A Behavioral Economics Approach to Team Performance, Risk Taking, and Discrimination: Experimental Evidence

Stefan Grimm

München 2018

A Behavioral Economics Approach to Team Performance, Risk Taking, and Discrimination: Experimental Evidence

Inaugural-Dissertation

zur Erlangung des Grades

Doctor oeconomiae publicae (Dr. oec. publ.)

an der Volkswirtschaftlichen Fakultät

an der Ludwig-Maximilians-Universität München

2018

vorgelegt von Stefan Grimm

Referent:	Prof. Dr. Martin G. Kocher
Korreferent:	Prof. Dr. Florian Englmaier
Promotionsabschlussberatung:	11. Juli 2018
Berichterstatter:	Prof. Dr. Martin G. Kocher
	Prof. Dr. Florian Englmaier
	Prof. Dr. Simeon Schudy
Mündliche Prüfung:	02. Juli 2018

Acknowledgements

While ultimately I am the single author of this dissertation, I would not have accomplished nearly so much without the help of others. Contributions including guidance on the general direction of my research, shared work on projects, support beyond what is visible at first sight, and welcome distractions made this thesis possible.

First, I would like to thank my supervisor Martin Kocher. By initiating a joint project early-on with me, he provided a jump start into the field of experimental economics. Since then he has been extremely valuable in always finding time to discuss and provide feedback on research ideas and projects. In addition, I very much appreciated the enriching and cooperative working environment in our team over the last years.

Further, I am very much indebted to Florian Englmaier and Simeon Schudy. Florian is co-author of one chapter of this dissertation, agreed to become my second advisor and has repeatedly lent an ear to discussions of research ideas, projects and future plans. I am much obliged for his investment of time and effort and his encouragement. Similarly, I can't thank Simeon enough for the many discussions, questions he answered and suggestions he made. Not only did we spend days and weeks on our joint project, but he also provided extensive feedback on many other issues and agreed to serve on my PhD committee.

Basically no work can be done without administrative and technical support. Therefore, special thanks go to Manuela Beckstein, Silke Englmaier, Manuela Firmin, Angelica Schmidt and last but not least Julia Zimmermann, who not only made my life much easier, but without whom this dissertation would look vastly different.

I gratefully acknowledge financial support from the German Research Foundation (DFG) through GRK 1928 and CRC TRR 190. This gratitude extends to many members of our economics faculty for creating an excellent research environment. I would like to highlight Carsten Eckel and Klaus Schmidt for their key roles in establishing our graduate research training group and collaborative research center, respectively. This provided funding for experiments and trips to many international conferences, workshops, and summer schools. It also supported my research stay in San Diego, for which I am very thankful to Uri Gneezy, who received me into his group. I greatly appreciated the opportunity to work in a different environment and the time spent in California.

I am grateful to my co-authors Michał Krawzcyk, Fabrice Le Lec and David Schindler for their effort and input on the joint projects, to all former and present members of the Chair for Behavioral and Experimental Economics for scientific advice and good times, and to all the research assistants at our laboratory MELESSA and the many others involved in Chapter 2 of this dissertation for their meticulous help. Special mentions go to Felix Klimm, with whom I shared an office for three years and who became the co-author of Chapter 4, as well as Dominik Grothe and Julia Rose who self-reliantly put many hours into the success of Chapter 2.

Last but not least, I would like to thank my parents and Teresa. Thank you, Irmgard and Gebhard, for always being there and for often delaying your own needs, for supporting and trusting my choices and for always providing a reflective perspective on the important topics in life. Your nurture made me the person that I am today and enabled me — among other things — to complete this thesis. Thank you, Teresa, for always believing in me, for your constant encouragement, for putting things into perspective and for lending your thoughts on so many experimental designs, results and draft versions.

Contents

List of Figures

List of Tables

Preface

The linguistic origin of "economics" lies in the Greek term "*οικονομικός*" ("oikonomikós") and denotes the laws ("nomos") of the house/home ("oikos"). In one of the earliest accounts of the term — in the dialogue between Xenophon, Socrates and Kritoboulos in roughly 362 B.C. — reference is made to the management of the household and property (see, e.g., the commentary by Pomeroy, 1996). While, since then, many definitions of economics have been put forward, the early reference still echoes in prominent modern definitions. These describe economics as the study of subjects and environments where resources are scarce and need to be efficiently or optimally employed to best satisfy unlimited wants (see, e.g., Backhouse and Medema, 2009, for a discussion). This is true for household management, where income necessarily is constrained but desires unbounded, and it is also true for societies, where available input is limited and collective demand infinite.

This understanding of economics invites economic research to target a wide range of topics, and many economists have employed their expertise in seemingly exotic fields of application. As a consequence, today, there are articles on soccer, smoking bans, discrimination, the neurological foundations of economic decision making, and many more topics in economic journals.

The field of behavioral economics has further increased the bandwidth of economic research by relaxing the strict assumptions made with respect to the traditional subject of economic analyses — 'homo oeconomicus'. By developing new ideas and applying concepts and insights from other disciplines (mainly psychology, but also sociology and anthropology), behavioral economics has enriched the model of a decision maker with 'clearly defined' and 'egoistic' preferences who has complete information and strictly maximizes utility. The seemingly redundant modifier "behavioral" describes the field's roots in psychology's behavioral decision research, where the clear distinction to purely rational decision making originated. From being a small field until the 1990s, behavioral economics today — with Nobel Prize Laureates in 2017 (Richard Thaler) and 2002 (Daniel Kahnemann and Vernon L. Smith) receiving awards for their work in behavioral economics, and Laureates Alvin E. Roth (2012) and Robert J. Shiller (2013) having worked in related domains — has matured, developed new models of decision making and has strongly improved our understanding of human behavior in a wide range of environments.

Besides evidence on, for example, biased beliefs and biased decision making, major progress has been made in the establishment and structured modeling of non-classical preferences (for a review, see, e.g., DellaVigna, 2009). Among other things, ample evidence shows that the utility of an action or choice depends on reference points. This represents a deviation from neoclassical assumptions and implies that individuals might make seemingly inconsistent choices when facing the same choices over time simply because of changing (sometimes arbitrary) reference points. A related concept is loss aversion and risk-seeking in the loss domain. Relative to a certain reference point, worse outcomes (i.e. losses) loom larger than relative gains and because of a decreasing sensitivity to ever larger losses, individuals become risk-seeking in the loss domain (Kahneman and Tversky, 1979). These findings have strong implications for understanding, for example, trading behavior, insurance markets and firm pricing strategies. Likewise, many theoretical and empirical contributions highlight the notion that people strongly care about the payoff and well-being of others; not only in absolute terms, but also in relative terms of inequity (e.g., Fehr and Schmidt, 1999). Hence, individuals' actions can be motivated by the desire to affect others' payoffs. Similarly, and in contrast to the neoclassical assumption of purely material self-interest, people are driven by non-monetary concerns. They care about how they are perceived by others and have a certain concept of self that they want to adhere to (Akerlof, 1980; Akerlof and Kranton, 2000). Among other things, this has consequences for how firms should design incentive schemes. People can be motivated not only by monetary rewards, but also, for example, by relative ranking or by an intrinsic nature per se.

While the methods applied in behavioral economic research span the universe of methods used in economics in general, experimental methods play a particular role for the field. Likewise, while experimental methods are used in many fields of economics, most studies in the field of experimental economics are behavioral in nature. Broadly speaking, experimental economics is the branch within economics that uses experiments (mostly with human subjects) to test, refine and suggest models of behavior (see, e.g., Croson and Gächter, 2010, for an overview). This approach is used for questions for which naturally occurring data are not available, and also in general when controlled data generating processes are particularly important for answering research questions. In experimental setups, the researcher has control over all (or most) factors that might affect behavior, and varies only one factor at a time. This controlled variation allows causal inferences. By this definition, experiments can take place in many settings: in a laboratory, in classrooms, at company sites, with pedestrians in the street, or even in the fMRI scanner (neuroeconomic studies).

All chapters of this dissertation make use of experimental methods and apply behavioral concepts and ideas. In particular, all chapters recognize that preferences are non-standard. Chapter 1 acknowledges that people can be intrinsically motivated to perform well on specific tasks. It further provides a test for loss aversion in this setting. Chapter 2 shows that the earnings of another person strongly affect one's own behavior, and do so in a specific pattern. Chapter 3 assumes that people care about how they are perceived by others and Chapter 4 suggests an explanation for experimental behavior that builds on self-image and identity concerns. Further, all chapters see economic decision makers as social beings in non-isolated environments. Team behavior (Chapter 1), interdependencies between individual payoffs (Chapter 2), non-private choices (Chapter 3) and behavior towards ingroup or outgroup members (Chapter 4) are the focus of this dissertation.

In the spirit of a modern definition of economics, this behavioral economics approach is applied to a range of different topics. Chapter 1 investigates the determinants of team productivity in an increasingly important task setting. This is essential for effectively designing incentive schemes to increase productivity and employee well-being in modern economies. In the bigger picture, it thereby helps to understand prerequisites for growth and innovation. Chapters 2 and 3 discuss how social decision environments shape individual risk taking. It is crucial to grasp such determinants of risky decision making with almost all decisions involving some type of uncertainty. This, for example, closely relates to individuals' retirement planning, firms' investment decisions, and even employees' career planning. Lastly, Chapter 4 concerns discrimination behavior in how people attribute responsibility for positive or negative shocks. It acknowledges that interactions between people from different cultural or societal groups are becoming increasingly prevalent in an ever-more globally interconnected world. Performance of teams and companies on a small scale, but also economic development of societies on a large scale, therefore crucially depend on the utilization of all available potential and the effective collaboration between different groups of people.

As for the overall spectrum of economic research, a unifying theme over all chapters is the assessment of how available resources and options are (optimally) employed to satisfy complex structures of needs. By using a behavioral economics approach to the different sub-fields, the chapters of this dissertation enrich the traditional framework in which these topics have been studied. Analyzing the impact of these important deviations from the classical setting systematically improves our understanding of behavior in these fields of application.

Chapter 1, which is joint work with Florian Englmaier, David Schindler and Simeon Schudy, investigates the determinants of performance in non-routine analytical team tasks. In particular, we analyze the effectiveness of bonus incentive schemes. Non-routine analytical team tasks are becoming increasingly important in modern economies (Autor et al., 2003; Autor and Price, 2013). Among other things, such tasks involve solving complex and previously unknown problems, and solutions often require information collection and (re-)combination and are fostered by innovative ideas. Nevertheless, while there is evidence regarding the effects of incentives in routine, manual and mechanical tasks (see, e.g., Bandiera et al., 2005), evidence is scarce for non-routine tasks.

Whether the reported positive effects of bonus incentives in routine tasks extend to non-routine analytical tasks is unclear. Incentives may discourage exploration and hinder creativity (e.g. Amabile, 1996). Further, if more intrinsically motivated people work on such tasks, extrinsic incentives potentially crowd out intrinsic motivation (e.g. Deci et al., 1999).

We make use of a unique field setting by cooperating with a real-life escape game provider. In escape games, teams have to solve a wide range of tasks to ultimately succeed (i.e. “escape”) in a given time limit. As for non-routine analytical team tasks, many problems faced are complex and unknown, require thinking outside the box, information collection and recombination, and coordination at the team level.

In a field experiment with more than 3000 participants — regular customers of our cooperation partner —, we document a positive effect of bonus incentives on performance. In the main treatment, where we provide teams with a monetary incentive for solving the task within 45 minutes (the time limit given by the provider is 60 minutes), teams perform significantly better than in a control treatment where no incentives are provided. This performance effect does not depend on how the incentive is framed: both a loss frame (teams receive the money before the task and have to hand it back in case of failure) and a gain frame (teams know they can earn the bonus and are paid after the task) result in roughly the same performance shift. We further show that it is not the reference point (performance threshold at 45 minutes) provided, but indeed the monetary component of the bonus that affects behavior. In a treatment only indicating 45 minutes as a very good finishing time and not offering a bonus, we do not see a performance effect. In contrast, in a treatment providing a bonus for solving the task in 60 minutes and hence not establishing a reference point at 45 minutes, we again do see an effect.

We find similar results for our main treatment with a sample of student participants. These students are exogenously assigned to teams and, by being invited to an unknown experiment, could not self-select into the specific task. Hence, they are presumably less intrinsically motivated. While this sample also shows a performance increase induced by the bonus, the exploration behavior of teams is different. Presumably less intrinsically motivated teams show a reduction in their willingness to explore original solutions under a bonus scheme. The student sample further allows us to shed light on team organization. An ex-post survey suggests that under the bonus schemes leadership structures emerge and coordination efforts increase.

This chapter addresses a widespread belief among practitioners that financial incentives reduce team performance in non-routine and creative tasks (see, e.g., Pink, 2009). Our findings lessen these concerns by providing evidence to the contrary.

Chapter 2 is a joint project with Martin Kocher, Michał Krawczyk and Fabrice Le Lec. The basic idea underlying this chapter is that decisions — including decisions under risk — are mostly embedded in a social context. From a vast literature on social preferences, we know that individuals oftentimes react very sensitively to others' situations and to differences in individual outcomes (e.g. Fehr and Schmidt, 1999). In economic research on decision making, however, we usually abstract from such contexts. This is true for most laboratory experiments on risk taking, too, in which individuals usually make choices in isolation.

In contrast to that practice, our experiment specifically investigates whether risk taking is affected by the social context. In particular, we look at the effect of social comparisons. Our focus is on situations where some fixed resource must be allocated between two parties. The decision maker (one of the two parties) can either share the resource deterministically, or allow a random device to allocate the entire resource to one of the parties. The ex-ante expected fraction of the resource to the decision maker is the same for both options. By varying the initial "power position" of the decision maker, i.e. her expected fraction, we can systematically investigate changes in risk attitudes by the nature of social standing.

While the existing evidence regarding social comparison effects on risk taking is inconclusive (see, e.g., Bolton and Ockenfels, 2010; Linde and Sonnemans, 2012), we find that the social context of the decision matters strongly and asymmetrically. When participants are in a disadvantaged initial position compared to the other party, they select the risky option much more often than in a purely individual decision, identical in all other respects. In contrast, the fraction of participants choosing the risky option is not different between the individual and social decision context when the social context involves a favorable relative standing, i.e. the decision maker initially has more claims on the resource.

We favor two distinct explanations for our data patterns. First, the (expected) payoff of the other party functions as a (social) reference point. Below that level of income, the decision maker finds herself in the loss domain and is risk seeking (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992). Second, individuals care about the social ranking and about being ahead and winning per se.

The findings clearly highlight the importance of the social context of decisions. More specifically, we show that relative standing has strong effects on risk taking. This suggests, for example, that managing information on relative standing can be an effective tool in shaping organizational risk taking and, e.g., in preventing escalations of excessive risk taking.

Chapter 3, similar to Chapter 2, aims at better understanding the social contexts of risky decision environments. In particular, the chapter focuses on choices being observed (non-anonymous) as one crucial element of many such contexts.

The non-private nature of choices can affect decision makers via social image concerns and social norms. Individuals care about how they are perceived by others and are motivated by adherence to social norms, especially when their behavior is publicly visible (Akerlof, 1980; Elster, 1989). These concerns have been documented to affect revealed social preferences in contexts with strong behavioral norms (e.g., Andreoni and Bernheim, 2009). There is some evidence that clear normative prescriptions exist in risk taking, too. Bem (1974) and follow-up studies have shown that high risk taking is rated desirable, and particularly so for males. Hence, observability of risk choices should increase risk taking, and should do so differently for males compared to females. I test this prediction and explicitly elicit gender-specific norms in risk taking in a laboratory experiment.

In one treatment, participants know that their investments in a risky asset will be revealed to a matched participant, while choices in the control treatment are anonymous. After investment decisions are made, I elicit participants' beliefs about the choice of the matched participants (descriptive norms) and behavior deemed appropriate (injunctive norms) using an incentivized eliciation procedure. I consequently link these norms to actual risk taking.

Interestingly, investments are not affected by the observability of choices. Nor do investments follow elicited norms for risk taking more closely when observed. These null findings hold for both males and females separately. However, I find strong evidence for gender-specific norms in risk taking. Males are expected to invest more than females. While these norm differences explain part of the gender gap in risk taking, males still “overshoot” by investing even more than the norm dictates. This is particularly pronounced for males matched with females.

Further, there is some evidence that the attractiveness of the matched participant is important for the effects of observability. Participants seem to react differently (more risk taking) to the choice being observed when matched with an attractive participant, compared to when matched with a less attractive participant.

The chapter clearly establishes that a gender gap exists not only in actual risk taking, but also in injunctive norms for risk taking. While I do not observe an overall effect of observability on risk taking, males’ “overshooting” beyond norms, their sensitivity to the gender of the matched participant and the potential impact of the attractiveness of observers are important findings, particularly for understanding organizational and group decision making in varying group compositions.

The last chapter of this dissertation, Chapter 4, is joint work with Felix Klimm. It studies discrimination in the context of responsibility attribution. More specifically, with a new experimental paradigm, we ask whether individuals attribute responsibility for positive or negative events to ingroup or outgroup members differently. While there is extensive evidence on discrimination and its consequences in a wide range of contexts (see, e.g., Chen and Li, 2009; Goldin and Rouse, 2000), it is unclear whether discrimination patterns persist in the attribution of responsibility such that, for example, negative shocks and events are blamed on outgroup members.

In the main experiment, we invite Arabic refugees to the laboratory. Regular (native) participants are matched with either a refugee or with another regular participant. They then experience a positive or negative income shock, which is with equal probability caused by a random draw or by the matched participant’s performance in a real effort task. Responsibility attribution is measured by beliefs about whether the shock is due to the other participant’s performance or the random draw.

In contrast to the vast majority of studies in the literature, we find evidence of reverse discrimination. Natives attribute responsibility more favorably to refugees than to other natives. That is, refugees are less often held responsible for negative, and more likely held responsible for positive income shocks. Moreover, natives with negative implicit associations towards Arabic names attribute responsibility less favorably to refugees than natives with positive associations. This speaks towards implicit associations having some predictive power for explicit discrimination behavior.

Since neither actual performance differences in the real effort task nor beliefs about natives' and refugees' performance can explain our finding of reverse discrimination, we rule out statistical discrimination as the driving force. Instead, we favor an explanation based on theories of self-image and identity concerns (Akerlof and Kranton, 2000). Tolerance is presumably part of our young and rather liberal participants' identity. To avoid being associated with xenophobic attitudes, participants may refrain from "blaming" the refugees, while this is not a concern with ingroup members. Further support for this hypothesis comes from our second experiment, in which we assign regular (native) participants to artificial ingroups and outgroups. In this experiment, we find no evidence of reverse discrimination.

The findings from our abstract design are informative for many settings in the field where responsibility for outcomes cannot be attributed with certainty. Hiring and promotion decisions involve many such examples in the labor market. On a societal level our design can also be related to attribution of aggregate developments to specific groups of people. One such example is the recent debate regarding the responsibility of refugee inflows for crime and cultural changes.

All four chapters of this dissertation are self-contained and can be read independently from one another. There is a separate appendix for each chapter. All of these are placed after Chapter 4. The bibliography including all references appears at the end of the document.

Chapter 1

The Effect of Incentives in Non-Routine Analytical Team Tasks — Evidence from a Field Experiment[*]

1.1 Introduction

Until the 1970s, a major share of the workforce performed predominantly manual and repetitive routine tasks with little need to coordinate in teams. Since then, we have witnessed a rapidly changing work environment. Nowadays, work is frequently organized in teams (see, e.g., Bandiera et al., 2013) and a large share of the workforce performs tasks that require much more cognitive effort rather than physical labor. Autor et al. (2003) analyze task input in the US economy using four broad task categories: routine manual tasks (e.g. sorting or repetitive assembly), routine analytical and interactive tasks (e.g. repetitive customer service), non-routine manual tasks (e.g. truck driving) and non-routine analytical and interpersonal tasks (e.g. forming and testing hypotheses) and document a strong increase in non-routine analytical and interpersonal tasks between 1970 and 2000. Autor and Price (2013) reaffirm the importance of these tasks in later years.

One main feature of non-routine analytical tasks is that they confront work teams with complex and previously unknown problems. Teams are supposed to come up with innovative solutions and, in order to succeed, they need to build up and recombine knowledge (Nelson and Winter, 1982). Examples range from teams of innovative product developers to management consultant teams who have to gather, evaluate, and recombine information about their clients' problems. While

[*] This chapter is based on joint work with Florian Englmaier, David Schindler and Simeon Schudy.

this idea of recombinant innovation goes back at least to Schumpeter (1934) and has been formalized in growth theory as "recombinant growth" by Weitzman (1998), it is also central in management research. The concept of the recombination of ideas is at the core of the study of innovation, and research has repeatedly found evidence for various forms of recombination as the main mechanism producing breakthroughs (see, e.g., Fleming, 2001; Gittelman and Kogut, 2003; Hall et al., 2001; Rosenkopf and Nerkar, 2001).

Given the pervasiveness of these tasks in modern economies and their importance for innovation and growth, understanding the determinants of performance in these tasks is crucial. One core question is how incentives affect teams working on these cognitively demanding, interactive and diverse tasks. In many modern work environments, contracts specify performance-related bonus payments as an important part of compensation. While there is well-identified evidence about the behavioral effects of monetary incentives on performance in mechanical and repetitive routine tasks such as fruit picking, tea plucking, tree planting, sales, or production (see, e.g., Bandiera et al., 2005, 2013; Delfgaauw et al., 2015; Englmaier et al., 2017; Erev et al., 1993; Friebel et al., 2017; Hossain and List, 2012; Jayaraman et al., 2016; Lazear, 2000; Shearer, 2004), evidence on the effects of bonus incentives is lacking for non-routine analytical tasks in which teams jointly solve a complex problem.

In this chapter, we exploit a unique field setting to measure the incentive effects for joint team performance in a non-routine analytical task. We study the performance of teams in a real-life escape game in which teams have to solve a series of cognitively demanding tasks in order to succeed (usually by escaping a room within a given time limit using a key or a numeric code). These games provide an excellent setting to study non-routine analytical and interactive team tasks: teams face complex and novel problems, have to solve analytical and cognitively demanding tasks, need to collect and recombine information which requires thinking outside the box. The task is also interactive, since members of each team have to collaborate with each other, discuss possible actions, and develop ideas jointly. At the same time, real life escape games allow for an objective measurement of joint team performance (time spent until completion), as well as for exogenous variation in incentives for a large number of teams. Our particular setting allows us to vary the incentive structure for more than 900 teams (with more than 4,000 participants) under otherwise equal conditions and thus enables us to isolate how bonus incentives affect team performance.

Whether bonus incentives positively affect performance in such tasks is an open question as the production technology as well as the selection of workers performing such tasks may differ compared to mechanical and routine tasks. Non-routine analytical and interactive tasks require information acquisition, information recombination, and creative thinking. There is thus room for incentives to discourage the exploration of new and original approaches (e.g. Amabile, 1996; Azoulay et al., 2011; Ederer and Manso, 2013; McCullers, 1978; McGraw, 1978).[1] Further, non-routine analytical tasks are more likely to be performed by people who are intrinsically motivated (see, e.g., Autor and Handel, 2013; Delfgaauw and Dur, 2010; Friebel and Giannetti, 2009). Extrinsic incentives could negatively affect team performance by crowding out such intrinsic motivation (e.g. Deci et al., 1999; Eckartz et al., 2012; Gerhart and Fang, 2015; Hennessey and Amabile, 2010).

Recent evidence from related strands of the literature on incentives for idea creation (Gibbs et al., 2017) and creativity (e.g. Bradler et al., 2014; Charness and Grieco, 2018; Gibbs et al., 2017; Laske and Schroeder, 2016; Ramm et al., 2013), however, do not indicate negative, but mostly positive incentive effects. While these studies provide interesting insights into how certain types of incentives can affect idea creation and creative performance, they almost exclusively measure individual production, instead of team production (i.e. workers may face team incentives but work on individual tasks).[2] One rare exception is the small scale laboratory experiment by Ramm et al. (2013), which investigates the effects of incentives on the performance of two paired individuals in a creative insight problem, in which the subjects are supposed to solve the candle problem of Duncker (1945). The study find no effects of tournament incentives on performance in pairs but it is unclear whether this effect is robust, as the authors achieve rather low statistical power.

Our unique field setting allows us to substantially advance the literature on incentives for non-routine tasks. We can study the causal effect of incentives on team performance as well as on teams' willingness to explore original solutions in a non-routine analytical team task in two very distinct samples. First, we conducted a series of field experiments with regular teams (customers of our cooperation partner) who were unaware of taking part in an experiment.[3] These

[1] Takahashi et al. (2016) further argue that incentive effects may also depend on whether the task is perceived as interesting.

[2] Bradler et al. (2014), Charness and Grieco (2018), and Laske and Schroeder (2016) study individual production. In Gibbs et al. (2017) team production is potentially possible but submitted ideas have fewer than two authors on average. Similarly, recent studies on the effectiveness of incentives for teachers (Fryer et al., 2012; Muralidharan and Sundararaman, 2011), who perform at least to some extent a non-routine task, find positive effects of performance incentives, but it remains unclear if and to what extent complementarities in individual teacher performance may be regarded as features of joint team production.

[3] Harrison and List (2004) classify this approach as a "natural field experiment". The study was approved by the Department of Economics' IRB at LMU Munich (Project 2015-11) and excluded customer teams with minors. Customers gave written consent that their data was to be shared with third parties for research purposes.

teams self-selected into the task and were intrinsically motivated to solve it. Second, we investigate whether our main treatment effects are also observed in a sample of student participants in which the teams did not self-select into the task and were exogenously formed.[4] Further, by using survey responses from the student participants, we provide some initial tentative insights on how incentives affect team organization.

To identify the effect of providing incentives, we implemented a between-subjects design, in which teams were randomly allocated to either a treatment condition or a control condition. For the main treatment, we offered a team bonus if the team completed the task within 45 minutes (the regular pre-specified upper limit for completing the task was 60 minutes). In the control condition, no incentives were provided. In both samples, we find that bonus incentives significantly and substantially increased performance in an objectively quantifiable dimension. Teams in the incentive treatment were more than twice as likely to complete the task within 45 minutes. Moreover, bonus incentives did not only have a local effect around the threshold for receiving the bonus but improved the performance over a significant part of the distribution of finishing times.

We leverage the advantages of our setting to study in depth the most important aspects of the incentive scheme for generating the treatment effect. We implemented the bonus incentive framed either as a gain or a loss, and find no significant differences in performance between these conditions. In contrast to earlier findings on bonus incentives for individually performed tasks (e.g., by Fryer et al., 2012; Hossain and List, 2012), our results suggest that framing might play a smaller role in non-routine, jointly solved team tasks. In addition, we implemented two treatments in the customer sample that allow us to disentangle whether bonus incentives are effective due to the performance threshold (the reference point) or the reward provided. A treatment in which we made the bonus threshold (i.e., 45 minutes) a salient reference point without providing incentives did not affect performance, whereas paying a bonus for completing the task in the regular pre-specified time of 60 minutes had a significant positive effect. Hence, the reward component seems to be key to bringing about the positive treatment effect, as opposed to merely a salient reference performance.

In order to understand what moderates the main treatment effects, we study different possible channels. Answers to our ex-post survey of the student sample suggest that incentives affect team organization in the sense that they promote the emergence of leadership and lead to a more focused and coordinated approach

[4] According to Harrison and List (2004), the student sample can be considered a framed field experiment as students are non-standard subjects in the context of real life escape games.

to solving the problem. Second, our findings (for the customer teams, who self-selected into the task) highlight that introducing incentives does not lead to a strong reduction in a team's willingness to explore innovative solutions. However, such discouragement is apparent among student teams, which were exogenously assigned to the task.

Our results provide important insights for researchers as well as practitioners in charge of designing incentive schemes for non-routine analytical team tasks. In particular, we speak to the pressing question of many practitioners, whether monetary incentives impair team performance in tasks that are non-routine and require creative thinking. This idea has recently been strongly promoted in the public, for instance by the best selling author Daniel Pink, in his famous TED talk with more than 19 million views and his popular book *Drive* (Pink, 2009, 2011). Our results alleviate most of these concerns, since we provide novel and robust evidence that bonus incentives are a viable instrument to increase performance in such tasks. The incentives in our experiment did not reduce performance but instead affected teams' outcomes positively across two distinct samples. Second, we show that it was indeed the reward component of the bonus, and not the reference point of good performance which improved teams' outcomes. The latter findings complement recent research on non-monetary means of increasing performance, in particular research referring to workers' awareness of relative performance (for a review of this literature see Levitt and Neckermann, 2014). Third, we add novel and interesting insights to the discussion of whether incentives discourage the exploration of new approaches. The answer to this question hinges crucially on the characteristics of the underlying sample. We observe such discouragement only among the student sample, in which, presumably, less intrinsically motivated teams worked on the task. This result substantially extends recent laboratory findings by Ederer and Manso (2013), who show that pay-for-performance schemes can discourage the exploration of new approaches, as it informs us about when and how incentives may result in unintended consequences. Finally, we discover a novel and interesting potential channel through which incentives may improve team performance as student teams facing incentives tended to be more likely to express a desire for leadership and to report being better led.

The rest of this chapter is organized as follows: Section 1.2 presents the field setting and the experimental design. Section 1.3 provides the results from both experiments. We provide a discussion in Section 1.4 and Section 1.5 concludes.

1.2 Experimental Design

1.2.1 The Field Setting

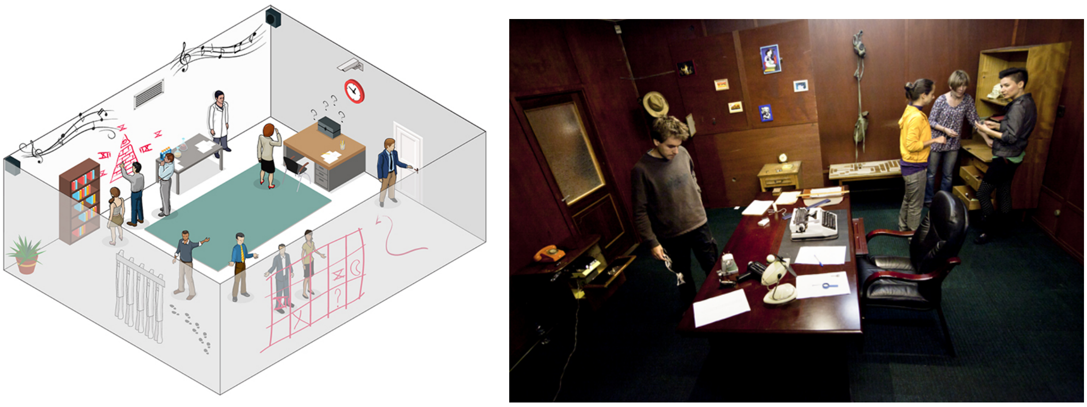
We cooperate with the company *ExitTheRoom*[5] (ETR), a provider of real-life escape games. In these games, teams of players have to solve, in a real setting, a series of tasks that are cognitively demanding, non-routine, and interactive, in order to succeed (usually by escaping from a room within a given time limit). Real-life escape games have become increasingly popular over the last years, and can now be found in almost all major cities around the globe. Often, the task is embedded in a story (e.g., to find a cure for a disease or to defuse a bomb), which is also reflected in the design of the room and how the information is presented. The task itself consists of a series of quests in which teams have to find cues, combine information, and think outside the box. They make unusual use of objects, and they exchange and develop innovative and creative ideas to solve the task they are facing within a given time limit. If a team manages to solve the task before the allotted time (one hour) expires, they win—if time runs out before the team solves all quests, the team loses.

Figure 1.1 ilustrates the idea and the setup of such escape rooms and shows an actual example from a real-life escape game room. The left panel is an illustration of a typical room, which contains several items, such as desks, shelves, telephones, books, and so on. These items may contain information needed to eventually solve the task. Typically, not all items will contain helpful information, and part of the task is determining which items are useful for solving the quests. The right panel shows a picture of participants actively trying to escape from their room. They already have opened drawers and closets to collect potential clues, and now jointly sort, process, and deliberate on how to use the retrieved information.

To illustrate a typical quest in a real-life escape game, we provide a fictitious example.[6] Suppose the participants have found and opened a locked box that contains a megaphone. Apart from being used as a speaker, the megaphone can also play three distinct types of alarm sounds. Among the many other items in the room, there is a volume unit (VU) meter in one corner of the room. To open a padlock on a box containing additional information, the participants will need a three digit code. The solution to this quest is to play the three types of alarms on the megaphone and write down the corresponding readings from the VU meter to obtain the correct combination for the padlock. The teams at ETR solve quests similar to this fictitious example. The tasks at ETR may further include finding

[5] See `https://www.exittheroom.de/munich`.

[6] Our partner ETR asked us to not present an actual example from their rooms.



Notes: The left panel shows typical layout of such a room, including items that might provide clues needed for a successful escape. Source: http://www.marketwatch.com/story/the-weird-new-world-of-escape-room-businesses-2015-07-20. The right panel shows a picture of participants actively searching their room for hints and combining the discovered information. Source: http://boredinvancouver.com/listing/escape-game-room-experience-vancouver/.

Figure 1.1: Examples of real-life escape games

hidden information in pictures, constructing a flashlight out of several parts, or identifying and solving rebus (word picture) puzzles (see also Erat and Gneezy, 2016; Kachelmaier et al., 2008).

We conducted our experiments at the facilities of ETR in Munich. The location offers three rooms with different themes and background stories.[7] Teams face a time limit of 60 minutes and can see the remaining time on a large screen in their room. A room will be declared as solved if the team manages to escape from the room (or defuse the bomb) within 60 minutes. If a team does not manage to do so within 60 minutes, the task is declared unsolved and the game ends. If a team gets stuck, they can request hints via radio from the staff at ETR. As they can only ask for up to five hints in all, a team needs to state explicitly that they want to receive a hint. The hints never state the direct solution to a task, but only provide vague clues regarding the next required step.

The setting at ETR reflects many aspects of modern non-routine analytical team tasks. First, finding clues and information very much matches the activity of research that is often necessary before collaborative team work begins. Second, combining the discovered information is not trivial, and requires ability for creative

[7] *Zombie Apocalypse* requires teams to find the correct mix of liquids before time runs out (the anti-Zombie potion). In *The Bomb*, a bomb and a code to defuse it has to be found. In *Madness*, teams need to find the correct code to open a door so as to escape (ironically) before a mad researcher experiments on them. For the sake of the reader, we refrain from presenting the regression specifications with room fixed effects in the main text. We provide these specifications in the Appendix. Adding room fixed effects does not change our results (see Table A.1).

problem solving. The subjects are required to process stimuli in a way that transcends the usual thinking patterns, or are required to make use of objects in unusual ways. Third, to solve the task, the subjects must effectively cooperate as a team. As in actual work environments, where the individuals in a team are supposed to provide additional angles on the problem at hand, different approaches to problem solving will enable a team to solve the task more quickly. Lastly, participants who self-select into the task have a strong motivation to succeed as they have spent a non-negligible amount of money to perform the task (participants pay between €79 (for two-person groups) and €119 (for six-person groups) for a one-hour game). We interpret the fact that many teams opt to write their names and finishing times on the walls of the entrance area of ETR as evidence for such a strong motivation. Another, more objective, reason to solve the task quickly is the fact that at any given point in time, teams do not know how many quests are left to solve the task in its entirety. That is, if a team wants to succeed, they have an incentive to succeed quickly.

