims, Christopher A. 2003. "Implications of rational inattention". *Journal of monetary Economics* 50 (3): 665–690.

— . 1998. "Stickiness". *Carnegie-Rochester Conference Series on Public Policy* 49:317–356.

Smith, Vernon L. 1976. "Experimental Economics: Induced Value Theory". *American Economic Review* 66 (2): 274–279.

Storey, Hannah. 2016. *Animal Pictograms and Sleep Infographic*. Visited on 05/28/2021. `https://hannah-storey.co.uk/post/160132182299/sleep-infographic`.

Sudman, Seymour, Norman M. Bradburn, and Norbert Schwarz. 1996. *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. 304 pp. San Francisco: Jossey-Bass.

Sutter, Matthias, and Martin G. Kocher. 2007. "Trust and trustworthiness across different age groups". *Games and Economic Behavior* 59 (2): 364–382.

Swensson, Richard G. 1972. "The elusive tradeoff: Speed vs accuracy in visual discrimination tasks". *Perception & Psychophysics* 12 (1A): 16–32.

Appendix 4.A Proofs

Proof of Lemma 4.1. Following the result from Equation 4.2, the optimal report for any given level of precision and signal is given by

$$r = \frac{\bar{\theta} + \tau\, x}{1 + \tau}\,.$$

Plugging into Equation 4.2.2 yields that the utility of the agent given the optimal response above equals

$$u_\theta(r, \tau) = -\frac{m}{(1+\tau)^2}\left[\bar{\theta} - \theta + \tau\,(x - \theta)\right]^2 - \frac{\tau}{a}$$

and, consequently, the agent chooses her precision τ to maximize

$$\begin{aligned}\mathbb{E}[u_\theta(r, \tau)] &= -\frac{m}{(1+\tau)^2}\left(\mathbb{E}\Big[\big(\bar{\theta} - \theta\big)^2\Big] + \tau^2 \mathbb{E}\Big[(x - \theta)^2\Big]\right) - \frac{\tau}{a} \\ &= -\frac{m}{(1+\tau)^2}\left(\sigma^2 + \tau^2 \frac{\sigma^2}{\tau}\right) - \frac{\tau}{a} = -\frac{m\,\sigma^2}{1+\tau} - \frac{\tau}{a}\,.\end{aligned}$$

Since $\mathbb{E}[u_\theta(r, \tau)]$ is strictly concave in r, the first-order condition yields the optimal level of effort for an interior solution.

$$0 \overset{!}{=} \frac{\partial\, \mathbb{E}[u_\theta(r, \tau)]}{\partial\, \tau} = \frac{m\,\sigma^2}{(1+\tau)^2} - \frac{1}{a} \quad \Rightarrow \quad \tau^* = \sqrt{m\,a}\,\sigma - 1$$

□

Proof of the theorem. We will prove the result in the more general setting with subjective self-knowledge and scale use as introduced in Sections 4.2.2 and 4.2.2, respectively. The case without subjective self-knowledge and scale use stated in the basic version of the model corresponds to the special case where $\tilde{\tau}_i = \tau_i$ and $\phi_i = 1$.

Throughout the proof, we fix $\tau_i, \tilde{\tau}_i > 0$ and $\phi_i \in (0, 1]$. The answer of agent i when asked for the t^{th} time about the k^{th} characteristic is given by

$$r_{ikt} = (1 - \phi_i)\, c + \phi_i\, \frac{\bar{\theta} + \tilde{\tau}_i\, x_{ikt}}{1 + \tilde{\tau}_i}\,.$$

By assumption, there exist independent, standard normally distributed random variables $\epsilon_{ikt}, \eta_{ik}$ such that

$$\begin{aligned} x_{ikt} &= \theta_{ik} + \frac{\sigma}{\sqrt{\tau_i}}\epsilon_{ikt}\,, \\ \theta_{ik} &= \bar{\theta} + \sigma\,\eta_{ik}\,. \end{aligned}$$

Plugging into the equation for the agent's responses yields that

$$r_{ikt} = (1 - \phi_i)\, c + \phi_i\, \left(\bar{\theta} + \frac{\tilde{\tau}_i}{1 + \tilde{\tau}_i}\sigma\left[\eta_{ik} + \frac{\epsilon_{ikt}}{\sqrt{\tau_i}}\right]\right)\,. \tag{4.A.1}$$

Denote agent i's average answer for question k by $\bar{r}_{ik} = \frac{1}{T}\sum_{t=1}^{T} r_{ikt}$, her average answer over all questions by $\bar{r}_i = \frac{1}{K}\sum_{k=1}^{K} \bar{r}_{ik}$, and similarly $\bar{x}_{ik} = \frac{1}{T}\sum_{t=1}^{T} x_{ikt}$, $\bar{\epsilon}_{ik} = \frac{1}{T}\sum_{t=1}^{T} \epsilon_{ikt}$, $\bar{x}_i = \frac{1}{K}\sum_{k=1}^{K} \bar{x}_{ik}$, $\bar{\epsilon}_i = \frac{1}{K}\sum_{k=1}^{K} \bar{\epsilon}_{ik}$, and $\bar{\eta}_i = \frac{1}{K}\sum_{k=1}^{K} \bar{\eta}_{ik}$. We have that

$$\frac{r_{ikt} - \bar{r}_{ik}}{\phi_i} = \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}(x_{ikt} - \bar{x}_{ik}) = \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\frac{\sigma}{\sqrt{\tau_i}}(\epsilon_{ikt} - \bar{\epsilon}_{ik})\,. \tag{4.A.2}$$

Similarly, we get that

$$\begin{aligned}\frac{\bar{r}_{ik} - \bar{r}_i}{\phi_i} &= \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}(\bar{x}_{ik} - \bar{x}_i) = \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\left(\left(\theta_{ik} + \frac{\sigma}{\sqrt{\tau_i}}\bar{\epsilon}_{ik}\right) - \left(\bar{\theta}_i + \frac{\sigma}{\sqrt{\tau_i}}\bar{\epsilon}_i\right)\right) \\ &= \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\left(\left(\theta_{ik} - \bar{\theta}_i\right) + \frac{\sigma}{\sqrt{\tau_i}}\left(\bar{\epsilon}_{ik} - \bar{\epsilon}_i\right)\right) \\ &= \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\left(\sigma\left(\eta_{ik} - \bar{\eta}_i\right) + \frac{\sigma}{\sqrt{\tau_i}}\left(\bar{\epsilon}_{ik} - \bar{\epsilon}_i\right)\right)\,.\end{aligned} \tag{4.A.3}$$

We first show that

$$A \coloneqq \frac{(1+\tilde{\tau}_i)^2}{\tilde{\tau}_i^2\sigma^2}\tau_i \sum_{k=1}^{K}\sum_{t=1}^{T}\left(\frac{r_{ikt} - \bar{r}_{ik}}{\phi_i}\right)^2$$

is χ^2 distributed with $K(T-1)$ degrees of freedom. It follows from Equation 4.A.2 that

$$A = \sum_{k=1}^{K}\sum_{t=1}^{T}\left(\epsilon_{ikt} - \bar{\epsilon}_{ik}\right)^2\,.$$

We have that $A_k \coloneqq \sum_{t=1}^{T}\left(\epsilon_{ikt} - \bar{\epsilon}_{ik}\right)^2$ is χ^2 distributed with $T-1$ degrees of freedom as it equals the sum of the squared distance of i.i.d. normals from the mean. As $A_k, A_{k'}$ are independent for $k' \neq k$ and $A = \sum_{k=1}^{K} A_k$, it follows that A is χ^2 distributed with $\sum_{k=1}^{K}(T-1) = K(T-1)$ degrees of freedom.

We next argue that

$$B \coloneqq \frac{(1+\tilde{\tau}_i)^2}{\tilde{\tau}_i^2\sigma^2}\frac{1}{1+\frac{1}{T\tau_i}}\sum_{k=1}^{K}\left(\frac{\bar{r}_{ik} - \bar{r}_i}{\phi_i}\right)^2$$

is χ^2 distributed with $K-1$ degrees of freedom. It follows from Equation 4.A.3 that

$$B = \sum_{k=1}^{K}\left(\lambda_{ik} - \bar{\lambda}_i\right)^2$$

where $\lambda_{ik} = \frac{1}{\sqrt{1+\frac{1}{T\tau_i}}}(\eta_{ik} + \frac{1}{\sqrt{\tau_i}}\bar{\epsilon}_{ik})$. As

$$\mathrm{var}(\lambda_{ik}) = \frac{\mathrm{var}(\eta_{ik}) + \frac{1}{\tau_i}\mathrm{var}(\bar{\epsilon}_{ik})}{1+\frac{1}{T\tau_i}} = \frac{1 + \frac{1}{\tau_i}\mathrm{var}\left(\frac{1}{T}\sum_{t=1}^{t}\epsilon_{ikt}\right)}{1+\frac{1}{T\tau_i}} = 1\,,$$

the random variables $(\lambda_{ik})_{k\in\{1,\ldots,K\}}$ are i.i.d. standard normal random variables. Again, as $\lambda_{ik}, \lambda_{ik'}$ are independent for $k \neq k'$, it follows that B is χ^2 distributed with $K-1$ degrees

of freedom.

Next, recall that for the Normal distribution, the sample variance $\frac{1}{T-1}\sum_{t=1}^{T}(\epsilon_{ikt}-\bar{\epsilon}_{ik})^2$ is independent of the sample mean $\bar{\epsilon}_{ik}$. As η is independent of ϵ it follows that $\sum_{t=1}^{T}(\epsilon_{ikt}-\bar{\epsilon}_{ik})^2$ and $\lambda_{ik}=\frac{1}{\sqrt{1+\frac{1}{T\tau_i}}}(\eta_{ik}+\frac{1}{\sqrt{\tau_i}}\bar{\epsilon}_{ik})$ are independent. This implies that A and B are independent. As A and B are independently χ^2 distributed it follows that

$$F_i \coloneqq \frac{\frac{1}{K-1}B}{\frac{1}{K(T-1)}A}$$

follows an F-distribution with parameters $K-1$ and $K(T-1)$.[18] Recall that in Equation 4.7, we defined $\hat{\tau}_i$.

$$\hat{\tau}_i = \frac{\frac{1}{K-1}\sum_{k=1}^{K}(\bar{r}_{ik}-\bar{r}_i)^2}{\frac{1}{K(T-1)-2}\sum_{k=1}^{K}\sum_{t=1}^{T}(r_{ikt}-\bar{r}_{ik})^2}-\frac{1}{T}$$

Plugging in the definition of A and B yields that

$$\begin{aligned}
\hat{\tau}_i+\frac{1}{T} &= \frac{K(T-1)-2}{K(T-1)}\frac{\frac{1}{K-1}\sum_{k=1}^{K}\left(\frac{\bar{r}_{ik}-\bar{r}_i}{\phi_i}\right)^2}{\frac{1}{K(T-1)}\sum_{k=1}^{K}\sum_{t=1}^{T}\left(\frac{r_{ikt}-\bar{r}_{ik}}{\phi_i}\right)^2} \\
&= \frac{K(T-1)-2}{K(T-1)}\frac{\frac{1}{K-1}B\;\frac{\tilde{\tau}_i^2\sigma^2}{(1+\tilde{\tau}_i)^2}\left(1+\frac{1}{T\tau_i}\right)}{\frac{1}{K(T-1)}A\;\frac{\tilde{\tau}_i^2\sigma^2}{(1+\tilde{\tau}_i)^2}\frac{1}{\tau_i}} \\
&= \frac{K(T-1)-2}{K(T-1)}\times\tau_i\left(1+\frac{1}{T\tau_i}\right)\times\frac{\frac{1}{K-1}B}{\frac{1}{K(T-1)}A} \\
&= \frac{K(T-1)-2}{K(T-1)}\times\left(\tau_i+\frac{1}{T}\right)\times F_i\,.
\end{aligned}$$

This establishes the first part of the theorem, i.e., Equation 4.8. Part 2 of the Theorem follows as $\mathbb{E}[F_i]=\frac{K(T-1)}{K(T-1)-2}$.[19] Part 3 follows as

$$\operatorname{var}(F_i)=\mathbb{E}[F_i]^2\,\frac{2((K-1)+K(T-1)-2)}{(K-1)(K(T-1)-4)}\,.$$

To prove Part 4, observe that Equation 4.9 is decreasing in T, and thus an upper bound is given by setting $T=2$.

$$\begin{aligned}
\sqrt{\mathbb{E}[(\hat{\tau}_i-\tau_i)^2\,|\,\tau_i]} &\leq \left(\tau_i+\frac{1}{2}\right)\sqrt{\frac{2((K-1)+K-2)}{(K-1)(K-4)}}=\left(\tau_i+\frac{1}{2}\right)\sqrt{\frac{4K-6}{(K-1)(K-4)}} \\
&\leq \left(\tau_i+\frac{1}{2}\right)\sqrt{\frac{4}{K-4}}=(2\tau_i+1)\,\frac{1}{\sqrt{K-4}}\,.
\end{aligned}$$

This establishes the result. Finally, we note that this result immediately extends to the

[18] See `https://en.wikipedia.org/wiki/F-distribution#Characterization` (accessed on June 17, 2021).

[19] See `https://en.wikipedia.org/wiki/F-distribution` (accessed on June 17, 2021).

case of endogenous effort introduced in Section 4.2.2, where for agent-specific ability a_i and incentives m_i, the precision is endogenously chosen as $\tau_i = \sqrt{m_i\, a_i}\,\sigma - 1$. □

Appendix 4.B Robustness of the Estimator

4.B.1 Characteristics with Different Averages and Variances

The estimator introduced in Section 4.3 assumes that the population means and variances of types are identical for all of the K characteristics that are being used. Empirically, however, this is usually not the case (at least not exactly). For this reason, we next describe a generalization of the estimator derived in Section 4.3 to the case where the population mean $\bar{\theta}_k$ and variance σ_k^2 of each characteristic k is potentially different. We make no assumption about the distribution of these population means and variances, but maintain the assumption that the agent's prior belief equals the distribution of characteristics in the population and that characteristics are independent. Throughout, we maintain the assumption of no scale use, i.e., $\phi_i = 1$.

Fix an infinite sequence of levels of perceived and objective self-knowledge of the respondents, $\tau_1, \tau_2, \ldots$ and $\tilde{\tau}_1, \tilde{\tau}_2, \ldots$, respectively. We denote by

$$C := \frac{1}{I}\sum_{i=1}^{I}\left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^2\left(1+\frac{1}{T\tau_i}\right)$$

and note that C is a non-negative constant independent of any specific characteristic. Throughout, we assume that each agent's self-knowledge τ_i is bounded from below by $\underline{\tau}$ which implies that C is bounded by $C \leq 1 + \frac{1}{T\underline{\tau}}$. There exist i.i.d. standard normally distributed random variables $(\epsilon_{ikt})_{ikt}$ and $(\eta_{ik})_{ik}$ such that

$$\begin{aligned} x_{ikt} &= \theta_{ik} + \frac{\sigma_k}{\sqrt{\tau_i}}\epsilon_{ikt}\,, \\ \theta_{ik} &= \bar{\theta}_k + \sigma_k\eta_{ik}\,. \end{aligned}$$

We get that (without scale use) the agent's response when asked for the t^{th} time about characteristic k is then given by

$$r_{ikt} = \frac{\bar{\theta}_k + \tilde{\tau}_i\, x_{ikt}}{1+\tilde{\tau}_i} = \bar{\theta}_k + \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}(x_{ikt} - \bar{\theta}_k) = \bar{\theta}_k + \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\sigma_k\left(\eta_{ik} + \frac{1}{\sqrt{\tau}_i}\epsilon_{ikt}\right).$$

We define the average response by agent i to question about characteristic k as $\bar{r}_{ik} = \frac{1}{T}\sum_{t=1}^{T} r_{ikt}$ and as $\bar{r}_k = \frac{1}{I}\sum_{i=1}^{I}\bar{r}_{ik}$ the average response to question k.