While these features provide an excellent framework for studying the effect of incentives on team performance, the setting is also extremely flexible. The collaboration with ETR allows implementing different incentives for more than 700 teams of customers and studying whether incentives increase performance also in a sample of presumably less motivated and exogenously formed teams of student participants. In particular, it affords a unique opportunity to compare incentive effects for teams who have self-selected into the task (regular customers) and incentive effects for teams who were confronted with the task by us, i.e., teams who perform the task as part of their paid participation in an economic experiment.

1.2.2 Experimental Treatments and Measures of Performance

We conducted the field experiment with 3308 customers (722 teams) of *ExitTheRoom* Munich and implemented a between-subjects design. Our main treatments included 487 teams who were randomly allocated to either the control condition or a bonus incentive condition. In the bonus condition, *Bonus45* (249 teams), a team received a monetary team bonus if they managed to solve the task in less than 45 minutes. In the *Control* condition (238 teams), teams were not offered any bonus. We framed the bonus either as a gain (125 teams) or as a loss (124 teams). In *Gain45*, each team was informed that they would receive the bonus if they managed to solve the task in less than 45 minutes. In *Loss45*, each team received the bonus in cash up front, kept it during their time in the room, and were informed that they would have to return the money if they did not manage to solve the task in less than 45 minutes.[8]

Additionally, we ran two experimental treatments that allow us to test whether bonus incentives were effective because of the monetary benefits or because the 45-minute threshold worked as a salient reference point. In the first additional treatment (*Reference Point*, 147 teams), we explicitly mentioned the 45 minutes as a salient reference point before the team started working on the task. However, we did not pay any bonus. We said: "In order for you to judge what constitutes a good performance in terms of remaining time: If you make it in 45 minutes or less, that is a very good result." In treatments *Gain60* (42 teams) and *Loss60* (46 teams), we provided a monetary bonus but did not provide the reference point of 45 minutes: Teams received the bonus if they solved the task within 60 minutes.

We collected observable information related to team performance and team characteristics, which include time needed to complete the task, number and timing of requested hints, team size, gender and age composition of the team[9], team language (German or English), experience with escape games[10], and whether the customers came as a private group or were part of a company team building event. Our primary outcome variable is team performance, which we measure by i) whether or not teams solved the task in 45 minutes and by ii) the time left upon completing the task. Comparing the incentive treatments with the control condition allows us to estimate the causal effect of bonus incentives on these objective performance measures. The difference between performance in *Loss45* and *Gain45* allows us to determine whether there is a benefit from providing incentives in a loss frame compared to a gain frame. Differences in performance between *Reference Point* and *Control* reveal whether the reference point of 45 minutes increased the

[8] The bonus amounted, on average, to approximately €10 per team member. Teams in the field experiment received a bonus of €50 (for the entire team of between two and eight members, on average about five). To keep the per-person incentives constant in the student sample with three team members (described below), the student teams received a bonus of €30. The treatment intervention (i.e. the bonus announcement) was always implemented by the experimenter present on-site. For that purpose, he or she announced the possibility for the team to earn a bonus and had the teams sign a form (see Appendix A.2) indicating that they understood the conditions for receiving (in *Gain45*) or keeping (in *Loss45*) the bonus. The bonus incentive was described as a special offer and no team questioned that statement. The experimenter also collected the data. We always made sure that the experimenters blended in with the ETR staff.

[9] In order to preserve the natural field experiment, we did not interfere with the standard procedures of ETR. Thus we did not explicitly elicit participants' ages. Instead, the age of each participant was estimated based on appearance to be either 1) below 18 years, 2) between 18 and 25 years, 3) between 26 and 35 years, 4) between 36 and 50 years, 5) 51 years or older. Teams with members estimated to be minors were excluded from the experiment (following the request by the IRB).

[10] ETR staff ask teams whether they have ever participated in an escape game irrespective of our experiment.

performance of the teams even if a monetary bonus was absent. The performance in *Gain60* and *Loss60* as compared to *Control* allows an additional test of whether the monetary component of the bonus was effective even when there was no change in the reference point as compared to the control.[11]

Further, we replicated our main treatments (*Gain45*, *Loss45* and *Control*) in a framed field experiment at ETR in which we randomly allocated student participants from the subject pool of the social sciences laboratory at the University of Munich (MELESSA) to teams (804 participants in 268 teams). The additional sample allows us to study whether bonuses affect team performance in similar ways when the team composition was exogenous and the teams did not themselves choose to perform the non-routine task. Further, it enables us to collect additional data on task perception and team organization.

1.2.3 Procedures

Natural Field Experiment (Customer Sample)

We conducted the field experiment with customers of *ExitTheRoom* during their regular opening hours from Monday to Friday.[12] We implemented the main treatments of the field experiment (*Gain45*, *Loss45* and *Control*) in November and December 2015 and from January to May 2017. In the second phase of data collection we further ran the additional treatments *Loss60*, *Gain60* and *Reference Point*. We randomized on a daily level to avoid treatment spillovers between different teams on-site (as participants from one slot could potentially encounter participants arriving early for the next slot, and overhear, e.g. the possibility of earning money). Further, we avoided selection into treatment by not announcing treatments ex ante and randomly assigning treatments to days after most booking slots had already been filled.[13]

Upon arrival, ETR staff welcomed teams of customers as usual and customers signed ETR's terms and conditions, including ETR's data privacy policy. Then, the

[11] Note that in *Control*, roughly ten percent of the teams solved the task within 45 minutes, whereas roughly 70 percent did so within 60 minutes. Hence, the treatments which paid a bonus for solving the task in 60 minutes reveal also whether bonuses worked even if they did not refer to extraordinary performance.

[12] ETR offers time slots from Monday through Friday from 3:45 p.m. to 9:45 p.m., and Saturday and Sunday from 11:15 a.m. to 9:45 p.m., with the different rooms shifted by 15 minutes to avoid overlaps and congregations of teams in the hallway.

[13] All slots in November and December 2015 were fully booked before treatment assignment. According to the provider, fewer than five percent of their bookings are made on the day of an event after the first time slot has ended.

staff explained the rules of the game. Afterwards, the teams were shown to their room and began solving the task. Teams were not informed that they were taking part in an experiment. The only difference between the treatment conditions and the control was that in the bonus conditions, the bonuses were announced as a special offer to reward particularly successful teams, while in the reference point treatment, the finishing time of 45 minutes was mentioned saliently before the team started working on the task.

Framed Field Experiment (Student Sample)

For the framed field experiment, we invited student participants from the social sciences laboratory at the University of Munich (MELESSA). Between March and June 2016, and January and May 2017, a total of 804 participants (268 groups) took part in the experiment. To avoid selection into the sample based on interest in the task, we recruited these participants using a neutrally framed invitation text that did not explicitly state what activity participants could expect. The invitation email informed potential participants that the experiment consisted of two parts, of which only the first part would be conducted on the premises of MELESSA whereas the second part would take place outside of the laboratory (without mentioning the escape game). They were further informed that their earnings from the first part would depend on the decisions they made and that the second part would include an activity with a participation fee that would be covered by the experimenters (as part of participants' compensation for taking part in the experiment).[14]

Upon arrival at the laboratory, the participants were informed about their upcoming participation in an escape game. The participants had the option to opt out of the experiment, but no one did so. In the first part of the experiment, i.e. on the premises of MELESSA, we elicited the same control variables as for the customer sample (age, gender, and potential experience with escape games). In addition, the participants took part in three short experimental tasks and answered several surveys. As the main focus of this chapter is to analyze the robustness of the incentive effects across the two samples, we relegate the discussion of the results from these additional tasks to another essay.[15] After completion of the laboratory

[14] Section A.3 in the Appendix provides a translation of the text of the invitation.

[15] These tasks included an elicitation of the willingness-to-pay for a voucher of *ExitTheRoom*, an experimental measure of loss aversion (based on Gächter et al., 2007) and a word creation task (developed by Eckartz et al., 2012). The participants also answered questionnaires regarding creativity (Gough, 1979), competitiveness (Helmreich and Spence, 1978), status (Mujcic and Frijters, 2013), a big five inventory (Gosling et al., 2003), risk preferences (Dohmen et al., 2011) and standard demographics. On average, the subjects spent roughly 30 minutes to complete the experimental tasks and questionnaires.

part, the experimenters guided the participants to the facilities of ETR which are located a ten-minute walk (0.4 miles / 650 meters) away from the laboratory. At ETR, each participant was randomly allocated to a team of three members, received the same explanations from ETR staff that were given in the field experiment, and, depending on the treatment, was informed about the possibility of earning a bonus. For the student sample, we randomized the treatments on the session level (stratifying on rooms), as student teams in different sessions on a given day could not talk to each other at the facilities of ETR. During the performance of the task, the same information about the team performance as in the field experiment was collected. On completion of the task, the participants answered questions about the team's behavior, organization, and their perception of the task individually, on separate tablet computers. At the end, we paid the earnings individually in cash. In addition to the participation fee for ETR, which we covered (given the regular price, this corresponds to roughly €25 per person), participants earned on average €7.53, with payments ranging from €3.50 to €87.[16]

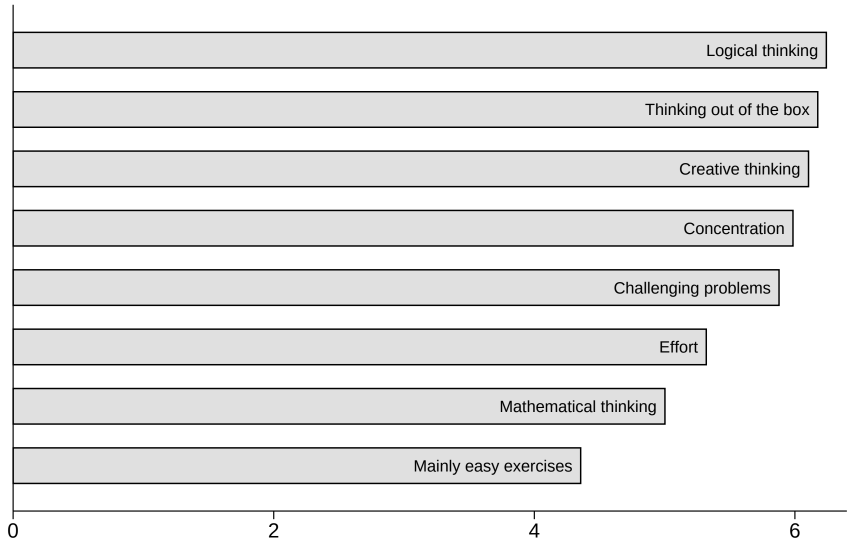
1.3 Results

We organize the presentation of our findings as follows. We begin our analysis by establishing the internal validity of our experimental approach. We show that the student participants perceive the task at *ExitTheRoom* as non-routine and analytical, i.e. involving more cognitive effort and creative thinking than easy, routine exercises. Then, we analyze our main research question, whether bonuses improve team performance. As our findings are affirmative, we explore next the channels through which bonus incentives operate. We disentangle which elements of the bonus (framing, monetary reward, reference point) are most relevant for bringing about the performance effect and investigate whether the observed effects of bonuses on performance are robust. We study whether the effects of bonuses on the teams that self-selected into the task differ from those on the teams that we confronted with the task, and whether the bonuses affect team organization. Finally, we highlight how bonus incentives affect a team's willingness to explore new approaches, and evaluate whether incentives affect this exploratory behavior differently for teams in the natural versus the framed field experiment.

[16] In one of the laboratory tasks, the student participants further had the chance to win a voucher for ETR worth roughly €100. Twenty-six participants actually won such a voucher, implying an average additional earning from this task of roughly €3.23. Adding up all these earnings assuming market prices as valuations, the participants on average earned an equivalent of €35.76 for an experiment lasting two hours.

1.3.1 Task Perception and Randomization

We have previously argued that real-life escape games offer the opportunity to study a class of tasks that is highly relevant to modern workplaces, as teams face a non-routine, analytical, and interactive challenge that requires thinking outside the box and logical thinking rather than easy repetitive chores. In order to not interfere with the standard procedures at ETR, we could not run extensive surveys and, e.g., ask regular customers about their perception of the task. However, we asked the student participants from the framed field experiment ($n = 804$) to what extent they agree that the team task exhibits various characteristics (using a seven-point Likert scale). Figure 1.2 shows the mean answers of our participants. Participants strongly agreed that the task involves logical thinking, thinking outside the box, and creative thinking, in particular as compared to mathematical thinking and easy exercises (signed-rank tests reject that the ratings have the same underlying distribution, all p-values < 0.01 except for *Thinking outside the box* vs. *Logical thinking*, $p = 0.16$ and *Thinking out of the box* vs. *Creative thinking* $p = 0.02$).



Notes: The figure shows mean answers of $N = 804$ student participants to eight questions concerning attributes of the task. Answers were given on a 7-point Likert scale.

Figure 1.2: Task perception

Table 1.1 provides an overview of the properties of the sample in the main treatments of the natural field experiment with ETR customers. The table highlights that our randomization was successful, based on observables such as the share of males, group size, experience, whether teams were taking part in a private or company event, and whether the team was English-speaking.

Table 1.1: Sample size and characteristics

	Control (n=238)	*Bonus45 (n=249)*
Share males	0.52 (0.29) [0,1]	0.51 (0.29) [0,1]
Group size	4.53 (1.18) [2,7]	4.71 (1.05) [2,8]
Experience	0.48 (0.50) [0,1]	0.48 (0.50) [0,1]
Private	0.69 (0.46) [0,1]	0.63 (0.48) [0,1]
English-speaking	0.12 (0.32) [0,1]	0.08 (0.28) [0,1]
Age category $\in$ {18-25;26-35;36-50;51+}	{0.29;0.45;0.21;0.05}	{0.18;0.42;0.33;0.07}***

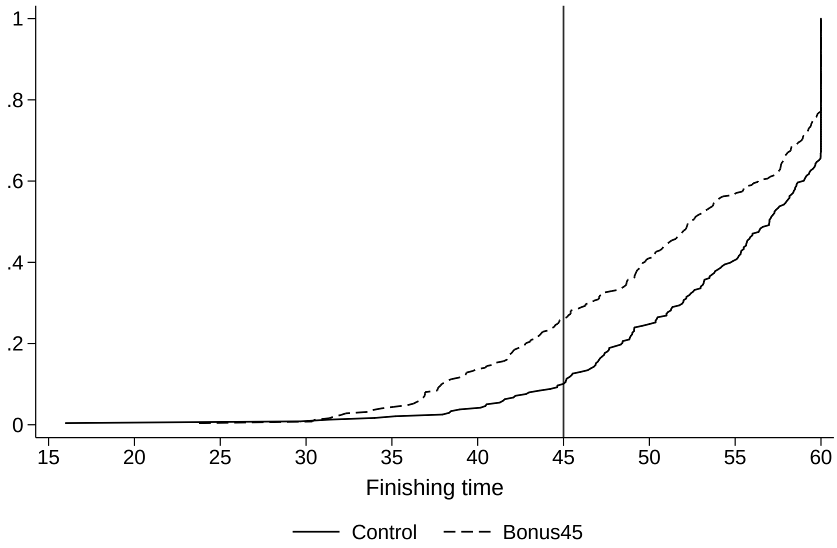
Notes: All variables except age category refer to means on the group level. Experience refers to teams that have at least one member who experienced an escape game before. Private refers to whether a team is composed of private members (1) or whether the team belongs to a team building event (0). Standard deviations and minimum and maximum values in parentheses; (std.err.)[min, max]. Age category displays fractions of participants in the respective age category. Stars indicate significant differences to *Control* (using χ^2 tests for frequencies and Mann–Whitney tests for distributions), and *** p<0.01, ** p<0.05, * p<0.1.

The only characteristic which differs significantly across treatments is the distribution of participants over the age categories guessed by our research assistants (χ^2 test, p-value < 0.01). We therefore provide results from both the regression specifications without controls and the regression specifications in which we control for the estimated age ranges (and other observables).

1.3.2 Bonus Incentives and Team Performance

We now turn to our primary research question, whether providing bonus incentives improves team performance. As mentioned earlier, our objective outcome measure of performance is whether teams manage to solve the task within 45 minutes and more generally how much time teams need to solve the task. Figure 1.3 shows the cumulative distribution of finishing times with and without bonus incentives in the field experiment. The vertical line marks the time limit for the bonus. The figure indicates that bonus incentives induce teams to complete the task faster and that the positive effect is not only prevalent around the bonus threshold but over a large part of the support of the distribution.

In *Control*, only 10 percent of the teams manage to finish the task within 45 minutes whereas in the bonus treatments more than twice as many teams (26.1 percent) do so (χ^2 test, p-value < 0.01). The remaining time upon solving also differs significantly between *Bonus45* and *Control* (p-value < 0.01, Mann–Whitney test). In *Bonus45*, teams are on average about three minutes faster than in *Control*. The positive effect of bonuses on performance is also reflected in the fraction of teams finishing the task within 60 minutes. With bonuses, 77 percent of the teams finish the task before the 60 minutes



Notes: The figure shows the cumulative distributions of finishing times with and without bonus incentives. The vertical line marks the time limit for the bonus.

Figure 1.3: Finishing times in *Bonus45* and *Control* in the field experiment

expire, whereas in *Control* this fraction amounts to only 67 percent (χ^2 test, p-value $= 0.01$, see also Table 1.4).

In addition to our non-parametric tests, we provide regression analyses which allow us to control for observable team characteristics (gender composition of the team, team size, experience with escape games, private vs. team building, English-speaking, and the estimated age of team members). Table 1.2 presents the results from a series of probit regressions that estimate the probability of solving the task within 45 minutes. We cluster standard errors at the day level (at which we varied the treatment) throughout. Column (1) includes only a dummy variable for the bonus treatments *Bonus45*. Bonus incentives are estimated to increase the probability of solving the task in less than 45 minutes by 16.5 percentage points. In Column (2), we add observable characteristics (see also Table 1.1). Here, and in the following analysis, group size and experience with escape games have a positive effect on performance whereas English-speaking groups perform slightly worse.[17] In Column (3) we add fixed effects for the ETR staff members on duty and in Column (4) we add week fixed effects. Across all specifications, the coefficients of the bonus treatments are positive and highly significant. Paying bonuses to teams solving a non-routine task strongly enhances their performance. We also estimate the effects of bonuses on the time remaining upon solving the task, which largely confirms both the results from the non-parametric tests on the remaining time as

[17] See also Table A.2 in the Appendix. Table A.2 further shows that the treatment effect does not strongly interact with the observable team characteristics. Only the interaction of incentives and experience (model (4) in A.2) turns out to be significantly positive at the ten percent level.

Table 1.2: Probit regressions: Solved in less than 45 minutes

	Probit (ME): Solved in less than 45 minutes				
	(1)	(2)	(3)	(4)	(5)
Bonus45	0.165***	0.164***	0.188***	0.151***	
	(0.024)	(0.022)	(0.025)	(0.041)	
Gain45					0.125***
					(0.037)
Loss45					0.174***
					(0.046)
Fraction of control teams solving the task in less than 45 min	0.10	0.10	0.10	0.10	0.10
Control Variables	No	Yes	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes	Yes
Week Fixed Effects	No	No	No	Yes	Yes
Observations	487	487	487	487	487

Notes: The table displays average marginal effects from probit regressions of whether a team solved the task within 45 minutes on our treatment indicators (with *Control* as base category). Control variables added from column (2) onwards include team size, share of males in a team, a dummy whether someone in the team has been to an escape game before, dummies for median age category of the team, a dummy whether the group speaks German and a dummy for private teams (opposed to company team building events). Staff fixed effects control for the employees of ETR present on-site and week fixed effects for week of data collection. All models include the full sample, including weeks that perfectly predict failure to receive the bonus (Table A.3 in Section A.4 of the Appendix reports regressions from a sample excluding weeks without variation in the outcome variable). Robust standard errors clustered at the day level reported in parentheses, and *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

well as the results from the probit models in Table 1.2, although the results are not statistically significant in all specifications (see Table A.4 in Appendix A.4.3).

We can look in more detail at the effectiveness of incentives depending on time elapsed since the beginning of the task. Since the incentive only rewards completing the task in the first 45 minutes, it should theoretically lose its effect in the last 15 minutes of the task. In addition, if incentives crowd out out intrinsic motivation to solve the task, we should see a decrease in performance after 45 minutes compared to *Control*. To test this hypothesis, we run a Cox proportional hazard model, where we define the hazard as completing the task. If our prior was true, we should observe the treatment to have a strong effect on the hazard in the first 45 minutes, and no or even a negative effect in the last 15 minutes, conditional on covariates.

Table 1.3 shows the hazard ratios using our usual set of controls and employing robust standard errors. Columns (1) through (3) estimate the effect on the hazard rate for the first 45 minutes and columns (4) through (6) for the last 15 minutes. In columns (1) and (4) we present the baseline effect of the treatment without any covariates. These are added in columns (2) and (5) respectively. Columns (3) and (6) also include week and staff fixed effects. The treatment clearly increases the hazard rate of completing the task in the first 45 minutes. All coefficients are significantly

Table 1.3: Influence of main bonus treatment on hazard rates

	Cox Proportional Hazard Model: Finishing the Task					
	First 45 min (1)-(3)			Last 15 min (4)-(6)		
	(1)	(2)	(3)	(4)	(5)	(6)
Bonus45	2.853***	2.947***	2.914***	1.178	1.251	0.841
	(0.446)	(0.474)	(0.844)	(0.189)	(0.248)	(0.180)
p-value for prop. haz. assumption	0.830	0.748	1.000	0.800	0.686	0.995
Control Variables	No	Yes	Yes	No	Yes	Yes
Staff Fixed Effects	No	No	Yes	No	No	Yes
Week Fixed Effects	No	No	Yes	No	No	Yes
Observations	487	487	487	487	487	487

Notes: Hazard ratios from a Cox proportional hazard regression of time elapsed until a team has completed the task on our treatment indicator *Bonus45*. Control variables, staff and week fixed effects as in Table 1.2. Robust standard errors clustered at the day level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1. Significant coefficients imply that the null hypothesis of equal hazards (i.e. ratio = 1) can be rejected. The proportional hazard assumption is tested against the null that the relative hazard between the two treatment groups is constant over time.

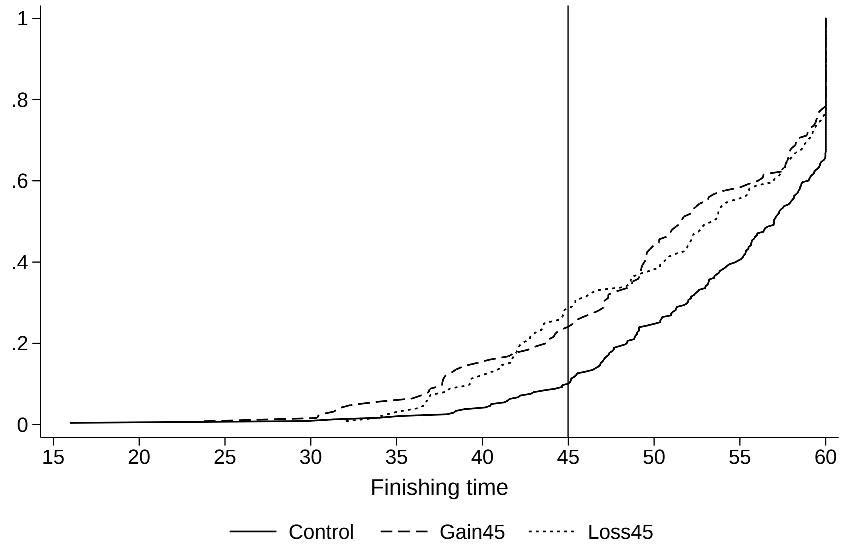
different from 1 and large in magnitude. Adding controls and fixed effects doesn't change the estimates by much, and the p-values of the proportional hazard assumption test do not indicate any reason to doubt our specification. In the last 15 minutes (columns (4) to (6)), however, the effect has almost completely vanished. The coefficient on our treatment switches from far above one to around one, and is not significantly different from 1 in any specifications. Again, the proportional hazard assumption cannot be rejected. Thus our data reflects two important aspects. First, the treatment indeed increases the likelihood of completing the task in the first 45 minutes, but much less so in the last 15 minutes. Second, incentives are unlikely to crowd out intrinsic motivation in our setting. We conclude:

Result 1 *Bonus incentives increase team performance in the non-routine task.*

1.3.3 Elements of Bonus Incentives: Framing, Rewards and Reference Performance

Framing of Bonus Incentives

As explained in the section on the experimental design, for roughly one-half of the teams in *Bonus45* we framed the bonus incentives as gains, while the other half faced a loss frame. Figure 1.4 shows the cumulative distributions of finishing times for both frames separately.



Notes: The figure shows the cumulative distribution of finishing times with bonus incentives framed as either gains, losses, or without bonuses. The vertical line marks the time limit for the bonus.

Figure 1.4: Finishing times in *Gain45*, *Loss45*, and *Control* in the field experiment

While somewhat ambiguous, we find that the framing of the bonus is of minor importance for team performance. A Mann–Whitney test fails to reject the null hypothesis that the finishing times for the two framings come from the same underlying distribution (p-value $= 0.70$). Also, the fractions of teams solving the task within 45 minutes does not differ significantly (in *Gain45*, 24 percent of teams finish within 45 minutes, in *Loss45* 28 percent of teams do so, χ^2-test, p-value $= 0.45$). Further, the fraction of teams solving the task in 60 minutes (78 percent in *Gain45* and 77 percent in *Loss45*) does not differ significantly (χ^2-test, p-value $= 0.85$) and no statistically significant differences are observed for the remaining times across frames. In *Gain45*, teams have on average 36 seconds more left than in *Loss45*, and the successful teams in *Gain45* have on average 37 seconds more left than in *Loss45* (Mann–Whitney test, p-value $= 0.71$). Table 1.4 summarizes these different performance measures.

Table 1.4: Task performance for main treatments

	Control	*Bonus45*	*Gain45*	*Loss45*
Fraction of teams solving task in 45 min	0.10	0.26***	0.24***	0.28***
Fraction of teams solving task in 60 min	0.67	0.77**	0.78**	0.77*
Mean remaining time (in sec)	345	530***	548***	512***
Mean remaining time (in sec) if solved	515	688***	707***	669***

Notes: This table summarizes key variables and their differences across our three treatments *Control*, *Gain45*, and *Loss45*, and the pooled bonus incentive treatment (*Bonus45*). Stars indicate significant differences from *Control* (using χ^2 tests for frequencies and Mann–Whitney tests for distributions), and *** p<0.01, ** p<0.05, * p<0.1.

In addition to the non-parametric analyses we report results from a regression of the probability of solving the task within 45 minutes on a separate dummy for each framing of the bonus and our control variables in Column (5) of Table 1.2. Incentives significantly increase the probability of solving the task within 45 minutes under both frames (as compared to the control condition). However, the average marginal effect for the *Loss45* treatment is estimated to be 5 percentage points larger and a post-estimation Wald test for the equivalence of the coefficients *Gain45* and *Loss45* in Column (5) of Table 1.2 identifies a statistically significant difference across the two frames (Wald test, p-value < 0.05). However, the same test fails to achieve significance at the ten percent level in alternative specifications that either exclude staff and week fixed effects (Wald test, p-value $= 0.26$) or use Huber-White standard errors instead of clustering standard errors at the day level (Wald test, p-value $= 0.38$). Furthermore, the results in Table A.4 show that framing bonuses as losses does not seem to have any additional effect on the time remaining (Wald test, p-value $= 0.98$). We thus summarize our findings as follows in Result 2.

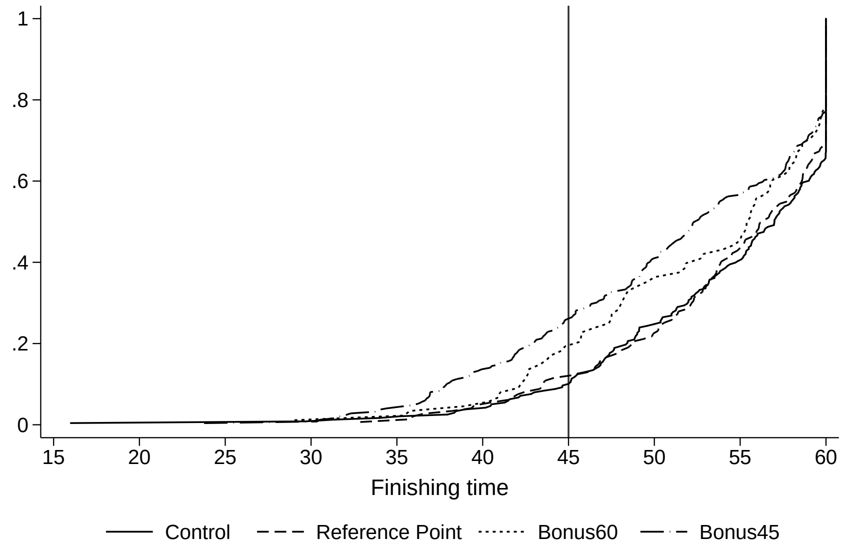
Result 2 *The framing of bonuses plays a minor role.*

Reference Points vs. Monetary Rewards

To understand whether bonus incentives work due to the monetary reward or due to the fact that the bonus also created a salient reference point at the 45-minute mark, we conducted two additional treatments. In *Reference Point* we introduce the 45-minute threshold as a salient reference point but do not pay a reward. In *Bonus60* we pay a bonus (again framed as a gain or a loss) for solving the task in 60 minutes.[18] Figure 1.5 shows the cumulative distribution of finishing times in *Control*, *Reference Point*, *Bonus60* and *Bonus45* and indicates that monetary rewards reduce the amount of time teams need to finish the task (*Bonus60* vs. *Control*, Mann–Whitney test, p-value $= 0.05$; *Bonus45* vs. *Control*, Mann–Whitney test, p-value < 0.01, with *Bonus45* vs. *Bonus60*, Mann–Whitney test, p-value $= 0.24$), whereas the cumulative distribution of remaining times in *Reference Point* almost perfectly overlaps with the cumulative distribution function in *Control* (Mann–Whitney test, p-value $= 0.78$). Hence, this is strong evidence that it is not the provision of a salient reference performance, but rather the reward component of the bonus incentives which generates the performance increase.

Lastly, we provide a regression analysis for the full sample in Table 1.5. We regress the probability of finishing within 45 minutes on the three treatment

[18] We do not differentiate between the gain and the loss frame of *Bonus60* in the following. As for *Bonus45*, no difference between the frames emerged.



Notes: The figure shows the cumulative distribution of finishing times of *Bonus45* (pooled), *Bonus60* (pooled), *Reference Point* and *Control*. The vertical line marks the time limit for the bonus in the *Bonus45* condition.

Figure 1.5: Finishing times for all treatments in the field experiment

indicators *Reference Point*, *Bonus60* and *Bonus45*. Column (1) includes only the treatment dummies. In Column (2), we add our set of control variables. In Column (3) we add staff fixed effects and in Column (4) we add week fixed effects. The regressions show that monetary incentives significantly increase the probability of finishing within 45 minutes, whereas the reference treatment does not.[19] It also becomes apparent that this finding is robust to the addition of covariates and fixed effects. Moreover, a post-estimation Wald test rejects the equality of coefficients of *Bonus60* and *Reference Point* in all specifications (models (1) to (4), p-values < 0.1). Similarly, the coefficient of *Bonus45* is significantly larger than the coefficient of *Reference Point* in all specifications (p-value $= 0.07$ in model (4) and p-value < 0.01 in all other specifications). Equality of coefficients of *Bonus60* and *Bonus45* can only be rejected for one of the specifications (model (2), p-value $= 0.095$). We summarize this finding in Result 3:

Result 3 *Bonuses increase performance due to the monetary reward they provide. Introducing a salient reference performance (indicating extraordinary performance) is not sufficient to induce a performance shift.*

[19] Table A.5 in Appendix A.4 confirms these findings for remaining time as dependent variable.

Table 1.5: Probit regressions: Solved in less than 45 minutes (all treatments)

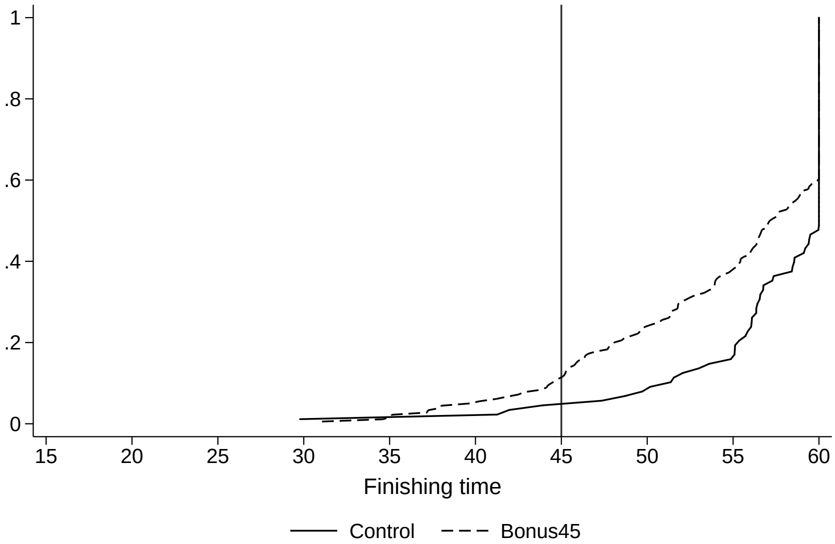
	Probit (ME): Solved in less than 45 minutes			
	(1)	(2)	(3)	(4)
Bonus45	0.160***	0.157***	0.164***	0.108***
	(0.023)	(0.022)	(0.026)	(0.035)
Bonus60	0.105**	0.102***	0.105***	0.127**
	(0.041)	(0.038)	(0.039)	(0.051)
Reference Point	0.025	0.023	0.011	0.020
	(0.032)	(0.035)	(0.039)	(0.039)
Fraction of control teams solving the task in less than 45 min	0.10	0.10	0.10	0.10
Control Variables	No	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes
Week Fixed Effects	No	No	No	Yes
Observations	722	722	722	722

Notes: The table shows average marginal effects from probit regressions of whether a team solved the task within 45 minutes on our treatment indicators *Bonus45* (pooled), *Bonus60* (pooled), and *Reference Point* with *Control* being the base category. Control variables, staff and week fixed effects as in Table 1.2. Robust standard errors clustered at the day level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

1.3.4 Robustness of the Bonus Incentive Effect: Results from the Framed Field Experiment

We have shown that bonus incentives increase performance in our non-routine team task in a sample of self-selected and motivated teams of ETR customers. To test whether the performance enhancing effect of bonus incentives in non-routine analytical team tasks is also present in demographics other than the self-selected ETR customer sample, we repeated our main treatments in a student sample. Student participants may react differently to bonus incentives than the teams from our natural field experiment for several reasons. Most importantly, the process by which the sample is drawn is different across the two experiments. While regular teams of *ExitTheRoom* customers self-select into the task and are likely to be intrinsically motivated to perform well (as they pay for it), student teams from the laboratory subject pool are confronted by us with the task, do not pay for it, and hence are less likely to be intrinsically motivated to solve the task. Teams in the field experiment are also formed endogenously and vary in size, whereas we randomly assign students to teams of three participants. Finally, our student participants differ along several observable dimensions, such as age, gender and experience with the task.[20]

[20] The students are on average younger (23.03), slightly less likely to be male (44 percent) and less experienced in escape games (36 percent of the student teams had at least one member with experience in escape games).



Notes: The figure shows the cumulative distributions of finishing times. The vertical line at 45 minutes marks the time limit for the bonus.