Lemma 4.2. *The average response to question k is normally distributed with mean $\bar{\theta}_k$ and variance*

$$var(\bar{r}_k) = \frac{\sigma_k^2}{I}C\,.$$

Furthermore, $\lim_{I\to\infty}\bar{r}_k = \bar{\theta}_k$ *almost surely.*

Proof. As η and ϵ are normally distributed with mean zero it follows that $\bar{r}_k$ is normally distributed and has mean $\bar{\theta}_k$. We are thus left to compute the variance of $\bar{r}_k$. We define $\bar{\epsilon}_{ik} = \frac{1}{T}\sum_{t=1}^{T} \epsilon_{ikt}$ as the average signal shock of agent i for characteristic k. As η_{ik} and $\bar{\epsilon}_{ik}$ are independent across agents, we have that

$$\begin{aligned}
\operatorname{var}(\bar{r}_k) &= \frac{1}{I^2}\sum_{i=1}^{I} \operatorname{var}(\bar{r}_{ik}) = \frac{1}{I^2}\sum_{i=1}^{I} \operatorname{var}\left(\bar{\theta}_k + \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\sigma_k\left(\eta_{ik} + \frac{1}{\sqrt{\tau}_i}\bar{\epsilon}_{ik}\right)\right) \\
&= \frac{\sigma_k^2}{I^2}\sum_{i=1}^{I}\left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^2 \operatorname{var}\left(\eta_{ik} + \frac{1}{\sqrt{\tau}_i}\bar{\epsilon}_{ik}\right) \\
&= \frac{\sigma_k^2}{I^2}\sum_{i=1}^{I}\left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^2 \left(1 + \frac{\operatorname{var}(\bar{\epsilon}_{ik})}{\tau_i}\right) = \frac{\sigma_k^2}{I^2}\sum_{i=1}^{I}\left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^2 \left(1 + \frac{\frac{1}{T^2}\sum_{t=1}^{T}\operatorname{var}(\epsilon_{ikt})}{\tau_i}\right) \\
&= \frac{\sigma_k^2}{I^2}\sum_{i=1}^{I}\left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^2 \left(1 + \frac{1}{T\tau_i}\right).
\end{aligned}$$

The almost sure convergence follows from Kolmogorov's strong law of large numbers for independently but not identically distributed random variables. □

Similarly, we define the variance in responses to question k as

$$s_k^2 = \frac{1}{I-1}\sum_{i=1}^{I}(\bar{r}_{ik} - \bar{r}_k)^2 .$$

Lemma 4.3. *We have that the expected sample variance converges almost surely*

$$\lim_{I\to\infty} s_k^2 = \sigma_k^2 C .$$

Proof. As $\lim_{I\to\infty} \bar{r}_k = \bar{\theta}_k$ a.s., the sample variance a.s. satisfies

$$\begin{aligned}
\lim_{I\to\infty} s_k^2 &= \lim_{I\to\infty} \frac{1}{I-1}\sum_{i=1}^{I}\left[(\bar{r}_{ik} - \bar{\theta}_k)^2 + (\bar{\theta}_k - \bar{r}_k)^2 + 2(\bar{r}_{ik} - \bar{\theta}_k)(\bar{\theta}_k - \bar{r}_k)\right] \\
&= \lim_{I\to\infty} \frac{1}{I-1}\sum_{i=1}^{I}\left[(\bar{r}_{ik} - \bar{\theta}_k)^2 + (\bar{\theta}_k - \bar{r}_k)^2\right] \\
&= \lim_{I\to\infty} \frac{I}{I-1}\left[(\bar{\theta}_k - \bar{r}_k)^2 + \frac{1}{I}\sum_{i=1}^{I}(\bar{r}_{ik} - \bar{\theta}_k)^2\right].
\end{aligned}$$

As $I/(I-1)$ converges to 1 and $(\bar{\theta}_k - \bar{r}_k)^2$ converges to zero almost surely, we get that almost surely

$$\lim_{I\to\infty} s_k^2 = \lim_{I\to\infty} \frac{1}{I}\sum_{i=1}^{I}(\bar{r}_{ik} - \bar{\theta}_k)^2 .$$

Note that $\bar{r}_{ik} - \bar{\theta}_k$ is independently normally distributed with mean zero and variance

$$\sigma_k^2\left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^2\left(1 + \frac{1}{T\tau_i}\right).$$

Thus, we get that

$$\mathbb{E}[(\bar{r}_{ik} - \bar{\theta}_k)^2] = \sigma_k^2 \left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^2 \left(1 + \frac{1}{T\tau_i}\right)$$

and

$$\operatorname{var}\big((\bar{r}_{ik} - \bar{\theta}_k)^2\big) = 2\sigma_k^4 \left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^4 \left(1 + \frac{1}{T\tau_i}\right)^2 \leq 2\sigma_k^4 \left(1 + \frac{1}{T\underline{\tau}}\right)^2 .$$

As the variance of $\left(\bar{r}_{ik} - \bar{\theta}_k\right)^2$ is bounded, we can apply Kolmogorov's strong law of large numbers and get that

$$\lim_{I\to\infty} s_k^2 = \lim_{I\to\infty} \frac{1}{I}\sum_{i=1}^{I} (\bar{r}_{ik} - \bar{\theta}_k)^2 = \lim_{I\to\infty} \frac{1}{I}\sum_{i=1}^{I} \sigma_k^2 \left(\frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\right)^2 \left(1 + \frac{1}{T\tau_i}\right) = \sigma_k^2 C\,. \qquad \square$$

We define the normalized response n_{ikt} as the difference between agent i's response and the average response, divided by the standard deviation of agents' average responses for the given characteristic k, i.e.

$$n_{ikt} = \frac{r_{ikt} - \bar{r}_k}{s_k}\,.$$

Together Lemma 4.2 and 4.3 imply the following result.

Lemma 4.4. *The normalized responses times $\sqrt{C}$ almost surely converge in the number of agents to*

$$\lim_{I\to\infty} \sqrt{C}\, n_{ikt} = \frac{\tilde{\tau}_i}{1+\tilde{\tau}_i}\left(\eta_{ik} + \frac{1}{\sqrt{\tau}_i}\epsilon_{ikt}\right) \tag{4.B.1}$$

We observe that the above asymptotic distribution for $I \to \infty$ of the normalized responses multiplied by $\sqrt{C}$ does not depend on scale use or the means and variances of characteristics. Moreover, the comparison of Equations 4.B.1 and 4.A.1 shows that the normalized responses are distributed exactly as if the respondents' scale use parameters ϕ_i equaled one, all means $\bar{\theta}_k$ were zero, and the variances σ_k^2 of characteristics all took the value of $1/C$. We define the population-based estimator as

$$\hat{\tau}_i^{POP} = \frac{\frac{1}{K-1}\sum_{k=1}^{K} (\bar{n}_{ik} - \bar{n}_i)^2}{\frac{1}{K(T-1)-2}\sum_{k=1}^{K}\sum_{t=1}^{T} (n_{ikt} - \bar{n}_{ik})^2} - \frac{1}{T}\,. \tag{4.B.2}$$

The proof given for the theorem now yields the following result:

Proposition 4.1. *For every K, T that satisfy $K(T-1) > 4$.*

1. *The estimator $\hat{\tau}_i^{POP}$ satisfies almost surely*

$$\lim_{I\to\infty} \hat{\tau}_i^{POP} = \left(\tau_i + \frac{1}{T}\right)\frac{K(T-1)-2}{K(T-1)}F_i - \frac{1}{T} \tag{4.B.3}$$

   for some random variable F_i that is F distributed with $K-1, K(T-1)$ degrees of freedom for every fixed vector of parameters $\tau_i, \sigma, \bar{\theta}$.

2. *$\hat{\tau}_i^{POP}$ is a consistent estimator for τ_i^{POP}, i.e., $\lim_{I\to\infty} \mathbb{E}\big[\hat{\tau}_i^{POP} \,\big|\, \tau_i\big] = \tau_i$ almost surely.*

Table 4.B.1: Accuracy of estimates with different means and variances

	(1)	(2)	(3)	(4)	(5)
I (respondents)	100	10,000	100	100	100
K (characteristics)	15	15	50	15	50
T (waves)	3	3	3	10	10
Correlation	0.68	0.68	0.87	0.76	0.91
Rank correlation	0.76	0.77	0.90	0.82	0.93
Median split	79%	80%	88%	83%	90%

3. *The standard error of the estimator* $\hat{\tau}_i^{POP}$ *in large populations is given by*

$$\lim_{I\to\infty} \sqrt{\mathbb{E}\big[(\hat{\tau}_i^{POP} - \tau_i)^2 \,\big|\, \tau_i\big]} = \left(\tau_i + \frac{1}{T}\right) \sqrt{\frac{2((K-1) + K(T-1) - 2)}{(K-1)(K(T-1)-4)}}\,. \quad \text{(4.B.4)}$$

4. $\hat{\tau}_i^{POP}$ *converges to* τ_i *at the rate* $1/\sqrt{K}$ *in the number of attributes, and for all* $K > 4$ *satisfies the following upper bound independent of the number of repeated observations* T

$$\lim_{I\to\infty} \sqrt{\mathbb{E}\big[(\hat{\tau}_i^{POP} - \tau_i)^2 \,\big|\, \tau_i\big]} \leq \frac{2\tau_i + 1}{\sqrt{K-4}}\,.$$

The properties of the population-based estimator are now asymptotic and do not necessarily hold in small samples. However, the only dimension of the sample size that is relevant for convergence is the number I of respondents. While, in most applications, the number of characteristics and waves (K and T, respectively) will probably be limited, the number of respondents is usually fairly large. The asymptotic properties might, therefore, be a realistic approximation of the actual behavior of the population-based estimator in many relevant contexts, as we illustrate with the simulation results below.

The table replicates Table 4.1, aside from that the means of the characteristics that are assumed. The means $\bar{\theta}$ are independently drawn from a Normal distribution with a mean of 5 and a standard deviation of 1. The standard deviations of characteristics, θ, are drawn from a log-normal distribution with the parameters $-1/2$ and 1, such that the expected standard deviation still equals one. A comparison of the result shows that the performance is almost identical to the case with equal means. This even holds for the cases where the simulated number of respondents is just 100, a sample size that most studies exceed.

4.B.2 Correlated Characteristics

We choose the Big Five inventory for estimating self-knowledge because, by design, the five measured traits are close to statistical independence. However, the five traits are each measured with a set of three survey items, which among each other are correlated. This does not impede the logic behind our estimator: subjects with high self-knowledge should give similar answers over time to the same questions, and they should give different answers to questions about different traits. What does not hold here is that estimates are

necessarily unbiased. In the stylized experiment presented in Section 4.4, all assumptions of the estimator were fulfilled, and yet unbiasedness was not the important property that we used for the results in Table 4.3. Instead, we relied on sample splits, i.e., our aim was to sort subjects according to how much information about the true type was entering their reports. Our interest here is the same, and the estimator remains informative. To gain a better understanding of how correlations in characteristics influence our estimates, we replicate the simulation results from Table 4.1 with the following modifications: we impose that characteristics are correlated in the same way as answers to the 15 Big Five questions in the 2009 wave of the SOEP, and we replicate all the columns that use 15 characteristics.

Table 4.B.2: Accuracy of estimates with correlated characteristics

	(1)	(2)	(3)
I (respondents)	100	10,000	100
K (characteristics)	15	15	15
T (waves)	3	3	10
Correlation	0.65	0.64	0.72
Rank correlation	0.74	0.74	0.80
Median split	78%	78%	81%
Bias	-0.19	-0.19	-0.19

The results are reported in Table 4.B.2, whose columns are identically constructed as Columns 1, 2, and 4 in Table 4.1. The main result is that the fraction of respondents who are correctly classified as having below- or above-median self-knowledge decreases only by about two percentage points, i.e., the informativeness of the median-splits remains.

Appendix 4.C Implications for OLS Estimates

In the analyses presented in the chapter, we concentrate on OLS regressions, where the relevant self-report serves either as the dependent or as an independent variable. To facilitate understanding of our results, we first summarize the effects that we would expect from τ in the light of our model. Table 4.C.1 provides a schematic overview of the effects that our model of survey responses predicts for regression coefficients estimated with OLS, formulated in terms of attenuation (bias towards zero; $-$) and amplification (bias away from zero; $+$). The two columns of the table differentiate between the cases of the report being used as the dependent variable (left-hand side of the equation) or as an independent variable (right-hand side of the equation). The respective other variable is assumed to be measured without error. In the upper panel, we distinguish between two channels through which a decrease in τ affects estimates: first, increased zero-mean noise around the expected answer, and second, bias in answers towards the population mean due to reduced confidence in one's signals. The lower panel presents the total effects for the three cases of $\tilde{\tau} < \tau$, $\tilde{\tau} = \tau$, and $\tilde{\tau} > \tau$ (see Section 4.2.2).

Table 4.C.1: Effect of reduction in self-knowledge τ on OLS estimates

Report as:	dependent variable	independent variable
Effect through:		
increased noise	none ($\circ$)	attenuation ($-$)
decreased $\tilde{\tau}$	attenuation ($-$)	amplification ($+$)
Overall effect with:		
$\tilde{\tau} < \tau$	$--$	$+$
$\tilde{\tau} = \tau$	$-$	$\circ$
$\tilde{\tau} > \tau$	$-/\circ$	$-$

4.C.1 Self-reports as the Dependent Variable

For the report as the dependent variable, it is well known that increased noise per se does not introduce any bias, as stated in the respective table cell. However, in our context, reduced confidence leads to attenuation bias, as we have already seen in the experimental results (see Figure 4.3a). Formally, assume that we want to estimate the following equation:

$$\theta_i = \beta_0 + \beta_1 \, y_i + \epsilon_i \, ,$$

where y_i is the respective realization of the independent variable and ϵ_i is an i.i.d. error term with an expected value of zero that is independent of y_i and the signals that subjects receive. Crucially, the value θ_i is not observable and instead replaced with the response r_i. To gain a deeper insight into the forces behind the composite effect, we use the notation involving subjective self-knowledge (see Section 4.2.2). The asymptotic result of the standard OLS estimator is derived below.

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\widehat{\operatorname{cov}(r_i, y_i)}}{\widehat{\operatorname{var}(y_i)}} \xrightarrow{p} \frac{\operatorname{cov}(r_i, y_i)}{\operatorname{var}(y_i)} = \frac{\mathbb{E}[(r_i - \bar{r})\,(y_i - \bar{y})]}{\mathbb{E}\Big[(y_i - \bar{y})^2\Big]} = \frac{\mathbb{E}\left[\frac{\tilde{\tau}(x_i - \bar{\theta})}{1+\tilde{\tau}}\,(y_i - \bar{y})\right]}{\mathbb{E}\Big[(y_i - \bar{y})^2\Big]} \\
&= \frac{\tilde{\tau}}{1+\tilde{\tau}} \, \frac{\mathbb{E}\big[\big(x_i - \theta_i + \theta_i - \bar{\theta}\big)\,(y_i - \bar{y})\big]}{\mathbb{E}\Big[(y_i - \bar{y})^2\Big]} = \frac{\tilde{\tau}}{1+\tilde{\tau}} \, \frac{\mathbb{E}[(x_i - \theta_i + \beta_1\,(y_i - \bar{y}) + \epsilon_i)\,(y_i - \bar{y})]}{\mathbb{E}\Big[(y_i - \bar{y})^2\Big]} \\
&= \frac{\tilde{\tau}}{1+\tilde{\tau}} \, \frac{\mathbb{E}[\beta_1\,(y_i - \bar{y})\,(y_i - \bar{y})]}{\mathbb{E}\Big[(y_i - \bar{y})^2\Big]} = \frac{\tilde{\tau}}{1+\tilde{\tau}} \, \beta_1
\end{aligned}$$

$$\hat{\beta}_0 \xrightarrow{p} \bar{\theta} - \beta_1 \, \bar{y} = \beta_0 + \beta_1 \, \bar{y} - \frac{\tilde{\tau}}{1+\tilde{\tau}} \, \beta_1 \, \bar{y} = \beta_0 + \left(1 - \frac{\tilde{\tau}}{1+\tilde{\tau}}\right) \beta_1 \, \bar{y}$$

Thus, as long as a decrease in τ is accompanied by a decrease in $\tilde{\tau}$, the overall effect on the absolute value of the slope parameter β_1 is strictly negative.

An Estimator for τ Based on Known True States. Suppose we know that τ is constant in the relevant population, or, alternatively, that all answers were given by the same individual. Suppose also that we know the true states, and we use them as the

independent variable, i.e., $y_i = \theta_i$ for all i. It follows that $\beta_0 = 0$, $\beta_1 = 1$, and $\bar{y} = \bar{\theta}$. For predicted answers, it follows that

$$\hat{r}_i \xrightarrow{p} \bar{\theta} + \frac{\tilde{\tau}}{1+\tilde{\tau}} \left(\theta_i - \bar{\theta}\right) .$$

For the model fit, it holds that

$$\begin{aligned} R^2 &= 1 - \frac{\sum_{i=1}^{I} \left[(r_i - \hat{r}_i)^2 \right]}{\sum_{i=1}^{I} \left[\left(r_i - \frac{1}{I} \sum_{i=1}^{I} r_i \right)^2 \right]} \xrightarrow{p} 1 - \frac{\mathbb{E}\left[\left(r_i - \bar{\theta} - \frac{\tilde{\tau}}{1+\tilde{\tau}} \left(\theta_i - \bar{\theta}\right) \right)^2 \right]}{\mathbb{E}\left[(r_i - \bar{r})^2 \right]} \\ &= 1 - \frac{\mathbb{E}\left[\left(\bar{\theta} + \frac{\tilde{\tau}}{1+\tilde{\tau}} \left(x_i - \bar{\theta}\right) - \bar{\theta} - \frac{\tilde{\tau}}{1+\tilde{\tau}} \left(\theta_i - \bar{\theta}\right) \right)^2 \right]}{\mathbb{E}\left[\left(\bar{\theta} + \frac{\tilde{\tau}}{1+\tilde{\tau}} \left(x_i - \bar{\theta}\right) - \bar{\theta} \right)^2 \right]} \\ &= 1 - \frac{\left(\frac{\tilde{\tau}}{1+\tilde{\tau}} \right)^2 \mathbb{E}\left[(x_i - \theta_i)^2 \right]}{\left(\frac{\tilde{\tau}}{1+\tilde{\tau}} \right)^2 \mathbb{E}\left[\left(x_i - \bar{\theta}\right)^2 \right]} = 1 - \frac{\frac{\sigma^2}{\tau}}{\sigma^2 + \frac{\sigma^2}{\tau}} = \frac{\tau}{1+\tau} . \end{aligned}$$

Rearranging yields that $R^2/1-R^2$ is a consistent estimator for τ.

4.C.2 Self-reports as the Independent Variable

For the report as an independent variable, noise in the sense of classical measurement error is well known to induce attenuation bias. However, reduced subjective self-knowledge works as a counter-force, inducing amplification, i.e., making the slope of the regression line *steeper*. To see the intuition, consider a regression line fitted through just two data points with coordinates (r_1, z_1) and (r_2, z_2). The point estimate for the regression coefficient is then given by $(z_2 - z_1) / (r_2 - r_1)$. Reduced subjective self-knowledge attenuates the absolute difference between r_1 and r_2, thereby increasing the estimate. Formally, assume that we want to estimate the unknown coefficients of the following equation:

$$z_i = \gamma_0 + \gamma_1 \, \theta_i + \eta_i \, ,$$

where z_i is the respective realization of the dependent variable and η_i an i.i.d. error term with an expected value of zero that is independent of both θ_i and the signals that subjects receive. Again, the unknown true values θ_i are replaced with reports r_i, and the asymptotic

result of the standard OLS estimator is derived below.

$$\begin{aligned}
\hat{\gamma}_1 \xrightarrow{p} \frac{\widehat{\operatorname{cov}(z_i, r_i)}}{\widehat{\operatorname{var}(r_i)}} &= \frac{\operatorname{cov}(z_i, r_i)}{\operatorname{var}(r_i)} = \frac{\mathbb{E}[(z_i - \bar{z})\,(r_i - \bar{r})]}{\mathbb{E}\Big[(r_i - \bar{r})^2\Big]} \\
&= \frac{\mathbb{E}\Big[\big(\gamma_0 + \gamma_1\,\theta_i + \eta_i - \gamma_0 - \gamma_1\,\bar{\theta}\big)\,\Big(\frac{\bar{\theta} + \tilde{\tau}\,x_i}{1+\tilde{\tau}} - \bar{\theta}\Big)\Big]}{\mathbb{E}\Big[\Big(\frac{\bar{\theta} + \tilde{\tau}\,x_i}{1+\tilde{\tau}} - \bar{\theta}\Big)^2\Big]} \\
&= \frac{\mathbb{E}\Big[\big(\gamma_1\,(\theta_i - \bar{\theta}) + \eta_i\big)\,\Big(\frac{\tilde{\tau}}{1+\tilde{\tau}}\,(x_i - \theta_i + \theta_i - \bar{\theta})\Big)\Big]}{\Big(\frac{\tilde{\tau}}{1+\tilde{\tau}}\Big)^2\,\mathbb{E}\Big[(x_i - \bar{\theta})^2\Big]} \\
&= \frac{\gamma_1\,\frac{\tilde{\tau}}{1+\tilde{\tau}}\,\operatorname{var}(\theta)}{\Big(\frac{\tilde{\tau}}{1+\tilde{\tau}}\Big)^2\,[\operatorname{var}(\theta) + \operatorname{var}(x \,|\, \theta)]} = \frac{\gamma_1\,\sigma^2}{\frac{\tilde{\tau}}{1+\tilde{\tau}}\,\Big(\sigma^2 + \frac{\sigma^2}{\tau}\Big)} = \frac{1+\tilde{\tau}}{\tilde{\tau}}\,\frac{\tau}{1+\tau}\,\gamma_1
\end{aligned}$$

$$\hat{\gamma}_0 \xrightarrow{p} \bar{z} - \gamma_1\,\bar{\theta} = \gamma_0 + \gamma_1\,\bar{\theta} - \hat{\gamma}_1\,\bar{\theta} = \gamma_0 + \left(1 - \frac{1+\tilde{\tau}}{\tilde{\tau}}\,\frac{\tau}{1+\tau}\right)\gamma_1\,\bar{\theta}$$

The overall effect of a reduction in τ for the report as the independent variable is thus ambiguous. As it turns out, for subjects that are correctly specified about their self-knowledge as assumed in our benchmark model, the effects cancel out exactly. If a reduction of τ results in an excess of subjective self-knowledge, estimates are attenuated. In the opposite case, the reverse applies and estimates are amplified.

In sum, contrary to economists' typical understanding of the effects of measurement error in the context of OLS, our model suggests that for responses from surveys, error in an independent variable might not always induce "innocent" attenuation bias but perhaps no bias at all or even amplification and that it always induces attenuation bias when reports are used as the dependent variable.

Appendix 4.D Robustness tests

4.D.1 Accounting for Selection

Table 4.D.1: Predictive power of domain-specific attitudes towards risk, with inverse probability weighting

	Without controls			Including controls		
Sample:	*pooled*	*below*	*above*	*pooled*	*below*	*above*
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable: *Risky financial securities*						
Risk attitude	0.0698***	0.0570***	0.0826***	0.0523***	0.0425***	0.0616***
	(0.00264)	(0.00380)	(0.00385)	(0.00280)	(0.00409)	(0.00412)
(Partial) R^2	0.0827	0.0597	0.108	0.0498	0.0351	0.0652
Observations	9095	4548	4547	7472	3736	3736
ΔR^2		81%, $p < 0.001$			86%, $p < 0.01$	
Dependent variable: *Performance pay*						
Risk attitude	0.0128***	0.0108***	0.0165***	0.00977***	0.00736**	0.0144***
	(0.00174)	(0.00245)	(0.00262)	(0.00199)	(0.00274)	(0.00309)
(Partial) R^2	0.00870	0.00670	0.0132	0.00487	0.00295	0.00966
Observations	5758	2879	2879	4464	2232	2232
ΔR^2		97%, $p = 0.20$			227%, $p = 0.15$	
Dependent variable: *Smoking*						
Risk attitude	0.0199***	0.0161***	0.0239***	0.0125***	0.00982***	0.0152***
	(0.00154)	(0.00222)	(0.00218)	(0.00175)	(0.00253)	(0.00248)
(Partial) R^2	0.0119	0.00755	0.0182	0.00481	0.00291	0.00723
Observations	15162	7581	7581	11652	5826	5826
ΔR^2		141%, $p < 0.01$			148%, $p = 0.12$	

Note: The table reports OLS estimates, with binary dependent variables taking the values zero and one. If not stated otherwise, all the data refer to the year 2009. The regressions use inverse probability weights that come from probit regressions of group assignment on gender, a second-order age polynomial, and years of education. If values are missing, we assume probabilities of 1/2. Regressions are based only on respondents who are 18 years or older, and those for performance pay include only respondents up to the age of 66 who work full-time and receive wages or salaries. *Risky financial securities* are, in the SOEP, a residual category of securities without a fixed interest rate, like stocks or options ("other securities"). Since the relevant question was asked on the household level in 2010, the units of observation in the respective regressions are households in that year. *Performance pay* indicates that an employee receives payments from profit-sharing, premiums, or bonuses. Smoking refers to 2010. The variable *risk attitude* in each of the panels refers to the respective domain-specific question asked in the SOEP. The contexts are *financial matters* for holding risky financial securities, *career* for performance pay, and *health* for smoking. The controls used in Columns 4–6 are gender, age, squared age, body height in 2010, years of education, parental education (whether mother and father each have either *Abitur* or *Fachabitur*), log net household income, and log wealth and log debts of the current household in 2007. The last three variables are calculated as ln(euro amount + 1). For the regressions involving *risky financial securities*, all variables are averaged on the household level, and we base our data only on respondents for whom all information is available individually. The p-values for the sizes of ΔR^2 are each based on 10,000 permutations. Heteroskedasticity-robust standard errors in parentheses. $^{*}\, p < 0.05$, $^{**}\, p < 0.01$, $^{***}\, p < 0.001$.

4.D.2 Big Five

Table 4.D.2 replicates Table 4.6, analyzing differences in the Big Five traits instead of differences in risk attitudes. The results are qualitatively similar to those observed for risk attitudes and quantitatively even stronger.

Table 4.D.2: Differences in Big Five

(a) Agreeableness, conscientiousness, and extraversion

	Unweighted			Weighted	
Sample:	*pooled*	*below*	*above*	*below*	*above*
	(1)	(2)	(3)	(4)	(5)
			Domain: Agreeableness		
Female	0.345***	0.274***	0.415***	0.271***	0.436***
	(0.0155)	(0.0207)	(0.0229)	(0.0208)	(0.0236)
R^2	0.0297	0.0212	0.0386	0.0209	0.0419
Observations	16359	8180	8179	8180	8179
ΔR^2		83%, $p < 0.001$		100%, $p < 0.001$	
Height (in '10)	-0.0172***	-0.0128***	-0.0215***	-0.0126***	-0.0229***
	(0.000862)	(0.00118)	(0.00126)	(0.00118)	(0.00129)
R^2	0.0262	0.0162	0.0369	0.0158	0.0413
Observations	14846	7423	7423	7423	7423
ΔR^2		127%, $p < 0.001$		161%, $p < 0.001$	
			Domain: Conscientiousness		
Female	0.143***	0.0760***	0.210***	0.0757***	0.210***
	(0.0157)	(0.0209)	(0.0233)	(0.0208)	(0.0239)
R^2	0.00510	0.00163	0.00986	0.00164	0.00976
Observations	16359	8180	8179	8180	8179
ΔR^2		506%, $p < 0.001$		494%, $p < 0.001$	
Height (in '10)	-0.00876***	-0.00512***	-0.0121***	-0.00513***	-0.0118***
	(0.000883)	(0.00118)	(0.00131)	(0.00117)	(0.00136)
R^2	0.00677	0.00261	0.0116	0.00265	0.0111
Observations	14846	7423	7423	7423	7423
ΔR^2		344%, $p < 0.001$		319%, $p < 0.001$	
			Domain: Extraversion		
Female	0.197***	0.122***	0.272***	0.127***	0.255***
	(0.0156)	(0.0194)	(0.0244)	(0.0196)	(0.0247)
R^2	0.00967	0.00481	0.0150	0.00527	0.0132
Observations	16359	8180	8179	8180	8179
ΔR^2		212%, $p < 0.001$		151%, $p < 0.01$	
Height (in '10)	-0.00254**	0.0000471	-0.00469***	-0.0000377	-0.00343*
	(0.000877)	(0.00109)	(0.00137)	(0.00111)	(0.00139)
R^2	0.000569	0.000000254	0.00159	0.000000163	0.000863
Observations	14846	7423	7423	7423	7423
ΔR^2		625381%, $p < 0.01$		529774%, $p < 0.01$	

Note: The table continues on the next page.

(b) Neuroticism and openness

	Unweighted			Weighted	
Sample:	*pooled*	*below*	*above*	*below*	*above*
	(1)	(2)	(3)	(4)	(5)
			Domain: Neuroticism		
Female	0.435***	0.359***	0.511***	0.360***	0.516***
	(0.0152)	(0.0198)	(0.0232)	(0.0199)	(0.0237)
R^2	0.0471	0.0386	0.0559	0.0391	0.0564
Observations	16359	8180	8179	8180	8179
ΔR^2		45%, $p < 0.01$		44%, $p < 0.01$	
Height (in '10)	-0.0204***	-0.0169***	-0.0241***	-0.0171***	-0.0244***
	(0.000862)	(0.00114)	(0.00130)	(0.00116)	(0.00133)
R^2	0.0366	0.0299	0.0444	0.0306	0.0456
Observations	14846	7423	7423	7423	7423
ΔR^2		49%, $p = 0.02$		49%, $p = 0.02$	
			Domain: Openness		
Female	0.129***	0.126***	0.132***	0.120***	0.122***
	(0.0156)	(0.0208)	(0.0232)	(0.0209)	(0.0239)
R^2	0.00415	0.00447	0.00393	0.00407	0.00331
Observations	16359	8180	8179	8180	8179
ΔR^2		-12%, $p = 0.79$		-19%, $p = 0.68$	
Height (in '10)	0.00107	0.00126	0.000595	0.00107	0.00205
	(0.000877)	(0.00116)	(0.00132)	(0.00117)	(0.00137)
R^2	0.000100	0.000157	0.0000283	0.000113	0.000335
Observations	14846	7423	7423	7423	7423
ΔR^2		-82%, $p = 0.61$		197%, $p = 0.73$	

Note: The table reports OLS estimates. All regressions only use respondents who are 18 years or older. The dependent variables are standardized among the respondents who enter the corresponding regression in Column 1. Columns 4 and 5 use inverse probability weights that come from probit regressions of group assignment on gender, a second-order age polynomial, and years of education. If values are missing, we assume probabilities of 1/2. Except for height, all data refer to the year 2009. The p-values for the respective sizes of ΔR^2 are each based on 10,000 permutations. Heteroskedasticity-robust standard errors in parentheses. $^{*}\, p < 0.05$, $^{**}\, p < 0.01$, $^{***}\, p < 0.001$.

Appendix 4.E Experimental Instructions

The instructions have been translated from German. Horizontal lines are used to separate screens.

Welcome

Welcome and thank you for participating in today's study!
For your participation, you will receive a flat fee of €5, which is going to be paid out in cash at the end of the study. During the study, you will respond to estimation tasks. Depending on the quality of your answers, you can additionally earn up to €10. On the following screens, everything will be explained in detail.
During the study, communication with other participants is not allowed and the curtain of your cubicle has to remain closed. Your cellphone has to be switched off and no aids are permitted. On the computer, only use the designated functions and use the mouse and keyboard to make inputs. If you should have any questions, please stick your hand out of the cubicle. One of the experimenters is then going to approach you.
Please now click on "Continue" to proceed.

Your Task

Generally, your task in this experiment is to estimate the height of stylized depictions of men. The more precisely you estimate, the more money you can earn. For that, you will, later on, see a series of pictures with men of different heights.
More precisely, the men are going to be depicted as "stick figures." An example is shown below.
[Picture of a male stick figure]
The men are split into eleven categories, depending on their body heights:

at most 1.55m	1.56m–1.60m
1.61m–1.65m	1.66m–1.70m
1.71m–1.75m	1.76m–1.80m
1.81m–1.85m	1.86m–1.90m
1.91m–1.95m	1.96m–2.00m
at least 2.01m	

Body Heights

As you know, very short and very tall men are found rather infrequently. Most common are men of around 1.78m. Exactly the same holds for the pictures that you are going to see later on. The pictures are informed by the actual height distribution among men in Germany. For that, the data from a large, representative sample of more than 20,000

people in Germany were used. The frequency of observing men of a given height is depicted in the image below.

For your orientation, we have also printed this image for you. It is lying on your desk.

[Subjects saw a figure here that contained two panels in vertical order. The top panel was a bar diagram showing the relative frequencies of the eleven different height categories. The bar heights for the categories corresponded to 26.1% for 1.76m–1.80m, 21.1% for 1.71m–1.75m and 1.81m–1.85m, 11.1% for 1.66m–1.70m and 1.86m–1.90m, 3.8% for 1.61m–1.65m and 1.91m–1.95m, 0.8% for 1.56m–1.60m and 1.96m–2.00m, and 0.1% for under 1.56m and above 2.00m. The bottom panel showed eleven male figures of different sizes. Each size was chosen as the center of one of the categories (e.g., 1.78m for 1.76m–180m; the extremes were 1.53m and 2.03m) and shown below the respective bar. Underneath each male figure, its size in meters was stated. The figure described here had also been printed and was available at each subject's desk.]

Body Heights

[Description of the figure on the previous screen, with the numeric values.]

It is important that you understand the relative frequencies of heights since the pictures that will be shown later are drawn from the displayed distribution. Thus, it is considerably more likely that you will see a man with a body height of 1.75m or 1.81m than a man with a body height of 1.58m or 2.03m.

To make the estimation of the body heights easier for you, every picture that will be displayed is accompanied by either a cat or an elephant. The cat has a height of 40cm, and the elephant is 3.50m tall (each at its highest points). In the picture below, you see an average man with a height of 1.78m next to the cat and the elephant, respectively.

[Two example images here, as described; animal pictograms adapted with permission from Storey (2016)]

Procedure

You will be shown a series of 60 pictures. For this purpose, we will randomly draw 15 different heights from the distribution in the population. Every drawn height will be shown to you four times in total. The accompanying animal and the position on the screen may change.

You will first be shown a countdown in seconds. After the countdown has finished, you will be shown a picture for [0.5/7.5] seconds. Afterward, the following question will be asked:

How tall was the displayed person?