Figure 1.6: Finishing times in *Bonus45* and *Control* in the framed field experiment (student sample)

In all, we randomized 268 teams of three students into the treatments *Control* (88), *Gain45* (90) and *Loss45* (90). Despite the assignment to the treatment being random and balanced across weeks, the average share of males in teams is lower in *Gain45* (0.39) than in *Control* (0.46) (Mann–Whitney test, *Gain45* vs. *Control*, p-value $= 0.08$) or *Loss45* (0.47) (Mann–Whitney test, *Loss45* vs. *Gain45* p-value $= 0.10$, *Loss45* vs. *Control*, p-value $= 0.97$), and the share of teams with at least one team member with experience in escape games is higher in *Loss45* (0.42) than in *Gain45* (0.29) (χ^2-test, p-value $= 0.06$). Age does not significantly differ by treatment (Mann–Whitney test, *Gain45* vs. *Control* p-value $= 0.47$, *Loss45* vs. *Control*, p-value $= 0.92$ and *Loss45* vs. *Control*, p-value $= 0.38$). Although the differences between treatments are not very pronounced, we will nevertheless control for these differences in our regression analyses.

Analogously to the analysis in the customer sample, we study treatment effects on team performance by analyzing the fraction of the teams solving the task in 45 and 60 minutes, respectively, as well as the remaining times of teams in general and among successful teams. Figure 1.6 shows the performance of teams in the framed field experiment and is the student sample analogue to Figure 1.3. While student teams perform on average worse than the ETR customer teams, the bonus incentives turn out to be similarly effective for the student teams.

Again, the fraction of teams finishing within 45 minutes is more than twice as high when teams face bonus incentives. In the incentive treatments, 11 percent

Table 1.6: Task performance for main treatments (student sample)

	Control	*Bonus45*	*Gain45*	*Loss45*
Fraction of teams solving task in 45 min	0.05	0.11*	0.13**	0.09
Fraction of teams solving task in 60 min	0.48	0.60*	0.54	0.66**
Mean remaining time (in sec)	169.90	327.97***	321.28*	334.67***
Mean remaining time (in sec) if solved	355.98	546.62***	590.10**	510.50***

Notes: This table summarizes key variables and their differences across our three treatments *Control*, *Gain45* and *Loss45*, as well as the combined *Bonus45* (pooled) for the student sample. Stars indicate significant differences from *Control* (using χ^2 test for frequencies and Mann–Whitney tests for distributions), and *** p<0.01, ** p<0.05, * p<0.1. P-values of non-parametric comparisons between *Gain45* and *Loss45* exceed 0.10 for all four performance measures.

of teams manage to solve the task within 45 minutes whereas only 5 percent do so in *Control* (χ^2-test, p-value $= 0.08$). The fraction of teams finishing the task within 60 minutes is also significantly larger under bonus incentives. With bonuses, 60 percent of the teams finish the task before the 60 minutes expire whereas in *Control* this fraction amounts to 48 percent (χ^2-test, p-value $= 0.06$). Further, with bonus incentives teams are on average about three minutes faster than in *Control*, and Mann–Whitney tests reject that finishing times in the control condition come from the same underlying distribution as finishing times under bonus incentives (Mann–Whitney test, p-values < 0.01). Table 1.6 summarizes these findings.

In addition to the non-parametric tests, we run regressions analogously to the analyses for the customer sample. As before, we control for the share of males in a team, average age and experience with escape games.[21] Table 1.7 reports the results from probit regressions on the probability of solving the task within 45 minutes. Column (1) only uses the treatment dummy and shows that bonus incentives significantly increase the probability of solving the task in 45 minutes. The positive effect of the bonus incentives is robust to controlling for background characteristics (Column (2)), for staff fixed effects (Column (3)), and week fixed effects (Column (4)). Overall, the probit regression results reinforce our non-parametric findings. Offering bonuses increases team performance. Running a regression separately for gain and loss frames yields qualitatively very similar results (Column (5)), as the coefficients for *Loss45* and *Gain45* are again both positive. However, only the coefficient for the gain frame turns out to be statistically significant. A post-estimation Wald test cannot reject equivalence for the coefficients of *Gain45* and *Loss45* at the ten percent level. Also for the student sample, the positive effect of bonus incentives is reflected qualitatively in the analyses of the time remaining (see Table A.6 in Appendix A.5).

[21] In contrast to the ETR customer sample all teams speak German and consist of three team members. Hence, we do not need to control for language or group size.

Table 1.7: Probit regressions: Solved in less than 45 minutes (student sample)

	Probit (ME): Solved in less than 45 minutes				
	(1)	(2)	(3)	(4)	(5)
Bonus45	0.075*	0.073*	0.075*	0.079**	
	(0.042)	(0.041)	(0.039)	(0.036)	
Gain45					0.101**
					(0.039)
Loss45					0.051
					(0.041)
Fraction of control teams solving the task in less than 45 min	0.05	0.05	0.05	0.05	0.05
Control Variables	No	Yes	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes	Yes
Week Fixed Effects	No	No	No	Yes	Yes
Observations	268	268	268	268	268

Notes: The table shows average marginal effects from probit regressions of whether a team solved the task within 45 minutes on our treatment indicators (with *Control* as base category). Control variables added from column (2) onwards include share of males in a team, a dummy whether someone in the team has been to an escape game before and average age of the team. Staff fixed effects control for the employees of ETR present on-site and week fixed effects control for week of data collection. All models include the full sample, including weeks that perfectly predict failure to receive the bonus (Table A.7 in Section A.5 of the Appendix reports regressions from a sample excluding weeks without variation in the outcome variable). Robust standard errors clustered at the session level reported in parentheses, and *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

1.3.5 Performance and Team Organization

In addition to establishing the robustness of the positive incentive effect, our student sample allows us to explore whether bonus incentives also affect team motivation and organization. We conducted two post-experimental questionnaires to analyze potential mechanisms through which the treatment effect could operate. In Questionnaire 1, we asked our student participants to agree or disagree (on a seven-point Likert scale) with a number of statements that might capture aspects of team motivation and organization. In Questionnaire 2 (which was conducted for a subsample of 375 participants), we use an additional set of questions based on the concept of team work quality by Hoegl and Gemuenden (2001). Table 1.8 reports the results from Questionnaires 1 and 2.

The upper panel of Table 1.8 shows that incentives in general do not strongly affect agreement with the statements we provided. However, it reveals some interesting insights about the channels through which incentives might potentially operate. First, teams appear to be notably more stressed when facing incentives than teams in *Control* (Mann–Whitney test, p-value $=$ 0.01). At the same time, similar to teams in *Control*, treated teams strongly agree with the statement "I would like to participate in a similar task again" (Mann–Whitney test, p-value $=$ 0.88),

Table 1.8: Answers to post-experiment questionnaires

	Control	*Bonus45*	*p*-value
Questionnaire 1 (n=804)			
"The team was very stressed."	3.57	4.13***	0.00
"One person was dominant in leading the team."	2.60	2.86**	0.03
"We wrote down all numbers we found."	5.64	5.50**	0.04
"I was dominant in leading the team."	2.64	2.87**	0.05
"We first searched for clues before combining them."	4.58	4.39	0.11
"We exchanged many ideas in the team."	5.87	5.74	0.12
"When we got stuck we let as many team members try as possible."	5.43	5.28	0.14
"The team was very motivated."	6.14	6.26	0.22
"We communicated a lot."	5.78	5.88	0.23
"All team members exerted effort."	6.23	6.37	0.24
"Our notes were helpful in finding the solution."	5.50	5.43	0.41
"I was able to present all my ideas to the group."	5.95	5.93	0.41
"We were well coordinated in the group."	5.73	5.80	0.61
"I was too concentrated on my own part."	2.88	2.83	0.76
"We made our decisions collectively."	5.51	5.58	0.87
"I would like to perform a similar task again."	6.30	6.28	0.88
"Our individual skills complemented well."	5.65	5.68	0.89
"The mood in our team was good."	6.30	6.36	0.93
"All team members contributed equally."	5.97	6.00	0.96
Questionnaire 2 (n=375)			
"How much did you wish somebody would take the lead?"	2.67	3.32***	0.00
"How well led was the team?"	3.85	4.21**	0.04
"How much did you think about the problems?"	6.00	5.79	0.11
"How much did you follow ideas that were not promising?"	5.02	4.79	0.17
"How much team spirit evolved?"	5.54	5.80	0.17
"How much coordination was there of individual tasks and joint strategy?"	3.28	3.51	0.18
"How much exploitation was there of individual potential?"	5.14	4.94	0.22
"How much helping was there when somebody stuck?"	5.70	5.58	0.22
"How much did you search the room for solutions?"	6.31	6.22	0.51
"How much exertion of effort was there by all the members?"	5.98	5.96	0.60
"How much communication was there about procedures?"	5.30	5.35	0.88
"How much was there of accepting the help of others?"	5.80	5.85	0.89

Notes: This table reports answers to our post-experiment questionnaires from the framed field experiment by treatment (*Control* and *Bonus45*), and p-values of the differences between the treatments. The scale ranges from not at all agreeing to the statement (=1) to completely agreeing (=7) in Questionnaire 1 and from very little (=1) to very much (=7) in Questionnaire 2. Stars indicate significant differences from *Control* using Mann-Whitney tests, and *** p<0.01, ** p<0.05, * p<0.1.

suggesting that incentives caused positive rather than negative stress among the team members. Second, participants in the incentive treatment are more likely to report that one team member was dominant in leading the team (Mann–Whitney test, p-value $= 0.03$), and also agree significantly more with the statement "I was dominant in leading the team" (Mann–Whitney test, p-value $= 0.05$). Additionally, we observe several differences in items relating to a more focused and directed approach within a team (although some of them fail to be statistically significant at the 10 percent level). With bonus incentives, participants tend to agree less with the statements "We wrote down all numbers we found." (Mann–Whitney test, p-value $= 0.04$), "We exchanged many ideas in the team." (Mann–Whitney test, p-value $= 0.12$) and "When we got stuck we let as many team members try as possible." (Mann–Whitney test, p-value $= 0.14$).

The results from Questionnaire 2 in the lower panel of Table 1.8 mirror the answers from Questionnaire 1. Teams facing incentives report more demand for leadership (Mann–Whitney test, p-value < 0.01), while they also report that teams were better led (Mann–Whitney test, p-value$= 0.04$). Further, also in Questionnaire 2 we observe several tendencies suggesting a potentially more focused and directed approach within the teams under incentives. Teams tend to be less likely to spend a long time thinking about problems (Mann–Whitney test, p-value $= 0.11$) and tend to follow ideas that were not promising less frequently (Mann–Whitney test, p-value $= 0.17$). Also, teams facing bonus incentives tend to be more likely to report the emergence of team spirit (Mann–Whitney test, p-value $= 0.17$) and the coordination of individual tasks and joint strategy (Mann–Whitney test, p-value $= 0.18$). Although these statistically insignificant results can serve as suggestive evidence only, we nonetheless believe that they highlight a potentially relevant channel through which bonus incentives for teams may increase performance. With an incentive, teams demand more leadership, individual team members are more likely to take the initiative and teams become more focused and better coordinated.

1.3.6 Bonus Incentives and the Willingness to Explore

The effectiveness of bonus incentives in the long run depends on whether monetary incentives crowd out intrinsic motivation, thereby inhibiting creativity and innovation. In fact, previous research has suggested that performance-based financial incentives may do just that, and thereby affect workers' willingness to explore in an experimentation task (see, e.g., Ederer and Manso, 2013). Our setup allows us to shed light on whether such behavioral reactions are also present in the context of non-routine analytical team tasks. We interpret the request for external help (hint taking) as a proxy for a team's unwillingness to explore on their own, and thus analyze how many out of the five possible hints teams request under the different treatment conditions, as well as whether they are more likely to take hints earlier in the presence of incentives.

Table 1.9: Hints requested in the field experiment and the framed field experiment

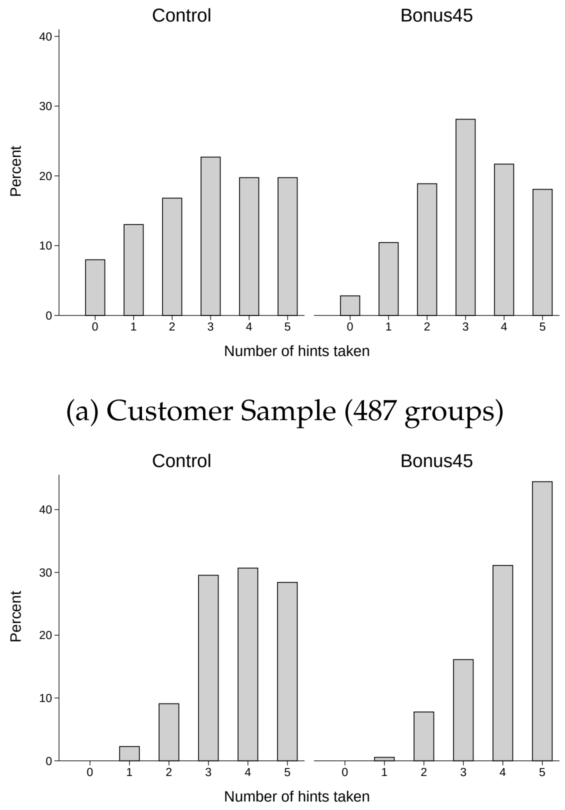
	Control	*Bonus45*	*Gain45*	*Loss45*
within 60 minutes				
Field Experiment (487 groups)	2.92(1.55)	3.10(1.34)	3.05(1.40)	3.15(1.29)
Framed Field Experiment (268 groups)	3.74(1.04)	4.11(0.98)***	4.10(0.98)**	4.12(0.98)**
within 45 minutes				
Field Experiment (487 groups)	1.97(1.22)	2.36(1.15)***	2.30(1.19)**	2.41(1.10)***
Framed Field Experiment (268 groups)	2.33(0.93)	3.17(1.04)***	3.07(1.04)***	3.28(1.04)***

Notes: This table summarizes mean number of hints taken across treatments in the field experiment and the framed field experiment (standard deviations in parentheses). Stars indicate significant differences from *Control* (using Mann–Whitney tests), and *** p<0.01, ** p<0.05, * p<0.1. *P*-values of non-parametric comparisons between *Gain45* and *Loss45* are larger than 0.10 for both the field experiment and the framed field experiment.

Table 1.9 shows the number of hints taken across samples and treatments. For teams who self-selected into the task (customer sample), we do not find a statistically significant difference in the number of hints taken within 60 minutes. These teams take on average about three hints in both the bonus treatment and the control condition. In contrast, for teams confronted by us with the task (the student sample), we observe (economically and statistically) significantly more hint taking in the bonus treatments than in *Control*, suggesting that incentives reduce these student teams' willingness to explore original solutions. To capture potential heterogeneity across teams, we report the fractions of teams requesting 0, 1, 2, 3, 4 or 5 hints for the customer sample in panel (a) and for the student sample in panel (b) of Figure 1.7. The figure reinforces our earlier findings: Bonus incentives have, if at all, a minor effect on the number of hints taken in the customer sample. These teams' willingness to explore original solutions fails to differ statistically significantly across treatments (χ^2-test, p-value=0.114). Panel (b) of Figure 1.7 depicts the same histogram for the framed field experiment with student participants. It becomes apparent that teams who did not self-select into the task are much more likely to take hints when facing incentives (χ^2-test, p-value=0.029). Roughly 75 percent of these teams take four or five hints when facing incentives, as compared to 59 percent doing so in *Control*. Regression analyses for hint taking including additional controls (see Table 1.10, models (1), (2), (5), and (6)) confirm these results.[22]

Focusing only on hints taken within the first 45 minutes, non-parametric tests indicate significant differences across treatments for both samples, but again, the effect is much stronger for student teams who were confronted by us with the non-routine task. Regression analysis implies that these teams take on average 0.84 more hints within the first 45 minutes when facing incentives, whereas customer

[22] An ordered probit regression yields qualitatively similar results, see Table A.8 in the Appendix.



(a) Customer Sample (487 groups)

(b) Student Sample (268 groups)

Notes: The figure shows histograms of hints taken across samples. Panel (a) depicts the fractions of customer teams choosing 0, 1, 2, 3, 4 or 5 hints in *Control* (left graph) and *Bonus45* (right graph). Panel (b) shows the fractions for student teams.

Figure 1.7: Hints requested across samples and treatments

teams take on average only 0.39 more hints (columns (3) and (7) of Table 1.10). When we add additional controls and fixed effects (columns (4) and (8) of Table 1.10), the results for the student sample remain largely unchanged, whereas the positive coefficient of the incentive condition becomes smaller and statistically insignificant in the customer sample.

Taken together our results are in line with the conclusion that intrinsic motivation and incentives interact in an interesting way when teams can choose whether or not to explore original and innovative solutions on their own. Customer teams who themselves chose to perform a task are presumably more intrinsically motivated to work on the task, and thus less likely to seek external help—even when facing performance incentives. In contrast, incentives strongly reduce the willingness to explore original solutions of teams that did not self-select into the task. While we are aware that the two samples differ along several other dimensions (such as exogenous versus endogenous team formation, age or educational

Table 1.10: OLS regressions: Number of hints requested

	OLS: Number of hints requested							
	Field experiment (1)-(4)				Framed Field Experiment (5)-(8)			
	within 60 min		within 45 min		within 60 min		within 45 min	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bonus45	0.172	0.098	0.387**	0.186	0.372**	0.343**	0.843***	0.808***
	(0.167)	(0.183)	(0.152)	(0.134)	(0.145)	(0.136)	(0.128)	(0.122)
Constant	2.924***	4.037***	1.971***	1.770***	3.739***	5.449***	2.330***	4.236***
	(0.130)	(0.442)	(0.109)	(0.469)	(0.126)	(0.650)	(0.102)	(0.698)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Staff FE	No	Yes	No	Yes	No	Yes	No	Yes
Week FE	No	Yes	No	Yes	No	Yes	No	Yes
Observations	487	487	487	487	268	268	268	268

Notes: Coefficients from OLS regressions of the number of hints requested within 60 minutes or 45 minutes regressed on our treatment indicator *Bonus45* (pooled). Controls and fixed effects (FE) identical to previous tables. Robust standard errors clustered at the day (for the field experiment) or session(for the framed field experiment) level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

background), it is less clear to what extent these other differences (as compared to differences in intrinsic motivation) are likely candidates to explain the differential reactions to incentives across samples. We summarize our findings in Result 4.

Result 4 *Bonus incentives reduce student teams' exploration behavior but affect exploration behavior of customer teams (if at all) to a much smaller extent.*

1.4 Discussion

Our results demonstrate that bonus incentives have sizable effects on team performance. Importantly, these effects are present throughout all our incentive treatments, and emerge in both the natural and the framed field experiments. The performance-stimulating effect of incentives therefore seems to be ubiquitous in the non-routine analytical team task in our setting, and not simply driven by a specific choice of subjects or certain treatment parameters. The framing of incentives turns out to be of minor importance. A loss frame did not generally outperform a gain frame but the effect only turns out to be statistically significantly larger when considering whether customer teams finished before 45 minutes (in some specifications). This result is in line with much of the literature, where significant framing effects have been observed in some environments (e.g. Fryer et al., 2012; Hossain and List, 2012; Muralidharan and Sundararaman, 2011), but not in others (DellaVigna and Pope, 2017).

Further, we find that bonus incentives do not lead to strong performance decreases once teams fail to meet the time limit to receive the bonus. Instead, the proportional hazard model analysis suggests that incentives (if anything) increase the likelihood of solving the task within 60 minutes even if teams do not meet the bonus threshold of 45 minutes. Teams facing incentives (for solving the task in 45 minutes) that eventually do not obtain the bonus perform at least as well as teams not facing incentives that do not solve the task in 45 minutes. This is particularly striking as the former are presumably more (adversely) self-selected: The incentive effect presumably boosts some relatively good teams towards finishing just before the 45-minutes threshold, while they would have barely missed the cutoff without incentives.

What is driving the observed performance increase? With respect to hint-taking behavior, we have several reasons to believe that changes in hint taking are not responsible for the observed performance effects. First, an increase in performance will mechanically make subjects request hints earlier, as they reach difficult stages earlier. Second, in our natural field experiment, overall hint taking behavior is not significantly different across treatments. Third, when studying at what point in time teams achieve an intermediate step early in the task and how many hints teams have taken before that step, we observe significantly better performance by teams facing incentives but no significant differences in hint taking (see Appendix A.7 and Table A.9).

An alternative possible explanation for how bonuses improve performance is that incentives enhance learning about the essentials of the production function, i.e. how combinations of different kinds of effort (e.g. searching, deliberating, combining information) map to performance. While we primarily designed our experiment with the goal of causally identifying the effect of bonus incentives, the richness of our data also allows us to shed some light on the importance of learning. We expect teams with prior experience in escape games to have acquired more knowledge on how combinations of different kinds of effort map to performance. Hence, if incentives increase performance due to learning, incentives should in particular increase the performance of inexperienced teams. However, we observe that, if at all, incentives have a stronger effect on performance of teams with prior experience (see model (4) in Table A.2), suggesting that incentives do not increase performance because of this kind of learning. While both hint taking and learning seem unlikely to be responsible for the performance increase, we provide suggestive evidence that teams facing incentives are more likely to wish for a leader and that leaders appear to emerge endogenously when teams face incentives. This renders changes in team organization a more likely explanation for why incentives improve a team's performance.

1.5 Conclusion

According to Autor et al. (2003) and Autor and Price (2013), non-routine, cognitively demanding, interactive tasks are becoming more and more important in the economy. At the same time we know relatively little about how incentives affect performance in these tasks. We provide a comprehensive analysis of incentive effects in a non-routine, cognitively demanding, team task in a large scale field experiment that allows us to study the causal effect of bonus incentives on the performance and exploratory behavior of teams. Together with our collaboration partner, we were able to implement a natural field experiment with more than 700 teams and to replicate our main findings in an additional student sample of more than 250 teams. We find an economically and statistically significant positive effect of incentives on performance. Teams in both samples are more than twice as likely to solve the task in 45 minutes under the incentive condition than under the control condition, and we observe a positive performance effect not only around the bonus threshold, but for a significant part of the distribution of finishing times.

By exploiting a number of additional treatment variations in our natural field experiment, we shed more light on the drivers and moderators of the treatment effect. First, we implement the bonus incentives both in a gain and in a loss frame and find that framing team bonuses as a loss at most has a modest additional effect on performance, and only does so for a subset of our data. Second, we complement the recent literature on how the provision of information about individuals' relative performance affects behavior. When providing teams with a reference point of good performance in an experimental treatment without monetary incentives, teams' finishing times do not improve compared to those in the control condition. Hence, the explicit incentives seem to be key to bringing about the positive treatment effect in our experiment. Third, we find that teams tend to be less likely to explore on their own when facing bonus incentives. However this was mainly true for those teams that were mandated to perform the task. These findings extend earlier work on the (negative) relationship between incentives and the exploration of new approaches (Ederer and Manso, 2013), by highlighting a potential relationship between the consequences of incentives for exploratory behavior and the intrinsic motivation to solve a task. The fact that incentives do not always crowd out intrinsic motivation also complements recent evidence on incentive effects in meaningful routine tasks

(Kosfeld et al., 2017). Finally, answers to our ex-post survey tentatively suggest that incentives may lead to the emergence of leadership within teams in non-routine team tasks and may result in more focused approaches to work.

Our study constitutes, to the best of our knowledge, the first systematic investigation into incentive effects in non-routine analytical team tasks. The results raise interesting questions for future research. For instance, it may be promising to study explicitly how team performance in non-routine tasks changes when leadership is exogenously assigned as compared to endogenously determined. As our findings only provide an initial glimpse at the incentive effects in these kinds of tasks, systematically varying incentive structures within teams could create additional insights into the functioning of non-routine team work. Looking beyond the question of incentives, the setting of a real-life escape game may be used to study other important questions such as goal setting, non-monetary rewards and recognition, the effects of team composition, team organization, and team motivation. Studies in this setting are in principle easily replicable, many treatment variations are implementable, and large sample sizes are feasible.

Chapter 2

Sharing or Gambling? On Risk Attitudes in Social Contexts[*]

2.1 Introduction

Many – if not most – economic decisions take place in a social context. People observe other people's choices, they are themselves observed when making decisions, they affect other people through their decisions, and they reflect on other people's situations when making decisions. This is also true for decisions under uncertainty. It is difficult to come up with examples of decisions under risk taken in situations that are totally free of a social context. Risky choices by managers have consequences for other organizational members; financial decisions within the family have an impact on all family members and will be influenced by similar decisions of peers; even at the roulette table or when playing lotteries social influences are very often present.

Models of decision under risk usually abstract from the social environment in which decisions are made. The typical situation studied by decision theory is one where the individual makes a choice with neither any influence on others nor any information about others' situations, choices, or outcomes. However, it may very well be that decisions under risk in social environments differ from the equivalent decisions taken in purely individual contexts. If that is the case, standard models would lack consideration for the social drivers of such risky decisions and ultimately lead to inaccurate behavioral predictions in many economic circumstances. At least two phenomena suggest an important role of the social context in risky decisions. First, broadly speaking, it has been shown that

[*] This chapter is based on joint work with Martin Kocher, Michał Krawczyk and Fabrice Le Lec.

preferences depend heavily on theoretically 'irrelevant' aspects of the environment or context (Tversky and Simonson, 1993). Second, there is ample evidence that individuals are sensitive in many ways to others' situations (Fehr and Schmidt, 1999; Frank, 2005).

Despite its potential relevance, the effect of the social context in risk taking has only recently received attention in the empirical/experimental literature in economics. Following the burgeoning of studies on other-regarding preferences that focused on deterministic outcomes, empirical research has started to explore the issue of the interaction of risk and social concerns.[1] Some of the relevant studies explicitly focus on peer effects in decision making under risk (Bursztyn et al., 2014; Cai et al., 2015; Cooper and Rege, 2011; Lahno and Serra-Garcia, 2015), while others have primarily looked at decision making about risk borne by others (Chakravarty et al., 2011; Vieider et al., 2016).

Most relevant for this chapter is the literature on risk taking with payoff implications for oneself *and* another person. Brennan et al. (2008) point towards only a small effect of another person's risk per se on own risky decision making. Bolton et al. (2015), on the other hand, indicate that individuals might become more risk averse when also being responsible for other people. Adam et al. (2014) show a decrease in risk taking if outcomes of lotteries of coupled participants in laboratory experiments are asymmetric, i.e. one player wins and the other one loses, compared to independent lotteries. Friedl et al. (2014) observe lower insurance take-up when risks are positively correlated (albeit this was not replicated by Krawczyk et al., 2017). Krawczyk and Le Lec (2010) report lowest giving in a dictator game with probabilistic, negatively correlated payoffs. Overall, although there are no unambiguous conclusions, the existing literature shows that individual differences in payoffs from lottery choices, and hence social comparisons, can often play a role in decision making under uncertainty.

However, very few empirical papers have explicitly focused on *the effects of social comparison on elicited risk attitudes*. The existing evidence, again, is not fully conclusive. Linde and Sonnemans (2012) find that decision makers are more risk averse when in a socially unfavorable situation (that is, when they are disadvantaged compared to another person that serves as a natural reference point) than in a socially favorable one. In contrast, Bault et al. (2008), Bolton and Ockenfels (2010), and Fafchamps et al. (2015) observe that decision makers are less risk averse when the situation is unfavorable. Dijk et al. (2014) find that investors on experimental asset markets performing below average favor positively skewed

[1] An early survey is provided in Trautmann and Vieider (2012). Another approach is to study the correlation of risk and social preferences on the individual level (e.g. Müller and Rau, 2016).

portfolios (those that have a small chance for very high returns), while those performing above average prefer negatively skewed portfolios. These effects occur independently of whether others' outcomes are payoff-relevant (tournament-based incentives) or not.

In this chapter, we want to shed more light on risky decision making within a strong social comparison context. Does risk taking depend on whether somebody else's payoff is affected and on the relative position towards that other person? And how can risk taking patterns be defined, depending on how unequal the initial positions are? Our specific setting that we will look at is resource allocation. Consider a decision maker who can either implement a certain allocation of the resource between herself and a second individual (the 'receiver') or use a random device to allocate the entire resource to either herself or to the receiver. More specifically, the choice is between splitting the resource (dividing the pie into shares of x% for the decision maker and 100-x% for the receiver) or using a random draw to allocate it in one piece (whereby the chances to get the entire pie are again x% and 100-x%, respectively). In our experimental protocol, x is varied across different decision tasks, allowing us to test changes in risk taking related to the relative social situation of the decision maker. Such a setup reproduces, in a simplified manner, important aspects of many situations that involve risk. A decision maker can either go for a given allocation (of financial resources, power, or positions) or gamble for the entire pie. For instance, a manager can accept the proposed split of available funding between her and another manager's project or argue that the company should focus on just one of them; a political leader may have a choice between accommodating the current division of power between herself and a party rival or go for a shootout that will leave just one of the two standing; a poker player in a cash game can leave the table with current possessions or continue playing until all is lost or won. In short, we capture a situation of competition for a resource, and by systematically varying the given x, we are able to analyze how risk attitudes are affected by the initial division of claims on the resource.

The different ranges for x correspond directly to the social standing of an individual. Socially favorable or advantageous situations are those where x is greater than 50. For instance with x = 70, the decision maker has to choose between (OPTION A) a deterministic division giving 70% of the resource to herself and 30% of the resource to the receiver, and (OPTION B) the gamble involving a 70% chance of receiving the entire resource for herself and the remaining 30% chance of losing the entire resource to the receiver. Symmetrically, unfavorable or disadvantageous situations are those in which x is smaller than 50.

The protocol is built in a way such that not only a risk-averse decision maker would choose the deterministic option for any x, but also that social preferences will either be neutral or reinforce this tendency. We consider two types of social considerations. First, for ex ante comparisons (that is, having social consideration for the allocations of expected payoffs as in Saito, 2013; Trautmann, 2009; Trautmann and Wakker, 2010), the two options are equal in terms of expected payoffs. Hence, if people have other-regarding preferences over expected outcomes, they should not play a role in the decision between the deterministic and the risky alternative. The second type of social consideration is to focus on ex post situations (the allocation of payoffs between individuals). In this case, the effect depends on the type of social consideration, but all of those discussed in the literature either have no role, or reinforce the preference for the deterministic option over the risky option. Indeed, the gamble option always creates the maximal inequality (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999), indicating that inequity aversion would rather lead to a preference for the deterministic option. A similar argument applies to a potential maximin motivation (Charness and Rabin, 2002): Choosing the risky option always worsens the situation of the less advantaged individual. Lastly, efficiency concerns do not play any role in the decisions because total payoff is fixed across the two options. Hence, after consideration of these different types of preferences we would not expect people to choose the gamble.

Moreover, if the social context, in the sense of the possibility of social comparisons, has no influence on risk taking, then a typical decision maker should choose the same option (deterministic or risky) when faced with the social lottery decision described above and when faced with the equivalent decision situation devoid of any opportunity for social comparisons. Typically, an individual that prefers splitting x% of the resource to herself and 100-x% of the resource to the receiver rather than gambling for it with the same odds proportion would be expected to choose x% of the resource for sure rather than x% chance of winning the resource in a purely individual context. Any systematic change of choice in our main task (the social lotteries) and the individual control lottery can then be attributed to a change of elicited risk attitudes when social comparisons are possible.

Our main finding is that the fraction of risky choices is strongly affected by the social context. Subjects seem to be more risk-seeking when the deterministic option involves unfavorable inequity in comparison to the same task in isolation. In contrast, a favorable social context (when the deterministic option corresponds to favorable inequity) does not increase the willingness to take risks. The analysis of individuals' behaviors suggests that most of this asymmetry is driven by about two thirds of the subjects who very strongly exhibit this pattern of choices. This pattern is robust to various controls and sensitivity checks, and the data suggests that a competitive element is at play in the participants' choices. Two specific explanations are compatible with our data. Either the other participant's payoff plays the role of a (social) reference point, below which the decision maker is risk-seeking (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992), or individuals are attentive to social ranking in a way that being ahead ('gloating') is more intensively sought than standing behind is avoided.

The remainder of the chapter is organized as follows. The next section presents the experimental design and procedures, Section 2.3 shows our results, and Section 2.4 discusses the results in light of the existing literature and existing theories of decision making under risk. Section 2.5 concludes the chapter.

2.2 Experimental Design and Procedures

The experiment consisted of five short parts: a series of risky choices in a social context, two dictator game decisions, a series of tasks to elicit individual risk preferences and potential loss aversion (in decisions without a social context), a series of risky choices without a social context (but different than in the part before), and the so-called ring test (the incentivized social value orientation questionnaire) to measure social value orientation.

2.2.1 Decision Making under Risk in Varying Decision Contexts

We use a within-subject design that allows us to compare decision making under risk in a social context with decision making under risk in a purely individual context. Since the decisions are identical with respect to the decision maker's payoffs and probabilities, differences in decision making between the individual and social context can be attributed to the context in which the decisions took place.

In **part 1** of the experiment, subjects faced tasks where €10 had to be allocated between the decision maker and an anonymous receiver, with both being present in the laboratory. Two options were available. The first one (OPTION A) was the deterministic (safe) option, which was the plain division of €10, i.e. the allocation (x, 10-x) for a given x. The second one was the risky option (OPTION B), which was the social lottery where the decision maker had a chance of $x/10$ of obtaining the entire €10 and the receiver obtaining €0, and the receiver had a chance of (10-x)/10

Table 2.1: Part 1 — Social context tasks

Task	Safe option (in €)	Risky option (in chances of winning €10)
T1	1 for chooser, 9 for receiver	10% for chooser, 90% for receiver
T2	2 for chooser, 8 for receiver	20% for chooser, 80% for receiver
T3	3 for chooser, 7 for receiver	30% for chooser, 70% for receiver
T4	4 for chooser, 6 for receiver	40% for chooser, 60% for receiver
T5	5 for chooser, 5 for receiver	50% for chooser, 50% for receiver
T6	6 for chooser, 4 for receiver	60% for chooser, 40% for receiver
T7	7 for chooser, 3 for receiver	70% for chooser, 30% for receiver
T8	8 for chooser, 2 for receiver	80% for chooser, 20% for receiver
T9	9 for chooser, 1 for receiver	90% for chooser, 10% for receiver

of getting the €10 and the decision maker getting €0. The chances of winning €10 were mutually exclusive between the decision maker and the receiver. The amount x was systematically varied to obtain nine different tasks, with x ranging from 1 to 9 in steps of 1. Table 2.1 displays all tasks subjects faced in part 1. Participants were asked whether they preferred Option A (henceforth also referred to as 'the safe option') or Option B (henceforth also 'the risky option'). They could also indicate indifference (Option C). For that case, they were told that Option A or Option B would be implemented randomly with equal probability, realized through a draw of the computer. Each subject was asked to make one choice in each row of the table.

In **part 4** of the experiment subjects faced a task equivalent to part 1 of the experiment. However, now the choice was individual. That is, they had to decide between a safe payoff of €x and a lottery with probability $x/10$ of receiving €10 and probability (10-x)/10 of receiving nothing. There was no other participant that was affected from the decisions taken in part 4.

By comparing decisions in part 1 and part 4 of the experiment, we can isolate attitudes towards risk in the social context and compare these to risk taking in the individual context. Social context here simply means that another participant's earnings were determined by the choices of the decision maker.