You can provide your answer on the following scale:

The height of the displayed person was ...

below average						above average				
. . . – 1.55m	1.56m– 1.60m	1.61m– 1.65m	1.66m– 1.70m	1.71m– 1.75m	1.76m– 1.80m	1.81m– 1.85m	1.86m– 1.90m	1.91m– 1.95m	1.96m– 2.00m	2.01m– . . .
○	○	○	○	○	○	○	○	○	○	○

Your Payoff

For each shown picture, there is exactly one correct answer (an interval). For example, if the height of the shown man should be 1.78m, then this would be the answer "1.76m–1.80m." You always have to select exactly one answer. **At the end of today's study, one of the shown pictures will randomly be selected for you.** Your answer for this picture then determines the payoff that you receive on top of the €5 flat fee.
If you have chosen exactly the correct option, you will additionally receive €10. The further away you were from the correct answer (how much further to the left or right you should have clicked), the more is deducted from the €10. For this, the deviation (steps to the left or right) is squared and multiplied by 10 cents. The maximal deviation is ten steps (e.g., if you have answered "2.01m–. . . " but ". . . –1.55m" would have been correct). In this case, the entire €10 would be deducted.
You receive more money, the fewer steps are between your selected answer and the correct answer. The table gives you an overview of the possible deductions and the resulting additional payments. A printed version of this table is also available at your desk.

Deviation (steps)	0	1	2	3	4	5	6	7	8	9	10
Deduction (€)	0.00	0.10	0.40	0.90	1.60	2.50	3.60	4.90	6.40	8.10	10.00
Additional payment (€)	10.00	9.90	9.60	9.10	8.40	7.50	6.40	5.10	3.60	1.90	0.00

Control Questions

Please respond to a few questions regarding your comprehension. Feel free to use the printout at your desk as an aid.

- In each case, which of the two is more likely: the picture depicts a man with a height of . . .
  - 1.76m–1.80m [correct]
    2.01m–. . .
  - 1.81m–1.85m
    1.76m–1.80m [correct]
  - 1.76m–1.80m [correct]
    1.71m–1.75m

  - 1.66m–1.70m
    1.81m–1.85m [correct]

- How much money would be deducted from the additional €10?
  - Correct would be "1.76m–1.80m." You responded "2.01m–. . . ." [€2.50]
  - Correct would be "2.01m–. . . ." You responded ". . . –1.55m." [€10.00]
  - Correct would be "1.76m-1.80m." You responded "1.81–1.85m." [€0.10]
  - Correct would be "1.86m-1.90m". You responded "1.76m–1.80m." [€0.40]
- Suppose you have missed the picture of the man, but you nonetheless must give an estimate. What is the best answer? [1.76m–1.80m]

Thank you for your responses! Please wait.

Trial Run

Before you see the 60 pictures and estimate the heights, there will first be a trial run. You will see ten pictures and subsequently have to estimate the height of the respective man you saw. Unlike later, you are afterward informed about the correct answer.
This trial run is unrelated to the final payout and is meant to introduce you to the task. The pictures will be displayed for [0.5/7.5] seconds, exactly as in later rounds.
When you are ready, click on "Begin".

Practice task [n]/10

[Countdown]

[Picture]

How tall was the shown person?

The height of the displayed person was . . .

below average						above average				
. . . – 1.55m	1.56m– 1.60m	1.61m– 1.65m	1.66m– 1.70m	1.71m– 1.75m	1.76m– 1.80m	1.81m– 1.85m	1.86m– 1.90m	1.91m– 1.95m	1.96m– 2.00m	2.01m– . . .
○	○	○	○	○	○	○	○	○	○	○

Nine more practice rounds.

Correct answer: [e.g., 1.71m–1.75m]
Your answer: [e.g., 1.81m–1.85m]

Thank you for your responses! Please wait.

Beginning of the Main Part

Thank you for completing the trial rounds.
You can now begin with the main part of the study. At the end of the study, one of your following responses will be chosen and determine how much additional money you earn.

Task [n]/60

---

60 rounds like the practice rounds but without feedback.

Further Questions

Thank you for completing the main part.
Please now also respond to a few more additional questions.
How difficult did you feel was the task? [very easy – very difficult; seven-point scale]
How sure were you about your responses? [very unsure – very sure; seven-point scale]

Further Questions

Big Five questionnaire (BFI-S; Gerlitz and Schupp, 2005)
Scale-use module
Bayesian updating question

Personal Details

Your gender: female male diverse
Your age (in years):
Your body height (in cm):

Do you have any final comments?

Thank you for your participation in this study!

You will receive a flat fee of €5.
In addition, answer no. [n] was chosen to determine your additional payoff. Due to the deviation of your answer from the correct answer you will additionally receive [X] euros and [Y] cents.
We will soon begin with the payouts. Please wait at your seat and keep the curtain of your cubicle closed until your cabin number is called. Then, please enter the adjoining room and remember to take the card on which your cabin number is printed with you and return it.

Chapter 5

State Institutions and the Evolution of Patience*

Joint work with Thomas Dohmen

Abstract

The degree to which people behave patiently is a crucial determinant of various economic outcomes at both the individual and the aggregate level. This chapter contributes to our understanding of this important economic concept by studying the persistent effect that statehood during the last two millennia has had on patience around the globe. We show that state history and individuals' levels of patient behavior exhibit a hump-shaped relationship, consistent with recent findings for the association between historical statehood and economic development. The relationship is robust to various controls, including contemporary institutions and even economic development. We then turn towards identifying the geographically portable component of the effect by comparing migrants from different origins that now reside in the same country. The analysis suggests that the effect of home countries' state history on patient behavior is negative. It is shown that our results are consistent with a model where state history has a persistently positive effect on patient behavior through the emergence of patience-promoting norms, which are substitutes for intrinsic patience but not portable. The overall effect of state history on present-day patient behavior masks partial crowding-out of intrinsic patience.

*An earlier version of this chapter has been submitted under the same title as my master thesis, presented to the Department of Economics at the Rheinische Friedrich-Wilhelms-Universität Bonn in partial fulfillment of the requirements for the degree of Master of Science (M.Sc.) in September 2018.

5.1 Introduction

It is widely accepted among economists that institutions play a decisive role in countries' economic development. Further, there is broad evidence that *ancient* institutions have an effect on present-day prosperity beyond the influence of current institutions. At the same time, research in the areas of economics and political science has shown that there exists a close link between culture and institutions and that the effect of one cannot be fully understood without taking into account the other. On this, Alesina and Giuliano (2015) note: "While much progress has been made in isolating the importance of culture and institutions, we need to do more to fully understand their complementarities and how they jointly affect development" (p. 938). This chapter examines the long-run effect of state institutions on an economically exceptionally important aspect of culture, namely patience as measured in the Global Preference Survey (Falk et al., 2018). We find a hump-shaped relationship between populations' historical exposure to state institutions and contemporary levels of patience. This finding melds in well with the hump-shaped relationship between historical statehood and economic development documented by Borcan, Olsson, and Putterman (2018) and the positive relationship between patience and economic development established by Sunde et al. (2020). Thus, our findings suggest that individual patience could be an important micro-level transmission channel for the persistent effect that historical institutions have until today.

To shed further light on the underlying mechanism, we use migrants to separate *intrinsic* and therefore portable components of the aggregate effect from *extrinsic* components, which vanish with migration. Our evidence suggests that the effect of state institutions on internal factors which determine patient behavior is either very weak or even reversed. We argue that those two seemingly contradicting results can be reconciled by a crowding-out effect between internal and external forces.

The literature on the importance of institutions for economic development was started by North and Thomas (1973) and North (1981). In more recent years, a lot of attention has been concentrated on colonial institutions. Acemoglu et al. (2001) have shown how the impact of European colonization has had a lasting impact on development, which they claim operates through institutions. Variation in implemented institutions has even reversed the ordering of economic success for American countries (Acemoglu, Johnson, and Robinson, 2002). More recently, the focus has also shifted towards pre-modern institutions. Arias and Girod (2010) argue that indigenous institutions in the Americas have mattered for optimal strategies chosen by colonizing powers and hence are among the deep causes for today's differences in economic performance. The relevance of ethnic traditions for modern institutions is also supported by Giuliano and Nunn (2013), who find that having a tradition of "democratic" succession for local leaders leads to having more democratic institutions. Michalopoulos and Papaioannou (2013) use ethnicity level variation in traditional political centralization and show its positive association with light density in Africa, suggesting a positive effect on economic development. Depetris-Chauvin (2015) shows for Sub-Saharan Africa that sub-national state history has a negative effect on the likelihood of civil conflict.

The long-run impact of statehood that places have experienced has also been studied on the country level, starting with Bockstette, Chanda, and Putterman (2002), who have developed the state antiquity index to measure the historical presence of state institutions. They find that its explanatory power is limited for levels of income but highly significant for growth, even when controlling for other standard determinants used in the literature.

While economists have traditionally been very open towards studying the impact of institutions, it has only been rather recently that attention has also turned towards the role of culture. The discussion was importantly restarted in political science by Putnam (1993). When he studied the adaption of an institutional reform in Italian regions starting in 1970, he found strong heterogeneity, which he traced back to differences in *civic traditions* or *social capital*, measured using voting behavior, referendum turnout, newspaper readership, and density of sports and other associations (see Putnam, 1993, p. 96). Tabellini (2010) regresses economic outcomes of European regions on measures of locus of control, trust, obedience, and respect, finding support for the relevance of culture. To distill exogenous variation in the cultural variables, Tabellini (2010) draws on two instruments: the literacy rate in 1880 and constraints on the executive during the years 1600 to 1850. The latter is certainly an institutional feature, which underscores that culture and institutions are interwoven.

As has recently been stressed by Alesina and Giuliano (2015), culture and institutions strongly interact, with causality potentially running in both directions. An influential early example of an argument where causality is running from culture to institutions is Greif (1994). He argued that the "collectivist" societies of the Muslim world as opposed to the "individualist" societies of the Latin world did not produce the same necessity for efficient formal institutions and hence dropped behind in terms of development. Most research, however, has been concerned with the other direction of causality, namely with the impact of institutions on culture. In a theoretical paper, Tabellini (2008) argues that initial differences in conditions—e.g., in institutions—can have a decisive impact through the endogenous evolution of culture and possibly lead to different unique equilibria. The implied understanding of culture seems influenced by the work of Boyd and Richerson (1985, 2005), who view culture as a set of *decisions heuristics* applied in situations where fully rational reasoning is too costly or plainly impossible. Returning to the Italian setting that had inspired the work of Putnam (1993) and using similar measures of social capital, Guiso, Sapienza, and Zingales (2016) explain differences in these variables between cities with experiences of political independence that these places have or have not had. To avoid selection bias, they instrument free city experiences with historical factors. Notably, the validity of the exclusion restriction relies on the long temporal distance and the fact that whether or not an attempt to conquer was successful or not depended to a large extent on chance.

This chapter adds to the above literature by shedding light on a new dimension of the cultural transmission channel between institutions and economic development. Sunde et al. (2020) have laid out various channels through which patience is important for economic development and have empirically shown a clear positive relationship between the two.

Regarding the roots of differences in patience, Galor and Özak (2016) have theoretically and empirically argued that contemporary levels of patience positively depend on historical returns to agricultural investment as determined by agro-climatic conditions. But to the best of our knowledge, the effect of state institutions on patience has remained unexplored. The main challenge in this endeavor is that patience and institutions are jointly determined. To come closer to causal identification, we use historical institutions instead of modern institutions, reaching back at least as far as 5,500 years. The analysis of migrants is even more robust towards concerns about reverse causality since, controlling for modern characteristics, it is unlikely that modern migrants are selected according to countries' historical institutions.

The remainder of the chapter proceeds as follows: Section 5.2 lays out our theoretical considerations regarding the effect of state history on patient behavior, distinguishing between intrinsic patience and social norms. Section 5.3 presents our dataset on patience and discusses the state history index based on Borcan, Olsson, and Putterman (2018). Section 5.4 shows our empirical results for country comparisons. Section 5.5 presents our results for immigrants within countries and discusses concerns about selective migration, which seems unable to explain our findings. Section 5.6 concludes.

5.2 Theoretical Framework

Patient behavior is important for success in various domains of life such as education, health, and career progress (Sutter et al., 2013; Golsteyn, Grönqvist, and Lindahl, 2014). Its positive effects are not limited to the individual alone and not even to its immediate social environment. If a patient investor and a persevering entrepreneur team up to found a new and successful enterprise, they create value for society at large. More specifically, the individuals that will profit most from others' patience are those acting themselves patiently, e.g., by climbing up the career ladder in a successful company. In economic terms, patient behavior induces a positive externality in the form of complementarities (for an empirical backing of this claim, see Appendix 5.A). Thus, rational individuals exhibit less patient behavior than would be socially optimal. It is therefore conceivable that a shared goal of societies throughout history has been to promote patient behavior. We argue that state institutions have been playing a vital role in this domain through at least two main channels: (i) Governance structures above the tribal level act as coordination and commitment devices at the same time. They allow agreeing on common norms and on sanctions against those who do not obey them. (ii) Large-scale state institutions provide for a more stable and certain living environment and thereby help to transmit and to accumulate norms over generations.

With state institutions, formal rules and laws are introduced that are enforced under the threat of violence. These can include such different things as taxes to support a ruling class or charges for fraudulent behavior towards strangers. Indeed according to Mayshar et al. (2015), the need for protection constituted the demand side of factors that led to the emergence of social hierarchy following the Neolithic Revolution. With a

lower risk of expropriation, the expected returns to investments are higher, and individuals invest more (cf. Acemoglu, Johnson, and Robinson, 2002). The costs associated with prohibited behaviors become prohibitively large, and individuals will abstain from involving in them. What is important about state institutions is that they affect all individuals that reside within the state's territory. Therefore, every single individual that abstains from certain punishable actions can be confident that others will do the same, which would not be true if that decision was due to individual preferences. Many patient behaviors are complementary to patient behavior by others. Formal institutions increase the scope of cooperation and thereby create new investment opportunities (cf. Greif, 1994). Due to these complementarities, the payoffs associated with patient behaviors demanded by the state increase. The combination of prohibitively high costs and low desirability of impatient actions, such as theft, make them so unattractive that individuals effectively remove them from the set of options that they use for making heuristic decisions (cf. Boyd and Richerson, 1985, 2005). Rules become norms. The effect is particularly strong if the government is responsive to people's needs rather than interested in rent extraction and when states are large enough to make it plausible that children stay within its borders throughout their lives.

The presence of norms changes the optimization problem of parents who want to ensure patient behavior in their children. Individuals constantly learn from others' behavior, and thus parents can expect their children to adhere to the present norm irrespective of their own parenting decisions. Consider the extreme case where norms restrict the choice set for any decision to exactly one element. Then parenting would be entirely powerless, and the costly forging of children's preferences would disappear. For realistic cases, the presence of patience-promoting norms partially crowds out the forging of intrinsic patience. Given that parents' own patience determines the emphasis that they put on their children's patience, the effect accumulates over generations, and the population's level of intrinsic patience decreases. A related empirical example is presented by Lowes et al. (2015), who provide evidence that for the historical African Kuba Kingdom, formal institutions reduced rule-following.[1]

In the model that we are going to develop, patient behavior in any given situation depends on two factors: The evaluation of the relevant decision problem based on a norm recalled from associative memory and on intrinsic patience. Both factors act as substitutes in the generation process of patient behavior, but while patience-promoting norms are utility-neutral for any fixed decisions, increasing patience comes at a psychic cost for the child (cf. Doepke and Zilibotti, 2017). On the other hand, parents are free in setting their children's patience parameters, but the maximum patience level of the norm depends on society. This reflects the above arguments that norms need to be agreed upon by society as a whole and that situations need to feature sufficient similarities to episodes stored in memory for norms to be recalled. The model borrows importantly from Doepke and

[1] Heldring (2016) does a very similar analysis for the historical Nyiginya kingdom in today's Rwanda and finds the opposite result. He reconciles the conflicting findings with reference to the very specific characteristics of the Nyiginya state.

Zilibotti (2017) in its treatment of (im-)patience and leans on Kőszegi and Rabin (2006) for modeling norms as reference points.

5.2.1 Formal Model

We develop a two-period overlapping generations (OLG) model, where each old individual is the parent of one young individual. We assume that young individuals do not behave like fully rational lifetime utility maximizers but instead use a decision heuristic to evaluate opportunities as "good" or "bad." Whenever an offer that they receive is evaluated as "good," they will accept, and, conversely, they will reject any offer that they evaluate as "bad." Individuals' intrinsic level of impatience in the sense of a character trait is reflected by the parameter a, which is typically larger than one. As an additional tool for inference about the attractiveness of an offer, children draw on a societally determined reference point r, which is activated with intensity η. They receive offers $x \in X$ to exchange one unit of young-age consumption against $1 + i$ units of old-age consumption.

Following the representation proposed by Kőszegi and Rabin (2006), the heuristic evaluation of offers by children takes the following form:[2]

$$U(x \,|\, a, r) = -a + 1 + i + \eta\,(i - r)$$

Offers are accepted if and only if they yield positive utility. We assume that the bounds for i are such that strictly positive fractions of offers are accepted as well as rejected. We are interested in the cut-off level i^* above which young individuals accept offers.

$$i^* = I(a, r) = \frac{a + \eta r - 1}{1 + \eta} \tag{5.1}$$

Let us consider the partial derivatives of the investment cut-off with respect to intrinsic patience and the relevant norm.

$$\frac{\partial I}{\partial(-a)}(a, r) = -\frac{1}{1+\eta} < 0, \qquad \frac{\partial I}{\partial(-r)}(a, r) = -\frac{\eta}{1+\eta} \in (-1, 0]\,.$$

A marginal increase in either intrinsic patience or the patience implied by the activated norm decreases the cut-off level i^* . Note that the cross derivatives are equal to zero. This shows that both instruments' efficacies in inducing patient behavior are independent. This leads us to our first central observation.

Proposition 5.1. *Intrinsic patience and patience-promoting norms are substitutes in generating patient behavior.*

Let us now consider the consumption profiles implied by children's heuristic decisions. Assume that children are endowed with one unit of potential consumption at present and offered a continuum of investment opportunities that allows them to transfer exactly

[2]We here abstract from the "kink" of the utility function at the reference point, since it does not add in a meaningful way to our particular argument.

their entire endowment to adulthood. Returns of offers are distributed uniformly over the interval $\left[\underline{i}, \bar{i}\right]$ and we have assumed that i^* lies in the interior. Then an individual's consumption at young age c^y is simply the fraction of offers that it rejects. In old age, the child consumes c^o, which is the initial value of all accepted offers and their respective private returns.

$$c^y = \int_{\underline{i}}^{i^*} \frac{1}{\bar{i} - \underline{i}} \, \mathrm{d}i = \frac{i^* - \underline{i}}{\bar{i} - \underline{i}} \,,$$

$$c^o = \int_{i^*}^{\bar{i}} \frac{1+i}{\bar{i} - \underline{i}} \, \mathrm{d}i = \left[1 + \frac{1}{2}\left(\bar{i} + i^*\right)\right] \frac{\bar{i} - i^*}{\bar{i} - \underline{i}} \,.$$

So far, we have studied the heuristic decisions of young individuals and the implied consequences for lifetime consumption. We now turn to the decision problem of parents, who try to steer their children in the right direction and thereby act as lifetime utility maximizers on their behalf.

Parents maximize the perceived welfare of their children by choosing a subjectively optimal level of intrinsic patience for them. For simplicity—as well as in accordance with children's own decision heuristics—we assume linear period utility from consumption in young and old age. Parents are not fully altruistic in a Beckerian sense but engage in various forms of paternalism. Regarding impatience, they act according to a convex combination of some exogenous anchor A for patience and their own patience a (with weights $1-\lambda_1$ and λ_1, respectively), not taking into account their child's impatience. In their choice of the impatience parameter a', parents are constrained by societal norms, which are modeled as quadratic costs associated with deviation of the implied cut-off i^* from the norm r, multiplied with the strength λ_2 of this motive. This reflects three main classes of effects. First, there will certainly be technological constraints that prevent parents from freely choosing their children's level of patience. For example, children might acquire personal characteristics by imitation of other members of society, and shielding a child from such influences might become exceedingly costly after some point (for a similar argument, see Bisin and Verdier (2001)). Second, society might hold ready a collection of explicit or implicit punishments for deviant parents. And third, parents might have an intrinsic desire for their children to adhere to norms that they themselves hold dear in the sense of increased consumption value from interacting with their child.

$$V\left(a' \,\middle|\, a, r\right) = \left[\left(1-\lambda_1\right) A + \lambda_1 a\right] \, c^y\left(a', r\right) + c^o\left(a', r\right) - \lambda_2 \left(i^*\left(a'\right) - r\right)^2$$

The first-order condition of the maximization problem is given by

$$\left[\left(1-\lambda_1\right) A + \lambda_1 a\right] \, c^y_{a'} + c^o_{a'} - 2\lambda_2 \left(i^* - r\right) i^*_{a'} \equiv 0 \,.$$

To understand the dynamic development of patience over generations and its interplay with institutions, we need to understand how the optimal solution to the parents' maximization problem depends on both r and a. The solution to the maximization problem is

given by

$$a^* = \frac{(1+\eta)\left[(1-\lambda_1)A + \lambda_1 a\right] + \left[2\lambda_2\left(\bar{i}-\underline{i}\right) - \eta\right](1+r)}{1+2\lambda_2\left(\bar{i}-\underline{i}\right)}. \tag{5.2}$$

We see that $\lim_{\lambda_2\to\infty} a^* = 1+r$, in conjunction with Equation 5.1 implying that for this case $i^* = r$. If parents only care about norm obedience, they will mold their child's personality such that it will perfectly adhere to the norm. We find the following effects:

$$\frac{\partial a^*}{\partial(-a)} = -\frac{(1+\eta)\lambda_1}{1+2\lambda_2\left(\bar{i}-\underline{i}\right)} \leq 0, \qquad \frac{\partial a^*}{\partial(-r)} = -\frac{2\lambda_2\left(\bar{i}-\underline{i}\right)-\eta}{1+2\lambda_2\left(\bar{i}-\underline{i}\right)} \lesseqgtr 0\,.$$

Note that if $\lambda_1 \geq \left[1+2\lambda_2\left(\bar{i}-\underline{i}\right)\right]/(1+\eta)$, the derivative $\partial a^*/\partial a$ is bigger or equal to one. This means that if parents project too much of their own impatience upon their children, the effect of an initial impatience shock is not dampened across generations. The supremum of values for λ_1 which assure that the dynamic response of intrinsic patience follows an ergodic and thereby stationary AR(1) process depends positively on λ_2 and negatively on η. For a higher value of λ_2, parents are less inclined to follow their own impatience since they put a higher weight on the conformity of their children. On the other hand, a higher value of η does not change the motives of parents but renders their choice of a' less effective. Consequently, they are making more extreme choices and divergence of intrinsic patience increases in likelihood. Notwithstanding this discussion, we make the below observations.

Proposition 5.2.

1. *Patience-promoting norms reduce the amount of intrinsic patience that parents desire for their children.*

2. *Intrinsic patience persists over generations.*

Before proceeding, we need to exclude the previously discussed case of an initial change in a triggering increasing changes over the following generations.

Assumption 5.1. *Parents' inclination of projecting their own impatience upon their children is sufficiently low to ensure ergodicity in the response to a change in intrinsic patience.*

$$\lambda_1 < \frac{1+2\lambda_2\left(\bar{i}-\underline{i}\right)}{1+\eta}$$

We are now able to analyze the dynamics of patience over generations. For this, we first take Equation 5.2 and look at the change $\dot{a} \equiv a^* - a$ between any two adjacent generations within a dynasty.

$$\dot{a} = \frac{\left[(1+\eta)\lambda_1 - 1 - 2\lambda_2\left(\bar{i}-\underline{i}\right)\right]a + (1+\eta)(1-\lambda_1)A + \left[2\lambda_2\left(\bar{i}-\underline{i}\right)-\eta\right](1+r)}{1+2\lambda_2\left(\bar{i}-\underline{i}\right)}$$

The expression above implies that a has a unique fixed point $\bar{a}$ at which the level of intrinsic patience that parents choose for their children exactly coincides with their own patience

parameter value.

$$\bar{a} = \frac{(1+\eta)(1-\lambda_1)A + \left[2\lambda_2\left(\bar{i}-\underline{i}\right)-\eta\right](1+r)}{1+2\lambda_2\left(\bar{i}-\underline{i}\right)-(1+\eta)\lambda_1}$$

In line with the behavior of a^*, it also holds that $\lim_{\lambda_2\to\infty}\bar{a} = 1+r$.

We can also learn about the dynamics around the equilibrium.

$$\frac{\partial \dot{a}}{\partial a} = -\frac{1+2\lambda_2\left(\bar{i}-\underline{i}\right)-(1+\eta)\lambda_1}{1+2\lambda_2\left(\bar{i}-\underline{i}\right)} < 0$$

This shows that $\dot{a}$ is strictly positive for any $a < \bar{a}$ and strictly negative for any $a > \bar{a}$. Intrinsic patience thus dynamically converges towards $\bar{a}$ from both sides, making the latter a stable equilibrium.

$$\frac{\partial \bar{a}}{\partial(-r)} = -\frac{2\lambda_2\left(\bar{i}-\underline{i}\right)-\eta}{1+2\lambda_2\left(\bar{i}-\underline{i}\right)-(1+\eta)\lambda_1}$$

Proposition 5.3. *The effect of tightened norms on intrinsic patience is distinguished by two cases.*

A: $2\lambda_2\left(\bar{i}-\underline{i}\right) > \eta$*: The discipline on norm obedience is strong enough to ensure a positive long-term response of intrinsic patience with respect to a tightening of patience-related norms.*

B: $2\lambda_2\left(\bar{i}-\underline{i}\right) < \eta$*: The direct behavioral impact of norms on children's behavior is strong relative to the importance that parents attribute to those norms. In the long-run, this has the effect of patience-promoting norms crowding out intrinsic patience.*

Let us now study the implications of tightened norms on patient behavior, first in the short run and then in the long run. It suffices to derive the effects on i^* since any increase in the cut-off increases young-age consumption while decreasing old-age consumption, and effects are reversed for decreases in i^*.

We analyze the first-generation effect of a decrease in r.

$$\frac{\mathrm{d}i^*}{\mathrm{d}(-r)} = \frac{\partial i^*}{\partial(-r)} + \frac{\partial i^*}{\partial a}\frac{\partial a^*}{\partial(-r)} = -\frac{2\lambda_2\left(\bar{i}-\underline{i}\right)}{1+2\lambda_2\left(\bar{i}-\underline{i}\right)} < 0$$

The effect on the cut-off interest rate is negative, meaning that consumption is shifted from young age to old age.

To analyze the long-run, we define $\bar{i^*} = I(\bar{a}(r), r)$, i.e., the equilibrium interest rate in terms of r that is reached once a has fully adjusted.

$$\bar{i^*} = \frac{(1-\lambda_1)(A-1) + \left[2\lambda_2\left(\bar{i}-\underline{i}\right)-\eta\lambda_1\right]r}{1+2\lambda_2\left(\bar{i}-\underline{i}\right)-(1+\eta)\lambda_1}$$

We now simply take the partial derivative with respect to $-r$.

$$\frac{\partial \bar{i^*}}{\partial(-r)} = -\frac{2\lambda_2\left(\bar{i}-\underline{i}\right)-\eta\lambda_1}{1+2\lambda_2\left(\bar{i}-\underline{i}\right)-(1+\eta)\lambda_1}$$

Again, the limit case of $\lambda_2 \longrightarrow \infty$ is consistent with our previous results showing, in this case, a one-to-one relationship between the norm and patient behavior.

Proposition 5.4. *A tightening of patience-promoting norms has the following dynamic effects:*

1. *In the short run, patient behavior increases.*

2. *For the long run, also the indirect effects of norms through the endogenous evolution of patience need to be considered. Our model implies three distinct cases.*

   A: $2\lambda_2\left(\bar{i}-\underline{i}\right) > \eta$*: The effect on intrinsic patience is positive. Since the direct effect of a more patient norm on patient behavior is always positive, the overall effect is also positive.*

   B1: $\eta > 2\lambda_2\left(\bar{i}-\underline{i}\right) > \eta\lambda_1$*: The effect on intrinsic patience is negative but outweighed by the direct effect of norms. Therefore, individuals become intrinsically less patient but still behave more patiently in response to tightened norms.*

   B2: $\eta\lambda_1 > 2\lambda_2\left(\bar{i}-\underline{i}\right)$*: The effect on intrinsic patience is negative and this indirect effect on patient behavior weighs heavier than the direct effect. In the long-run response to more patient norms, individuals become intrinsically less patient and also behave less patiently, i.e., the initial effect of the norm is fully reversed.*

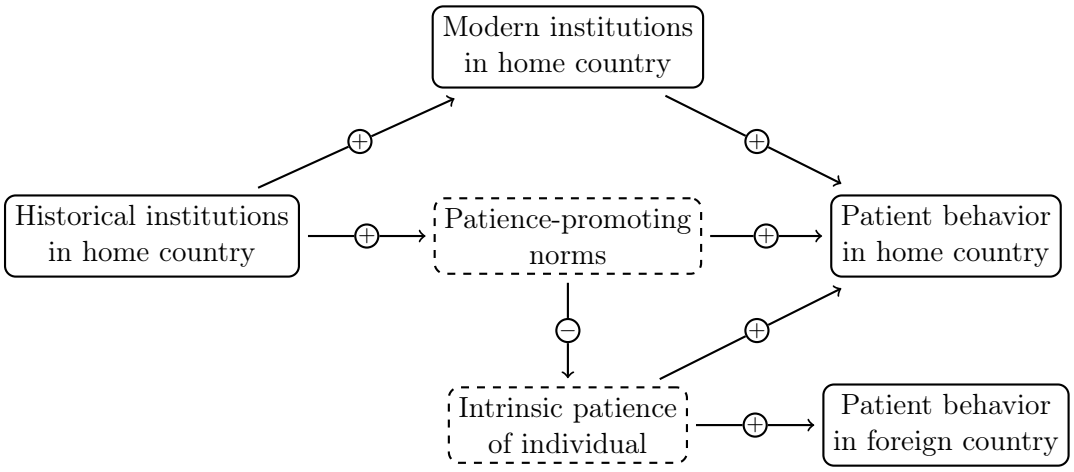
In principle, our model is agnostic as to which case should be considered more realistic. It should, however, be noted about case B2 that in contrast to the other cases, the provided condition does not imply Assumption 5.1, although it is compatible. The proximity to the case of divergence makes this case seem less likely than the others, as do the counterintuitive implications: Individuals in countries with more patient norms would have to behave less patiently. But how could the patient norms then credibly be communicated from generation to generation? The other two cases are both theoretically plausible, and it is an empirical question which one is better at capturing reality. As it will turn out, the evidence is more in favor of B1 than of A, which is why we will focus on this case.

This concludes our model. One should, however, keep in mind that if we talk about the long run, this might indeed refer to periods as long as hundreds or even thousands of years. Also, the presence of state institutions is probably better understood if we do not think about it as a one-time shock to norms but rather a device for speeding up the accumulation process of norms. The dynamics are thus more complex than described above, but we have still gained insights that will prove valuable for our analysis.

5.2.2 Empirical Implications

Before deriving the hypotheses to be empirically tested, let us carefully structure the conjectured causal mechanisms that are at work. There are multiple reasons to believe that the effect of historical institutions on patient behavior should be persistent. First, there is clear evidence that institutions themselves are persistent. For example, Bockstette,

Chanda, and Putterman (2002) "show that the state antiquity index is correlated with indicators of current institutional capacity" Bockstette, Chanda, and Putterman, 2002, p. 348. Second, institutions can have long-run effects on culture (Putnam, 1993; Guiso, Sapienza, and Zingales, 2016). The emphasis of this chapter is clearly on the latter, making it important to control for modern institutions. Figure 5.1 summarizes the presented discussion. Besides the effect going through institutional persistence, we have argued that



Note: Observed variables are represented by solid boxes and latent variables by dashed boxes.

Figure 5.1: Causal channels

historical statehood has had a positive impact on patience-promoting norms. Further, we expect norms to have partially crowded-out intrinsic patience. Finally, both norms and intrinsic patience increase the patient behavior of non-migrants. The positive effect of historical statehood through norms prevails over the negative effect caused by partial crowding-out.

Hypothesis 5.1. *Individuals living in countries with a longer history as a state behave more patiently today.*

Things look different regarding migrants. We have argued that patience-promoting norms should, in the long-term, crowd out intrinsic patience. But while the latter is portable, the former is not. The portability of intrinsic patience should imply the following.

Hypothesis 5.2. *Among migrants living in the same country, those coming from a more patient country behave more patiently.*

When individuals come to a new place with a different culture, they experience the presence of different norms, while the norms of their home country do not prove successful in an environment where others are not following them. While institutional persistence and norms that have emerged do not have a (strong) influence on the patient behavior of emigrants, decreased intrinsic patience does.

Hypothesis 5.3. *Among migrants living in the same country, those coming from a country with more historical state experience behave less patiently.*

These hypotheses will be tested in Sections 5.4 and 5.5. None of the hypotheses allows for direct validation of the theory that we are proposing. Yet the combination of the hypotheses is not trivial, and, in particular, Section 5.5.2 will argue that it is hard to think of alternative stories consistent with all three hypotheses and the presented data at the same time.

5.3 Data

For the empirical analysis that will follow, we combine the patience data from the Global Preference Survey introduced by Falk et al. (2018) with the extended version of the state antiquity index developed by Borcan, Olsson, and Putterman (2018). To account for migration, which is particularly relevant for New World countries, we use the matrix of post-1500 migration flows from Putterman and Weil (2010). We complement these combined data with various controls from other sources to add robustness to our analyses.

5.3.1 Patience

The data on patient behavior comes from the Global Preference Survey (Falk et al., 2018), which was implemented as part of the Gallup World Poll 2012. Interviews were either conducted face-to-face or via telephone, and the full dataset covers more than 80,000 individuals across 76 strongly heterogeneous countries.