2.2.2 Experimental Controls

The remaining parts of the experiment (parts 2, 3, and 5) aim at measuring social preferences, as well as risk attitudes and potential loss aversion. More precisely, in **part 2** of the experiment, subjects had to play two dictator games (Bolton et al., 1998; Forsythe et al., 1994). The first was a regular dictator game with €10 to be divided

Table 2.2: Part 3 — Risk and loss aversion choices

Task	Option A	Option B
R1	50%: €5, 50%: €4	50%: €9.50, 50%: €0.25
R2	60%: €5, 40%: €4	60%: €9.50, 40%: €0.25
R3	70%: €5, 30%: €4	70%: €9.50, 30%: €0.25
L1	€0 for sure	30%: €-2.50, 70%: €2.50
L2	€0 for sure	40%: €-2.50, 60%: €2.50
L3	€0 for sure	50%: €-2.50, 50%: €2.50

between the decision maker (the dictator) and the receiver. The second allocation decision consisted of dividing chances to win €10 (the 'competitive probabilistic dictator game' of Krawczyk and Le Lec, 2010). For example, the dictator could decide that with a probability of 70% she would win €10 and the other participant would win nothing, and that otherwise (with a 30% probability) the opposite would be implemented. Finally, participants had to indicate which of the two 'games' they preferred. The first game provides us with a control for outcome-based social concerns, while the second game speaks to preferences regarding procedural social concerns. Thus, we have a measure of subjects' concerns for others to potentially identify the role these concerns may have played in part 1 of the experiment.

In **part 3** of the experiment participants received a truncated and adapted Holt and Laury (2002) multiple choice list to estimate subjects' risk attitudes with stakes comparable to the ones used in the main part of our experiment. This three-question version of the standard choice list contains the choices in which the vast majorities of experimental subjects usually switch from safe to risky lotteries. We also included three decisions that aim at measuring potential loss aversion. Table 2.2 lists all choices in part 3 of the experiment.

Part 5 elicited the subjects' social value orientation with the so-called ring test (Brosig, 2002; Offerman et al., 1996; Van Dijk et al., 2002; Van Lange et al., 1997). In this fully incentivized test, subjects have to make binary choices in 24 different allocation tasks (see Appendix B.1 for details). In each task, a subject has to choose among two allocations that give money to herself and another (anonymous) recipient. The recipient stays the same in all 24 allocation tasks, and all 24 tasks are paid. Adding up the 24 decisions yields a total sum of money allocated to oneself (x-amount) and to the recipient (y-amount). Using the ratio (x/y) one can assign a subject to one of eight categories of social orientation (individualism, altruism, cooperation, competition, martyrdom, masochism, sadomasochism, and aggression).

2.2.3 Experimental Procedures

The experiment was conducted at the Munich Experimental Laboratory for Economic and Social Sciences (MELESSA) in summer 2015. The experiment was programmed and conducted using z-tree (Fischbacher, 2007), and recruitment of participants was done with "ORSEE" (Greiner, 2015). We ran six sessions with a total of 144 subjects; mainly students from the University of Munich. Subjects were allowed to participate in only a single session.

All subjects were asked to take all the choices described above. Their role in parts 1 and 2 of the experiment – either decision maker or receiver – was determined after the experiment, using the strategy method (Brandts and Charness, 2011; Selten, 1967). Resolution of uncertainty was implemented and outcome information was given only at the very end of the entire experiment. Instructions for parts 1 to 4 were distributed and read aloud at the beginning of the experiment. Upon finishing part 4, instructions for part 5 were read and distributed. Subjects knew that there were exactly five parts from the beginning of the experiment.

To determine payoffs, subjects were randomly matched with another participant at the end of the experiment. Always one subject in these matched pairs was randomly selected for the role of the decision maker; the other was the receiver. For all pairs of participants separately, a random mechanism decided the payoff-relevant part from parts 1 to 4. If part 1 or 2 was chosen, another random mechanism decided which specific task within the part was to be implemented for both participants. If part 3 or 4 was chosen, the specific task to be implemented was determined for both participants separately. In addition to the payoff from this single decision out of parts 1 to 4, all subjects received their earnings from the ring test in part 5. This consisted of their payoff from their own choices and the payoff from the choices of the matched participant. Matching of participants in part 5 of the experiment was independent of the matching in parts 1 and 2. On top of these earnings, participants received a fixed payment of €4 for showing up on time. On average, participants earned €13.40, and a session took about 50 minutes.

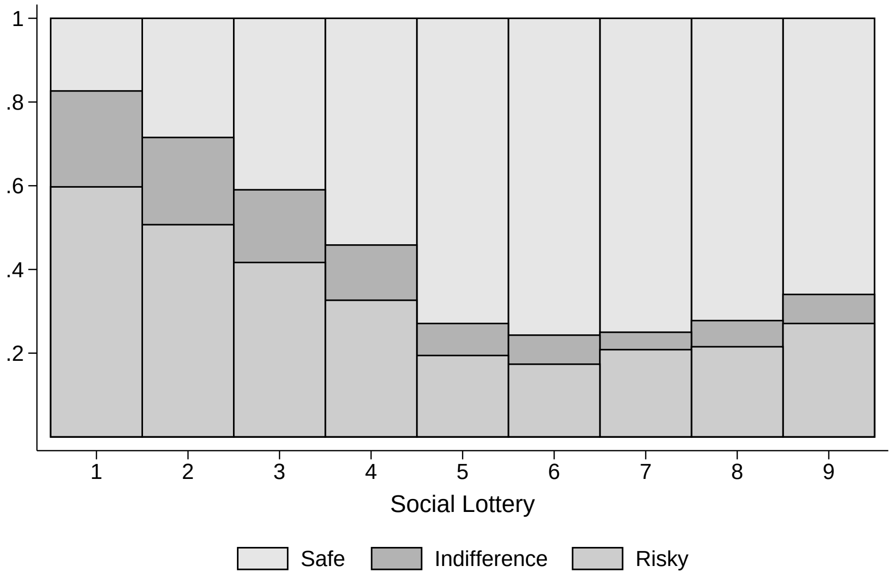
Participants were also asked to fill out a questionnaire after part 5 including a short description of their motivation underlying decisions in the experiment and questions regarding sociodemographic characteristics. All design details and the procedural details were common knowledge among participants (see the instructions for all parts in Appendix B.2).

2.3 Experimental Results

We will first have a look at aggregate results (Section 2.3.1), before taking into account the heterogeneity in responses (Section 2.3.2). Section 2.3.3 reports the results of a pilot experiment that provides further support for the robustness of our results.

2.3.1 Aggregate Results

An overview of the results from decision making under risk in the social context (part 1 of the experiment) is shown in Figure 2.1. The aggregate pattern of risk taking is roughly L-shaped, with subjects willing to take considerably more risk in unfavorable tasks where the expected payoff to the decision maker is smaller than the expected payoff to the matched participant. The level of risk taking reaches its lowest value just above the equal split.



Notes: The y-axis denotes the fraction of subjects that choose a certain option for a given lottery. The x-axis represents the different types of lotteries from T1 to T9 with unfavorable lotteries to the left (T1 to T4) and favorable one to the right (T6 to T9).

Figure 2.1: Distribution of choices in the social context

The asymmetry between favorable and unfavorable situations is statistically significant. Leaving the case of the equal split aside for the moment, all comparisons between tasks corresponding to sure payoffs adding up to 10 (T1 vs. T9, T2 vs. T8, T3 vs. T7, and T4 vs. T6) suggest that the risky option is relatively more appealing when the safe option implies unfavorable inequity. These differences are significant according to a Stuart-Maxwell test at the 1%-level.[2] If we pool indifference with the risky option or with the safe option, McNemar's tests remain significant at the 1%-level for either pooling version and for all comparisons.[3] Looking at the unfavorable situations only, statistical tests support increasing risk taking from the equal split towards the more unfavorable tasks. For all binary comparisons between the tasks in the unfavorable domain, choices move strongly towards more risk taking, the more unfavorable and risky the tasks become ($p < 0.01$ for Stuart-Maxwell tests for all comparisons).

This pattern of choice is not necessarily in itself indicative of a change in behavior in the favorable and unfavorable social domains compared to the individual context. It is overall equally compatible with an inverted-S transformation of probabilities as in cumulative prospect theory (Tversky and Kahneman, 1992). As is well established (Wakker, 2010, for instance), low probabilities of the good outcome are typically overweighed from 0 to roughly one third, while intermediate and large probabilities of the good outcome are usually underweighed. Hence, the asymmetry of choices could result from the typically observed probability transformation.

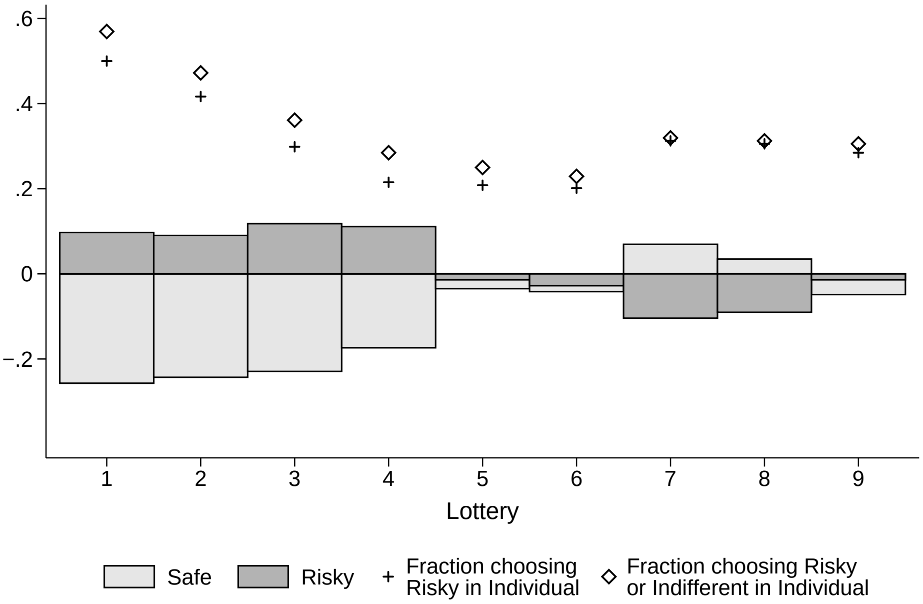
To test whether individuals' choices are actually driven by the social context of the decision, we compare the social tasks from part 1 with choices from part 4 of the experiment. The nine tasks in part 4 (henceforth T1i to T9i, where i stands for "individual") were the exact counterparts of T1 to T9 from part 1 in terms of payoffs and probabilities for the decision maker, but stripped from the social context, as there was no receiver. If choices are influenced by the social context, we should observe differences in the frequencies of risky choices between the individual task and the social task. Comparisons are displayed in Figure 2.2, indeed suggesting systematic differences between decisions in social and individual contexts.

These differences are large in the unfavorable range. For T1 vs. T1i, T2 vs. T2i, T3 vs. T3i, and T4 vs. T4i, individuals take more risk when facing the social lottery than in the equivalent individual task, and this difference is highly significant ($p < 0.01$ for all four Stuart-Maxwell tests).[4] We observe, for T1-T4, that roughly 20% fewer subjects choose the safe option in the social context and about 10% more choose the risky one. The percentage of changes is relatively constant for all the

[2] The Stuart-Maxwell marginal homogeneity test is applied the same way as the McNemar test for testing marginal homogeneity. It is used for variables with more than two categories ("safe", "indifference", and "risky"). For two categories, the two tests are equivalent.

[3] In the remainder we will only indicate the results of McNemar's tests grouping indifference with either safe or risky choices if they differ from the respective Stuart-Maxwell test.

[4] Here, if we pool indifference and safe option choices, McNemar's tests of marginal homogeneity result in differences that are significant only at the 10%-level for T1 vs. T1i and T2 vs. T2i and that are significant at the 5%-level for T3 vs. T3i and T4 vs. T4i. If we pool indifference with risky choices, all tests are significant at the 1%-level.



Notes: Bars denote the change in the fraction of subjects choosing the the risky (safe) option when going from the individual to the social context. Positive values indicate that a higher fraction of subjects chooses the respective option in the social context (part 1) in the specific lottery. Bars do not add up to zero, since the fraction of indifferent subjects changes simultaneously. The dots correspond to the fraction of subjects choosing risky (plus) and risky OR indifference (rectangular) in the individual lotteries to allow for a direct comparison to Figure 2.1. The area above the rectangular dots (y-axis cut off here at 0.7) consequently refers to the fraction of subjects choosing safe in the individual lotteries.

Figure 2.2: Difference in choices between the individual and social context

unfavorable social situations. The fact that already a large share of subjects chooses the risky option for low probabilities in the individual tasks (T1i-T4i) partly hides the magnitude of the change of choice between the individual and the social context. As an illustration, consider T1: While already only 43% of the subjects choose the safe option in the individual task (T1i), only 17% do so in the social context. Hence, in the social context 60% less subjects choose the safe option. This share of subjects moving away from the safe option in the social context ranges from 24% to 60% from T4(i) to T1(i). At the same time, the percentage increase in subjects choosing the risky option ranges from 19% to 52%. This is a substantial shift in choice patterns in the unfavorable domain.

The pattern is less clear for the favorable range. For T7 vs. T7i and T8 vs. T8i, there is more risk taking in the individual tasks (p = 0.01, Stuart-Maxwell tests). However, this result is not robust to using McNemar's tests and pooling indifference with risky choices, since many subjects simply switch from the risky choice in the individual to indifference in the social context. Other possible comparisons do not yield significant differences.

Overall, we observe that decision makers seem to be affected by the social context when making a risky decision, but not in a symmetric way. They unambiguously take more risk when the situation is unfavorable, but display similar choices when it is favorable to them compared to a risk-equivalent individual context.

To test the robustness of our results, we ran an ordered probit model (column 1 of Table 2.3) on the choices made in all 18 lotteries (with and without social context). We use indicator dummies for the nine types of lotteries (*Type 1* to *Type 9*) without separating the social and individual tasks, and dummies *Social 1*, *Social 2*, etc. for the task being social (T1, ... T9) or not (T1i, T9i). Hence, coefficients on *Type 1* up to *Type 9* correspond to the average risk taking in the individual lottery tasks, while coefficients for *Social 1*, *Social 2*, etc. correspond to the additional risk taking in the social context (relatively to the individual one). The results are displayed in Table 2.3.[5]

The regression results confirm the findings based on non-parametric tests. Decision makers indeed take on average more risk in unfavorable social situations compared to the equivalent individual situations. All terms indicating the social context are positive and significant in the unfavorable domain (at the 1%-level). This is not true for favorable situations. It seems like – if anything – individuals reduce risk taking in the social context for favorable situations compared to situations without such context. These results are robust to using an ordinary probit model. Both when taking an indicator variable for risky choices only (column 2) and when taking an indicator variable for risky or indifferent choices (column 3) as dependent variable, we obtain the same pattern of results.

In sum, we observe some variability in the proportion of risky choices in the individual tasks, possibly related to a non-linear treatment of probabilities, but more relevantly here, we observe a strong effect of social comparisons in the unfavorable domain. In such tasks, participants take much more often the risky option than in the individual task. To the contrary, very little, if any, effect is found in the favorable domain.

[5] Our findings from Table 2.3 and Table 2.5 (see below) are fully robust to using linear (probability) models. This also addresses concerns regarding the interpretation of interaction terms in non-linear models (Ai and Norton, 2003; Greene, 2010). Further, inference based on manually calculated correct marginal effects does not yield different insights compared to the coefficients reported in Table 2.3 and Table 2.5.

Table 2.3: Regression analysis: Risky choices in the social vs. individual context

	Ordered probit (1)	Probit baseline (2)	Probit indifference (3)
Type 1	0.848***	0.812***	0.849***
	(0.146)	(0.150)	(0.146)
Type 2	0.616***	0.602***	0.605***
	(0.144)	(0.149)	(0.142)
Type 3	0.311**	0.284**	0.319**
	(0.136)	(0.142)	(0.134)
Type 4	0.079	0.024	0.106
	(0.117)	(0.120)	(0.118)
Type 5	- ref. -	- ref. -	- ref. -
Type 6	−0.054	−0.024	−0.067
	(0.107)	(0.107)	(0.108)
Type 7	0.257**	0.323**	0.205
	(0.129)	(0.126)	(0.129)
Type 8	0.237*	0.304**	0.186
	(0.139)	(0.136)	(0.138)
Type 9	0.199	0.243*	0.166
	(0.145)	(0.145)	(0.144)
Social 1	0.431***	0.246*	0.765***
	(0.113)	(0.126)	(0.139)
Social 2	0.411***	0.228*	0.639***
	(0.107)	(0.120)	(0.120)
Social 3	0.457***	0.318**	0.584***
	(0.120)	(0.132)	(0.128)
Social 4	0.411***	0.338**	0.464***
	(0.116)	(0.136)	(0.121)
Social 5	0.027	−0.049	0.064
	(0.123)	(0.136)	(0.123)
Social 6	−0.002	−0.103	0.045
	(0.141)	(0.146)	(0.143)
Social 7	−0.257*	−0.323**	−0.205
	(0.136)	(0.133)	(0.136)
Social 8	−0.171	−0.280**	−0.101
	(0.131)	(0.130)	(0.132)
Social 9	0.045	−0.041	0.097
	(0.132)	(0.135)	(0.130)
Constant (Cut1)	0.644***	−0.812***	−0.674***
	(0.115)	(0.118)	(0.114)
Constant Cut2	0.875***		
	(0.117)		
Observations	2,592	2,592	2,592
Pseudo R-sq.	0.058	0.054	0.092

Notes: As dependent variable in the ordered probit regression in column 1 we use an ordinal scale for risk taking (0=safe choice, 1=indifference, 2=risky choice) in the respective lottery. Columns 2 and 3 display results from probit regressions using an indicator variable for the choice being risky (column 2), and risky or indifferent (column 3), respectively. As independent variables we use dummies for all nine types of lotteries in general (Type 1 to Type 9), as well as nine indicator variables for the social context lotteries (Social 1 to Social 9). Hence, coefficients of the first nine dummies refer to risk taking in the individual context, while the latter nine indicate whether for a given lottery type, risk taking is higher or lower in the social context. Standard errors are robust and clustered at the subject level, and *** $p<0.01$, ** $p<0.05$, * $p<0.1$. Constant Cuts in column 1 are threshold parameters of the ordered probit model to differentiate low risk takers from indifferent and risky subjects (Cut1) and low risk takers and indifferent subjects from risky subjects (Cut2). They are estimated such that Pr(Safe)=Pr(Xb+u<Cut1), Pr(Indifferent)=Pr(Cut1<Xb+u<Cut2), and Pr(Risky)=Pr(Cut2<Xb+u). Results are robust to using OLS.

2.3.2 Individual Heterogeneity

We now turn to heterogeneity with three aims in mind. The first one is to check the robustness of our finding when taking into account possible sampling variations in other characteristics (social preferences, risk attitudes, etc.). The second one is that these characteristics can be correlated with the strength of the effect we observe and shed light on its psychological drivers. And the last one is to simply establish the heterogeneity of the sample with respect to the effect of social comparisons on risk taking. For that purpose, we can look at the individual characteristics elicited in parts 2, 3 and 5 of our experiment. Table 2.4 provides an overview of these characteristics.

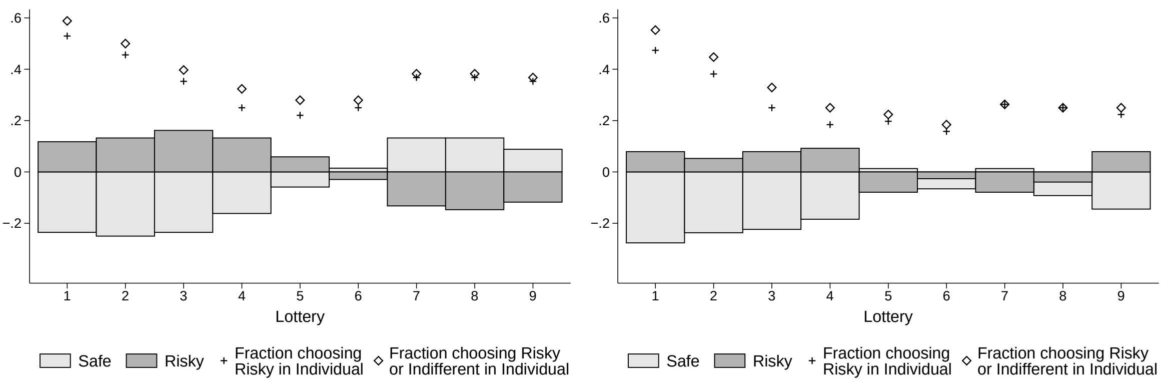
Table 2.4: Summary statistics of parts 2, 3, and 5

Dictator game	mean	median	10th	25th	75th	90th
Standard	1.99	2	0	0	3	5
Probablistic	14.04	10	0	0	27.5	40
Part 3 Lotteries	3 safe	2 safe	1 safe	0 safe		inconsist.
Risk aversion	31.94	43.75	11.11	12.5		1 obs
Loss aversion	22.92	40.28	14.58	19.44		4 obs
Ring test	Competitive	Individualist	Cooperative		(-) angle	(+) angle
Sample fraction	1.39%	71.53%	27.08%		45.14%	54.86%

One aspect in which subjects potentially differ is whether they are socially oriented, i.e. other-regarding (inequity averse, altruistic, etc.). Categorizing selfish and pro-social subjects on the basis of a median split in their offer in the dictator game in part 2 of the experiment provides us with additional insights.[6] Figure 2.3 shows the differences in choices for the two groups (for reasons of elucidation, call them "egoists" and "altruists") when going from the individual to the social context.

For both groups, decision making in the unfavorable range changes strongly from the individual to the social context. Still, the pattern of changes is slightly different: Altruists in the dictator game (right panel) strongly switch from the safe option to mainly indifference (and some risky), while self-interested subjects (left panel) switch more to the risky option and less to indifference. In the favorable range, the difference between the groups becomes even more apparent. Selfish participants switch from risky to safe from the individual to the social context.

[6] Roughly 47% of the dictators give nothing or €1, while 53% give €2 or more.



Notes: Same as Figure 2.2, now only comparing the effects in two subsamples. The left part of the figure refers to subjects with below-median dictator giving (in part 2 of the experiment), while the right part describes risk taking of above-median dictator giving subjects.

Figure 2.3: Choices by dictator giving (left: egoists, right: altruists) in the social vs. individual context

Altruists, however, show a less clear-cut pattern of change. They often switch from safe and risky to indifference for T8i vs. T8 and from safe to risky and indifference for T9i vs. T9. These results remain roughly unchanged if we use generosity in the probabilistic dictator game for the sample split. This suggests that the effect of social comparisons is very widespread in the unfavorable domain while in the favorable one, it may only concern self-interested participants, and to a weaker extent.

To see how these effects depend on other personal characteristics and to check their robustness, we ran ordered probit models similar to the one in Table 2.3, now including interaction terms with the different types of personal characteristics. For that purpose, in contrast to Table 2.3, we now only use three dummies for the different types of lotteries. This limits the number of interaction terms and makes the interpretation of the results more straightforward. *Unfavorable* is a dummy indicating that the lottery has an expected value below five (T1(i) to T4(i)); *Equal Split* indicates the equal split lottery (T5(i)); and *Favorable* stands for lotteries with an expected value for the decision maker larger than 5 (T6(i) to T9(i)). As before, we also include interaction terms for these lottery types with a dummy indicating a social context (*Social*). Column 1 shows results for this baseline specification with fewer dummy indicators than in Table 2.2. In columns 2-4, we then interact the six baseline variables with an indicator variable for below median dictator giving as in Figure 2.3 (column 2 of Table 2.5), for a negative angle in the ring test for social value orientation (column 3), and for low loss aversion (column 4) from part 3 of the experiment. This indicator variable is denoted X. A negative angle in the ring test implies that the decision maker in part 5 of the experiment chose such that the matched participant received a negative payoff from these choices. This is only possible if, at least at one point for the 24 tasks, the decision maker preferred to take money away from the matched participant, with no monetary benefit or possibly even at a cost for him- or herself.[7] However, as argued above, being classified as individualistic with a negative angle already implies some form of competitive preferences. Low loss aversion (column 4) means that subjects at least in all but one of the loss aversion decisions chose the option involving the chance of a loss. This is true for 49 of the subjects. The results are provided in Table 2.5.

The baseline regression results (column 1) again confirm the pattern found for the finer-grained lottery definitions in Table 2.3. Compared to the individual context, average risk taking increases for unfavorable tasks in the social context. If we now look at the regression results including interaction terms, interesting patterns emerge. For altruists – according to our dictator giving measure (upper part of column 2) – the increase in risk taking due to the social context in unfavorable tasks is still positive and significant. The interaction term for unfavorable lotteries in the social context for selfish participants (*Unfavorable* × *Social* × *X*) is small and not significant at conventional levels. This also holds for other specifications of the altruism indicator: None of the differences in the social context between altruistic and selfish participants are statistically significant if we consider positive (non-zero) transfers in the standard or probabilistic dictator game as altruistic or if we define above-median giving in the probabilistic dictator game as altruistic behavior. Overall, pro-sociality as measured by generosity in a dictator game does not seem to be related to the tendency to take more risk in unfavorable social contexts.

The differences are more clear-cut for the split based on ring test choices (column 3 in Table 2.5). For less competitive types in the upper part of the table, as for altruists in column 2, choices in the unfavorable range are affected by context. In this case, however, the more competitive types are clearly more strongly affected by the social context (significant at the 5%-level). Remember that the two measures for social preferences capture potentially different behavioral inclinations. Dictator giving is a proxy for altruism, whereas the ring test puts cooperative individuals against competitive individuals. The latter category cannot be captured by standard dictator giving decisions. It seems as if more competitive individuals show the strongest reaction to unfavorable situations in the social context. This line of reasoning is robust to a sample division into cooperative versus individualistic or competitive individuals, strictly based on the classifications

[7] In our preferred specification, we refrain from using the strict classification into types (individualistic, competitive, cooperative, etc.) described in Appendix B.1, since we only have two subjects classified as purely competitive and a vast majority in the individualistic category.

Table 2.5: Heterogeneity in the effects of the social context

	Ordered probit: Risk choice			
	(1)	(2)	(3)	(4)
Unfavorable	0.474***	0.463**	0.553***	0.649***
	(0.122)	(0.189)	(0.164)	(0.169)
Favorable	0.163	0.062	0.100	0.010
	(0.114)	(0.173)	(0.167)	(0.141)
Unfavorable × Social	0.417***	0.394***	0.255**	0.317***
	(0.084)	(0.117)	(0.119)	(0.105)
Equal Split × Social	0.027	−0.128	−0.033	−0.093
	(0.123)	(0.175)	(0.157)	(0.170)
Favorable × Social	−0.096	0.095	0.003	0.009
	(0.109)	(0.149)	(0.150)	(0.139)
		X: Low dict. giving	X: Non-cooperative	X: Low loss av.
Unfavorable × X		0.172	−0.212	−0.110
		(0.169)	(0.169)	(0.178)
Equal Split × X		0.146	−0.033	0.361
		(0.229)	(0.230)	(0.238)
Favorable × X		0.350*	0.106	0.736***
		(0.189)	(0.190)	(0.192)
Unfavorable × Social × X		0.056	0.366**	0.292*
		(0.168)	(0.163)	(0.171)
Equal Split × Social × X		0.304	0.133	0.292
		(0.247)	(0.250)	(0.254)
Favorable × Social × X		−0.388*	−0.221	−0.251
		(0.219)	(0.218)	(0.236)
Constant Cut1	0.646***	0.717***	0.631***	0.780***
	(0.115)	(0.164)	(0.155)	(0.147)
Constant Cut2	0.871***	0.943***	0.857***	1.011***
	(0.117)	(0.164)	(0.158)	(0.144)
Observations	2,592	2,592	2,592	2,592
of which interaction	-	1,224	1,170	822
Pseudo R-squared	0.041	0.046	0.043	0.058

Notes: The dependent variable is an ordinal scale measure for risk taking (as in Table 2.3). As in column 1 of Table 2.3, the columns report results from an ordered probit regression with robust and clustered standard errors, and *** p<0.01, ** p<0.05, * p<0.1. Constant Cuts are threshold parameters of the ordered probit model to differentiate low risk takers from indifferent and risky subjects (Cut1) and low risk takers and indifferent subjects from risky subjects (Cut2) (also see Table 2.3). Lottery types are now grouped in three blocks: Unfavorable (T1(i) to T4(i)), equal split (T5(i)), and favorable (T6(i) to T9(i)). Columns (2) to (4) use different sample splits for the interaction terms. For each column, results in the upper part of the table refer to risk taking of subjects for which statement X of the corresponding column does not hold. The equal split is omitted for the individual context here. The three interaction terms refer to the effect of the social context on these subjects. The lower part describes whether risk taking by subjects for which X holds is different from that behavior. The first three coefficients refer to the difference in risk taking in the individual context. The last three coefficients describe the difference in risk taking in the social context for the subjects for which X applies compared to what can be expected by the effect of the social context on subjects for which X does not apply and by the effects of X applying in the individual context. A significant effect here indicates a particularly strong effect of the social context on subjects for which X applies. Results are robust to using OLS.

described in Appendix B.1. In this regression, strictly cooperative types do not show an increase in risk taking in the unfavorable social domain. Instead, these subjects even take more risks in the favorable social domain. For the non-cooperative types based on this specification the reverse holds. They significantly take more risk in the unfavorable social domain while they take much less risk on average in the favorable social lotteries.

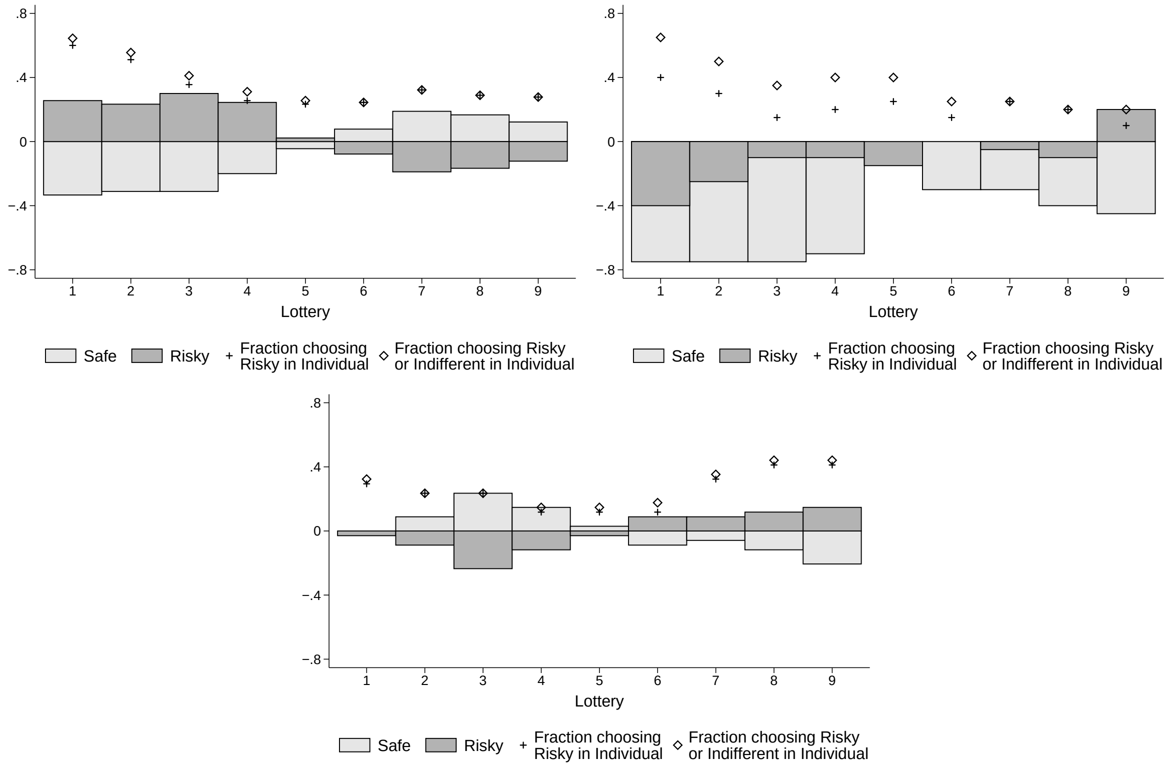
Column 4 looks at interactions with loss attitude. Loss averse subjects, according to our measure, are, as in the overall pattern, affected by the social context in the unfavorable domain, but low loss averse subjects seem, if anything, disproportionally more strongly affected (significant at the 10%-level). This result, however, is not robust to defining low loss aversion as choosing all three lotteries involving losses or as choosing only at least one lottery involving a potential loss. Further, the coefficient on *Unfavorable* × *Social* × *X* is insignificant if we consider indifference choices in the loss aversion tasks as taking the lottery involving the loss.

Finally, we ran a k-medians clustering analysis on all social lotteries dividing the subjects into three clusters.[8] This analysis results in the following characterization: the first and clearly largest class of decision makers is comprised of 90 (out of 144) individuals who exhibit a strongly domain-dependent pattern of risk attitudes in the social context (strongly risk-seeking in the unfavorable case, and risk-averse in the favorable one); the second cluster (20 subjects) is overall risk-averse and very often chooses indifference (especially for the unfavorable range); and the final cluster (34 individuals) shows increasing risk taking for the favorable range as well as for extremely unfavorable tasks.[9] The results from the cluster analysis by types are shown in Figure 2.4.

The categorization of subjects can help explain the aggregate pattern. The overall asymmetry in risky behavior across favorable and unfavorable situations seems to be mostly (but not only) driven by – the most prevalent – type-1 individuals. For these subjects the increase in risk taking in the unfavorable range when going from the individual to the social context is very pronounced, while they seem to reduce risk taking in the favorable range. Furthermore, the small surge in overall risk-seeking behavior when the situation becomes more and more socially favorable can almost entirely be accounted for by type-3 participants. These subjects, however, even reduce risk taking in the unfavorable range when they are in the social context. Type 2 individuals are most effectively characterized by their inclination towards

[8] The k-medians clustering partitions subjects into k groups by finding k centroids that minimize the overall distance between data points and the closest centroid. The use of medians rather than means ("k-means clustering") is more appropriate for discrete data as is the case here. Three clusters are chosen for the reduction in the within (cluster) sum of squares is large until the addition of the third cluster, but gets very small for a larger number of clusters.

[9] The exact clustering always depends on the random starting points for building the clusters. Hence, if we had chosen different starting points, we would have ended up with different clusters. However, varying the starting points leaves the overall conclusions largely unchanged.



Notes: See Figure 2.2 for an explanation of the bars and dots. The upper-left panel represents choices of type-1 individuals, the upper-right panel refers to type-2 subjects, and the lower panel represents type-3 individuals.

Figure 2.4: Choices by types (1-3) in the social vs. individual context

indifference in the social lotteries – especially in the unfavorable domain. Statistical tests confirm these first impressions (see Appendix B.3).

2.3.3 Additional Evidence from a Classroom Experiment

A classroom experiment was conducted prior to the laboratory experiment described in detail above, and it inspired most of the latter's design. In the classroom experiment, the social context tasks were the same as in the present study, even though stakes and payment procedures differed (only a subset of participants was selected for payments and the stake size was €50 instead of €10). Next to the social context tasks, subjects also worked on an equivalent risk and loss aversion elicitation task, as well as on three individual context tasks (as opposed to all nine tasks in the above laboratory study) for comparison to the social tasks.