The patience measure that is used here is based on two items. The first item elicits individuals' time discount rates by giving them a hypothetical choice between a fixed amount of money today or some amount of money in twelve months, instructing them to ignore any inflation that might arise. Their switching point is elicited by narrowing down the choices using a "staircase method". The other item asks respondents to self-assess their willingness to delay gratification.[3] The two items were shown to be particularly suited for predicting incentivized discounting choices in a controlled laboratory setting (Falk et al., 2016). They are aggregated using relative weights that were obtained from an OLS regression of observed behavior in a lab experiment on the two items implemented for the global sample. Therefore, the weights ensure optimal predictive power for actual behavior. In order to facilitate interpretation, the patience measure is normalized to a mean of 0 and a standard deviation of 1 on the individual level.

Besides the patience measure, the data also contains a number of useful individual-level control variables. Among those, we will use age, squared age, gender, years of education, household size, and a second-order polynomial of per capita household income. On the country level, the set of controls that we use consists of the average age of the population, the fraction of women, average years of education, linguistic diversity (Fearon, 2003), ethnic and religious fractionalization (Alesina et al., 2003), population density in 1500 (McEvedy and Jones, 1978) adjusted for migration (Putterman and Weil, 2010), longitude, latitude, average monthly temperature and precipitation (1961–1990) (Ashraf and Galor, 2013),

[3]For details see Online Appendix I.F.1. of Falk et al. (2018).

percentage of the population at risk of malaria (Gallup and Sachs, 2001), percentage of the area within the tropical or subtropical climatic zones (Gallup), and religion shares for Buddhists, Hindus, Catholics, Protestants, other Christians, Muslims, Jews, and followers of other religions (Barro, 2003). For summary statistics of all these variables, see Appendix 5.B.

5.3.2 State Antiquity

The concept of the state antiquity index was introduced by Bockstette, Chanda, and Putterman (2002). It measures the extent of countries' experiences with state-level institutions. In the original version, the index covered years from 0 CE to 1950 CE. In this chapter, we use the updated and extended version of the index developed by Borcan, Olsson, and Putterman (2018). Besides minor revision of the existing data, the index now covers the years from 3500 BCE until 2000 CE and thus goes back to the oldest known states of Mesopotamia.

Each 50-year period t from the years 1 to 1950 is first coded separately, with $t = 109$ denoting the most recent period (1951–2000) and $t = 0$ referring to the most distant one (3451 BCE–3500 BCE). Three dimensions of historical state presence in country i are coded in the following manner: (i) z_{it}^1 measures if there exists a government above the tribal level; (ii) z_{it}^2 reflects whether a country's government was foreign or locally based; and (iii) z_{it}^3 captures the degree to which a country's current territory was ruled by this government. Each component takes values between 0 and 1. The three components are aggregated multiplicatively for each period to a composite index s_{it}:

$$s_{it} = z_{it}^1 \cdot z_{it}^2 \cdot z_{it}^3 \cdot 50$$

Each of the components contains valuable information in the light of our theory outlined in Section 5.2. The more complex the structure of government, the more likely it is to produce elaborate sets of rules. Whether or not these rules evolve into norms critically depends on whether they cater to some needs of the population, which should rather be the case when the government is locally based. And last, parents will be more confident that their children will conform to current norms throughout their lives if a country has had a homogeneous history. Precisely, Borcan, Olsson, and Putterman (2018) calculate the state history index for each country as follows:

$$S_{i\tau} = \frac{\sum_{t=0}^{\tau}(1+\rho)^{t-\tau} \cdot s_{it}}{\sum_{t=0}^{\tau}(1+\rho)^{t-\tau} \cdot 50} \tag{5.3}$$

The time periods that are considered range from 0 (the most recent one) to τ. The discount rate is ρ. The denominator gives the maximum value that the numerator could hypothetically take and thereby restricts the range of possible values to the interval $[0, 1]$.

A conceptual problem that arises when using the state antiquity index is that it refers to territories while preferences refer to people. Since for many people, it is not true that their ancestors used to live in the same areas as they themselves, the state antiquity of

their country of residence does not properly reflect their inherited level of experience with state-level institutions. Putterman and Weil (2010) have developed a migration matrix that allows adjusting for migratory movements since the year 1500, which is also used by Borcan, Olsson, and Putterman (2018). The matrix has on the rows receiving countries as of the year 2000 and on the columns countries of origin, again in their borders of the year 2000. Each cell gives the share of the ancestry of a receiving country's population that in the year 1500 used to live within the territory of what is the present-day country of origin. Consider the vector of all $S_{i\tau}$'s in the order as countries $i = 1, \ldots, n$ appear in the migration matrix. Then we obtain the vector of migration adjusted $S^*_{i\tau}$'s of state history in period τ by multiplying with the migration matrix M.

$$(S^*_{1\tau} \cdots S^*_{n\tau})' = M \times (S_{1\tau} \cdots S_{n\tau})' \tag{5.4}$$

The values of the above index then represent the experience with state institutions that the ancestors of a given country's year-2000 population had in the year corresponding to period t. Note that the adjustment cannot be sensibly used for more recent periods than until 1500 CE since the migration matrix is silent about the precise time at which the migratory movements took place.

5.3.3 Timing of the Agricultural Transition

As an alternative proxy for the long-run exposure to state-like institutions, we use the time elapsed since the Neolithic Revolution, which occurred at vastly differing times around the globe. It is a long-standing hypothesis that surplus generated from transitioning from hunting and gathering towards agriculture was the prerequisite for sustaining an elite and thus for the emergence of more elaborate forms of social hierarchy and, ultimately, early states. Note that a more recent refinement of this hypothesis argues that, in fact, the agricultural surplus was neither a necessary nor sufficient condition for the emergence of an elite since, in a Malthusian world, any surplus would in equilibrium be eaten up by a growing population. Instead, Mayshar et al. (2018) argue that cultivation of *cereals*, as opposed to roots and tubers, stood at the beginning of increasingly complex social hierarchy, since cereals can easily be appropriated by either a formal authority or thieves. Thus, elites were then equipped with the technology to sustain a surplus, independent of the absolute level of productivity. Notwithstanding this differentiation, there is broad consensus that the transition to agriculture stood at the beginning of the emergence of states.

Putterman (2008) has comprised a dataset of country-specific estimates of the timing of the transition to agriculture for 170 modern countries denoted in 1,000 years counted since 2000. The estimates range between 10.5 for Israel, Jordan, Lebanon, and Syria and only 0.362 for Mauritius. For the overlap with the Global Preference Survey, the countries where the transition occurred first are Israel and Jordan, while in Australia, the transition did not take place before the arrival of Europeans about 400 years ago. Compared with the state history index, the agricultural transition data capture an even longer time horizon

but are noisier in terms of capturing exposure to state institutions. Still, they seem even more robust regarding potential concerns about endogeneity and add, as a second and distinct measure, to the robustness of our analysis. Again, we adjust for migration as already described in Section 5.3.2, where this is particularly relevant for countries in the New World, like, e.g., Australia.

5.4 Cross-Country Analysis

The empirical analysis establishes a hump-shaped relationship between historical exposure to state institutions and patient behavior on the country level, which we argue is driven by the emergence of patience-promoting norms. As a direct measure of historical state institutions, in Section 5.4.1 we use the extended state history index by Borcan, Olsson, and Putterman (2018) described in Section 5.3.2. To add further credibility to our results, in Section 5.4.2, we use the time elapsed since the Neolithic Revolution (Putterman, 2008) as a useful upper bound on time since the emergence of state structures. Again, a hump shape relationship emerges.

5.4.1 Results Using State History

In Section 5.3.2, different ways of calculating the state antiquity index were introduced. To make the analysis parsimonious, we will only use the version of the extended state history index covering the years 3500 BCE until 1500 CE, discounted at one percent per 50-year period.[4] This mitigates concerns about the endogeneity of institutions with respect to patience since they are measured with a time lag of more than 500 years. Also, the temporal distance makes sure that the measure of historical institutions that we use is clearly distinct from contemporary institutions, which we will further discuss below. The raw association between countries' average level of patience and the ancestry-adjusted state history index is depicted in Figure 5.2.

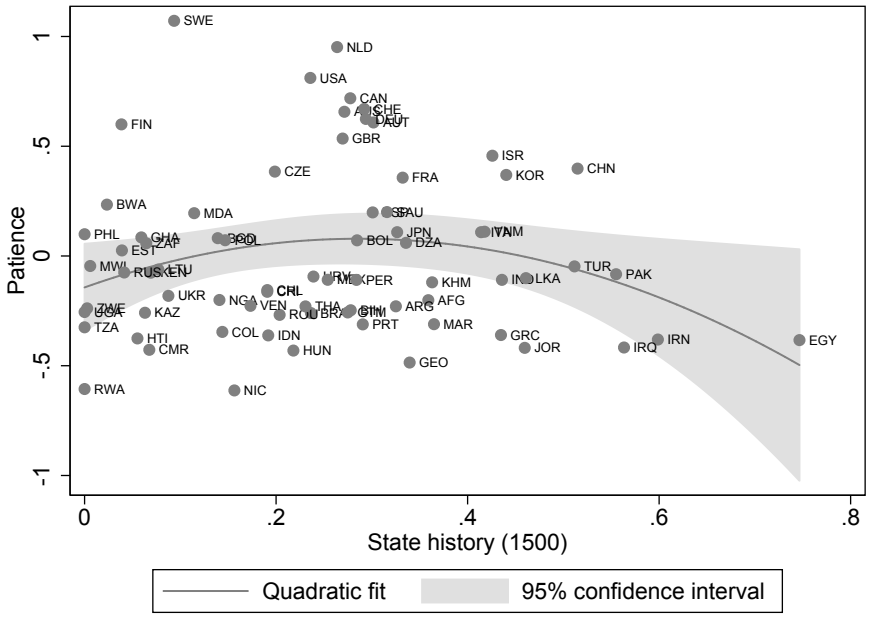
The solid line represents the fit of a simple regression of the form

$$\overline{\text{Patience}}_k = \beta_0 + \beta_1 \,\text{State history}_k + \beta_2 \,\text{State history}_k^2 + \epsilon_k \,,$$

for countries k. As we can see, the association between state history and contemporary patience seems hump-shaped. Indeed, the relationship does look rather similar to the one found by Borcan, Olsson, and Putterman (2018) for the relationship between state history and contemporary economic development, which makes our finding particularly appealing. For low to intermediate levels of state history, the pattern is consistent with Hypothesis 5.1. In the light of the finding that patience is associated with higher economic development, the empirical pattern points at a potential micro-level transmission channel.

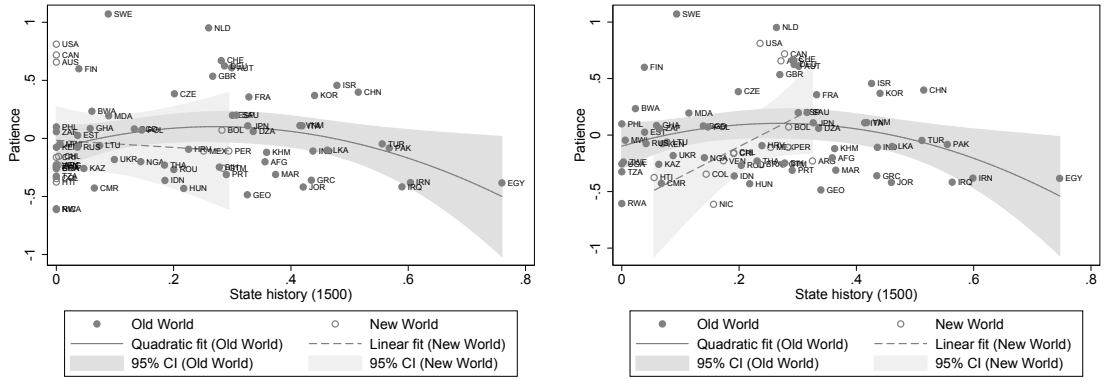
For the above finding, it is important that we adjust the state history measure for migration after 1500 CE. To see the relevance of adjusting for migration, consider Figure 5.3.

[4] This version of the index is also used by Borcan, Olsson, and Putterman (2018) and provides the largest explanatory power for contemporary economic development.



Note: State history is the extended state antiquity index (Borcan, Olsson, and Putterman, 2018) calculated for all years from 3500 BCE to 1500 CE, discounted at one percent per 50-year period. and adjusted for migration after the year 1500 CE (Putterman and Weil, 2010). The solid line represents a quadratic fit and the shaded area the respective 95% confidence interval.

Figure 5.2: State history and patience



(a) Unadjusted (b) Adjusted

Note: State history is the extended state antiquity index (Borcan, Olsson, and Putterman, 2018) calculated for all years from 3500 BCE to 1500 CE, discounted at one percent per 50-year period. Only Figure 5.3b, the state history index is adjusted for migration after the year 1500 CE (Putterman and Weil, 2010). The solid lines represent quadratic fits for old-world countries and dashed lines linear fits for new-world countries. Shaded areas indicate 95% confidence intervals.

Figure 5.3: Migration adjustment

The Figure again shows scatter plots of state history and average patience of individuals on the country level. As before, state history is measured from 3500 BCE to 1500 CE and discounted at one percent per 50-year period. Both panels show separately fitted lines for Old World and New World countries, respectively. For the latter, a linear fit is used instead of a quadratic one since no country in the New World exhibits levels of state history where we would expect a non-monotone relationship to occur. Importantly, however, Figure 5.3a is not adjusted for migration after 1500 CE. The Figure differentiates between the Old World (Africa, Europe, and Asia) and the New World (Americas and Oceania). While the relationship within the Old World looks very similar to the one seen in Figure 5.2, the regression line for the New World is slightly downward-sloping. This might be linked to the "reversal of fortunes" discussed by Acemoglu, Johnson, and Robinson (2002), which states that countries that were more highly developed in pre-Columbian times saw particularly exploitative and thus erosive institutions implemented after colonization. However, the relationship is insignificant and importantly driven by the so-called Neo-Europes, i.e., Australia, Canada, and the United States, which are highly patient today but only look back on a rather short history of statehood. From these examples, it is obvious that the background of immigrants should be taken into account and thus that the state antiquity index should be adjusted for migration. The result from this adjustment is shown in Figure 5.3b. The points exactly coincide with the ones in Figure 5.2 but as in Figure 5.3a, separately fitted lines are shown for the Old World and the New World. We now see that the regression line for the New World is upward-sloping, and no additional reversal seems to remain. This is in line with Chanda, Cook, and Putterman (2014), who find that the reversal of fortunes can be explained by the persistence of fortunes for people, and with Maloney and Valencia (2016), who use sub-national data to show that pre-colonial development differences within New World countries have not reversed but persisted until today. When using the adjusted measure of state history, our results hold uniformly across the Old and New World in the sense that the relationship is positive for low and intermediate levels of state history and then turns negative for very high levels of state history found exclusively in the Old World.

Table 5.1 shows the result from cross-country regressions of patience on state history. Note that we account for the quadratic functional form in a way that slightly differs from the one in Figures 5.2 and 5.3b. Instead of first adjusting the state history measure for migration and then including a second-order polynomial, we conduct the migration adjustment separately for the linear and the squared component. The idea behind this is that we would not expect a country consisting of two equally-sized populations of migrants with very low and very high state history to be comparable to a homogeneous country with intermediate state history.

Column 1 presents the results from a simple cross-country regression model without any controls or fixed effects. The analyses presented in Columns 2–7 do not use countries but individuals i as the units of observation, ignoring their migration status. This approach should be favored since it allows to control not just for country-level controls X_k but also

for individual-level controls $Z_{i,k}$. We estimate models of the following form.

$$\begin{aligned} \textit{Patience}_{\,i,k} = \beta_0 + \beta_1 \,\text{State history}_{\,k} + \beta_2 \,\text{Sq. State history}_{\,k} \\ + \beta_3 X_k + \beta_4 Z_{i,k} + \beta_5 \,\text{Continent}_{\,k} + \epsilon_{i,k} \end{aligned}$$

Individuals are weighted according to the weights provided by Gallup to restore representativeness of the national samples.[5] Furthermore, standard errors are clustered at the country level to allow for correlated error terms of individuals within the same country.

Table 5.1: Countries' patience and state history

	Dependent variable: Patience						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
State history (1500)	1.696***	1.707***	2.034***	1.427**	2.856***	2.848***	1.720***
	(0.605)	(0.587)	(0.599)	(0.544)	(0.489)	(0.537)	(0.477)
Sq. state history (1500)	-2.979***	-2.984***	-3.597***	-2.223**	-4.688***	-4.548***	-2.837***
	(0.918)	(0.922)	(1.057)	(0.979)	(0.675)	(0.820)	(0.750)
Democracy				-0.00413		-0.0101	0.00820
				(0.0123)		(0.0137)	(0.0118)
Property rights				0.00722***		0.00369	0.00146
				(0.00236)		(0.00250)	(0.00172)
French law					-0.172**	-0.157**	-0.143**
					(0.0696)	(0.0655)	(0.0542)
German law					-0.114	-0.158	-0.289***
					(0.106)	(0.118)	(0.103)
Scandinavian law					0.297**	0.263*	0.0685
					(0.129)	(0.142)	(0.131)
Socialist law					-0.554***	-0.514***	-0.213*
					(0.0966)	(0.112)	(0.110)
GDP pc							0.0181***
							(0.00424)
HH income pc							0.00380***
							(0.00101)
HH income pc^2							-0.00000146***
							(0.000000369)
Controls	No	No	Yes	Yes	Yes	Yes	Yes
Continent FEs	No	No	Yes	Yes	Yes	Yes	Yes
Observations	73	77220	62640	60686	62640	60686	59699
Clusters		73	63	61	63	61	60
R^2	0.0785	0.0113	0.118	0.125	0.134	0.135	0.142
Adj. R^2	0.0521	0.0113	0.117	0.125	0.133	0.134	0.141

Note: Standard errors are clustered on the country level and sampling weights are used; * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

The results in Column 1 confirm the finding from Figure 5.2. There exists a hump-shaped relationship between state history and patience, which is strongly significant and can account for about eight percent of the between-country variation. Column 2 shows the result of a regression like the one in Column 1 but on the level of individuals. The estimated coefficients for state history are very similar to the corresponding coefficients of the country-level regression seen in Column 1, which adds credibility to our analysis. Column 3 adds

[5]The same weights are also used for calculating the aggregate patience measure used for the analysis on the country level.

the individual and country-level controls that were introduced in Section 5.3.1, as well as continent fixed effects. The coefficients for state history increase in magnitude and remain statistically and economically significant.