Due to the design differences, results in the two studies should only be compared with caution. Nonetheless, the conclusions from the classroom experiment and the laboratory experiment are strikingly similar. As in the present study, we also found much stronger risk taking in the unfavorable domain compared to the favorable situations, and risk taking in the unfavorable range increases towards more unfavorable situations. Comparing this behavior to choices in the individual context within the classroom experiment leads to similar conclusions as in the laboratory experiment. Risk taking in the unfavorable decisions is clearly higher in the social context than in the individual one, while it is only weakly higher in the favorable range. The fact that it still is higher in the favorable range points towards the only difference between the two studies: In the earlier classroom experiment aggregate risk taking somewhat increases in the favorable range, too, such that the aggregate picture rather gives a U-shaped pattern of risk taking, whereas the laboratory experiment reveals a more L-shaped pattern. The details of the results from the classroom experiment are provided in Appendix B.4.

2.4 Discussion

Overall, our results suggest that individual attitudes towards risk are strongly affected by the social context. We observe systematic deviations in social situations from what decision makers decide in similar situations that do not allow for social comparisons. In the following, we deepen our discussion outlined in the introduction on potential explanations in the light of different utility functions or decision theories in turn.

A natural contender to explain a change in risk attitude in social situations is the role of (ex post) social preferences. For instance, it seems intuitive that more inequity averse individuals (or more spiteful subjects) would see the safe situation as more unattractive in the unfavorable domain than in the favorable domain and would consequently be willing to gamble rather than to go for the deterministic outcome in this disadvantageous situation. In fact, this intuition is erroneous. Choosing the risky option, for instance in T1, means to potentially end up (with a very high probability of 90% in case of T1) in a situation even worse from the point of view of these ex post preferences. More formally, what determines the attitude towards risk in our social tasks is not the type of social motives (altruism, spitefulness, competitiveness, inequity aversion, etc.) but, leaving aside probability transformations, the curvature of the utility on the linear segment [(0,10), (10,0)]. Our results suggest that a substantial share of subjects have a convex utility function from (0,10) to (5,5) and a concave one from (5,5) to (10,0) (see Appendix B.5 for the formal derivation). Said differently, the type of social ex post motivation that individuals have does not play an important role in determining their choice in our social lottery tasks (or if they do, they should favor the safe option). Likewise, ex ante (or procedural or process) fairness concerns cannot easily help in explaining our results (Trautmann, 2009; Trautmann and Wakker, 2010). Ex ante, both options provide the same expected payoff to both participants. Consequently, procedural inequity aversion preferences should not affect individual choices unless the decision maker has a preference for a stochastic allocation decision over a deterministic one. For instance, one could feel less responsibility for the stochastically implemented uneven distribution than for one that is implemented deterministically. Notice, however, that our subjects had the possibility to choose indifference and let a random mechanism decide. A part of the strong increase in indifference choices that we observe could be related to this, but that makes responsibility avoidance less of a plausible candidate for favoring the risky option.

Overall our data pattern is consistent with two explanations based on (i) a social reference point and (ii) a stronger willingness to being ahead of others compared to being behind. Regarding the first explanation, the other's payoff could play the role of a reference point in prospect theory, an idea developed by Linde and Sonnemans (2012). Gain and loss domains consequently would be defined through the earnings of the other participant, predicting more risk seeking in the loss domain (unfavorable situations) and more risk aversion in the gain domain (favorable situations). When in favorable situations, subjects in our experiment could mainly lose relative to the other participant when choosing the risky option. Instead, by selecting the safe option they can secure their relative social gain. In contrast to that, in unfavorable situations, subjects are not much affected by the prospect of getting (0,10) rather than (1,9) because of diminishing sensitivity (i.e. convex utility) in the (relative) loss domain. Gambling in this case means a large probability of a subjectively small loss but a small probability of a very large gain.

Such reasoning could also help to explain the data by Haisley et al. (2008), who show that, when reminded of their low status, low income individuals were more likely to engage in risky purchases such as buying lottery tickets. It is also the reasoning of Schwerter (2013), indicating that decision makers indeed experience social losses and gains in a risk task when exposed to another participant receiving a varying fixed payment.

The second explanation relates to the strength of gloating, i.e. the utility of being ahead of the other. If gloating is more important than envy, i.e. the disutility of being behind, we should observe the mirror effect in the social situation of what is observed under (individual) risk involving gains and losses because of loss aversion. Loss aversion, for lotteries involving gains and losses, implies a strong avoidance to risk in the individual situation (see Rabin, 2000). However, Bault et al. (2008) argue that attitudes to gains and losses reverse in a social context. Whereas in its standard version, the theory implies that losses are valued more in absolute terms than gains, it may be that the opposite holds in social contexts. That is, relative gains may be subjectively valued more strongly than relative losses. Attitudes to gains and losses reversing in a social context may also in part explain the discrepancy between, on the one hand, our results and those of Bolton and Ockenfels (2010) and, on the other hand, the findings in Linde and Sonnemans (2012). In the latter paper, the authors did not use social lotteries that gave the decision maker the opportunity to switch relative positions with the receiver (from being behind to being ahead), but at best the possibility to reach the same level of payoffs. If being ahead is what is really prized by subjects, there is little motivation in Linde and Sonnemans's tasks to take risks, since it is impossible to earn more than the matched participant by choosing the risky option. An alternative version of this interpretation is that 'winning' – that is, earning more than the counterpart, independently of the absolute payoff difference – generates a psychological bonus: What is prized is not really the favorable difference between the decision maker and the receiver, but simply whether the decision maker has 'won'. In this case, there is no reason any more to take risks in favorable tasks, and such an explanation is consistent with the general pattern we observe.[10]

Our findings concerning individual heterogeneity are in line with arguments in favor of a social reference point and a psychological bonus of winning. Those decision makers that reduce the other's payoff in the ring test exhibit the overall pattern more strongly than those that do not. Reducing the other's payoff can only be rationalized by making some form of relative comparisons with the matched participant and by a wish to earn more in relative terms (apart from pure forms of anti-social behavior). It is not surprising that these people are more strongly affected by the social context. The cluster analysis also helps rationalizing the patterns. Type-1 individuals, who drive the aggregate pattern described above, are not only disproportionally less often categorized as cooperative, they also explicitly state motivations based on a social reference point story. In the subjects' comment section at the end of the experiment, where participants were supposed to elaborate on their motivation behind choices in the social context task, one type-1 subject explained switching to the risky option in the unfavorable cases by stating that "as long as I earned more than the other, I chose the certain amount". Another explicitly wanted to "get a higher payoff than the other". These statements are a specific characteristic of type-1 individuals.

[10] Such reasoning can be seen as a social version of aspiration level theory, developed by Diecidue and Van De Ven (2008).

In contrast to the large group of type-1 individuals, there seems to be something else driving behavior of type-2 and type-3 decision makers. Responsibility aversion is one potential explanation. Type-2 individuals are characterized by a switch towards indifference in unfavorable tasks in the social context – and slightly less pronounced in the favorable context. This might in part be driven by responsibility avoidance, which could have also lead to the results in Sandroni et al. (2013). Subjects' comments provide some indication for such a conclusion. One subject, for example, explicitly stated that she did not want to make the decision herself, but rather leave it to luck.[11] Similar mechanisms could apply to type-3 subjects. Choosing the risky option more often in the favorable social decisions implies that in the end it is the random draw that establishes an uneven distribution and not the participant's choice directly. These subjects also more often state that they want to implement the probabilistic dictator game, instead of the deterministic version in part 2 of the experiment. Procedural fairness concerns are another potential explanation for type-3 subjects. Even though procedural concerns (Bolton et al., 2005; Saito, 2013; Trautmann, 2009) should play no role with equivalent expected outcomes for both options, it might still be that a subset of individuals perceives the risky lottery to be fairer in the favorable range. Giving a chance (even if small) to get the entire amount could be considered as more appropriate than implementing for sure a very unequal payoff structure. Participants' comments again are in line with both lines of reasoning. One subject explicitly stated that he or she chose out of fairness concerns and another said that probabilities reduce the responsibility and feeling of guilt. These motivations stand in stark contrast to type-1 individuals.

Is it possible that our results are driven partly by the experimental design? The order of the experimental parts (in particular between part 1 and part 4) was not randomized or varied, such that any within-subject treatment effects could potentially stem from this task sequence. However, it does not seem very plausible that the order of treatments would explain the change towards more risky choices in the socially unfavorable domain. First, uncertainty about the experimental payment as well as individual lotteries were only resolved at the very end of the experiment and no feedback of any sort was provided beforehand. Hence, there was no room for any type of income effects. Second, and most importantly, it is hard to see how order effects could explain the asymmetric effect observed from the social context in part 1 to the individual context in part 4. To stand as an explanation, order effects should have impacted the choices made by subjects in tasks T1i-T4i but not in tasks T5i-T9i. Even more so, the documented differences for subsamples of our subjects

[11] It also seems that these subjects are genuinely more altruistic. In both the deterministic and the probabilistic dictator game, on average, they give the most to the recipient.

and the contents of the comment section are very difficult to interpret based on the order of tasks.

2.5 Conclusion

Our data suggest that risk taking is influenced by the relative social situation of the decision maker. Compared to equivalent situations without a social context, more risk is chosen in unfavorable situations, while similar risk taking is observed for favorable social situations. A large share of our decision makers exhibits this pattern in a very pronounced way.

This observed behavioral pattern cannot be straightforwardly explained by extensions of models of outcome-based social preferences for stochastic environments. The overall asymmetric pattern rather points towards the importance of social reference points and/or a utility from winning, i.e. leaving the experiment with more money than the matched participant (the only available reference person).

Our experimental results suggest that the role of social context may be critical also in understanding organizational and financial risk taking. When subjects directly compete against each other (e.g., over resources or power), even without any explicit competition incentives such as tournament prizes, they might take excessive risks that they would not take absent information on outcomes of others. Information provision or the way this information is presented may affect managers and investors alike, and policy makers should take them into account when designing rules and regulations.

In a broader context, our study provides another piece of evidence for the idea that risk taking is strongly affected by the social environment in which decisions take place. Future studies could test specific theoretical models of excessive risk taking that embed the risky situation into a social environment. Further, the social situation could be varied in different dimensions (such as the level of competition, the size of the references group, the presentation of information, etc.), not only along the outcome dimension. We see our results as a first steps towards a better understanding of the influence of social comparisons in risk taking.

Chapter 3

Show What You Risk — Norms for Risk Taking

3.1 Introduction

Decision making under risk is ubiquitous. Almost all — also economic — decisions involve some consideration of possible states of the future. Likewise, all of our risky decisions are embedded in a certain context, mostly including social features: decisions are made jointly, individual decisions affect other people, the decision maker observes others before deciding or her decision is observed by other people.

One elementary part of many, if not most, social contexts is that the choice is observed by other people. This is true in team decision making where team members get to know each individual's choice (live or ex post), in household decision-making when family members might learn about decisions made, or in many seemingly individual decisions where at one point others find out about one's choices (e.g. smoking, sports, investing). The decision maker might care about her choice being revealed if she cares about the signal the decision might send and her social image the decision might affect.[1] Such social image concerns have been shown to have an effect on revealed social preferences when strong behavioral norms exist (Andreoni and Bernheim, 2009).[2] In the present chapter, I shed light on this

[1] See Brennan and Pettit (2004) for a detailed representation of how people care about how they are perceived. Other seminal work in economics by Akerlof (1980) and Holländer (1990) discusses theoretical models incorporating social image and reputation concerns into utility functions.

[2] Andreoni and Bernheim (2009) as well as Dana et al. (2007) provide evidence that dictator giving might stem from social image concerns instead of pure altruism. Bohnet and Frey (1999), Dufwenberg and Muren (2006), Filiz-Ozbay and Ozbay (2014), Gächter and Fehr (1999) and Rege and Telle (2004) show that identification can have strong effects on behavior in dictator and public good games. Ariely

issue by analyzing pure observability effects and eliciting norms in risky decision making.

Even though choosing safe or risky options can both be rational depending on risk preferences, there is some evidence that norms for risk taking actually exist and that these are gender-specific. Bem (1974) initiated research investigating desirability scales for personality traits (see also Auster and Ohm, 2000; Harris, 1994; Holt and Ellis, 1998; Prentice and Carranza, 2002). Subjects rate traits as desirable or not desirable — for females and males separately. The ratings are then compared to define a trait as female or male in relative desirability. Supported by the follow-up studies, Bem (1974) finds that 'willingness to take risks' is among the masculine characteristics. That is, this characteristic is significantly more desirable for males.[3] Further, there is ample evidence for actual gender differences in risk preferences with males being less risk averse than females (for example Charness and Gneezy, 2012; Eckel and Grossman, 2008, for reviews).[4] Hence, if descriptive norms (what people actually do) and injunctive norms (what people should do) are linked (as suggested, e.g., by Rudman and Phelan, 2008, p.63), we would expect to see a gender difference in injunctive norms for risk taking as well. Finally, Prentice and Carranza (2002) denote risk taking as a "gender-relaxed prescription", i.e. it is generally desirable — including females —, but only more so for males.

With people wanting to adhere to social norms (see, e.g., Elster, 1989; López-Pérez, 2008) and with observability of choices further increasing norm adherence preferences through reputation and social image concerns (see, e.g., Akerlof, 1980; Holländer, 1990), this evidence on social norms in risk taking implies that observability of risk choices should increase risk taking. This is particularly true for males, for whom desirability of risk taking seems much more pronounced.

To the best of my knowledge, I am the first to cleanly investigate the effect of observability of risk choices alone and hence to analyze a social image effect of the revealed risk preferences. With respect to potential channels, I provide first evidence on social norms in risk taking from an incentivized elicitation procedure.

and Levav (2000) and Ratner and Kahn (2002) show similarly strong effects for variety seeking in consumption.

[3] This assessment does not depend on the gender of the rating person. Farthing (2005) and Wilke et al. (2006) report somewhat different results. In Farthing (2005) only heroic risk taking is generally deemed desirable. Non-heroic risk takers are only preferred by males in same-sex friends. Wilke et al. (2006) report that social and recreational risk taking was rated attractive in a potential partner, while for example risk taking in investment was rated neutrally. They do not find pronounced gender differences in these ratings.

[4] Note that Filippin and Crosetto (2016) indicate that these gender differences might depend on the specific task used. They suggest that, e.g., the availability of a salient safe option is one element in risk tasks that induces gender differences (Crosetto and Filippin, 2017).

I implement a laboratory experiment in which participants are matched with another participant and make a risky investment choice (Gneezy and Potters, 1997). The matching includes visual identification. While choices for participants in the control condition are anonymous, participants in the treatment condition know that the matched participant will learn about their risk choices at the end of the experiment. To account for potential gender differences in treatment effects, I balance the sample on gender. This further allows me to analyze matched participant gender effects. After investment decisions are made, I elicit beliefs about the choice of the matched participants (descriptive norms) and behavior deemed appropriate (injunctive norms) using a procedure similar to Krupka and Weber (2013).[5] I consequently link these norms to actual risk taking.

Overall — and for both males and females separately — I do not find an effect of choices being observed on risk taking. However, both descriptive and injunctive norms strongly differ between males and females. This helps to understand the gender gap in risk taking, that is also prevalent in my data. While females on average do not strongly deviate from injunctive norms, males clearly "overshoot": They invest more than what they think other people deem appropriate. This pattern is driven by participants that indicate to care little about norm conformity.

Understanding pure observability effects and existing norms in risky choices is important for both modeling economic behavior and to comprehend biased measurement. First, policy makers, firms or other agents can better estimate individual and group decision making depending on the specificities of the social context. For example, financial industry firms can set up policies limiting or easing direct information flows between proprietary traders to affect signaling or social image effects on investment decisions. Policy makers can influence the observability of customer investment or insurance decisions in financial consulting procedures, but similarly in related domains (e.g. preventive health care or treatment choice). Furthermore, group decision making strongly hinges on norms and social image concerns. When forming teams for sensitive functions, supervisors need to know whether signaling concerns lead for example specific gender constellations to produce very different risky behaviors. Second, an "observer effect" in stated and revealed risk preferences — if existent — needs to be considered in survey designs when interviewers observe responses. Otherwise, measurement error in response behavior systematically impedes high data quality.

[5] Descriptive norms relate to "what most others do", while injunctive norms define "what most others approve or disapprove" (see, e.g., Cialdini et al., 1990). These concepts closely correspond to descriptive and prescriptive stereotypes (Gill, 2004).

In my design I exclude any potential channels that might confound the measurement of mere observability effects: learning by the observer, outcome-based preferences via observing outcomes as well, and signaling skill or superior information by the choice. Basically only information on the curvature of the utility function is observed. My setup further allows a systematic analysis of gender pairing effects and can directly link incentivized risk taking to incentivized beliefs about norms in risk taking. These norms have so far only been elicited in non-incentivized procedures and the evidence mostly relates to relative, not absolute, desirability by gender.

Other studies have looked at aspects of observability of risky choices, where other confounding concerns are present, too. Yechiam et al. (2008) run an experiment with choices between a risky lottery and a safe option and pairs of participants both observe the other's choice and outcome live on their screen. The authors find that the social exposure increases risk taking compared to a purely individual control group in one out of two tasks used.[6] Tymula and Whitehair (2018), however, find no effect of live observation on risky choices in the laboratory. While these studies allow to discuss observer behavior as well, I can disentangle the effects of merely being observed (social image concerns and norms) from the effects of possibly affecting the choice of the matched participant (expecting learning from the other) and mere consistency preferences between tasks. Further, by observing choices *and* outcomes, outcome-based social preferences might play a role in the findings of Yechiam et al. (2008). Curley et al. (1986) find that ambiguous choices are made less often if experimental participants were observed by a group of other participants. Other evidence comes from accountability studies. In these studies, participants have to explain and justify their choices to the experimenter after the experiment. Vieider (2009) finds that participants behave less loss averse when "held accountable". For choices under risk, Weigold and Schlenker (1991) report that experimental participants become more extreme in their revealed risk preference. With subjects facing rather complex lotteries represented as histograms this might stem from these more extreme but consistent choices being easier to justify in front of the experimenter.[7]

Lastly, there is a strand of literature predominately in psychology that provides mostly correlational evidence on the effects of being observed in various forms

[6] In a small second study including 32 participants, only one participant observes the matched participant's choices and outcomes. In this study those participants observing the other's choice and outcome choose the risky option more often.

[7] As reviewed by Patil et al. (2014), depending on whether there is a normatively correct choice or not, accountability can lead subjects to make choices that are simply more easily justified or to exert more cognitive effort to find the correct answer (see also Simonson, 1989; Simonson and Nye, 1992).

of risky behavior. Hamed (2001), Himanen and Kulmala (1988) and Pawlowski et al. (2008) assess road crossing behavior depending on group composition and bystanders, Chen et al. (2000), Ebbesen and Haney (1973), Jackson and Gray (1976) and Nuyts and Vesentini (2005) take car driving behavior in combination with proximity of other cars and passenger characteristics, Ronay and Hippel (2010) report from skateboard tricks, and Frankenhuis et al. (2010) take bridge crossing time in virtual reality as risk measure. These studies mostly indicate that males increase risk taking in the presence of females. While this is suggestive evidence for an observability effect, it remains unclear to what degree these findings arise from endogenous assignment to treatment, from subjects being able to signal more than mere risk preferences and being able to affect others with their choices, from the lack of incentives, or indeed from social image concerns regarding risk preferences. Baker and Maner (2008, 2009) and Frankenhuis and Karremans (2012) further explicitly relate risk taking in males to mating preferences and relationship status, arguing that single males use risky behaviors to attract attractive females. My design allows me to test these observer gender and observer characteristics effects, too. While I indeed see that males take more risk and "overshoot" norms more strongly when matched with a female, this is independent of the choice being observed and hence does not relate to social image concerns. However, I find some evidence for the attractiveness of the matched participant being important for the effects of observability. Participants react differently (more risk taking) to the choice being observed if matched with an attractive participant compared to being matched with a less attractive participant.

The remainder of the chapter is structured as follows. Section 3.2 describes the experimental design in detail. Section 3.3 presents the risk-taking results and Section 3.4 discusses gender-specific norms and norm following behavior. I discuss the results in Section 3.5 and conclude in Section 3.6.

3.2 Experimental Design and Procedures

3.2.1 Experimental Design

At the beginning of the experiment, subjects were randomly assigned seats in the laboratory. After reading of the instructions, subjects were informed on-screen that they were matched with the subject seated in the seat vis-à-vis their own (facing each other). For that purpose, the wooden wall of the cubicle usually shielding subjects facing each other was removed before the experiment. Hence, besides the screen blocking most of the view, matched subjects were able to see each other.[8] Further, the first experimental screen showed the picture of the matched participant and all following decision screens showed a small picture of the matched participant at the bottom of the screen. In a between-subjects design, one subject in a pair was assigned to the treatment condition and the other to the control condition.

The main and first part of the experiment was an investment task (Gneezy and Potters, 1997). Subjects received 100 Taler worth €5 and could invest any integer amount in a risky asset. The asset paid off 2.5 times the invested amount with a 50% probability. With the remaining 50% the investment was lost, implying an expected return of 25%. The amount not invested was kept with certainty, but did not pay any interest.

To investigate the effect of the choice being observed, before making their decision in the investment game, half of the subjects were told that their choice would be shown to their matched participant at the end of the experiment. Hence, choices of these subjects in *Treatment* were not anonymous to the matched partner. The other half of the subjects (*Control*) was told that their choice was anonymous. Revealing the choice at the very end of the experiment excludes any type of learning by the observer. Not revealing the outcome excludes an impact of outcome-based social preferences. Showing matched participant pictures and allowing visual identification of the matched participant in both *Treatment* and *Control* holds matched participant identification effects constant across treatments and enables me to measure a clean effect of only the choice itself being observed.

In the second part of the experiment, all subjects answered a non-incentivized risk questionnaire on stated willingness to take risks in general and in the domains of driving, finance, sport, trust, health and career (see, e.g., Dohmen et al., 2011). Here too, subjects in *Treatment* were informed that their choices would be shown to their matched participant at the end of the experiment. This allows me to provide some evidence on whether observability of choices affects decision making differently in the different domains. Comparing the results from these non-incentivized questions to incentivized investment behavior can further speak to the interaction between signaling and signaling costs.

In the third part, subjects' beliefs about choices of others and different types of norms were elicited. This part contained five elicitation procedures of which one was at the end of the experiment randomly chosen to be paid. However, before starting part three, subjects were debriefed about the treatment conditions. They were informed that half of the participants had made anonymous choices while the

[8] See Figure C.1 in Appendix C.1 for a picture of the seat arrangement.

other half of the participants had made choices that would be observed at the end of the experiment. It was also announced that always one subject from each treatment formed a pair. This is important for comparing elicited beliefs and norms across treatments. It further allows me to test whether subjects in fact expected treatment effects.[9]

In the first belief eliciation procedure, subjects were asked for beliefs about the matched participant's investment (*guess partner*). Subjects could earn €5 if they did not deviate by more than 10 Taler from the true value. This 10-Taler-deviation-based incentive scheme was the same for all following elicitation procedures. In the second procedure, subjects stated beliefs about the average investment in the experimental session (*guess all*).

The third, fourth, and fifth elicitation procedures measure injunctive norms for investment and are inspired by the procedure first used by Krupka and Weber (2013). Prior to that, all subjects were informed that their picture would be shown to four other participants at a later stage. These four participants would then have to indicate the appropriate amount that the person in the picture "should have invested". I label the average of these four statements as the injunctive norm for that person.

Consequently, in the third elicitation, each subject (e.g. subject A) had to anticipate this average norm (*perceived norm*), that was to be indicated by the four (unknown) other participants (when seeing A's picture). Subjects were then told that one of the four participants that would see their picture would be their matched participant.[10] For the fourth elicitation, subjects had to anticipate what appropriate amount the matched participant would indicate when seeing their picture. I denote this anticipation *perceived norm partner*. In the fifth and final elicitation procedure all subjects then actually saw — one by one — four pictures and indicated what they thought were appropriate investment amounts for the respective participants (*stated norm 1 - stated norm 4*). The last picture rated (*stated norm 4*) always showed the matched participant. Hence, this fifth elicitation procedure consisted of four choices: Every subject indicates this norm for four participants, and hence every participant's picture is seen by four other participants.[11] If this fifth elicitation procedure in the end was by the computer chosen to be payoff relevant for a participant, one of the

[9] Further, without debriefing it would have been very difficult to elicit beliefs about behavior of the entire group and the matched participant without deceiving subjects (by omission of information). Subjects would have wrongly expected the matched participant to have seen the same instructions.

[10] The first three pictures were actually chosen randomly — by randomly assigned seat numbers.

[11] The average of the four statements of participants seeing the same picture is what I labeled as injunctive norm above. This again is the value that the subject in the picture had to anticipate in *perceived norm*.

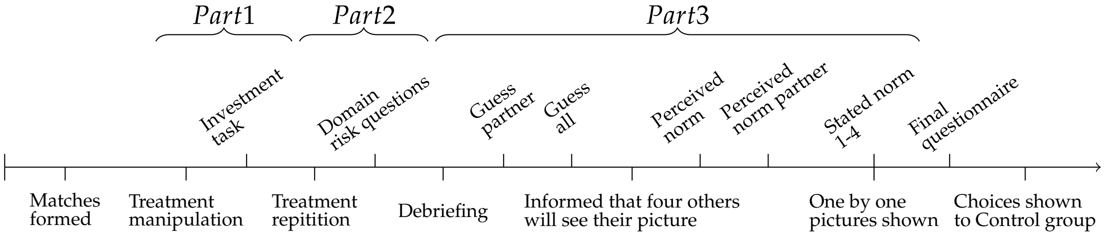


Figure 3.1: Timeline of the experiment

four stated norms was randomly selected for payment. The subject was then paid if she did not deviate by more than 10 Taler from the average answer of the three other participants rating the respective picture.[12] Figure 3.1 displays the timeline of the experiment.

After the elicitation tasks, all subjects answered a final questionnaire. Next to standard questions related to age, gender, field of study and mother tongue, the questionnaire also consisted of open-end questions that allowed subjects to make general comments regarding the experiment and to explain what they considered during the decision process. Further, Likert-scale questions asked for *norm conformity* ("How much do you usually conform to norms?"), *rule breaking* ("How much do you like to break rules?") and *social image* ("How much do you care about other people's perception of you?") preferences, as well as how risk-loving (risk-avoiding) subjects would want to be perceived (*ideal perception*). While I can control for standard observables when estimating the treatment effect, the survey questions allow me to analyze heterogeneity in the treatment effect.

Lastly, I collected data on picture characteristics. Four research assistants (RAs) independently coded each picture on whether the individual made "eye contact" with the camera and looked friendly, and rated attractiveness on a scale from one to ten.[13] This allows me to check whether treatment or matched gender effects depend on visual cues.

[12] This coordination game induced by the incentive scheme is used to identify beliefs regarding group perceptions of what people ought to do, i.e. injunctive norms. See Krupka and Weber (2013) for details.

[13] For the data analysis I use the average attractiveness rating of all four RAs. The binary variables *eye contact* and *friendly face* are one if more than two out of four RAs indicated so. RAs further guessed age, gender and ethnicity to account for subjects' looks potentially diverging from facts. No meaningful differences emerged (e.g. none at all for the assignment of participants to sex).

3.2.2 Procedural Details

I programmed and conducted the experiment with "z-Tree" (Fischbacher, 2007) and 428 subjects, 215 males and 213 females, were recruited using the online recruiting system "ORSEE" (Greiner, 2015).[14]

All 28 experimental sessions took place at the Munich Experimental Laboratory for Economic and Social Sciences (MELESSA) between May and October 2017. To ensure a gender balance for in expectation same-sized gender pairing cells, half of the planned number of subjects per session were of one gender. Upon arrival at the laboratory, subjects first had to sign a consent form that allowed pictures to be taken during the experiment.[15] Subjects were then randomly assigned seats in the laboratory and (portrait) pictures were taken individually upon entering the laboratory. While instructions were read out aloud, RAs copied the pictures via remote access from the camera to the local drives of the subjects, so that pictures could later be displayed onscreen.

All subjects were paid privately after the experiment and earned €12.68 on average (including a fix payment of €5 for showing up on time), ranging from €5 to €22.50. While the investment task was always paid, only one of the elicitation procedures in part 3 was at the end randomly and individually chosen for payment by the computer. The sessions lasted slightly less than 45 minutes on average.

3.3 Risk Taking

3.3.1 Manipulation Check and Sample Balance

To make sure that subjects indeed perceived the treatments differently — a necessary condition to observe a causal treatment effect — I asked subjects at the end of the experiment how much they felt being watched when making the investment decision (on a scale from one to ten). Figure 3.2 shows average answers to the question by treatment condition including 95% confidence intervals.

[14] The final questionnaire included a question whether subjects had heard of the experiment before participating. Since subjects were debriefed and future subjects could have been made aware of the treatment manipulation and research interest of the experiment, I exlude 12 (out of 440) subjects that had indeed heard of the experiment. This exclusion does not affect my results.

[15] See Appendix C.2 for the exact wording. The email invitation to sign in for an experimental session made clear that pictures would be taken during the experiment. Nobody objected to pictures being taken. Appendix C.3 further includes the exact wording of the instructions.

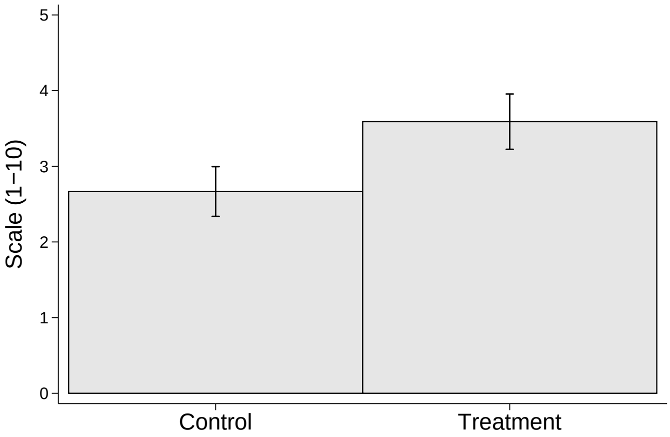


Figure 3.2: Manipulation check: Stated feeling of being watched

Subjects in *Treatment* clearly indicate that they felt being watched more strongly than subjects in *Control* (p-value < 0.01, Mann-Whitney test). Hence, the treatment effectively changed the decision environment of subjects.[16]

Table C.1 in the Appendix reports a randomization table between *Control* and *Treatment*. I do not observe any significant differences in socio-economic background variables and picture ratings between the treatment conditions.

3.3.2 Full Sample Results

Based on the literature discussed above, I expect higher investments when being observed for males, but possibly also a (weaker) positive treatment effect for females (see the weaker, but existing, norms for females in Prentice and Carranza, 2002). If willingness to take risks indeed is a desirable trait, social image concerns in *Treatment* should push subjects to riskier choices compared to *Control*.

In contrast to that, I do not see an overall difference in investment by treatment. Subjects in *Treatment* invest 51.14 Taler on average, while *Control* subjects' average investment is 52.29. This small difference is clearly insignificant (p-value $= 0.55$, Mann-Whitney test). Statistical power to detect treatment effects is not the reason for this null-finding. To detect the measured effect size as being significant, I would need roughly 16,000 observations and with the 428 observations and given the standard deviation in investment in my sample I would be able to detect an effect size of roughly 7 Taler (0.27 standard deviations) with a power of 80% (two-sided test, $\alpha = 0.05$).

[16] While power naturally decreases, this is directionally true for all gender and gender pairing subsamples.

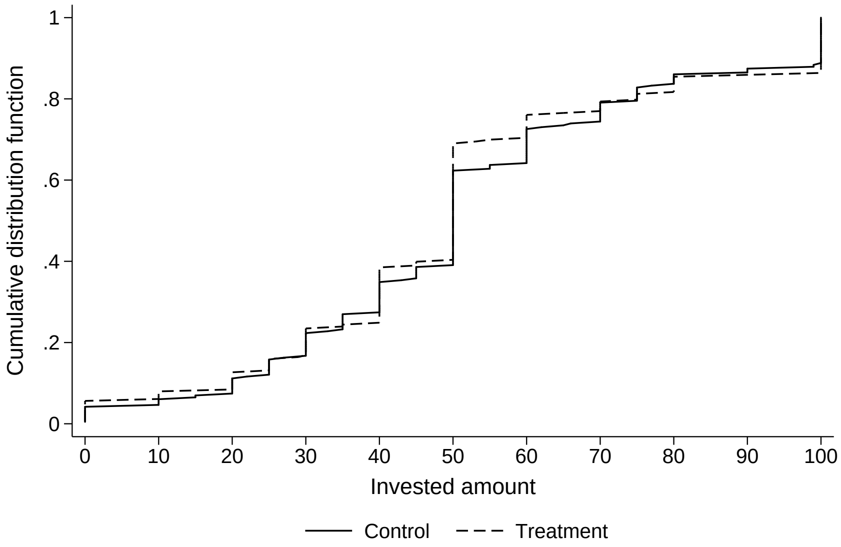


Figure 3.3: Cumulative distribution function of investment by treatment

For a complete representation of investment amounts, Figure 3.3 displays the cumulative distribution functions of investment by treatment condition. With the functions crossing multiple times and never strongly diverging, I clearly do not see large differences in the distributions (p-value = 0.61, Kolmogorov-Smirnov test).[17] Further, the distribution shows that I have sufficient variation in investments to potentially observe treatment effects, such that an overly strong focal point at 50 is not responsible for the null effect.

Even though the sample was balanced on observables across treatments, I check the robustness of the overall non-parametric null finding in a regression framework with additional controls. Table 3.1 displays Tobit regressions on investment. Model (1) explains investment solely with the treatment indicator and therefore is the parametric equivalent to the non-parametric test. Model (2) and (3) add gender and the gender of the matched participant as controls, respectively. The effect of *female* clearly shows a large gender gap in investments. This is in line with much of the literature suggesting gender differences in risk taking, particularly in this type of task. The gender of the matched participant has no significant overall effect. Model (4) adds standard observables, model (5) further includes information from the individual's picture and model (6) further incorporates survey responses on *norm conformity*, *rule breaking*, *social image* and *ideal perception*.

The regression table shows that the null effect of *Treatment* is very robust to controlling for all available information. Even when adding the survey measures

[17] There seem to be some differences by treatment in the fraction of subjects choosing round numbers (multiples of 10; i.e., 0, 10, 20,...). Table C.2 in the Appendix gives a more detailed overview of these patterns.