One channel through which state history could have an impact on contemporary patient behavior is institutional persistence. Countries that during the last 2,000 years have had stronger institutions have, on average, more stable institutions today (Bockstette, Chanda, and Putterman, 2002). In particular, the latter should have a positive effect on the expected returns to investment and thereby increase patience. In Column 4, we, therefore, add controls for these two dimensions of contemporary institutions. For democracy, we use the *Polity IV* index, averaged over the period 2003–2012, which ranges from one to ten. For property rights, we use the average of the property rights index from the Heritage Foundation for the years 2004–2013, defined between zero and 100. The measure of democracy does not enter the regression model significantly due to a very small point estimate. In contrast, property rights are meaningfully and significantly positively related to patience, which is consistent with our argument made about the impact of historical statehood.

Another dimension of institutions that has attracted attention in the literature is legal origins. La Porta, Lopez-de-Silanes, and Shleifer (2008) argue "that common law stands for the strategy of social control that seeks to support private market outcomes, whereas civil law seeks to replace such outcomes with state-desired allocations" La Porta, Lopez-de-Silanes, and Shleifer, 2008, p. 286. They partition countries according to whether their current legal system has its origin in British, French, German, Scandinavian, or socialist law. In Column 5, we include their classification into our model, where British (common) law is the omitted category. The coefficients for French, German, and socialist law are negative. This gives support to the idea that common law is particularly entrepreneurial, fostering investment and, therefore, patience. The positive coefficient for Scandinavian law is also interesting but should be interpreted with caution since, in this specification, only Finland and Sweden fall into this category, and the coefficient is not significantly different from zero. What is important to note, however, is that including the set of legal origin dummies strongly increases the size and significance of the effect associated with state history. In fact, one could argue that considering legal origins is adding a measure of institutional quality to our model, while the state antiquity index primarily captures intensity.

Column 6 combines all institutional measures in a single regression model. Coefficients do not change dramatically, and while contemporary protection of property rights does not come out significantly anymore, the effect of state history remains large and highly significant at the one-percent level. Column 7 further adds economic controls, namely per capita GDP (average over the years 2003–2012, 1000$ as of 2005) and a second-order polynomial of per capita household income (1000$, purchasing power parity). Note that this is clearly a case of over-controlling since patience also has a positive effect on income. Still, the effect of state history only decreases modestly and remains highly significant.

5.4.2 Results Using Timing of the Agricultural Transition

To support our findings, we show that similar results are retrieved when using the time elapsed since the Neolithic Revolution as an alternative proxy of experience with statehood (cf. Section 5.3.3). Mirroring the specifications in Section 5.4.1, we include both linear and squared time since the Neolithic Revolution, counted in 1,000 years before 2000 and separately adjusted for migration after 1500.

Table 5.2: Countries' patience and timing of the transition to agriculture

	Dependent variable: Patience						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Agricultural trans.	0.246**	0.243**	0.142	0.140*	0.288***	0.254***	0.230***
	(0.0962)	(0.0963)	(0.0952)	(0.0801)	(0.0881)	(0.0749)	(0.0668)
Sq. agricultural trans.	-0.0201**	-0.0198**	-0.0162**	-0.0157**	-0.0263***	-0.0243***	-0.0202***
	(0.00790)	(0.00808)	(0.00750)	(0.00592)	(0.00676)	(0.00570)	(0.00522)
Democracy				0.0134		0.0136	0.0299**
				(0.0148)		(0.0144)	(0.0113)
Property rights				0.00522*		0.00233	0.000828
				(0.00278)		(0.00271)	(0.00182)
French law					-0.207***	-0.170**	-0.124**
					(0.0717)	(0.0684)	(0.0555)
German law					-0.0548	-0.0130	-0.229**
					(0.109)	(0.117)	(0.102)
Scandinavian law					0.118	0.0762	-0.0317
					(0.196)	(0.198)	(0.143)
Socialist law					-0.583***	-0.437***	-0.0881
					(0.126)	(0.126)	(0.111)
GDP pc							0.0222***
							(0.00338)
HH income pc							0.00371***
							(0.000982)
HH income pc^2							-0.00000143***
							(0.000000359)
Controls	No	No	Yes	Yes	Yes	Yes	Yes
Continent FEs	No	No	Yes	Yes	Yes	Yes	Yes
Observations	73	77267	61718	59764	61718	59764	58777
Clusters		73	62	60	62	60	59
R^2	0.0806	0.0106	0.115	0.126	0.129	0.133	0.143
Adj. R^2	0.0543	0.0105	0.115	0.126	0.128	0.132	0.143

Note: Standard errors are clustered on the country level and sampling weights are used; * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

The columns of Table 5.2 exactly correspond to the ones previously described in Table 5.1. Consistent with the previous results, a hump-shaped relationship between time since the agricultural transition and contemporary patience emerges. The maximum is reached at a value of around six. This means that the highest values of patience are predicted for countries where the (ancestry-adjusted) agricultural transition took place at around 4000 BCE, i.e., well within the range of observed values.

Overall, the timing of the agricultural transition has comparable explanatory power for patience, as does state history. The significance of coefficients is, however, not quite as robust. In Column 3, which includes controls and continent fixed effects, the coefficient

for the linear term is not significantly different from zero, although still positive. However, when institutional controls are added, the hump-shaped relationship again becomes apparent. It should not come as a surprise that, overall, the results using the timing of the Neolithic Revolution are slightly worse than when using the state history index. After all, as also Borcan, Olsson, and Putterman (2018) note, state history is the more direct and, in that sense, better measure of historical exposure to state institutions. It is, however, reassuring to see that the pattern for state history can still be replicated.

5.5 Results for Migrants

The cross-country analyses have shown a marked effect of state history on current patient behavior, which is positive for low to intermediate levels of exposure and then diminishes. This holds even when controlling for a large set of controls, including current institutions. Yet, it is not clear from this analysis to which extent the observed effects are due to factors that are *internal* or *external* to the individual in the sense of equilibrium effects. To get a better understanding of this issue, we use that internal factors should also be portable, i.e., travel with the individual, while external factors are likely to vanish when an individual changes its living environment (for a review of studies using this epidemiological approach, see Fernández, 2010). Our dataset covers a sufficient number of immigrants with information about their country of birth to study these different components empirically.

5.5.1 Patience of Immigrants

Our identification strategy is to compare immigrants within a given country of residence with one another and to explain differences in their behavior using characteristics of their home country, notably the country of origin's level of state history. Technically, we simply include fixed effects for the countries of residence. Again, we use the version of the extended state antiquity index that only takes into account the years until 1500 CE, still adjusting for migration and discounting at 1 percent per 50-year period. The idea behind this is twofold: (i) Internal factors are probably shaped in the very long run and the effects of the more recent past probably more heavily work through factors external to the individual and (ii) using a country-of-origin characteristic from at least 500 years in the past reduces the risk of spurious results stemming from selection effects.

We estimate versions of the following equation:

$$\text{Patience}_{\,i,j,k} = \beta_1 \,\text{State history}_{\,j} + \beta_2 X_j + \beta_3 Z_{i,j,k} + \beta_4 \,\text{Country}_{\,k} + \epsilon_{i,j,k}$$

All models include fixed effects for the country of residence k, ruling out the inclusion of any further characteristics of that country. Instead, we are interested in the effects associated with characteristics X of the country of origin j, and in particular in that of state history. Individual-level controls are denoted by $Z_{i,j,k}$.

In Column 1 of Table 5.3 we see that the signs of the coefficients for the linear and the quadratic state history terms are reversed relative to the results for the cross-country

Table 5.3: Migrants' patience and state history

	Dependent variable: Patience						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
State history (1500)	-0.669			-2.406***	-2.683***	-2.301***	-1.161
	(0.664)			(0.705)	(0.540)	(0.435)	(1.061)
Sq. state history (1500)	0.464			2.796**	3.154***	2.800***	1.683
	(1.180)			(1.105)	(0.845)	(0.695)	(1.286)
Agg. patience		0.320***		0.463***	0.180*	0.180*	0.112
		(0.0725)		(0.0776)	(0.0920)	(0.0995)	(0.129)
Democracy			-0.0106	-0.00610	-0.00745	-0.00509	0.0125
			(0.0133)	(0.0156)	(0.0142)	(0.0118)	(0.0184)
Property rights			0.00433*	0.00180	-0.00197	-0.00189	-0.00236
			(0.00217)	(0.00246)	(0.00228)	(0.00254)	(0.00321)
French law			-0.0821	0.120*	0.0572	0.0203	0.0305
			(0.0909)	(0.0686)	(0.0655)	(0.0767)	(0.0689)
German law			-0.0439	-0.0852	-0.0916	-0.0538	0.0430
			(0.169)	(0.169)	(0.139)	(0.149)	(0.123)
Scandinavian law			-0.325**	-0.686***	-0.710**	-0.457*	-0.306
			(0.132)	(0.250)	(0.266)	(0.234)	(0.363)
Socialist law			0.0373	0.0381	0.0305	-0.0101	0.0941
			(0.127)	(0.131)	(0.127)	(0.122)	(0.158)
GDP pc					0.0137***	0.0103***	0.0176**
					(0.00374)	(0.00350)	(0.00874)
Ind. controls	No	No	No	No	No	Yes	Yes
Agg. controls	No	No	No	No	No	No	Yes
Country FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2308	1798	1953	1662	1662	1553	1513
Countries	52	50	51	50	50	48	46
Origins	134	72	97	68	68	68	65
R^2	0.196	0.186	0.194	0.204	0.208	0.241	0.251
Adj. R^2	0.178	0.163	0.171	0.175	0.179	0.204	0.208

Note: Standard errors are two-way clustered; * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

analysis presented in Section 5.4. However, the coefficients are rather small and not statistically significant. Columns 2 shows a consistency check to see if patience generally seems to be traveling with migrants, which is what we expected according to Hypothesis 5.2. Column 3 shows results for the other measures of institutions, of which property rights protection in an immigrant's country of origin seems to exhibit some positive effect on the level of patient behavior. However, this effect is not robust, as can be seen in the other columns. The coefficient for Scandinavian law should again be treated with caution due to the low sample size (cf. Section 5.4.1). When state history is included along with the other variables in Column 5, a clear U-shaped relation can be observed. This is also robust to including the country of origin's GDP and all individual controls that were also used in Section 5.4. According to Column 6, the lowest predicted level of patience is reached at a level of state history of around 0.4. Interestingly, the location of this minimum approximately equals that of the maximum seen in Figure 5.2. Thus, Hypothesis 5.3 is supported on about the same interval as Hypothesis 5.1. When we add all the country-level controls that were used in Section 5.4 (but now for the country of origin), the signs of the coefficients for state history remain unchanged, but results are not significant anymore.

However, neither of the other previously studied variables passes the significance threshold of ten percent either, i.e., not even patience in the country of origin. This indicates that the specification in Column 7 might be overly demanding. Overall, it seems that the positive effects from low to intermediate levels of state history found in the cross-country analysis seem to reverse for immigrants, which is compatible with our theory and suggests that the observed cross-country pattern is rather an equilibrium outcome driven by norms than the expression of intrinsic differences between populations of different countries.

The results for migrants using the adjusted timing of the Neolithic Revolution in their home countries are inconclusive. Table 5.4 is constructed identically to Table 5.3, only omitting the otherwise redundant Columns 2 and 3. In the first column, the empirical relationship takes the form of a hump-shape and does not show the reversal. It is, however, extremely weak and only significant without any further controls. When GDP is added as a control in Column 3, the signs switch towards a U-shape, and this pattern also remains when adding further controls on the level of the countries of origin and of individuals. However, this pattern is never significant.

Table 5.4: Migrants' patience and timing of the transition to agriculture

	Dependent variable: Patience				
	(1)	(2)	(3)	(4)	(5)
Agricultural trans.	0.000108*	0.00000453	-0.0000843	-0.0000858	-0.0000534
	(0.0000616)	(0.000154)	(0.000153)	(0.000133)	(0.000152)
Sq. agricultural trans.	-1.09e-08**	-3.84e-09	3.09e-09	3.28e-09	2.91e-09
	(4.71e-09)	(1.19e-08)	(1.22e-08)	(1.04e-08)	(1.24e-08)
Agg. patience		0.441***	0.163*	0.173*	0.0378
		(0.0914)	(0.0817)	(0.0868)	(0.140)
Democracy		0.00239	0.00296	0.00280	0.0153
		(0.0186)	(0.0174)	(0.0146)	(0.0160)
Property rights		-0.000391	-0.00389	-0.00341	-0.00324
		(0.00309)	(0.00427)	(0.00430)	(0.00369)
French law		0.0963	0.0414	0.0173	0.0554
		(0.0927)	(0.0812)	(0.0909)	(0.0767)
German law		-0.0261	-0.0295	0.00552	0.0802
		(0.155)	(0.129)	(0.132)	(0.135)
Scandinavian law		-0.525***	-0.537**	-0.355*	-0.246
		(0.185)	(0.217)	(0.201)	(0.357)
Socialist law		0.155	0.210	0.129	0.115
		(0.183)	(0.175)	(0.157)	(0.155)
GDP pc			0.0136***	0.00996**	0.0163**
			(0.00466)	(0.00485)	(0.00807)
Ind. controls	No	No	No	Yes	Yes
Agg. controls	No	No	No	No	Yes
Country FEs	Yes	Yes	Yes	Yes	Yes
Observations	2239	1592	1592	1485	1445
Countries	52	50	50	48	46
Origins	135	67	67	67	64
R^2	0.200	0.198	0.202	0.239	0.251
Adj. R^2	0.180	0.168	0.172	0.200	0.205

Note: Standard errors are two-way clustered; * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Overall, there is no evidence that increased patience induced by historical exposure to

state institutions travels with migrants, and we conclude that the intrinsic component of patience is not positively affected. Indeed, there is some evidence coming from the more precise measure of historical institutions—the state history index—that the hump-shaped relationship observed in the cross-country analysis reverses into a U-shape for migrants. This crowd-out effect, although it might seem surprising at first sight, is well consistent with our model in Section 5.2.

5.5.2 Selective Migration

The nature of our data only permits indirect testing of our theoretical consideration laid out in Section 5.2. As we have seen in Sections 5.4 and 5.5, our predictions seem consistent with the data. But since our theoretical approach is importantly driven by the differential effects of state history on individuals staying in the countries they were born in and on migrants, a key question is in how far migration itself can account for these observed differences. Concerns mainly fall into two classes of questions: (i) Could historical migration have been the driver of diverging evolution of cultures with respect to patience? (ii) And is it possible that the observed negative association between patience and state history for immigrants is a mere artifact of selective migration in our times?