Table 3.1: Tobit regressions on investment

	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	−1.033	−0.117	0.044	0.362	0.204	1.597
	(3.046)	(2.661)	(2.648)	(2.695)	(2.596)	(2.437)
Female		−18.493***	−18.065***	-16.972***	-17.199***	-16.197***
		(2.580)	(2.550)	(2.535)	(2.581)	(2.509)
Female Partner			3.280	3.166	2.982	2.571
			(3.249)	(3.237)	(3.312)	(3.133)
Constant	53.365***	62.169***	60.246***	39.648***	28.039	2.597
	(2.088)	(2.548)	(3.076)	(15.170)	(19.243)	(19.486)
Standard observables	No	No	No	Yes	Yes	Yes
Picture characteristics	No	No	No	No	Yes	Yes
Survey responses	No	No	No	No	No	Yes
Observations	428	428	428	428	428	428
Pseudo R^2	0.00	0.01	0.01	0.01	0.02	0.02

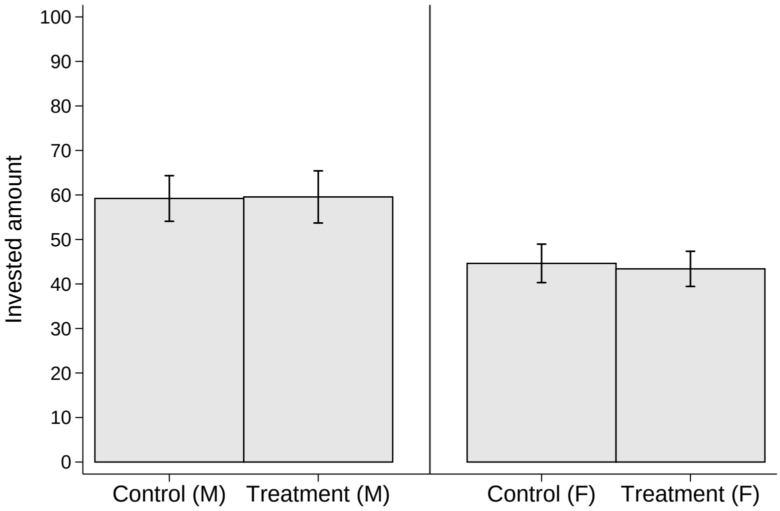
Notes: Two-limit (0-100) Tobit regressions on invested amount. Clustered (on experimental session level) standard errors in parentheses. Stars indicate significant coefficients, and *** p<0.01, ** p<0.05, * p<0.1. Standard observables include: age, last math grade in school (indicator variables), nationality (indicator for German, European, Non-European), relationship status, whether the subject studies economics or business and time spent on the first experimental screen. Picture characteristics were rated by RAs: friendly look, attractiveness and eyecontact. Survey responses include *norm conformity*, *rule breaking*, *social image* and *ideal perception* from the questions at the end of the experiment.

in model (6), the treatment effect remains very small and clearly insignificant. This is despite these measures potentially being endogenous to treatment and most likely biasing the treatment coefficient upwards.[18] Apart from the significant effect of gender, better math grades (for all specifications), *rule breaking* and *ideal perception* are significantly positively linked to investment.

3.3.3 Gender and Matched Gender

Looking at overall treatment effects possibly obscures differences by gender. While there is evidence that risk taking is desirable for females, too, the vast majority of papers considering desirability of risk taking highlights a strong asymmetry in desirability by gender. With such an asymmetry, the overall null effect might be the result of a negative treatment effect for females canceling out a positive treatment effect for males, for example. Similarly, gender pairing is a prime candidate for heterogeneity. The literature in psychology on mating preferences and risk taking (Baker and Maner, 2008, 2009; Frankenhuis and Karremans, 2012) for example is an indication for an asymmetric treatment effect on males. Males should

[18] Both *social image* (p-value = 0.081, Mann-Whitney test) and *ideal perception* (p-value = 0.086, Mann-Whitney test) are weakly significantly lower in *Treatment* compared to *Control*. They are also both overall positively related to investment amounts. Hence, controlling for these measures wrongly estimates lower predicted investments for subjects in *Treatment* if survey responses are affected by treatment, leading to a higher coefficient on *Treatment* to compensate for that effect.



Notes: Error bars indicate 95% confidence intervals. "M" and "F" refer to data for males and females, respectively.

Figure 3.4: Treatment effect for males and females separately

react particularly strong and positive when being matched with a female. In this subsection, I will consider these sources of treatment effect heterogeneity.

Contrary to expectations, I do not find a treatment effect for neither gender. Figure 3.4 shows that males and females both do not react to their choices being observed (p-value for females = 0.51; for males = 0.80; Mann-Whitney tests).[19] I only clearly see the overall gender differences in risk taking already seen in the regression results. Men on average (independent of treatment) invest 59.38, while females only invest on average 43.99 (p-value < 0.01, Mann-Whitney test).

Besides considering gender separately and showing again the overall treatment effect, Table 3.2 displays subgroup treatment effects for the four possible gender pairs. This shows whether potentially opposing treatment effects by matched gender cancel each other out when ignoring matched gender.

It indeed seems as if the effects within gender pairs cancel out for females, obscuring existing treatment effects. Females invest clearly less in *Treatment* when matched with males, while they directionally invest more in *Treatment* when matched with another female. The difference between these differences is large and significant (diff-in-diff of 16.11, p-value = 0.01, two-sided t-test) suggesting that females react differently to *Treatment* depending on the matched

[19] Similarly, there is no change of the distribution from *Control* to *Treatment* for neither females nor males considered separately. See Figure C.2 in the Appendix for cumulative distribution functions of investment by treatment conditional on gender.

Table 3.2: Treatment effects overall, by gender and by gender pairs

	Control (n)	Treatment (n)	Treatment effect	p-value
All Subjects	52.29 (215)	51.14 (213)	−1.15	0.55
Females	44.63 (102)	43.40 (111)	−1.23	0.51
Males	59.21 (113)	59.56 (102)	0.35	0.80
Females matched with females	38.40 (47)	46.11 (45)	7.71	0.14
Females matched with males	49.95 (55)	41.55 (66)	−8.40	0.02**
Males matched with females	61.48 (66)	64.36 (55)	2.88	0.54
Males matched with males	56.02 (47)	53.94 (47)	−2.08	0.94

Notes: Invested amounts for all subjects, males and females separately, and all four gender pairs separately, by treatment condition. Values in parenthesis denote the number of observations in a given cell. P-values for the treatment differences are based on Mann-Whitney tests, and *** p<0.01, ** p<0.05, * p<0.1.

participant gender.[20] While the treatment effect for females matched with males gets (weakly) insignificant when correcting for multiple hypothesis testing (conservative Bonferroni correction for four hypotheses tested, $\hat{p} = 0.09$), the difference in differences is pronounced and robust ($\hat{p} = 0.014$, correction for two hypotheses tested).

However, this does not seem to be a treatment effect per se: The difference arises from within *Control*, where investment depends on the gender of the matched participant. The average investment of females in *Treatment* does not depend on the matched gender (p-value = 0.55, Mann-Whitney test) and for both matched genders does not differ significantly from the average investment of females overall. In *Control* however, females' investment is much higher if they are matched with a male as compared to when being matched with a female (p-value < 0.01, Mann-Whitney test). Further, there is some evidence that these treatment effects on the gender pair level arise from failed randomization in some subsamples. I will discuss these results in Section 3.4.2 when relating investment behavior to norms.

3.3.4 Other Dimensions of Treatment Heterogeneity

Apart from the main subsamples by gender, I further check heterogeneity of treatment effects by personality traits. *Social image* and *ideal perception* could strongly affect the treatment effect. Social image concerns are important, because if a person does not care about how she is perceived, then behavior should be independent

[20] I can compare two empirical distributions non-parametrically. However, for difference in differences tests in my between-subjects design, I can only calculate the treatment effect on means and have to rely on parametric assumptions for testing using the t-test.

of treatment. Likewise, conditional on having social image concerns, the effect of *Treatment* should crucially depend on how subjects want to be seen. If subjects do not want to be perceived as willing to take risks, they should lower risk taking in *Treatment*. If they do want to be perceived as willing to take risks, they should increase risk taking. While I find *ideal perception* to have a positive and significant impact on investment overall, there is no significant interaction with the treatment condition. This holds for controlling for above median social image concerns.

The only weak heterogeneity in treatment effects relates to the attractiveness of the matched partner. While for both — being matched with an attractive or non-attractive participant (measured by below or above median rated attractiveness) — the treatment effect is insignificant, they clearly go in opposite directions: Those matched with an attractive partner increase risk taking in *Treatment* (insignificantly) and those matched with an unattractive partner reduce risk taking in *Treatment* (again insignificantly). This difference in treatment effects (TED) is significant (TED = 11, p-value = 0.04, two-sided t-test). I can look at specific subsamples separately. It seems that males (TED = 14, p-value = 0.08) and those matched with female partners (TED = 13, p-value = 0.09) show this pattern more strongly. The difference gets clearly larger (TED = 33; p-value = 0.01; n = 66) only considering single males matched with females. Further — among these — only considering those with above median social image concerns and above median ideal risk perception shoots up the TED to 72 Taler (p-value = 0.08; n = 12).[21]

3.3.5 Non-Incentivized Domain-Specific Risk Questions

The reaction to observability might generally depend on how choices are incentivized. On the one hand, if subjects want to signal a specific type or trait with their risky choice when being observed, this comes at a signaling cost if the choice is incentivized. This signaling is costless if the choice is not incentivized. On the other hand, there might be a relationship between the (perceived) informativeness of a signal and signaling costs. Signaling a specific type might only be effective if the signal itself is credible, i.e. when deviation from truth-telling is costly.

Further, there might be domain-specific effects of the risk choice being visible, since people may want to be perceived differently depending on what type of risk taking is considered.

I use the domain-specific risk questionnaire from the German Socio-Economic Panel (SOEP) that elicits willingness to take risks in general, in car driving, in

[21] See Figure C.3 in the Appendix for a graphical illustration of these patterns.

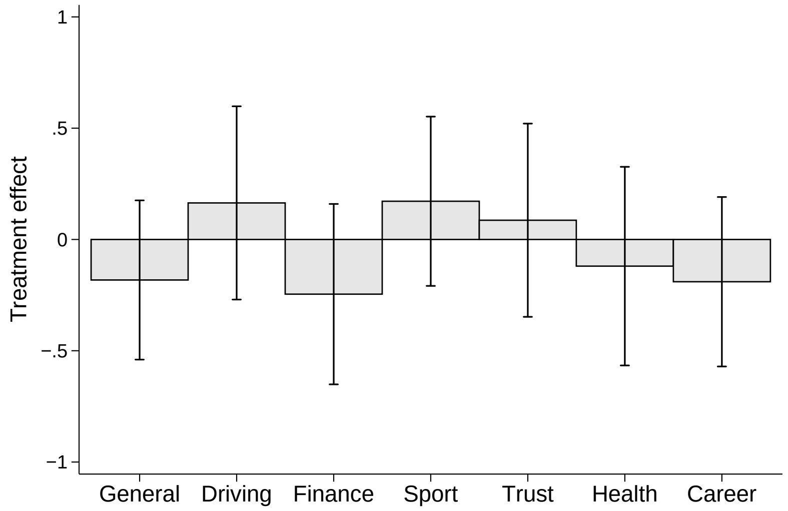


Figure 3.5: Treatment effect on non-incentivized domain-specific risk taking

personal finance, in sports, in trusting other people, in health and in one's career. With the main task in the experiment relating to financial risk taking, this can also shed light on how the null effect measured in the main part might translate into other domains.

Figure 3.5 shows basically no differences between the treatment effects in the different domains. All domain-specific treatment effects are small and insignificant. This again is no power issue. The 95% confidence intervals span only slightly more than a 0.5 treatment effect size allowing me to detect small effects on the questionnaire scale from 1-10 (with a power of 80%, $\alpha = 0.05$, and two-sided tests, I would be able to detect an effect size of roughly 0.25 for all domains). Figure C.4 in Appendix C.5 shows treatment effects by domain for all four gender pairs separately. Also there, none of the effects is significant (and none large).

3.4 Norms

With much of the experimental design focusing on different types of norms in the investment decision, I next discuss the overall gender-specific patterns in these norms and then relate these elicited norms to actual investment behavior. If not stated otherwise, I refer to overall norms — independent of treatment — and only distinguish between the treatment conditions where informative.

3.4.1 Gender-Specific Norms

In light of the large gender differences in actual risk taking, it is interesting to see whether subjects expect these gender differences. This is indeed the case. Subjects think their matched participant invested more if the matched participant was male (p-value $<$ 0.01, Mann-Whitney test) providing strong evidence for gender differences in descriptive norms in risk taking. The left panel of Figure 3.6 shows the effect of matched participant gender on *guess partner* (including 95% confidence intervals). This difference in beliefs is stronger for females (p-value $<$ 0.01, Mann-Whitney test), but directionally similar for males only (p-value = 0.16). The finding is robust to only using data from subjects that indicated to be at least somewhat confident in their guess. After every belief statement, I asked subjects to indicate confidence on a scale from one to five. Excluding subjects that stated one ("I am not at all sure about my answer — I basically guessed randomly") does not change the result — if anything, the (matched) gender differences become more pronounced. This is true for all of the following other norm statements.

While *guess all* does not allow to differentiate between norms for females and males, I observe that, despite there being large gender differences in actual investment, the average guess of females regarding the average session investment does not at all differ from the average guess of males. Subjects generally underestimate average investment by slightly more than 4.5 Taler (p-value $<$ 0.01, Wilcoxon signrank test).

As depicted in the middle panel of Figure 3.6, not only descriptive norms differ by gender, but *perceived norm* clearly depends on gender, too (p-value $<$ 0.01, Mann-Whitney test). Males think they should have invested 49.69, while females think they should have invested only 42.96.[22] This is strong support for the non-incentivized survey evidence on gender-specific desirability of risk attitudes (e.g. Bem, 1974) and shows that the difference is robust to incentivizing subjects for normative statements.

Lastly, also stated norms for investment are higher for male pictures (right panel of Figure 3.6). That is, subjects agree on males being supposed to invest more. This

[22] This is very similar for *perceived norm partner*, even though the difference is only weakly significant (p-value = 0.06, Mann-Whitney test). This difference becomes larger and significant at the 1%-level if I consider only subjects that indicated to not having basically randomly guessed the value (one fourth excluded). *Perceived norm partner* was elicited to detect potential differences in the perceived norm by gender of the rating person. This is not the case, neither overall nor for females or males separately. Since *perceived norm* corresponds to the more general notion of norms and is much more meaningful for subjects in *Control*, I refer to *perceived norm* in the main analyses. Results generally are very similar when using *perceived norm partner* and I indicate any difference where applicable.

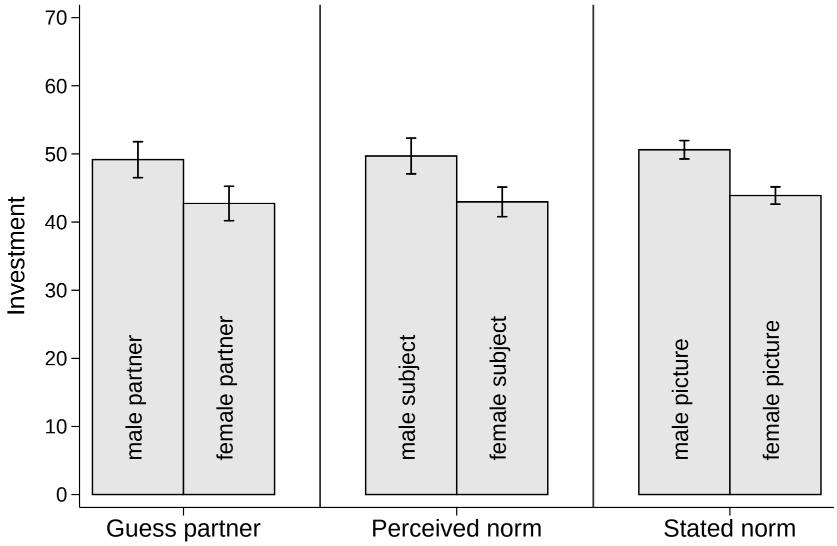


Figure 3.6: *Guess partner* by matched participant gender, *perceived norm* by subject gender, and *stated norm* by picture gender

difference by gender of the rated picture is highly significant over data from all four pictures rated combined (average of *stated norm 1* to *stated norm 4* by picture gender as displayed in Figure 3.6, p-value < 0.01, Mann-Whitney test), but also when I compare ratings for male and female pictures for *stated norm 1* to *stated norm 4* separately (p-value < 0.01 for all four pictures, Mann-Whitney test).[23] Both males and females hold these gender specific norms.

For a detailed overview of all indicated norms (including *perceived norm partner* and *guess all* which are not shown in Figure 3.6) by gender, matched participant gender and treatment cell, see Table C.3 in Appendix C.6.

In the next subsection, I discuss the relationship between norms and investment, both overall and by treatment. Independent of this relationship it is interesting to note that in line with there being no treatment effect, subjects did not expect investment differences based on treatment. *Guess partner* is independent of treatment.

[23] Having to state norms for different people might make subjects think there should be a difference in their assessment - even when there originally is not. This is not a concern when only considering the first picture. While subjects could in principle have inferred that they would have to rate more participants (they knew that their own norm would be based on four other participants) only 10% of the subjects in the follow-up questionnaire indicated that they expected to rate more than one picture.

3.4.2 Norms and Investment Behavior

When people care about social image that in turn depends on norm adherence, one should expect individuals' investment behavior to more closely track perceived norms in *Treatment* compared to *Control*. However, patterns of norm adherence in my experiment are independent of treatment condition, i.e. of the choice being observed or not. This is true overall and when looking at behavior of females or males separately. Consequently, when discussing the relationship between norms and investments in the following, I will abstract from the treatment condition and report results over both treatments combined.

Section 3.4.1 demonstrates large norm differences between males and females. These can explain (at least part of) the gender gap in choices under risk in my experiment. However, they do not explain the entire gap in investment. While descriptive and injunctive norms for males are roughly at 50 Taler, average actual investment of males lies at almost 60 Taler. That is, males clearly "overshoot" beyond their perceived norms leading to an even larger gender gap in actual investments.

The difference between actual investment and *perceived norm* (what should affect individual behavior) — which I use to describe "norm following" — is highly significant for males (p-value < 0.01, Wilcoxon signrank test). This is not the case for females. Figure 3.7 shows average norm following by gender and matched gender cells and plots 95%-confidence intervals. While the difference for males is significant for either matched participant gender, it is especially pronounced for males being matched with females. This difference in norm following by matched gender is weakly significant (p-value $= 0.05$, Mann-Whitney test).

The norm following results for males inform an interesting pattern. Ignoring the treatment conditions, men invest on average 62.79 when matched with a female and 54.98 when matched with males. This investment difference by the gender of the matched partner is significant at the 5% level (p-value $= 0.04$, Mann-Whitney test). As indicated in Figure 3.7, this cannot be explained by a difference in perceived norms depending on matched gender. It rather is indeed a (matched gender-specific) "overshooting" beyond perceived norms.

In contrast to males, females' investment behavior does overall not significantly differ from perceived norms. Only for those matched with males, the deviation is weakly significantly positive (p-value $= 0.09$, Wilcoxon signrank test).[24] As for

[24] Inference based on the confidence intervals (calculated with t-statistics) can lead to slightly different results (p-values) than when using the (correct) non-parametric Wilcoxon signrank test.

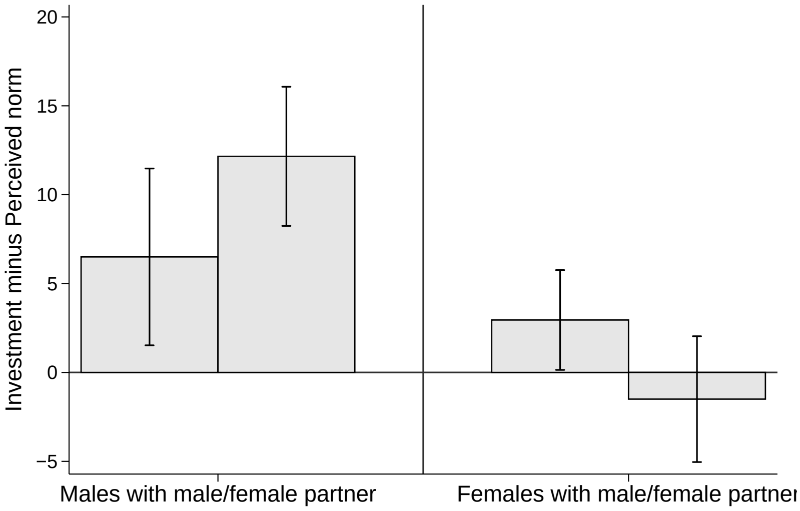


Figure 3.7: Norm following by gender pairs

males, the difference in norm following by matched gender is weakly significant for females, too (p-value = 0.09, Mann-Whitney test).[25]

Lastly, *rule breaking* and *norm conformity* should naturally be linked to norm following behavior. Those with strong norm conformity preferences can be assumed to follow perceived norms more closely (vice versa for rule breaking preference). This is borne out by the data, at least for males. For those with above median norm conformity preferences the norm "overshooting" is not significant, and small in magnitude. In contrast to that, the "overshooting" is very large and highly significant for males with below median norm conformity preferences. As expected, just the reverse — only less pronounced — is true for rule breaking preferences. For females not much of a difference emerges for *norm conformity*, while *rule breaking* is positively linked to norm "overshooting". See Figures C.7 and C.8 in the Appendix for details.

Explaining the Treatment Effect for Females Matched with Males

I use answers to belief and norm questions to look into and understand the significant treatment effect for investment of females matched with males. As discussed above, female investments do not strongly deviate from norms in neither

[25] These findings are similar if I use *perceived norm partner* instead of *perceived norm*. The diff-in-diff for males is then significant at the 1%-level. For females matched with males the deviation becomes insignficant, while the "undershooting" when matched with females becomes significant (p-value $<$ 0.01, Wilcoxon signrank test). The diff-in-diff for females is not significant. The "undershooting" of females matched with females, however, is entirely driven by females in *Control*, for which *perceived norm partner* is less applicable. This difference to Figure 3.7 using *perceived norm* should therefore not be overweighted. For norm following by treatment condition see Figure C.6 in the Appendix.

treatment when matched with males. This shows that the strong treatment effects for these females can to a large extend be explained by differences in perceived norms. The question then is, where these norm differences by matched gender and treatment come from.

One can put the norm differences by treatment for females differently: Given being in *Control* or *Treatment*, females' *perceived norm* depends on the gender of the matched participant. In *Control*, females perceive higher investments as the norm when matched with males compared to when matched with females (p-value = 0.05, Mann-Whitney test). In *Treatment*, however, females perceive lower investments as the norm when matched with males compared to when matched with females (p-value = 0.04, Mann-Whitney test). This is something I should not observe. Subjects know that four other — not mentioned — participants will state the appropriate investment amount. As far as subjects are concerned, they cannot infer anything from the matched participant about the four selected participants. Hence, *perceived norm* should be independent of the matched participant.

For stated norms something very similar applies. Independent of what picture they rated (random for pictures one to three), stated norms of females matched with males are on average lower in *Treatment* than in *Control* (p-value = 0.06, Mann-Whitney test).[26] If they indeed perceived different norms based on treatment, why would they also on average state different norms for random other people (some in *Treatment*, some in *Control*; some female, some male)? The pattern for descriptive norms is similarly unintuitive.[27]

This casts doubt on the notion that investment and norm differences for females matched with males are indeed induced by the treatment per se. The evidence speaks rather in favor of an unfortunate randomization leading females in *Treatment* and matched with a male to invest less and at the same time indicate lower descriptive and injunctive norms compared to those in the control group. Evidence from balancing tests supports this. Females matched with males have a significantly lower *ideal perception* in *Treatment* (p-value = 0.02, Mann-Whitney test), which can possibly explain the treatment difference for these females.[28]

[26] I only take average stated norms for pictures 1 to 3, since including picture 4 (the matched participant) would make the average again depend on the matched partner. Including the matched participant does not change the result.

[27] As can be seen in Table C.3, for descriptive norms, too, females matched with males in *Control* consistently state higher values than those in *Treatment*.

[28] Note that *ideal perception* of course can itself be endogenous to treatment. It seems rather unlikely, however, that *ideal perception* should only be depending on treatment for females matched with males. For all other gender pairs no differences exist.

This overall asymmetry in response behavior can help to explain the observed treatment effect for females matched with males arising from the large investment difference for females in *Control* depending on the matched participant gender. The remaining unexplained difference in (opposing) norm deviations in *Control* between those matched with males and those matched with females is only weakly significant (p-value = 0.09, Mann-Whitney test). That is, in *Control*, when matched with a male, females invest relatively more than their perceived norm compared to when matched with a female. The difference in the deviation from norms is insignificant for females in *Treatment*. This again points towards — if anything — *Control* inducing these behavioral differences.

3.5 Discussion

In this chapter I demonstrate a clear overall null effect of observability and hence signaling opportunities on choice under risk. That is, merely having somebody knowing the choice does not affect decision making. The experimental design eliminates other channels that could potentially affect decision making. As such, I exclude concerns regarding the influence one might have on others, outcome-based social preferences, the mere psychological pressure to decide with "live" audiences, opportunities to explain or justify choices and the chance to provide more than merely a signal regarding the curvature of the utility function. While the manipulation check demonstrates that subjects indeed were affected by the treatment, they overall did not change risk taking out of social image concerns when choices were observed.

This is a surprising null effect based on the literature. If willingness to take risks is deemed desirable — particularly for males — the opportunity to signal a risky type should lead subjects to invest more. Interestingly, however, injunctive norms are generally not very high in my experiment. Norms averaging at an investment of 50 Taler stand somewhat in contrast to the notion that willingness to take risks is generally deemed appropriate.[29] This could be one potential explanation for the null effect of the treatment manipulation. If the absolute norm level is not high, why should people — when observed and when caring about social image — increase risk taking? The norm levels do not explain norm following behavior though, which surprisingly does not depend on the treatment condition either. If people care about their social image, they should have a much stronger incentive to behave according

[29] Similar to investments, the average of 50 for perceived norms does not merely arise from a focal point at 50. Instead, there is large variation in these norms. See Figure C.5 in Appendix C.6 for the distribution of perceived norms and investment.

to prescriptions when observed compared to when making anonymous, purely individual decisions.

Gender and gender pairing are the most natural subsamples in my setting to consider in terms of heterogeneity in treatment effects. On these dimensions I find very little evidence for treatment effects, and the evidence on treatment effects for females matched with males seems to be mainly driven by randomization issues.

The attractiveness of the matched participant, however, seems to interact with the treatment effect. While the overall effect for the entire sample is weak, this interaction becomes very large for some subsamples. Considering the cell sizes of these ever smaller subsamples, I urge the reader to interpret these patterns cautiously. Nevertheless, these sometimes very pronounced asymmetries are striking and relate to the literature on mating preference induced behavior in psychology. Baker and Maner (2008) relatedly indicate that the mere exposure of males to pictures of attractive females leads to a positive relationship between "mating preferences" and risk taking, which was not observable for any other group. Similarly, males in Baker and Maner (2009) that expected to meet a female participant at the end of the experiment selected riskier experimental choices when that female participant was single, interested in seeing somebody and would learn about the outcome. Frankenhuis and Karremans (2012) show contrasting results for males in a relationship. They seem to not adjust own behavior to what they think females consider attractive, contrary to behavior of single males.[30]

Moving away from the incentivized investment task, I also do not find treatment effects on non-incentivized risk attitude statements in any domain. This is maybe even more surprising than the null effect in the investment task, since the signal here is basically free (if we abstract from truth-telling preferences). However, it is possible that signals have to be costly to be credible.

Lastly, and besides treatment effects, males — independent of treatment condition — clearly invest more when matched with females. Since this holds also for *Control* and cannot be explained by differences in injunctive norms, this is a remarkable effect. It relates to the findings by Carr and Steele (2010) and D'Acunto (2015) showing that males increase risk taking after a stereotype threat or gender identity priming, respectively. A similar effect might drive behavior in

[30] Evolutionary theory suggests some mechanisms why risk taking of males might indeed be perceived as attractive by females (see, e.g., Kelly and Dunbar, 2001, for a discussion of the arguments). Mate choice theory highlights resource availability and protection as elementary factors for female survival. These might be better provided by brave and risk tolerant males. Signaling good genes by risky behavior makes risk taking attractive based on sexual selection theory.

my experiment: Sitting vis-à-vis a female participant and seeing her picture could already prime males on their gender identity and induce more risk taking, clearly beyond the perceived individual norm.

3.6 Conclusion

This chapter provides first clean experimental evidence that observability of the choice alone in a decision under risk does not affect overall risk taking. That is, in my setting, risk taking is not strategically used as a signal to affect social image.

This directly relates to many settings of individual decision making without strong relationships between the decision maker and the observer. Considering survey interview responses, but also decisions for example in front of doctors or financial advisors have very much in common with the controlled environment of the experiment. In many other and related domains, next to the mere observability of the choice and the opportunity for signaling, other elements of social contexts are relevant. Disentangling the effect of this one basic element is crucial to understand these more complex environments.

One prominent setting in which knowledge of these effects is especially important is group decision making under risk where signals are immediate and oftentimes important. Understanding the signaling values of revealed risk preferences can potentially help to explain inconclusive findings regarding the transmission of individual risk preferences into group risk preferences and decision making (see, e.g., Kugler et al., 2012). The evidence on gender-specific effects of observability depending on the attractiveness of observers (i.e. for example team colleagues in a group setting) further highlights that the gender distribution in teams might have very specific effects. While very recent papers (e.g. Lamiraud and Vranceanu, 2017; Lima de Miranda et al., 2017) discuss the effects of gender composition per se on risky group decision making, more research is needed to understand the mechanisms behind gender and possibly attractiveness specific effects. The finding that males generally increase risk taking when being matched with a female further highlights the importance of understanding the effect of gender identities and matching.

Besides the analysis of treatment effects, my findings clearly establish a large gender difference in norms for risk taking. This closely relates to and helps to explain the usually observed gender differences in actual risk taking — with males taking more risk than females. At the same time, I clearly show that norms do not explain the entire difference in actual risk taking. Rather, males "overshoot" in their risky choices clearly beyond norms. Importantly, this pronounced asymmetry in revealed norm conformity — which surprisingly is independent of observability of choices — is robust to controlling for self-assessed norm conformity preferences. While I measure and establish endogenously emerged norms, it would be interesting to look at norm following behavior with respect to exogenously established norms. Exogenous variation allows to directly measure behavior as a function of different norms.

Interestingly, while the general finding of gender-specific norms in risk taking is very robust, the absolute norm levels provide an insight into the general desirability of risk taking. Norms in the experimental setting describe intermediate levels of risk taking and do not fully support the idea that risk taking overall is desirable. Further research in different domains of risk taking is needed to assess the robustness of this finding.

Chapter 4

Blaming the Refugees? Experimental Evidence on Responsibility Attribution[*]

"You know what a disaster this massive immigration has been to Germany and the people of Germany — crime has risen to levels that no one thought they would ever see."

U.S. president Donald Trump on refugees in Germany[1]

4.1 Introduction

Europe experienced a large inflow of refugees in 2015. As a consequence, a heated debate about whether to tolerate large refugee inflows or whether to instead close borders arose in both the U.S. and Europe. As reflected by the quote of U.S. president Donald Trump at the beginning of this chapter, this discussion focuses to a large extent on whether refugees are responsible for negative outcomes such as rising crime rates, adverse aggregate employment, or poor economic development. Some suggest such responsibility, while others argue against it and accuse their opponents of xenophobic attitudes.[2] Despite the relevance of discrimination against refugees for social and economic outcomes, surprisingly little is known about whether

[*] This chapter is based on joint work with Felix Klimm.

[1] `https://www.washingtonpost.com/news/worldviews/wp/2016/08/16/trump-says-german-crime-levels-have-risen-and-refugees-are-to-blame-not-exactly` (last accessed on March 8, 2018).

[2] Besides the article in The Washington Post referred to in footnote 1, see `https://www.nytimes.com/2016/12/09/world/europe/refugees-arrest-turns-a-crime-into-national-news-and-debate-in-germany.html` (last accessed on March 8, 2018).

natives indeed blame refugees for undesired events, and if so, whether this is caused by statistical discrimination.

We address these questions by implementing a laboratory experiment with refugees who are placed in Munich, Germany. German participants are randomly paired either with another German or a refugee. This allows us to provide clean evidence on differences in responsibility attribution and to shed light on mechanisms of discrimination in this context. More precisely, our subjects receive a positive or a negative income shock. This shock is either due to a random draw or the partner's performance in a real effort task, which took place before the main part of the experiment. If the partner actually is responsible for the shock — unbeknownst to the participant — and his performance was high enough to pass a certain threshold, a positive income shock occurs. In contrast, low performance implies a negative shock when the partner is responsible. After displaying the individual income shocks to the participants, we elicit beliefs about responsibility, i.e., whether the matched partner or the random draw was responsible — our core outcome measure. To investigate whether our results are driven by statistical discrimination, we further elicit beliefs about the partner's performance.[3]

This setup closely relates to many situations in which responsibility has to be assigned while there is uncertainty with respect to the actual cause. Consider, for example, employee evaluations. Increasing or decreasing sales can arise directly from the performance of an employee or be due to general shifts in demand. Layoff or promotion as well as bonus and raise decisions will crucially depend on the supervisor's assessment of this responsibility. However, responsibility attribution is not only essential for an individual's success once in a certain position, it can also critically affect the chances of being hired in the first place. The interpretation of a vita's quality signals — for example whether good performance evaluations refer to the individual's performance or merely to lenient HR policies — but also the assessment of late arrivals to interviews or sickness strongly affect hiring decisions. For all good and bad outcomes, many explanations for responsibility of either the candidate or "nature" are possible. Differing attribution behavior for refugees compared to natives can consequently have a major impact on refugees' labor market integration efforts. To the best of our knowledge, we are the first to investigate such discrimination in responsibility attribution, do so by inviting

[3] In the literature, the term statistical discrimination is most often used for discrimination based on actual differences in characteristics or behavior between different groups (e.g., Fershtman and Gneezy, 2001). Since our subjects have no information about average performances of Germans and refugees, we instead refer to discrimination based on (potentially inaccurate) *beliefs* about different performances as statistical discrimination.

refugees — a highly relevant group for that matter — to the laboratory and implement a new experimental paradigm.

We do not observe discrimination against the outgroup of refugees by blaming them for negative outcomes. Quite the contrary can be inferred from our data. Refugees are treated more favorably than Germans. They are held responsible relatively more often for positive and less often for negative shocks. Actual performance differences and beliefs about the performance of Germans and refugees cannot explain this difference. Hence, statistical discrimination does not explain our result of reverse discrimination. Furthermore, we measure implicit associations towards Arabic names and show that, despite our finding of reverse discrimination, Germans on average have negative implicit associations towards Arabic names. Indicating a positive relationship between implicit attitudes and explicit attribution behavior, subjects with positive implicit associations favor refugees more than subjects with negative associations. In addition, we do not find any evidence for reverse discrimination in a second experiment, in which we assign Germans to artificial in- and outgroups. This shows that our findings from the first experiment are driven by our natural outgroup of refugees and are not a result of our experimental design per se.

Discrimination affects a wide range of social and economic outcomes and comes in many forms and domains. For instance, discrimination can result in disadvantages for education and health related outcomes (e.g., Heckman, 1998; Shapiro et al., 2013; Krieger, 2014) as well as in obstacles to participate in the labor market (e.g., Goldin and Rouse, 2000; Carneiro et al., 2005; Lang and Manove, 2011). This chapter abstracts from these different domains and sheds light on a specific form of discrimination that has not been studied yet — responsibility attribution. Our design also allows us to distinguish between statistical and other types of discrimination and hence to talk about the channels for discriminatory behavior. Other experimental papers have specifically looked at a variety of underlying mechanisms, too.[4] Fershtman and Gneezy (2001) investigate trust and social preferences of ingroup and outgroup members in the Israeli society. Using the investment, dictator, and ultimatum game, they find clear stereotypes associated with different ethnic groups leading to discriminatory behavior. Ockenfels and Werner (2014) provide related evidence on ingroup favoritism. They show that people share more of their endowment in a dictator game when paired with an ingroup member, which indicates an explanation based on social preferences. Similarly, Chen and Li (2009) report increased altruism towards ingroup members in allocation games for different measures of social

[4] For a meta-study on economic experiments on discrimination, see Lane (2016).

preferences, e.g., punishment for misbehavior. In stark contrast to these papers, we do not observe ingroup favoritism or discrimination "against" the outgroup but document reverse discrimination.

We also contribute more generally to the understanding of how responsibility is attributed per se. Bartling and Fischbacher (2011) and Bartling et al. (2015) show that responsibility can be effectively shifted through the delegation of choice and not being pivotal. This evidence indicates that responsibility attribution is malleable and that there is scope for discrimination in attribution behavior.

The much more extensive literature on responsibility attribution in psychology focuses on whether individuals attribute explicit behaviors to internal characteristics or situational factors. Ross (1977) coined the term "fundamental attribution error", which presumes the tendency to underestimate the role of external circumstances when judging others' behavior. Jones and Harris (1967), the original paper to address this issue, investigate subjects' assessments of a writer's private opinion of Fidel Castro. Although subjects know that the writer was randomly told to either praise or criticize Castro in an essay, they rated the writer's opinion as more favorable towards Castro when he had written a pro-Castro text. Hence, subjects wrongfully attributed responsibility for the content of the text to the writer. Pettigrew (1979) relates this bias to ingroup favoritism and hence discriminatory behavior calling it "ultimate attribution error". Negative actions by an outgroup member will more likely be attributed to personal causes, whereas positive actions are more likely attributed to external factors (e.g., luck or "the exceptional case") compared to actions by an ingroup member (for an extensive review see Hewstone, 1990). In contrast to this literature, we do not study whether internal or external factors cause individual behavior. This would correspond, for example, to attributing responsibility for an employee's explicit action. That is, the supervisor knows that the sales manager hired an excellent sales rep but can either attribute this to excellent knowledge of human nature or to mere luck. Instead, we investigate whether an event where the true underlying cause is unknown — who hired the sales rep — is attributed to an individual or something else — the specific sales manager or someone else.

As our subjects are willing to sacrifice part of their payoffs in order not to blame refugees, our finding is not compatible with the standard economic model of purely self-interested agents. Instead, we interpret our results as being in line with theories of economics of identity and motivated beliefs. In such a framework, people care about a positive self-image or generally want to behave according to certain prescriptions pertaining to their identity (Akerlof and Kranton, 2000). These concerns can affect behavior and may lead to self-serving beliefs over behavior of other people (e.g., Di Tella et al., 2015). For our context, it is important that being open and tolerant towards minorities and refugees is part of the social identity of many people, presumably especially in our student sample. Hence, identity concerns might motivate our participants to attribute responsibility more positively towards refugees since blaming refugees is clearly associated with xenophobic attitudes.[5] We also favor this interpretation because in our anonymous laboratory setting, we rule out social image concerns as much as possible.

The remainder of this chapter is structured as follows. Section 4.2 describes the experimental design in detail. Section 4.3 presents our results on responsibility attribution. Section 4.4 is about a robustness experiment that we ran with artificially formed groups. Section 4.5 discusses our main finding and Section 4.6 concludes.

4.2 Experimental Procedures and Design

4.2.1 Procedural Details

We programmed and conducted the experiment with "z-Tree" (Fischbacher, 2007). Germans, 152 students from various fields of study, were recruited using the online recruiting system "ORSEE" (Greiner, 2015). Additionally, 43 refugees were recruited in Munich with leaflets at refugees camps, in front of local registration offices, and in cooperation with the NGO *Social Impact Recruiting* (SIR).[6] Figure D.1 in the Appendix shows an English version of the leaflet.

Because the vast majority of SIR clients and most of the refugees arriving in Germany were male, we decided to restrict the sample to male refugees.[7] Consequently, we also invited only male Germans to have single sex pairs in both ingroups and outgroups such that we did not have to control for potential gender effects. In addition, we wanted our refugee subjects to be of roughly the same age as our other participants. Hence, only refugees between the age of 18 and 29 were invited to participate in the experiment. To have a relatively homogeneous

[5] For instance, see http://www.independent.co.uk/voices/justin-welby-is-wrong-it-is-racist-to-blame-migrants-for-your-fears-about-jobs-and-wages-a6925106.html (last accessed on March 8, 2018).

[6] SIR supports refugees in finding a job by creating a German CV, preparing for interviews, and contacting employers. For further information see http://si-recruiting.org/ (last accessed on March 8, 2018).

[7] See page 21 of the German report of the German Federal Office for Migration and Refugees: http://www.bamf.de/SharedDocs/Anlagen/DE/Publikationen/Broschueren/bundesamt-in-zahlen-2015.html (last accessed on March 8, 2018).

outgroup that represents the majority of refugees in Germany, we only invited Arabic native speakers.[8] To also have a homogeneous ingroup, we only invited native participants with a German sounding name. This ensured that participants assigned to an ingroup member indeed regarded the matched participant as ingroup member.[9]

All 10 experimental sessions took place at the Munich Experimental Laboratory for Economic and Social Sciences (MELESSA) at the University of Munich from August to November 2016. The assignment to the seats in the laboratory made clear that there were two different groups in the experiment. Refugees had to draw a card with a seat number from a bag with the label "Arabic" (in Arabic letters) and Germans from a bag with the label "German" (in German). The cards ensured that the participants were seated in front of a computer screen with instructions in the respective language. Within each group, subjects were randomly assigned to a seat. An English version of the instructions is included in Appendix D.2. Refugees were invited to the experiment half an hour earlier than Germans to make sure they knew what to expect and to check reading and writing proficiency in Modern Standard Arabic.[10] Announcements before and during the experiment were repeated in Arabic by two student research assistants. If necessary, they answered questions by the refugees individually at the subjects' seats. Questions of Germans were answered by the experimenter.

For the main part of the experiment, we formed ingroup and outgroup pairs. As we do not focus on how refugees attribute responsibility, we denote Germans matched with another German as belonging to the *German* treatment (ingroup) and Germans matched with a refugee as belonging to the *Refugee* treatment (outgroup). In order to increase the number of decisions taken by Germans, we matched each refugee with up to two Germans. Group assignment of Germans was random conditional on assigning the same number of Germans to the treatments *German* and *Refugee*.[11] At the beginning of the main part of the experiment, subjects needed to

[8] German Federal Office for Migration and Refugees: `http://www.bamf.de/SharedDocs/Anlagen/EN/Publikationen/Migrationsberichte/migrationsbericht-2015-zentrale-ergebnisse` (last accessed on March 8, 2018).

[9] All refugees indeed had Arabic names. See Section D.1 in the Appendix for a complete list of first names of all participants. At the time of writing this chapter, only roughly 3% of our regular subjects registered for experiments at the Munich Experimental Laboratory for Economic and Social Sciences (MELESSA) had Arabic sounding names. It therefore should have been clear to our German participants that they were matched with a refugee when their partner's name was Arabic sounding.

[10] Some refugees could not participate in the experiment since they indicated that they were not sufficiently able to read and spell.

[11] Only even numbers of German subjects participated in the sessions. If dividing the number of German subjects into two groups of equal size resulted in an odd number, groups were formed such that there were two more Germans matched with a refugee than with another German. For instance, in a session with 18 Germans, 10 of them were matched with a refugee.

enter their first name, which was then shown to their matched partner and enabled all subjects to identify their partner's group affiliation.[12]

At the end of the experiment, the participants answered a questionnaire about socio-demographic characteristics. Thereafter, all subjects were paid privately and earned €12.3 on average, including a fixed payment of €6 for showing up on time. The sessions lasted between 60 and 75 minutes. Each subject participated in one session only.

4.2.2 Experimental Design

Our experiment consisted of two parts. In the first part, subjects received a flat fee of €3 for performing a real effort task. They solved up to eight simple (6×4) jigsaw puzzles (henceforth puzzles) within ten minutes. The puzzles were placed next to the keyboard and were covered by a sheet of paper at every seat. Subjects were asked not to touch the stack until the experimenter had indicated to begin. We chose puzzle motives to be culturally neutral (see Figure D.2 in the Appendix). This real effort task has the advantage of being familiar to participants from different parts of the world. We could not use a computer-based task because many of the refugees were not familiar with working with a personal computer.[13] Furthermore, many Germans arguably would have expected a large performance difference between refugees and Germans. Importantly, at the time of solving the puzzles, participants knew nothing about the content of the rest of the experiment. At the end of part one, the experimenter and student research assistants quietly counted the number of correctly solved puzzles at the subjects' seats.

For the second and main part of the experiment, subjects were randomly paired with another participant in the experiment into ingroup (both subjects Germans) and outgroup pairs (one German and refugee each). Prior to making any decisions in the second part of the experiment, subjects received an income shock. Figure 4.1 illustrates the income generating process. Player A faced a positive or negative income shock. He either received €5 or €5 were subtracted from his experimental earnings.[14] However, player A did not know how this shock came about. With an ex-ante probability of 50%, this shock was due to the performance of player B (the

[12] Loss of anonymity is not a concern despite identification via names. In the questionnaire at the end of the experiment, only 6% of German participants indicated that they knew another participant in their session. Further and more importantly, there is no pair of matched participants where both of the subjects indicated to know somebody else in the session.

[13] In the first three sessions, we asked refugees whether they are familiar with puzzles before the start of the experiment. All of them confirmed.

[14] Subjects knew that their total earnings from the experiment would be a positive amount.

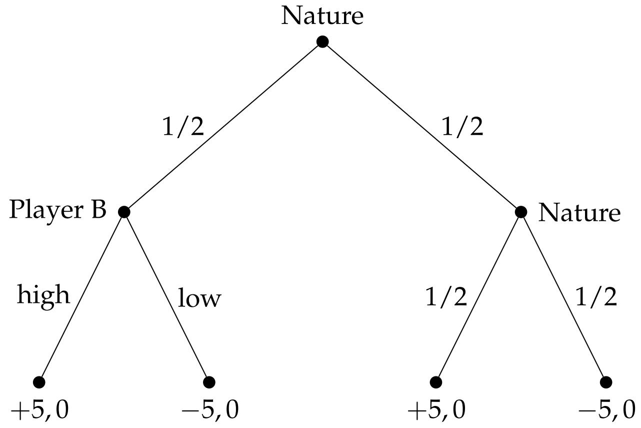


Figure 4.1: Income generating process

matched participant) and otherwise due to nature. If player B's performance was responsible for the income shock, the shock was positive if player B's number of correctly solved puzzles was at least four and negative otherwise. In the case of nature being responsible for the income shock, one of the two shocks was randomly chosen with equal probability. Furthermore, player B's payoff was not affected by whether player A received a positive or negative shock.

The income shock was independently generated for both subjects within each pair, i.e., every subject was player A and player B. Subjects were fully aware of the setup. All participants had to answer four control questions correctly before starting the main part of the experiment to make sure they fully understood the income generating process.

Subsequently, in the first belief elicitation, subjects guessed whether nature or player B's performance caused the income shock and received €5 if their guess was correct. This allows us to identify differences in responsibility attribution to Germans and refugees and is our main variable of interest. In order to get a more precise measure of responsibility attribution, we additionally asked for the participants' confidence in their own guess in a second belief elicitation. More specifically, participants filled out a 9-item choice list with two options (A and B) for each of the nine choices (based on Becker et al., 1964, henceforth BDM). If they chose option A and the respective choice became payoff relevant, they received €5 if their chosen mechanism (in the first belief elicitation) was indeed responsible for the shock (player B or nature). Option A was the same for all nine choices. Option B gave them the chance to receive €5 with probabilities ranging from 10% to 90% in 10% increments. If a participant, for example, expected player B to be responsible in the first elicitation and switched to option B in row seven, he assigned between 60% and 70% probability to the event that player B indeed was responsible.

In addition, we elicited binary beliefs about performance to see whether potential differences in responsibility attribution stem from statistical discrimination. We asked whether subjects believed that the matched player's performance passed the threshold of four solved puzzles or not (again incentivized with €5). Finally, we asked for the probability player A assigned to the matched participant having solved at least four puzzles. Again, subjects faced a (BDM-based) choice list with nine choices between option A, i.e., receiving €5 if the partner's performance was at or above the cutoff, and option B, i.e., receiving €5 with given probabilities ranging from 10% to 90%. Hence, in total, we elicited four incentivized beliefs. At the end of the experiment, in order to prevent hedging, one of these belief questions was randomly chosen for payment and either paid €5 or nothing.

The order of the four belief elicitations, however, was not the same in all sessions. In half of the sessions, we elicited performance beliefs before explaining the income generating process. Hence, in these sessions (henceforth *Uncond*), participants first worked on the puzzles, were then matched with a partner and directly asked for the two (unconditional) performance beliefs regarding the partner (binary choice and choice list). Only then the income generating process was explained and the shock realized. In the other half of the sessions (henceforth *Cond*), (conditional) performance beliefs were elicited after the income generating process had been explained, the shock had realized, and after subjects had attributed responsibility. This allows us — by comparing performance beliefs in the treatments *Uncond* and *Cond* — to examine whether subjects formed distorted or motivated beliefs after observing the shock and attributing responsibility. For instance, assume that a subject attributes responsibility to the partner after observing a negative shock. If this subject is asked about his performance belief, he could justify his attribution behavior by stating low performance beliefs, although he actually thinks that the partner passed the cutoff. Hence, we had a 2×2 treatment design along the dimensions group assignment and task order. Figure 4.2 provides an overview of task orders in the respective treatments.

After these two main parts of the experiment, participants performed the Implicit Association Test (IAT) to measure implicit associations towards Arabic names. Subjects had to assign positive (e.g., "appealing", "love", "cheer") or negative expressions (e.g., "selfish", "dirty", "bothersome") to Arabic or Caucasian names by pressing keys on their keyboard. The IAT score, which indicates positive or negative associations towards Arabic names, is calculated based on response times to sort names to expressions. If a subject needed more time to assign positive expressions and less to assign negative expressions to Arabic compared to Caucasian names, the IAT score is below zero indicating negative implicit attitudes

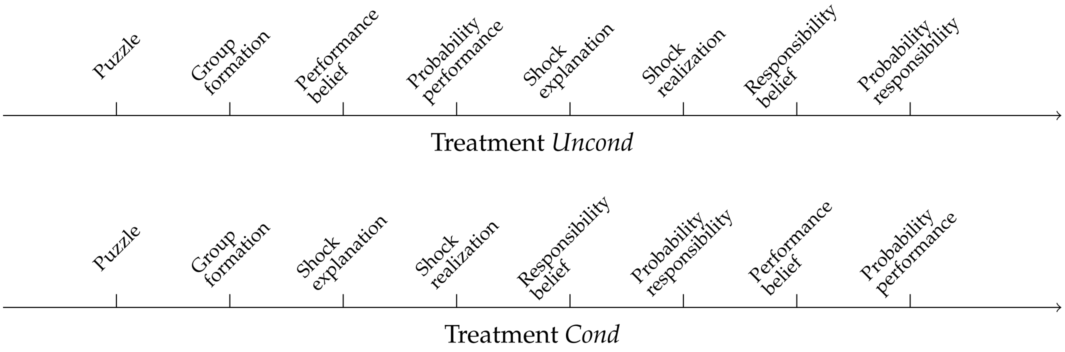


Figure 4.2: Timeline of the experiment

towards Arabic names. This task has been shown to relate to various dimensions of field behavior such as job recruitment (see Greenwald et al. (2009) for a meta study). We used FreeIAT, a free software to run IATs.[15] Subjects were paid €2 for completing the IAT.

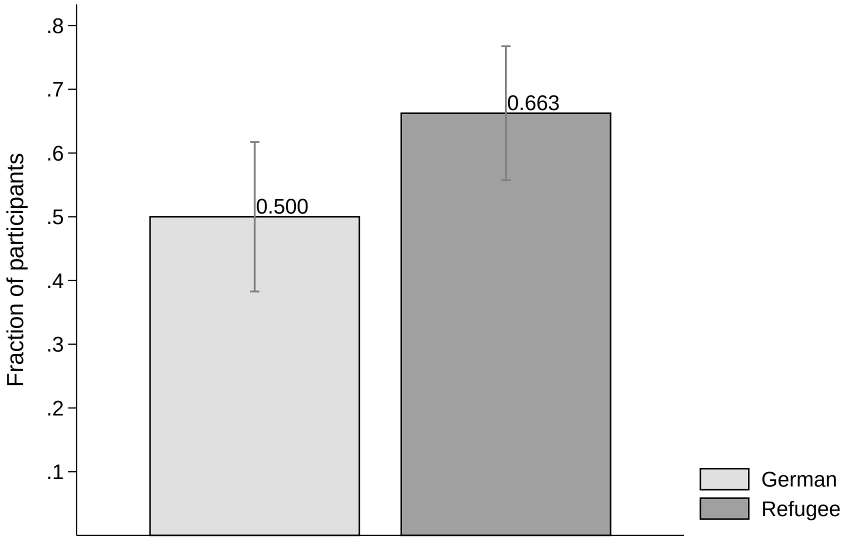
4.3 Results

Our main results on the comparison of responsibility attribution by group assignment over all sessions combined are reported in Section 4.3.1. This abstracts from potential systematic differences between *Uncond* and *Cond*, which we analyze in 4.3.2 separately. Section 4.3.3 presents evidence for heterogeneity using scores from the Implicit Association Test. Section 4.3.4 reports results using the BDM-based probability measures of our main outcome variable and performance beliefs. Unless stated otherwise, all our results in this section consider attribution behavior of our German participants only.

4.3.1 Favorable Responsibility Attribution

Since we test whether our subjects assign responsibility less, equally or more favorably to Germans or refugees, i.e., whether there is discrimination in attribution behavior, we define the binary variable *favorable attribution*. We denote responsibility attribution as favorable if a positive shock occurs and the matched partner is believed to be responsible for the shock. Attribution is also favorable if a negative shock is observed and responsibility is assigned to nature. In contrast, attributing responsibility to the matched partner after a negative shock or to nature after

[15] `http://www4.ncsu.edu/~awmeade/FreeIAT/FreeIAT.htm` (last accessed on March 8, 2018).



Notes: The figure shows *favorable attribution* for both treatments. Error bars indicate 95% confidence intervals.

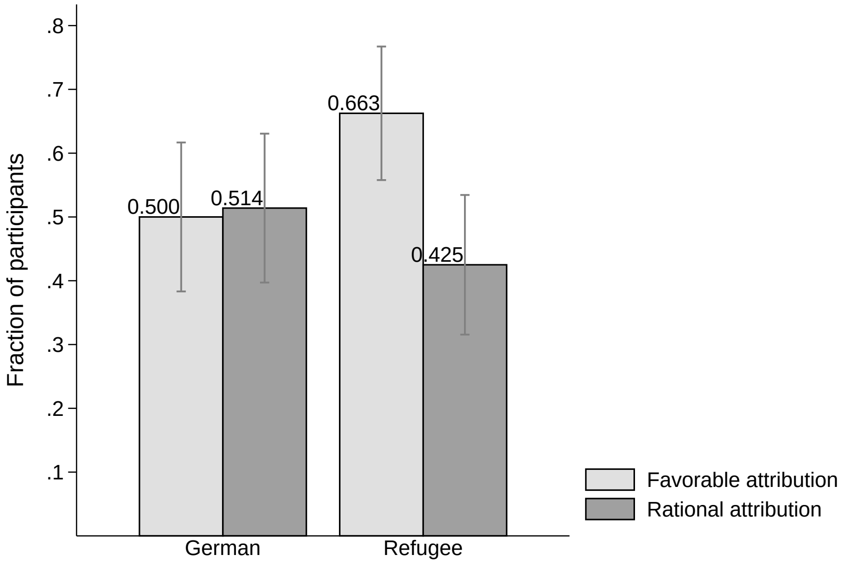
Figure 4.3: *Favorable attribution* depending on group affiliation

a positive shock implies unfavorable attribution.[16] This simplification ignores potential asymmetries in behavior after positive versus negative income shocks. We will show later that our results hold for both shock directions.

Figure 4.3 displays *favorable attribution* by group affiliation. Germans matched with another German ($n = 72$) equally often attribute responsibility favorably and unfavorably. In stark contrast to that, Germans matched with a refugee ($n = 80$) attribute responsibility favorably in roughly two thirds of the cases. This difference in attribution behavior is statistically significant ($p = 0.042$, χ^2-test, two-sided) and evidence for reverse discrimination, i.e., a positive bias towards the refugee outgroup.

Under bayesian updating, *favorable attribution* represents the belief about the matched partner having solved at least four puzzles. Hence, the results displayed in Figure 4.3 could be driven by performance beliefs depending on group affiliation. We would expect more favorable attribution in *Refugee* if subjects believed that refugees are better than Germans in solving puzzles. However, comparing performance beliefs reveals no significant difference. If anything, Germans expect

[16] The intuition underlying this distinction is rational behavior based on bayesian belief updating. Nature and the matched partner are ex-ante responsible with equal probability (*prior*). Given nature is responsible, positive and negative shocks occur with equal probability. Hence, if a participant expects the matched partner to having solved four or more puzzles and thus assigns a probability larger than 50% to this event, he should attribute responsibility favorably (*posterior*). Therefore, under the assumption of bayesian updating, *favorable attribution* captures underlying beliefs about the partner reaching the puzzle cutoff.



Notes: The figure shows *favorable attribution* and *rational attribution* implied by beliefs for both treatments. Error bars indicate 95% confidence intervals.

Figure 4.4: *Favorable attribution* and *rational attribution* implied by beliefs

refugees to perform slightly worse, which renders reverse discrimination even more pronounced. While 43% of Germans matched with a refugee expect the refugee to have solved at least four puzzles, 51% of Germans matched with another German have high performance beliefs ($p = 0.273$, χ^2-test, two-sided).[17] This indicates that the asymmetry in responsibility attribution cannot be rationally based on performance beliefs. In Figure 4.4, we compare actual favorable responsibility attribution (*favorable attribution*) and rational favorable responsibility attribution (*rational attribution*). We define *rational attribution* to be one if the German participant has high performance beliefs regarding the matched partner and zero otherwise. Figure 4.4 shows that while actual responsibility attribution is on average in line with performance beliefs for Germans matched with another German, attribution is clearly more favorable than dictated by performance beliefs for Germans matched with refugees.[18] The difference in *Refugee* is significant ($p < 0.01$, McNemar test, two-sided).[19]

[17] With our sample size, we have 80% power to detect an effect size on the 5% significance level that implies a belief difference of around 22 percentage points. Actual performance differences are much more pronounced. While 47% of the Germans solve four or more puzzles, only 2.3% of the refugees (1 out of 43) reached the performance cutoff. Therefore, statistical discrimination based on actual behavior would imply much more favorable attribution to Germans and thus cannot explain our results.

[18] We cannot analyze refugee behavior by group affiliation since refugees are only matched with Germans. While this is not the interest of this chapter and we do not have adequate power to detect patterns, 51.2% attribute responsibility favorably, whereas only 9.3% of them believe that their partner made the performance cutoff.

[19] These findings are robust to comparing attribution behavior with the individual's own performance. While own performance need not necessarily be a perfect proxy for beliefs regarding

Next, we control for the direction of the income shock. Since the actual performance of refugees was much worse than that of Germans, Germans in *Refugee* observe negative shocks much more often. Hence, more favorable attribution after negative shocks, independent of group affiliation, could explain our results. However, the shock direction does not drive our finding. For both negative and positive shocks, there is a clear asymmetry by group affiliation in terms of how performance beliefs translate into responsibility attribution (see Figure D.3 in the Appendix). Importantly, there is no evidence for blaming the refugees in case of negative shocks. We observe the contrary. Refugees are attributed responsibility much more favorably after a negative shock compared to rational attribution based on performance beliefs ($p < 0.01$, McNemar test, two-sided).

To verify the robustness of our non-parametric results, we run different regression models. The regression framework helps us to further understand attribution behavior by explicitly measuring the effects of beliefs and shock direction on *favorable attribution* while being able to control for observables, too. Table 4.1 reports marginal effects from probit regressions on our binary variable *favorable attribution*.

Column (1) is the parametric equivalent to Figure 4.3 replicating the significant positive effect of being matched with a refugee on *favorable attribution*. This is indicated by the binary variable *Refugee*, which is equal to one if a subject is matched with a refugee and zero otherwise. Column (2), equivalent to Figure 4.4, controls for performance beliefs with *belief high* as binary variable. *Belief high* is equal to one if a subject believes that the partner passed the cutoff and zero otherwise. The effect of group affiliation remains highly significant and sizable. Being matched with a refugee increases the likelihood to attribute responsibility favorably by 19.5 percentage points. The effect in model (2) is slightly larger than in model (1), which is in line with our non-parametric results. As performance beliefs are slightly worse for refugees, controlling for beliefs increases the effect of group affiliation. Reassuringly, high performance beliefs lead to more favorable responsibility attribution. Subjects who believe that the partner passed the cutoff are 37.2 percentage points more likely to exhibit favorable attribution. As motivated above, we include the shock direction in column (3) with *neg shock* as binary variable. It is equal to one if a negative shock occurs and zero otherwise. We find a significant positive effect of negative shocks indicating that participants attribute responsibility generally more favorably after a negative shock. However, this does not alter our

the performance of the other, performance is certainly orthogonal to treatment — unlike beliefs that could potentially be affected by treatment. We will extensively discuss this in Section 4.3.2.

Table 4.1: Favorable responsibility attribution

	Favorable attribution			
	(1)	(2)	(3)	(4)
Refugee	0.160***	0.195***	0.155***	0.146***
	(0.056)	(0.050)	(0.040)	(0.038)
Belief high		0.372***	0.369***	0.375***
		(0.067)	(0.070)	(0.068)
Neg shock			0.164**	0.158**
			(0.064)	(0.064)
Additional controls	No	No	No	Yes
Observations	152	152	152	152
Pseudo R^2	0.020	0.149	0.172	0.179

Notes: Probit regressions on *favorable attribution* reporting average marginal effects. Column (4) includes additional covariates from the questionnaire: age, semester, and number of experiments so far (all insignificant). Robust and clustered (on session level) standard errors in parentheses. Stars indicate significance on the levels: *** p<0.01, ** p<0.05, * p<0.1.

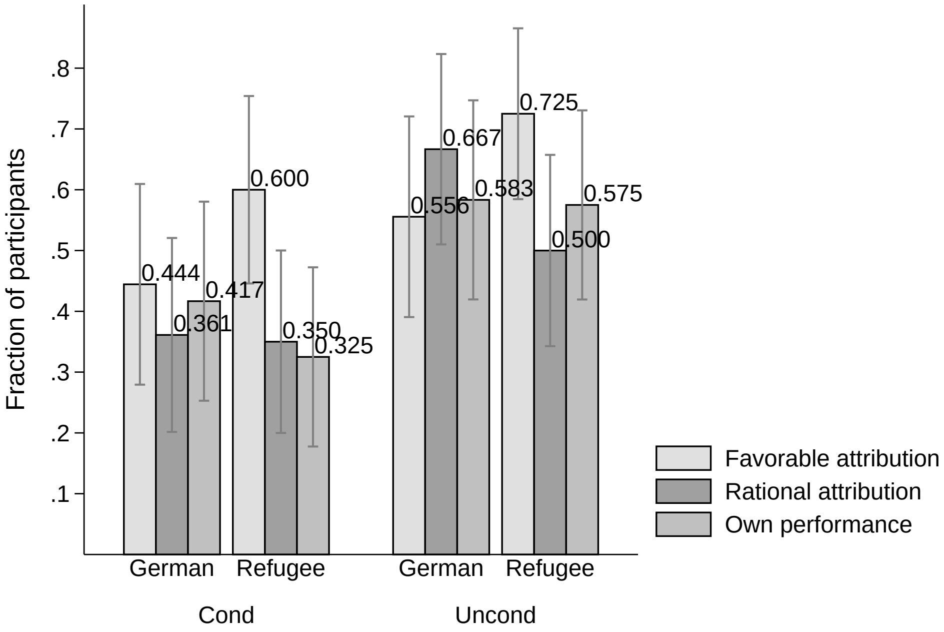
finding regarding group affiliation. Finally, our results are robust to controlling for personal background variables in column (4).

Result 1: *Germans attribute responsibility more favorably to refugees than to other German participants. This cannot be explained by differing performance beliefs and holds for behavior after both negative and positive shocks.*

4.3.2 Unconditional vs. Conditional Beliefs

Participants in our *Cond* treatment were asked to state their performance beliefs after observing the shock and after attributing responsibility. Hence, in order to justify attribution in front of themselves, participants may report distorted beliefs. To quantify this potential distortion, we ran half of the sessions with performance beliefs elicited before shock realization and responsibility attribution (*Uncond*).

To investigate whether performance beliefs are distorted, we relate these beliefs to *own performance* — measured by whether the individual solved at least four puzzles. *Own performance* serves as a benchmark for beliefs regarding others' performances and hence should be the main driver for performance beliefs. This hypothesis is supported by our data. In *German*, 50% pass the puzzle cutoff and 51% expect the matched partner to having done so. In *Refugee*, 45% of Germans solve at least four puzzles and 43% expect that from the matched partner. Only roughly one fourth of our subjects, both in *German* and *Refugee*, does not believe the matched participant to have performed in the same way as they did. Figure 4.5 displays average *own performance*, beliefs in the other's performance (i.e., *rational attribution*),



Notes: The figure shows *favorable attribution*, *rational attribution*, and the fraction of participants reaching the puzzle cutoff (*own performance*) by group affiliation for the treatments *Cond* (left panel) and *Uncond* (right panel). Error bars indicate 95% confidence intervals.

Figure 4.5: *Favorable attribution*, *rational attribution*, and *own performance*

and actual responsibility attribution (*favorable attribution*) by group affiliation and task ordering (*Uncond* vs. *Cond*) separately.[20]

Performance beliefs cannot be distorted by knowledge about our responsibility attribution task in *Uncond*. In this case, displayed in the right panel of Figure 4.5, Germans expect other Germans on average to perform slightly better than themselves and refugees to be slightly worse. Compared to that, performance beliefs seem distorted in *Cond*. Beliefs of ingroup members are slightly lower than *own performance*, while they are higher for Germans in *Refugee*. On average, Germans matched with a refugee in *Uncond* are 7.5 percentage points less likely to believe in the performance of their partner compared to their own performance. However, German outgroup participants in *Cond* are 2.5 percentage points more likely to believe in the performance of the refugee than in their own. Hence, the difference in the differences between *own performance* and performance beliefs over the two treatments for subjects in *Refugee* is 0.1. This corresponds to a positive belief distortion in favor of refugees once knowing the income generating process. Performing the same difference in differences calculation for subjects in *German*,

[20] This reveals that randomization was not successful with regard to puzzle performance. A significantly larger fraction of subjects in *Uncond* pass the performance cutoff than subjects in *Cond* ($p < 0.01$, χ^2-test, two-sided). Table D.1 in the Appendix shows the sample balance.

we find a difference in differences of 0.14 that shows worse performance beliefs in *Cond* (negative distortion against other Germans). While this 24 percentage points difference in distortion between *German* and *Refugee* is considerate, it is insignificant ($p = 0.151$, t-test, two-sided).[21]

Hence, under the assumption of unbiased beliefs in *Uncond* our findings from Section 4.3.1 provide a lower bound for the extent of reverse discrimination. The results from this section indicate that true underlying beliefs in *Cond* could actually be worse for refugees and better for other Germans than stated in the belief elicitation. This would increase the asymmetry between rational and actual responsibility attribution beyond what we measure in Section 4.3.1.

Result 2: *We find no significant evidence for subjects stating distorted beliefs. However, if anything, the results point towards favorably distorted beliefs with respect to refugees, suggesting that the results from the pooled sample (Section 4.3.1) constitute a lower bound for reverse discrimination.*

The assumption in this section is that beliefs in *Uncond* are unbiased. This seems reasonable since participants are unaware of the rest of the experiment in this treatment when stating their guess about their partner's performance. However, unconditional performance beliefs regarding refugees could already be distorted upwards such that true underlying performance beliefs would actually be lower. If this was the case, our overall finding of reverse discrimination would again be a lower bound of the true discrimination. Given true performance beliefs, the difference between these beliefs and responsibility attribution would be larger than the one we find with stated beliefs. In contrast to that, performance beliefs could also be biased downwards and explain our result of reverse discrimination. This, however, seems very unlikely because it would imply discrimination at the level of performance beliefs — by stating lower than actual beliefs about performance for refugees — and, to the contrary, reverse discrimination at the level of responsibility attribution. Furthermore, it is implausible that participants have such extremely inaccurate beliefs given that refugees actually perform very poorly in the real effort task.

To account for the possibility of biased performance beliefs, we substitute these beliefs by own performance to check the robustness of our main findings. Table D.2 in the Appendix reports results from regressions replicating Table 4.1

[21] This calculation is equivalent to regressing the individual difference between *rational attribution* (performance beliefs) and *own performance* in an OLS estimation on *Refugee*, *Cond*, and their interaction term *Refugee*×*Cond*. The interaction term shows the 24 percentage points distortion for Germans matched to refugees once they know the income generating process.

while using each participant's number of correctly solved puzzles as explanatory variable instead of his performance beliefs.[22] The results for *Refugee* from all models are strikingly similar to the ones from Table 4.1, which renders our finding of reverse discrimination robust to performance belief distortions.

4.3.3 Implicit Associations

The key personal characteristic that we elicit and correlate with attribution behavior relates to implicit associations. The IAT measures people's relative implicit associations towards a specific group compared to a baseline group. In our case, it is a measure of associations towards Arabic names relative to Caucasian names.[23] A positive test score implies relatively positive associations towards Arabic names, while a negative score indicates the opposite.

Overall, the results from the IAT are in line with ingroup favoritism. While 72% of Germans have a negative IAT and hence relatively more negative associations towards Arabic names, this is the case for only 12% of the refugees ($p < 0.01$, χ^2-test, two-sided).[24]

Importantly, implicit attitudes have predictive power for explicit discrimination behavior. People with negative IAT scores favor refugees less with regard to responsibility attribution. 83% of Germans with a positive IAT in *Refugee* attribute responsibility favorably, while only 59% with a negative IAT do so. This difference is significant ($p = 0.034$, χ^2-test, two-sided).

To test the correlation between implicit associations and *favorable attribution* when holding other variables constant, we further apply a regression framework. We control for own performance rather than for performance beliefs since beliefs might have been distorted, and this potential distortion is likely to be related to the IAT score. For instance, subjects who are in general favorable towards refugees are likely to have a positive IAT score *and* possibly upwards biased beliefs about a refugee's performance.

[22] Alternatively, using a binary variable for whether the respective participant solved at least four puzzles does not change the significance of the *Refugee* or *neg shock* indicators.

[23] Arabic names are Hakim, Sharif, Yousef, Wahib, Akbar, Muhsin, Salim, Karim, Habib, and Ashraf, and Caucasian Names are Ernesto, Matthais, Maarten, Philippe, Guillame, Benoit, Takuya, Kazuki, Chaiyo, and Marcelo. Positive associations are Excellent, Cheer, Delight, Joyous, Excitement, Cherish, Friendship, and Beautiful, and negative associations are Hate, Pain, Gross, Failure, Rotten, Humiliate, Sickening, and Horrible. The IAT for Arabic names can be taken online by visiting `https://implicit.harvard.edu/implicit/selectatest.html` and selecting "Arab-Muslim IAT".

[24] The same holds true for average values. The average IAT score for Germans is -0.199, while the average for refugees is 0.215. This difference is again highly significant ($p < 0.01$, Mann-Whitney U-test, two-sided).