Table 5.5: Determinants of stated intention to migrate

	Dependent variable: Intend to migrate				
	(1)	(2)	(3)	(4)	(5)
Patience	0.0152**	0.00937*	0.00958	0.0505**	0.0176
	(0.00635)	(0.00547)	(0.0127)	(0.0197)	(0.0184)
Patience × avg. patience		-0.0705***			-0.0524**
		(0.0249)			(0.0210)
Patience × state history (1500)			-0.0415		0.0488
			(0.0812)		(0.107)
Patience × sq. state history (1500)			0.186		0.0338
			(0.114)		(0.161)
Patience × democracy				-0.00376	-0.00217
				(0.00306)	(0.00290)
Patience × prop. rights				-0.000194	-0.0000749
				(0.000379)	(0.000356)
Constant	0.126***	0.130***	0.126***	0.128***	0.130***
	(0.000376)	(0.00143)	(0.000484)	(0.00102)	(0.00130)
Country FEs	Yes	Yes	Yes	Yes	Yes
Observations	10040	10040	9763	9662	9385
Clusters	52	52	50	50	48
R^2	0.00168	0.00353	0.00309	0.00290	0.00493
Adj. R^2	0.00158	0.00333	0.00278	0.00259	0.00429

Note: Standard errors are clustered on the country level; * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Both questions concern the issue of selective migration, historically and in modern times. We start with the question of whether historical migration could have led to divergence in patience between countries by studying whether contemporary patterns would make us expect such effects for the future. Table 5.5 presents the results of regression

Table 5.6: Patience of immigrants relative to native population

	Dependent variable: Patience				
	(1)	(2)	(3)	(4)	(5)
Immigrant	-0.0401	-0.0345	-0.0228	0.0873	0.127
	(0.0257)	(0.0251)	(0.0981)	(0.0577)	(0.0972)
Immigrant × agg. patience		-0.0635			0.0732
		(0.0538)			(0.115)
Immigrant × state history (1500)			-0.435		-0.243
			(0.592)		(0.612)
Immigrant × sq. state history (1500)			0.970		0.743
			(0.842)		(0.924)
Immigrant × democracy				-0.00642	-0.00430
				(0.00454)	(0.00625)
Immigrant × property rights				-0.00141	-0.00256
				(0.00129)	(0.00268)
Constant	-0.0569***	-0.0568***	-0.0557***	-0.0479***	-0.0463***
	(0.00125)	(0.00127)	(0.00119)	(0.00124)	(0.00114)
Country FEs	Yes	Yes	Yes	Yes	Yes
Observations	68100	68100	65592	65116	62608
Clusters	66	66	63	63	60
R^2	0.000901	0.00563	0.00186	0.00440	0.00220

Note: Standard errors are clustered on the country level; * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

models where the dependent variable is individuals' stated willingness to migrate, which serves as a proxy for future migration status. We compare individuals within countries today and study how their patience influences their willingness to migrate. Further, we interact individual patience with country-level variables to detect potential heterogeneity in selection patterns. Column 1 shows that, on average, individuals with higher patience are more willing to migrate relative to other people who live in the same country. This would be consistent with the idea that cross-national migration implies immediate costs, which can be more easily offset by later returns for more patient individuals. Column 2 reveals marked heterogeneity in this effect of individual patience. In impatient countries, patient individuals are more likely to state that they intend to migrate, whereas, in patient countries, they are less likely to migrate than other individuals. This hints at the possibility that migration patterns are assortative, meaning that those individuals leave a country that are not very similar to the general population (the pattern would also be consistent with a simple Roy model building upon the complementarity of individual and aggregate patience described in Appendix 5.A). This could point towards selective migration as the mechanism behind the historical divergence of patience. Such an argument has been made by Olsson and Paik (2016) with respect to individualism. Olsson and Paik (2016) argue that when regions experienced the Neolithic revolution, individualistic people left for formerly uninhabited land. This pattern repeated, leading to a negative association between the time elapsed since the Neolithic revolution and individualistic values across regions observed today. If such a mechanism was at work for patience, we would expect migrants from highly patient countries to behave less patiently than migrants from impatient coun-

tries. But as we have seen in Table 5.3, the opposite is true, namely that migrants from patient countries are more patient than could otherwise be expected.

Table 5.6 adds further insight by comparing the general migrant population within countries with the native population in terms of patience. Column 1 shows that, on average, migrants are less patient than the native population. In Column 2, individual migration status is interacted with countries' average level of patience. The results are insignificant, and the point estimates suggest that if any differential effects should exist, immigrants in patient countries are relatively less patient. This means that patient behavior in already patient countries is rather being decreased by immigration. Table 5.5 also shows that historical as well as modern institutions do not seem to cause differential selection effects with respect to patience. All interactions of individual patience with measures of institutions are insignificant. Similarly, in Table 5.6, all interactions of immigrant status with institutions are insignificant. Our data, therefore, offer no support for the idea that selective migration could be the cause of our cross-country results.

The same empirical results that have been presented also speak to the question of whether the reversed effect of state history on patience that we have seen for immigrants could be a mere artifact of selective migration. Table 5.5 has shown that countries' aggregate patience levels induce differential selection effects with respect to individual patience. Yet, as has already been discussed above, Table 5.3 has shown that there is no reversal of patience along with home countries' aggregate patience. Table 5.5 gives no indication of differential selection effects along the dimensions of historical or modern institutions. Hence, selection effects alone cannot account for the empirical findings.

5.6 Conclusion

We have shown that historical institutions exhibit a hump-shaped relationship with present-day levels of patient behavior across countries. The results are suggestive of a persistent long-term effect with potentially important implications for understanding differences in economic development and consistent with previous empirical findings. Our analysis of immigrants suggests that over the course of centuries, state institutions have partially crowded out intrinsic patience. Reversely, countries that have little experience with statehood and, according to the literature, suffer various disadvantages from it might not have less intrinsically patient populations. The latter point could be important for the growth perspectives of such countries but also for the assessment of economic impacts associated with migratory movements.

The results presented here are a first step towards understanding the long-term consequences of state institutions for patience. Future research should try to do similar analyses on the subnational level and also expand on the analysis of immigrants.

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2002. "Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution". *Quarterly Journal of Economics* 117 (4): 1231–1294.

Acemoglu, Daron, Simon Johnson, James A. Robinson, and David Albouy. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation". *The American Economic Review* 91 (5): 1369–1401.

Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg. 2003. "Fractionalization". *Journal of Economic Growth* 8 (2): 155–194.

Alesina, Alberto, and Paola Giuliano. 2015. "Culture and Institutions". *Journal of Economic Literature* 53 (4): 898–944.

Arias, Luz Marina, and Desha Girod. 2010. *Indigenous Origins of Colonial Institutions*. Mimeo.

Ashraf, Quamrul, and Oded Galor. 2013. "The "Out of Africa" Hypothesis, Human Genetic Diversity, and Comparative Economic Development". *American Economic Review* 103 (1): 1–46.

Barro, Robert J. 2003. *Religion Adherence Data*. http://scholar.harvard.edu/barro.

Bisin, Alberto, and Thierry Verdier. 2001. "The Economics of Cultural Transmission and the Dynamics of Preferences". *Journal of Economic Theory* 97 (2): 298–319.

Bockstette, Valerie, Areendam Chanda, and Louis Putterman. 2002. "States and Markets: The Advantage of an Early Start". *Journal of Economic Growth* 7 (4): 347–369.

Borcan, Oana, Ola Olsson, and Louis Putterman. 2018. "State History and Economic Development: Evidence from Six Millennia". *Journal of Economic Growth* 23 (1): 1–40.

Boyd, Robert, and Peter J. Richerson. 1985. *Culture and the Evolutionary Process*. Chicago, IL: University of Chicago Press.

— . 2005. *The Origin and Evolution of Cultures*. Oxford: Oxford University Press.

Chanda, Areendam, C. Justin Cook, and Louis Putterman. 2014. "Persistence of Fortune: Accounting for Population Movements, There Was No Post-Columbian Reversal". *American Economic Journal: Macroeconomics* 6 (3): 1–28.

Depetris-Chauvin, Emilio. 2015. *State History and Contemporary Conflict: Evidence from Sub-Saharan Africa*. Mimeo.

Doepke, Matthias, and Fabrizio Zilibotti. 2017. "Parenting with Style: Altruism and Paternalism in Intergenerational Preference Transmission". *Econometrica* 85 (5): 1331–1371.

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

Falk, Armin, Anke Becker, Thomas Dohmen, David Huffman, and Uwe Sunde. 2016. *The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences*. IZA Discussion Paper 9674. Bonn: Institute for the Study of Labor.

Fearon, James D. 2003. "Ethnic and Cultural Diversity by Country". *Journal of Economic Growth* 8 (2): 195–222.

Fernández, Raquel. 2010. *Does Culture Matter?* NBER Working Paper 16277. Cambridge, MA: National Bureau of Economic Research.

Gallup, John Luke. *Country Geography Data.* https://www.pdx.edu/econ/jlgallup/country-geodata.

Gallup, John Luke, and Jeffrey D. Sachs. 2001. "The Economic Burden of Malaria". *American Journal of Tropical Medicine and Hygiene* 64 (1, 2): 85–96.

Galor, Oded, and Ömer Özak. 2016. "The Agricultural Origins of Time Preference". *American Economic Review* 106 (10): 3064–3103.

Giuliano, Paola, and Nathan Nunn. 2013. "The Transmission of Democracy: From the Village to the Nation-State". *American Economic Review* 103 (3): 86–92.

Golsteyn, Bart H. H., Hans Grönqvist, and Lena Lindahl. 2014. "Adolescent Time Preferences Predict Lifetime Outcomes". *The Economic Journal* 124 (580): F739–F761.

Greif, Avner. 1994. "Cultural Beliefs and the Organization of Society: A Historical and Theoretical Reflection on Collectivist and Individualist Societies". *Journal of Political Economy* 102 (5): 912–950.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2016. "Long Term Persistence". *Journal of the European Economic Association* 14 (6): 1401–1436.

Heldring, Leander. 2016. *The Persistence of State Capacity in Rwanda.* Mimeo.

Kőszegi, Botond, and Matthew Rabin. 2006. "A Model of Reference-Dependent Preferences". *Quarterly Journal of Economics* 121 (4): 1133–1165.

La Porta, Rafael, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2008. "The Economic Consequences of Legal Origins". *Journal of Economic Literature* 46 (2): 285–332.

Lowes, Sara, Nathan Nunn, James A. Robinson, and Jonathan Weigel. 2015. *The Evolution of Culture and Institutions: Evidence from the Kuba Kingdom.* Mimeo.

Maloney, William F., and Felipe Valencia Caicedo. 2016. "The Persistence of (Subnational) Fortune". *The Economic Journal* 126 (598): 2363–2401.

Mayshar, Joram, Omer Moav, Zvika Neeman, and Luigi Pascali. 2015. *Cereals, Appropriability and Hierarchy.* CEPR Discussion Paper 10742. London: Centre for Economic Policy Research.

— . 2018. *The Emergence of Hierarchies and States: Productivity vs. Appropriability.* Mimeo.

McEvedy, Colin, and Richard Jones. 1978. *Atlas of World Population History.* Harmondsworth: Penguin Books.

Michalopoulos, Stelios, and Elias Papaioannou. 2013. "Pre-Colonial Ethnic Institutions and Contemporary African Development". *Econometrica* 81 (1): 113–152.

North, Douglass C. 1981. *Structure and Change in Economic History.* London: W. W. Norton.

North, Douglass C., and Robert Paul Thomas. 1973. *The Rise of the Western World. A New Economic History.* Cambridge: Cambridge University Press.

Olsson, Ola, and Christopher Paik. 2016. "Long-run cultural divergence: Evidence from the Neolithic Revolution". *Journal of Development Economics* 122:197–213.

Putnam, Robert D. 1993. *Making Democracy Work: Civic Traditions in Modern Italy.* Princeton, NJ: Princeton University Press.

Putterman, Louis. 2008. "Agriculture, Diffusion and Development: Ripple Effects of the Neolithic Revolution". *Economica* 75 (300): 729–748.

Putterman, Louis, and David N. Weil. 2010. "Post-1500 Population Flows and the Long-Run Determinants of Economic Growth and Inequality". *Quarterly Journal of Economics* 125 (4): 1627–1682.

Sunde, Uwe, Thomas Dohmen, Benjamin Enke, Armin Falk, David Huffman, and Gerrit Meyerheim. 2020. *Patience and Comparative Development*. Discussion Paper 035. Bonn and Cologne, Germany: ECONtribute.

Sutter, Matthias, Martin G. Kocher, Daniela Glätzle-Rützler, and Stefan T. Trautmann. 2013. "Impatience and Uncertainty: Experimental Decisions Predict Adolescents' Field Behavior". *America Economic Review* 103 (1): 510–531.

Tabellini, Guido. 2010. "Culture and Institutions: Economic Development in the Regions of Europe". *Journal of the European Economic Association* 8 (4): 677–716.

— . 2008. "The Scope of Cooperation: Values and Incentives". *Quarterly Journal of Economics* 123 (3): 905–950.

Appendix 5.A Returns to Patient Behavior

To analyze whether there is a complementarity between individual patience and aggregate patience in a country, we run separate regressions of log household income per capita on individual patience for each country.

$$\ln \text{household income p.c.}_i = \beta_0 + \beta_1 \text{individual patience}_i + \epsilon_i$$

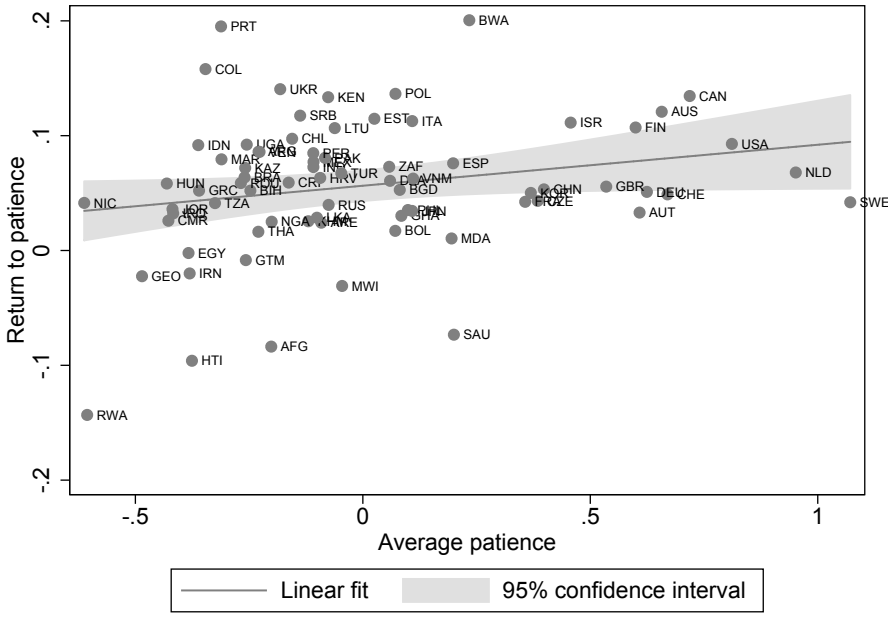


Figure 5.A.1: Return to individual patience and aggregate patience

In Figure 5.A.1, we plot the respective estimates for β_1, which are the percentage returns to a one-standard-deviation increase in patience against the average level of patience of countries. There exists a positive correlation of $r = 0.23$ with a p-value of 0.05. Thus,

individual patience and aggregate patience appear to be complements in the generation of per capita household income.

Appendix 5.B Descriptive Statistics

Table 5.B.1: Descriptive statistics

Variable	Mean	Minimum	Maximum	Standard deviation	Observations
Individual level					
Patience	.0017235	-1.313386	2.763126	1.001785	78712
Female	.546524	0	1	.497834	78712
Age	41.63164	15	99	17.42593	78454
Years of education	10.20976	0	96	4.785664	75641
Household size	4.006009	1	34	2.664378	78712
Buddhist	.0483341	0	1	.2144728	69992
Hindu	.0344325	0	1	.1823387	69992
Catholic	.3126929	0	1	.4635937	69992
Protestant	.1775632	0	1	.3821473	69992
Other christian	.1227426	0	1	.3281439	69992
Muslim	.193465	0	1	.3950171	69992
Jew	.0119299	0	1	.1085715	69992
Other religion	.011287	0	1	.1056398	69992
Household income p.c. (1,000 USD)	7.577	0	2620.499	20.95458	77218
Country level					
Patience	-.0016294	-.6125203	1.071452	.3718196	75
State history (1500)	.2126697	0	.7597628	.1919654	73
State history (1500, adj.)	.2454062	0	.7465726	.1670537	73
Sq. state history (1500)	.0815743	0	.5772395	.1096682	73
Sq. state history (1500, adj.)	.0926466	0	.56036	.1009376	73
Years (1,000) since ag. trans.	5240.541	400	10500	2473.25	74
Sq. years (1,000) since ag. trans.	3.35e+07	160000	1.10e+08	2.84e+07	74
Years (1,000) since ag. trans. (adj.)	5742.728	1480	10400	2071.029	73
Sq. years (1,000) since ag. trans. (adj.)	3.82e+07	2363800	1.09e+08	2.46e+07	73
Longitude	19.12231	-99.16666	151.1667	58.88809	75
Abs. latitude	30.62374	.3333333	60.13334	16.89833	75
Avg. precipitation	84.98907	2.910641	241.7184	58.11169	75
Avg. temperature	16.39466	-7.929411	27.36805	8.527968	75
% in (sub-)tropical zones	.319227	0	1	.4133007	75
% at risk by Malaria	.2050297	0	1	.3505378	75
Pop. dens. in 1500 (adj.)	10.15513	.3283102	46.63923	9.419302	74
Rel. fractionalization	.4342984	.0034627	.8602599	.2389034	75
Ethnic fractionalization	.4092448	.001998	.930175	.2477207	75
Linguistic diversity	.1965109	0	.6890886	.1976287	75
Democracy (*Polity IV*)	6.611872	0	10	3.521667	73
Property rights	48.60959	8.5	90.5	24.03834	73
French legal origin	.4133333	0	1	.4957477	75
German legal origin	.0666667	0	1	.2511236	75
Scandinavian legal origin	.0266667	0	1	.1621922	75
Socialist legal origin	.2133333	0	1	.4124198	75
GDP p.c. (1,000 USD)	11.3189	.2217485	53.41695	14.29949	75