Table 4.2: Favorable responsibility attribution depending on IAT

	Favorable attribution		
	Refugee (1)	*German* (2)	pooled (3)
IATneg	–0.272**	0.089	0.084
	(0.114)	(0.159)	(0.162)
# correct puzzles	0.077**	0.104***	0.092***
	(0.036)	(0.024)	(0.020)
Refugee			0.395***
			(0.146)
IATneg × Refugee			–0.343*
			(0.186)
Neg shock			0.123**
			(0.058)
Additional controls	No	No	Yes
Observations	80	72	152
Pseudo R^2	0.076	0.071	0.114

Notes: Probit regressions on *favorable attribution* reporting average marginal effects. Column (1) and (2) include only the sample of outgroup and ingroup participants respectively. Column (3) includes the entire sample and additional covariates from the questionnaire: age, semester, and number of experiments so far (all insignificant). Robust and clustered (on session level) standard errors in parentheses. Stars indicate significance on the levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 4.2 reports probit regressions of *favorable attribution* on $IATneg$, which is equal to one if the IAT score is negative (negative associations towards Arabic names) and zero otherwise (positive associations towards Arabic names), and own performance. Column (1) includes subjects in *Refugee* only. As indicated by our non-parametric results discussed before, we observe a large and significant correlation between having a negative IAT score and responsibility attribution for Germans matched with refugees. Those that have negative implicit association towards Arabic names are 27.2 percentage points less likely to attribute responsibility favorably to their matched Arabic partner. Column (2) shows that a negative IAT score has no effect on favorable responsibility attribution in *German*.[25] Column (3) reports regression results for the entire sample with additional controls and an interaction of the IAT score and our treatment. The marginal effect of the interaction term of –0.343 indicates that a negative IAT value has a more negative effect on *favorable attribution* for participants in *Refugee* compared to participants in

[25] Ex-ante, it is not obvious why the effect of implicit associations should be stronger in *Refugee* compared to *German*. The effects in the two different groups should go into opposite directions, but there is no apparent reason why positive implicit associations towards one's ingroup should not lead to more favorable attribution towards these ingroup members. We interpret this finding in the following way. First, it is plausible that associations regarding the more salient outgroup determine the IAT scores. In that case, the IAT score should not predict behavior towards the ingroup. Second, we used a standard version of the IAT measuring associations towards Arabic names. This version uses a wide range of Caucasian names in the baseline group. Hence, attitudes towards German participants might not be perfectly captured by this IAT. This again supports the idea that our IAT scores predominantly represent implicit associations towards Arabic names and not German names.

German. Further, we see that IAT scores (*IATneg*) do not affect *favorable attribution* in *German*. In contrast, having a negative IAT score decreases the likelihood to attribute responsibility favorably by 25.9 percentage points in *Refugee* ($p = 0.030$, F-Test for *IATneg* + *IATneg* x *Refugee*).[26] These results confirm our findings from column (1) and (2). In addition, the coefficient of *Refugee* shows that our result of reverse discrimination is mainly driven by participants with a positive IAT score since the treatment difference is insignificant for subjects with a negative IAT score ($p = 0.390$, F-Test for *Refugee* + *IATneg* x *Refugee*).

However, in nonlinear models including interaction terms, interpreting the marginal effect of the interaction term is flawed (Ai and Norton, 2003) and hypothesis testing can be misleading (Greene, 2010). This is due to the fact that, in nonlinear models, the marginal effect of the interaction term is not the same as the cross derivative with respect to both interacted variables (the interaction effect). In order to account for this problem, we compute the predicted values of *favorable attribution* split up along two dimensions — having a positive or negative IAT score as well as being in *Refugee* or *German*. We calculate the difference in differences of these four groups, which reflects the interaction effect in models including interaction terms with two binary variables. We find that the effect of a negative IAT score on *favorable attribution* is 36.19 percentage points lower in *Refugee* than in *German*.[27] Since this estimate is very close to the marginal effect of our interaction term in column (3), –0.343, the mistake induced by interpreting the marginal effect of the interaction term as interaction effect is negligible in our estimation.

Result 3: *Implicit associations directly relate to explicit behavior. Reverse discrimination is mainly driven by subjects with positive implicit association towards Arabic names.*

4.3.4 Alternative Measures of Responsibility Attribution and Performance Belief

By using the binary measure of responsibility attribution and by enforcing a choice, we treat more or less indifferent participants the same as those who have a clear opinion about responsibility. In this section, we want to check whether these indifferent people could be driving our results. For this purpose, we define two new

[26] All results from Table 4.2 are qualitatively unchanged if we use the continuous variable of the IAT instead of the binary version. Only the F-Test for *IAT* + *IAT* x *Refugee* in the interaction model becomes borderline insignificant ($p = 0.143$).

[27] Estimation of the difference in differences in predicted values can be found in Appendix D.5.

Table 4.3: Contingency table for binary vs. BDM choices

	1	2	3	4	5	6	7	8	9	10
Responsibility:										
(1) Binary favorable: Switchpoint	0	2	0	3	21	31	18	11	2	1
(2) Binary unfavorable: Switchpoint	3	2	7	14	16	14	2	4	1	0
Performance:										
(3) Binary positive: Switchpoint	0	0	0	3	10	14	23	12	7	2
(4) Binary negative: Switchpoint	3	5	12	21	22	11	2	2	3	0

variables called (i) *responsibility switchpoint* and (ii) *performance switchpoint* based on the two BDM belief elicitations. These variables indicate probabilistic confidence in (i) the partner being responsible for a positive shock (conditional on observing a positive shock) or the partner *not* being responsible for a negative shock (conditional on a negative shock) and (ii) the partner having solved four or more puzzles. A higher value of *responsibility switchpoint* hence indicates a more favorable attribution. A higher value of *performance switchpoint* indicates a higher confidence in the matched partner having solved four or more puzzles. Both variables, corresponding to the nine-item choice list, are measured in 10 percentage point steps. Thus, a switchpoint of one corresponds to assigning 0-10% probability to the event and a switchpoint of 10 corresponds to 90-100%.

The average of *responsibility switchpoint* by group affiliation highlights a clear difference to the findings from the binary measure. With an average switchpoint of 5.65 and 5.56 in *German* and *Refugee* respectively, there is no difference in responsibility attribution by group affiliation. Is this difference in response behavior driven by outliers, by indifferent participants, or do we observe other inconsistencies? To understand consistency between the binary and BDM belief elicitation, Table 4.3 displays a contingency table for these choices reporting combinations of binary choices and BDM choices. Row (1) and (2) refer to responsibility consistency, given that in the binary choice responsibility was assigned favorably (1) or unfavorably (2). Rows (3) and (4) display consistency for performance beliefs depending on the binary performance belief elicitation.

If consistent, row (1) subjects should have a *responsibility switchpoint* above five and thus assign more than 50% probability to the "favorable" event. Those around the threshold are close to indifference (highlighted in dark gray), while those in light gray choose clearly inconsistently. For instance, assigning only 30-40% probability to the matched partner being responsible for a positive shock but before indicating to believe the partner is responsible — as is the case for the three participants highlighted in row (1) in the fourth column — is not consistent. The table shows that a substantial fraction of participants reports probabilities around the indifference threshold of 5 and 6, indicating that indifference could help to explain our difference in non-parametric results between our binary and BDM responsibility measures.

Moreover, it seems that some subjects did not understand the BDM choice list. Twelve participants strongly violate consistency when asked about responsibility, and ten participants do so for the performance beliefs. In line with the notion of misunderstanding, it takes these participants also clearly longer to make these BDM choices. Those being inconsistent for the performance questions take on average 24 seconds longer (out of 90 seconds they have) for this BDM, while they are 2.5 seconds faster than the consistent subjects for the binary performance belief (both comparisons do not exceed a p-value of 0.037, Mann-Whitney U-test, two-sided). Directionally, the same is true for the responsibility questions. Participants that are inconsistent spend on average 3.5 seconds longer on answering the BDM version of the question, while they are almost 5 seconds faster for the binary responsibility question.[28] Hence, in the following regression analysis, we exclude those participants that misunderstood the elicitation procedure.

Table 4.4 reports results from regressions including the alternative measures of the responsibility and performance beliefs. Again, adding performance beliefs as controls is crucial since even same levels of responsibility attribution across group affiliations in the BDM can imply reverse discrimination. This would be the case if Germans had higher performance beliefs for other Germans than for refugees. The two-limit Tobit specification of column (1) includes *responsibility switchpoint* as dependent variable and the binary performance belief as control variable. We also control for the direction of shocks. The coefficient for *Refugee* is positive as before but now insignificant ($p = 0.393$), as opposed to in Table 4.1. Hence, also when controlling for beliefs and shock direction, we do not see a statistically significant positive effect of being matched with a refugee on responsibility attribution implied by the BDM elicitation. Using the binary responsibility measure and including non-binary performance beliefs in column (2), however, results in similar findings as in Table 4.1. The effect of *Refugee* is significantly positive. With both switchpoint variables instead of their binary counterparts in column (3), we again observe no significant reverse discrimination.

[28] When designing the experiment, we decided against including control questions to ensure understanding of the BDM — as is often done for these complex elicitation procedures. We did not want to treat refugees and Germans differently because that by itself could have induced a treatment effect, and explaining the BDM in depth to the refugees would presumably have taken very long.

Table 4.4: Favorable responsibility attribution with continuous measures

	Responsibility attribution		
	(1)	(2)	(3)
Refugee	0.216	0.119**	0.181
	(0.318)	(0.052)	(0.306)
Belief high	0.911***		
	(0.262)		
Switchpoint cutoff		0.113***	0.356***
		(0.011)	(0.090)
Neg shock	0.333	0.172***	0.339
	(0.226)	(0.047)	(0.258)
Constant	4.265***		2.590**
	(0.959)		(1.122)
Additional controls	Yes	Yes	Yes
Observations	140	142	131
Pseudo R^2	0.032	0.197	0.064

Notes: Column (1) and (3) report two-limit Tobit regressions on *responsibility switchpoint*. Column (1) includes the binary performance belief indicator (*belief high*), whereas column (3) uses *performance switchpoint*. Column (2) reports average marginal effects from a probit model explaining binary responsibility attribution (*favorable attribution*) with *performance switchpoint*. Subjects that clearly misunderstood the BDM elicitation are dropped. All columns include additional covariates from the questionnaire: age, semester, and number of experiments so far. Robust and clustered (on session level) standard errors in parentheses. Stars indicate significance on the levels: *** p<0.01, ** p<0.05, * p<0.1.

How can we explain the insignificant coefficients for the specifications using *responsibility switchpoint*? First, even when excluding inconsistent subjects, we still expect some misunderstanding in the BDM. Especially the BDM for responsibility attribution is rather difficult to grasp. This increases noise in the data and makes detecting the effect more difficult.

Second, indifference or only weak binary preferences are important. These weak inconsistencies, however, are still highly asymmetric. If only indifferent subjects were responsible for the different results of Table 4.1 and Table 4.4, a substantial fraction of Germans matched with a refugee would have to be indifferent and attribute favorably in the binary elicitation, while those in *German* and indifferent would attribute unfavorably. This still is a clear form of reverse discrimination — it would only be less costly than if it was not driven by indifference. Similarly, other types of inconsistencies and choice reversals that we cannot categorize could drive the difference in our findings. We do have some evidence for this type of strong asymmetry in inconsistencies for the responsibility beliefs. Of the twelve participants being strictly inconsistent (light grey in upper panel of Table 4.3), five are subjects in *German* and all of these switch from unfavorable binary attribution to favorable switchpoint attribution. In stark contrast to that, of the seven strictly inconsistent Germans in *Refugee*, five switch from favorable binary attribution to unfavorable probabilistic attribution. Despite the very low number of observations, this is a significant difference ($p = 0.028$, Fisher's exact test, two-sided). The same is true for weak inconsistencies. For this purpose, we define those with a switchpoint of 5 in row (1) of Table 4.3 and a switchpoint of 6 in row (2) as being weakly inconsistent. In *German*, 12 out of 19 inconsistent subjects change from unfavorable binary to favorable switchpoint attribution, while only 9 out of 28 do so in *Refugee*. This difference is again significant ($p = 0.043$, Fisher's exact test, two-sided).

Third, with the BDM it might be more vague what the "right" thing to do is. If reverse discrimination is driven by self-image and identity concerns, the BDM elicitation procedure might well not make the identity prescriptions as clear as the binary elicitation. For the binary responsibility attribution it is obvious what the subjects should do if they do not want to blame someone. With probabilities this is less clear.

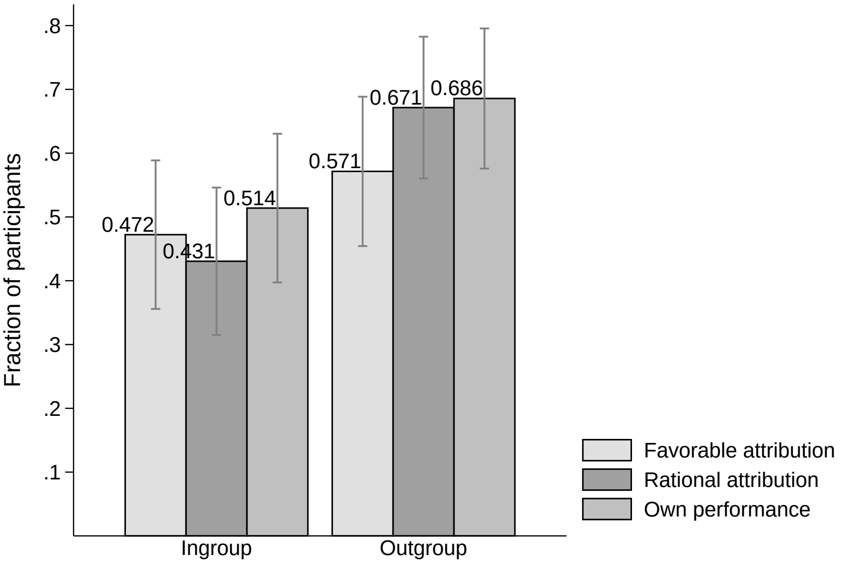
In summary, we get directionally very similar results with the non-binary belief elicitations. However, these results are weaker. Increased noise, indifference, systematic inconsistencies, and possibly increased opagueness of the normative prescription can help explaining this difference. While this provides some additional insights into individual decision making, it does not change our main message: We observe strongly asymmetric behavior leading to reverse discrimination and more favorable treatment of refugees.

Result 4: *The evidence for reverse discrimination is weaker when considering non-binary beliefs. The asymmetry in behavior explaining this difference, however, again points to strongly group-specific patterns.*

4.4 The *KleeKandinsky Experiment*

In an additional experiment, we only invited participants from the regular subject pool and applied a minimal group paradigm to analyze whether our result of reverse discrimination is a general result for in- and outgroups or whether it stems from our specific groups in the *Refugee Experiment*. Since groups were formed based on preferences for paintings of the artists Klee and Kandinsky, henceforth we call this experiment *KleeKandinsky Experiment* (and our main experiment *Refugee Experiment*). With a total of 142 subjects, we ran six sessions in August 2016. Subjects earned €13.85 on average, including a €6 fixed payment for showing up on time. Each subject participated in one session only.

Procedures differed only in dimensions explicitly catered to refugees mentioned in Section 4.2. Hence, there was no gender restriction for participation, no Arabic



Notes: The figure shows *favorable attribution*, *rational attribution*, and *own performance* for the *KleeKandinsky Experiment*. Error bars indicate 95% confidence intervals.

Figure 4.6: *Favorable attribution*, *rational attribution*, and *own performance* in the *KleeKandinsky Experiment*

announcements were made, participants only drew seat numbers from one bag, and group affiliation was communicated via group names (Klee or Kandinsky) instead of first names. Moreover, every subject is matched with only one other subject. Subjects in the *Ingroup* treatment ($n = 72$) are matched with a subject of the same group, while we match subjects of different groups with each other in the *Outgroup* treatment ($n = 70$).

We employ a modified version of the minimal group paradigm used by Chen and Li (2009). Subjects evaluate paintings of the artists Paul Klee and Wassily Kandinsky. Five pairs of paintings containing each a painting of Klee and Kandinsky are shown. For each pair and without knowing the artist of the paintings, participants have to decide which of the two paintings they prefer. Based on a median split in artist preferences, subjects are assigned to the Klee or Kandinsky group. This assignment procedure takes place at the very beginning of the experiment.

Contrary to the results of the *Refugee Experiment*, responsibility attribution is not affected by group affiliation of the matched partner in the *KleeKandinsky Experiment*. Figure 4.6 shows that attribution is more favorable in the *Outgroup* treatment (light gray bars), however, this can be explained by beliefs about performance. If anything, given rational attribution (dark gray bars), subjects in *Outgroup* should attribute responsibility even more favorably and subjects in *Ingroup* even less

Table 4.5: Favorable responsibility attribution (*KleeKandinsky Experiment*)

	Favorable attribution			
	(1)	(2)	(3)	(4)
Outgroup	0.099**	−0.006	0.023	0.010
	(0.038)	(0.057)	(0.061)	(0.056)
Belief high		0.392***	0.336***	0.345***
		(0.079)	(0.085)	(0.079)
Neg shock			0.258***	0.248***
			(0.057)	(0.055)
Additional controls	No	No	No	Yes
Observations	142	142	142	142
Pseudo R^2	0.007	0.141	0.206	0.224

Notes: Probit regressions on *favorable attribution* reporting average marginal effects. Column (4) includes additional covariates from the questionnaire: age, gender, semester, and number of experiments so far. Robust and clustered (on session level) standard errors in parentheses. Stars indicate significance on the levels: *** p<0.01, ** p<0.05, * p<0.1.

favorably. As can be seen from the intermediate gray bars at the very right, the difference in performance beliefs can be explained by differences in individual performances.[29]

Table 4.5 shows the same regression analysis as Table 4.1 does for the *Refugee Experiment*. As we already observed in Figure 4.6, in the baseline regression in column (1), it seems as if there is some form of reverse discrimination. This positive effect of being matched with an outgroup member is not robust to controlling for beliefs. The effect of group affiliation becomes a rather precise zero when we control for performance beliefs (see column (2)). In column (3), we include a dummy for the direction of the shock. As in the *Refugee Experiment*, we find that subjects assign responsibility more favorably after negative shocks. Since shocks were evenly distributed across group affiliation in the *KleeKandinsky Experiment*,[30] we did not expect to observe an effect on the *Outgroup* coefficient. This is confirmed by column (3). Adding more controls in column (4) does not alter the results. Also note that effect sizes of *belief high* and *neg shock* are quite similar to the ones from the *Refugee Experiment*. Overall, this demonstrates that our finding of reverse discrimination is a result of our natural group assignment in the *Refugee Experiment* and not a general result in our experimental design.

[29] Even though individual performances should be orthogonal to treatment assignment, we still see pronounced differences. Participants in *Outgroup* solve 4.06 puzzles on average, while participants in *Ingroup* only solve 3.36 puzzles on average. This difference is significant ($p < 0.01$, Mann-Whitney U-test, two-sided). Table D.3 in the Appendix reveals that the sample is balanced otherwise. There are no differences with respect to age, number of semester, and number of experiments so far.

[30] 57% of subjects in *Outgroup* and 51% in *Ingroup* receive a positive income shock.

Result 5: *There is no evidence for reverse discrimination with artificially assigned groups.*

4.5 Discussion

In this section, we discuss several explanations for why we find reverse discrimination in our setting. As we can rule out statistical discrimination, taste-based discrimination is a first natural candidate to look at. Subjects are willing to pay a price to attribute responsibility favorably towards refugees. In our context, taste-based discrimination would imply that this is the case because they have some sort of preference for this group. This explanation seems, however, unlikely. First, participants matched with refugees do not affect refugees' payments by attribution behavior. Hence, outcome based tastes cannot play a role for choices. Second, the same holds for tastes for interaction. Participants never interact with their matched partner, and responsibility attribution choices do not affect the degree of interaction. Third, the results of the IAT reveal that Germans on average have negative implicit associations towards Arabic names. Lastly, taste-based explanations also stand in stark contrast to the literature on ingroup favoritism.[31]

The finding of favoring refugees might also be caused by the desire to be seen as a good person by others. Social image concerns have been shown to be an important motivation for decisions in various settings where behavior is publicly observable (e.g., Andreoni and Bernheim, 2009; Ariely et al., 2009; Lacetera and Macis, 2010). In our setting, however, subjects take their decisions completely anonymously, which is common knowledge to our subjects.[32] Similarly, our experimental results could be affected by experimenter demand effects (EDE), that is, in our case, by norm conformity pressure. While we cannot completely rule out such effects, some considerations render an interpretation of our results predominately based on this pressure unlikely. Participants could indeed perceive favorable attribution towards refugees as the appropriate behavior in the eyes of the experimenter. However, EDE should have also affected behavior of our subjects in *German* (*Refugee Experiment*) and in the *KleeKandinsky Experiment*. This applies, in particular, to the *KleeKandinsky Experiment* because the minimal group paradigm is artificial (as opposed to a more natural identification based on first names). This should make EDE even more likely

[31] See, e.g., the literature review by Hewstone et al. (2002).

[32] At the beginning of the experiment, we guarantee our subjects that all of their decisions will be analyzed anonymously. The experimenter is not present in the laboratory while decisions are taken. In addition, it is not possible to infer decisions directly from the level of payoffs (which is observed by the research assistant privately handing out the earned money).

as subjects will think more about the purpose of the study in light of the artificiality (Zizzo, 2010). In these treatments though, beliefs about performance do not differ from favorable attribution. That is, behavior is in line with rational responsibility attribution leaving the *Refugee* treatment as the only biased sample.[33] Importantly, both social image concerns and norm conformity pressure — if they occurred in our experiment — are likely to more strongly occur in non-anonymous decision environments. Compared to actual behavior in the field, our results would then provide a lower bound.

In addition to being motivated by appearing as a good person in front of others, one could be motivated by appearing as a good person in front of oneself. Keeping up a certain identity, a person's self-view, oftentimes conflicts with profit maximizing behavior and explains departures thereof in different economic spheres (e.g., Akerlof and Kranton, 2000; Mazar et al., 2008). This can also lead to deliberately distorted beliefs, i.e., motivated beliefs (e.g., Di Tella et al., 2015; Gneezy et al., 2016; Grossman and Van Der Weele, 2017). Agents with such motivated beliefs have a positive willingness to pay for keeping up a specific self-image. We find that our subjects make choices that are in line with behaving "politically correct". Especially with regard to our student subject pool, it seems to be plausible that being open and tolerant towards minorities is part of our subjects' identity. In order to keep up a positive self-view, they seem to be reluctant to blame refugees. There is some evidence from psychology supporting such reasoning. Dutton (1973) finds that middle-class Canadian whites donate more when the solicitor is of black or Indian ethnicity as compared to when the solicitor is white. With donors perceiving black people and Indians to be targets of discrimination, the author interprets the results as supportive evidence for a specific type of revealed reverse discrimination. In addition, Byrd et al. (2015) show that liberal and moderate whites favor black over white politicians in an artificial setting. Participants read political speeches and saw a picture of either a black or a white person who was supposed to have given the speech. Among other outcome variables, more participants indicated that they would vote for a black politician. The evidence of these studies suggests that actively avoiding explicit discrimination might be part of the identity of politically liberal and moderate middle-class people to which the majority of our subjects should belong to. This explanation is also in line with the stronger results for the binary responsibility beliefs compared to the finer-graded probability beliefs. In the former elicitation, it is absolutely clear what the "good" or "bad" thing to do is. Hence,

[33] At the end of the experiment, we further ask for non-incentivized verbal explanations for behavior. We do not have a single statement that could be related to EDE.

our subjects try to avoid taking the bad action towards the refugees.[34] In contrast, "good" and "bad" is not as clearly defined for the latter elicitation procedure. We therefore argue that motivated belief formation is the most plausible explanation for our main result.

4.6 Conclusion

We experimentally study responsibility attribution for negative and positive income shocks. In particular, we ask whether there is asymmetric attribution of responsibility, depending on whether a German participant is matched with another German or a refugee. In our setting, there is imperfect information regarding the source of the shock. It can either be due to a random draw or due to the performance of the matched participant. This experimental paradigm is an abstract setting related to several environments in the field. Oftentimes, there is uncertainty with regard to what or who is responsible for a certain outcome. Group-specific behavior can thus strongly impact the lives of different societal groups. Prominent examples relate to labor market settings, where people that are discriminated against in responsibility attribution will be strongly disadvantaged. This might occur in the hiring process or at later stages in promotion, job assignment, or bonus decisions. Our study also relates on a more aggregate level to how developments and outcomes for the society as a whole might be related to groups of people. Recent examples are the strongly debated effects of refugees on crime, economic prospects of societies, and cultural developments. The negative shock of rising crime rates in some European countries might be indeed (in part) caused by the influx of refugees (as suggested by Donald Trump's quote at the beginning of this chapter) but could also be due to many other factors.

Surprisingly and contrary to the literature, which predominantly documents ingroup favoritism, we find no discrimination against refugees in responsibility attribution. Importantly, refugees are clearly not blamed for negative events but less often held responsible when a negative shock occurs. That is, we observe reverse discrimination. German participants generally attribute responsibility to refugees more favorably as compared to other Germans. We put forward an explanation based on identity concerns and motivated beliefs. Participants want to

[34] We further assume that there is a clear difference in moral prescriptions between stating performance beliefs and responsibility beliefs. While it should be perceived a good (bad) thing to praise (blame) for responsibility, there should be no such moral connotation to stating mere performance beliefs. This is why we expect to observe distorted (discriminating) responsibility attribution and rather unbiased performance beliefs.

view themselves as non-xenophobic and tolerant and hence distort attribution as to not conflict with this identity. This belief distortion consequently leads to reverse discrimination. Comparing these results to an experiment with artificial group assignment, we show that our results are not a general result for in- and outgroups but rather depend on our specific sample. This lends support to the idea that the refugee sample indeed induces identity concerns. Furthermore, implicit associations of our German participants towards Arabic names are negative, while responsibility attribution is irrationally favorable on average. This suggests that favoring refugees is a conscious choice in our experiment. Moreover, we find that subjects with more positive associations towards Arabic names attribute responsibility more favorably to them. Implicit associations — which are correlated with important field behavior such as hiring decisions — thus predict responsibility attribution in a meaningful way.

The evidence for reverse discrimination towards refugees together with our results on potential mechanisms provide fruitful avenues for future research. First, while we find strong evidence in the domain of responsibility attribution, our study cannot draw conclusions about whether our finding for the natural outgroup of refugees translates into other domains of discrimination such as trust or social preferences. Second, our sample of university students (in Munich) is not representative for the population (of Germany). This has implications for the generalizability of our results. Similar studies with more right-wing and less liberal subpopulations might yield different results. Hence, testing our findings with different subject pools can yield additional insights — especially with regards to the effect of identity concerns. Future research could also exogenously vary identity concerns by priming certain aspects of subjects' identities. This could help to establish a causal link between these concerns and discrimination behavior. Lastly, the difference between our findings in the binary versus the probability-scale responsibility attribution highlight a potentially mediating effect of moral prescriptions. Using a range of choice environments that differ in the strength of behavioral prescriptions could test this relationship.

Appendix A

The Effect of Incentives in Non-Routine Analytical Team Tasks — Evidence from a Field Experiment

A.1 Room Fixed Effects for the Natural and Framed Field Experiment

Table A.1: Main treatments probit and GLM regressions including room fixed effects

	Field experiment (1)-(2)		Framed field experiment (3)-(4)	
	Probit (ME)	GLM	Probit (ME)	GLM
	(1)	(2)	(3)	(4)
Bonus45	0.150***	0.266**	0.076**	0.655***
	(0.041)	(0.113)	(0.036)	(0.215)
Constant		3.706***		3.896***
		(0.488)		(0.834)
Fraction of control teams solving the task in less than 45 min	0.10		0.045	
Control Variables	Yes	Yes	Yes	Yes
Staff Fixed Effects	Yes	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes	Yes
Room Fixed Effects	Yes	Yes	Yes	Yes
Observations	487	487	268	268

Notes: The table shows average marginal effects from probit regressions of whether a team solved the task within 45 minutes (1) and (3) and coefficients of GLM regressions on the remaining time (2) and (4) for the customer and the student sample. The specifications are as in Table 1.2 (4), A.4 (4), 1.7 (4), and A.6 (4), but include in addition Room Fixed Effects. Robust standard errors clustered at the day (field experiment) and session (framed field experiment) level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

A.2 Treatment Form for Bonus Treatments

Bonus treatment teams had to sign the following form, indicating understanding of the treatment procedures. For teams in the loss frame, the form further included the obligation to give back the money in case the team did not qualify for the bonus. Only one member of each team signed the form and the forms differed between the customer and student sample only in the amount of the bonus mentioned (€50 for the customer sample and €30 for the student sample). Similarly, the forms of *Bonus45* and *Bonus60* only differed in the time set for receiving the bonus.

The form for *Gain45* said:

"As usual, you have one hour in total to escape from the room. Furthermore, we have a special offer for you today: If you escape from the room within 45 minutes, you will receive €50."

The form for *Loss45* said:

"As usual, you have one hour in total to escape from the room. Furthermore, we have a special offer for you today: You now receive €50. If you do not escape from the room within 45 minutes, you will lose the €50."

A.3 Text of the Invitation to Laboratory Participants

We added the following paragraph to the standard invitation to student participants in the framed field experiment:

"Notice: This experiment consists of two parts, of which only the first part will be conducted on the premises of the MELESSA laboratory. In Part 1 you will be paid for the decisions you make. Part 2 will take place outside of the laboratory. You will take part in an activity with a participation fee. Your compensation in Part 2 will be that the experimenters will pay the participation fee of the activity for you."

A.4 Additional Analyses for the Field Experiment

A.4.1 Bonus Incentives and Team Characteristics

Table A.2: Linear probability model: Solved in less than 45 minutes

	OLS: Solved in less than 45 minutes					
	(1)	(2)	(3)	(4)	(5)	(6)
Bonus45	0.172***	0.200***	0.023	0.120**	0.130**	0.169***
	(0.050)	(0.071)	(0.122)	(0.057)	(0.056)	(0.047)
Share males	0.102*	0.130**	0.102*	0.100*	0.105*	0.103*
	(0.055)	(0.048)	(0.055)	(0.054)	(0.056)	(0.058)
Group size	0.056***	0.056***	0.042**	0.057***	0.055***	0.056***
	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
Experience	0.125***	0.126***	0.126***	0.058*	0.124***	0.125***
	(0.031)	(0.031)	(0.032)	(0.032)	(0.031)	(0.031)
Private	0.040	0.039	0.039	0.036	−0.001	0.039
	(0.041)	(0.042)	(0.042)	(0.041)	(0.049)	(0.041)
English-speaking	−0.115*	−0.117*	−0.113*	−0.114*	−0.117*	−0.129***
	(0.060)	(0.062)	(0.062)	(0.060)	(0.059)	(0.044)
Bonus45 ...						
... × Share males		−0.055				
		(0.128)				
... × Group size			0.031			
			(0.025)			
... × Experience				0.132**		
				(0.051)		
... × Private					0.077	
					(0.056)	
... × English speaking						0.027
						(0.139)
Constant	−0.177	−0.192	−0.109	−0.179	−0.163	−0.172
	(0.132)	(0.151)	(0.142)	(0.132)	(0.133)	(0.138)
R-squared	0.155	0.156	0.157	0.162	0.157	0.156
Staff Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	487	487	487	487	487	487

Notes: Coefficients from a linear probability model. Dependent variable: Dummy for finishing within 45 minutes. All models include staff and week fixed effects as in Table 1.7. Robust standard errors clustered at the day level reported in parentheses, and *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table A.2 shows the results from linear probability models estimating a dummy for whether teams solve the task within 45 minutes. Model (1) includes no interactions and uses the same variables and fixed effects as model (4) in Table 1.2. The effect of bonus incentives is of a similar magnitude as the average marginal effect in the probit specification. In models (2) to (6) we add interactions with observable team characteristics. The findings from these models suggest that the treatment effect does not strongly interact with the observable team characteristics. Only the interaction of incentives and experience in model (4) turns out to be significant (at the five percent level) and positive, while at the same time the treatment dummy is still statistically significant and large in magnitude. Hence, the positive incentive effect is robust and slightly larger for teams with experience.

A.4.2 Probability of Solving the Task in 45 Minutes (Field Experiment)

Table A.3 reports the results for the regression columns (1) to (5) from Table 1.2 excluding those weeks where we do not observe variation in the outcome variable. This confirms our previous findings.

Table A.3: Main treatments probit regressions: Excluding weeks with no variation in the outcome variable

	Probit (ME): Solved in less than 45 minutes				
	(1)	(2)	(3)	(4)	(5)
Bonus45	0.150***	0.151***	0.183***	0.163***	
	(0.026)	(0.024)	(0.027)	(0.045)	
Gain45					0.134***
					(0.040)
Loss45					0.188***
					(0.050)
Fraction of control teams solving the task in less than 45 min	0.11	0.11	0.11	0.11	0.11
Control Variables	No	Yes	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes	Yes
Week Fixed Effects	No	No	No	Yes	Yes
Observations	451	451	451	451	451

Notes: The table reports average marginal effects from probit regressions of whether a team solved the task within 45 minutes on our treatment indicators (with *Control* as base category). Control variables, staff and week fixed effects as in Table 1.2. All models exclude weeks that perfectly predict failure to receive the bonus. Robust standard errors clustered at the day level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

A.4.3 Regression Analysis for Remaining Time as Dependent Variable (Field Experiment)

We also estimate the effects of bonuses on the remaining time in seconds. Because our outcome measure is strongly right skewed and contains many zeroes (as there is no time left for those not finishing the task at all), we estimate a GLM regression with

Table A.4: GLM regressions: Remaining time

	GLM: Remaining time in seconds				
	(1)	(2)	(3)	(4)	(5)
Bonus45	0.432***	0.447***	0.406***	0.257**	
	(0.088)	(0.096)	(0.094)	(0.116)	
Gain45					0.259**
					(0.108)
Loss45					0.256*
					(0.136)
Constant	5.842***	4.041***	4.251***	3.803***	3.803***
	(0.082)	(0.393)	(0.359)	(0.403)	(0.403)
Control Variables	No	Yes	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes	Yes
Week Fixed Effects	No	No	No	Yes	Yes
Observations	487	487	487	487	487

Notes: Coefficients from a generalized linear model regression with a log link of the remaining time on our treatment indicators (with *Control* as base category). Control variables, staff and week fixed effects as in Table 1.2. Robust standard errors clustered at the day level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

a log link, again employing cluster-robust standard errors (Table A.4). Column (1) starts out with our baseline specification which includes a dummy for the incentive treatments (pooled) only. Bonus incentives significantly increase performance (measured by the remaining time). Analogously to our analysis in Table 1.2, we add the set of observable controls in Column (2). In Column (3) we add staff fixed effects. In Column (4) we present the results from an estimation that also includes week fixed effects. Finally, in Column (5) we include two treatment dummies to test whether gain or loss frames affect performance differently. Both coefficients are of similar size and we cannot reject the equality of the coefficients for the *Loss45* and *Gain45* treatments (Wald test, *p*-value = 0.98).

Analogously to the probit regressions reported in Table 1.5, we also run GLM specifications with the remaining time as the dependent variable (Table A.5) for the full set of treatments. This confirms our findings that incentives that include rewards increase performance whereas only mentioning the reference performance does not.

Table A.5: GLM regressions: Remaining time (all treatments)

	GLM: Remaining time in seconds			
	(1)	(2)	(3)	(4)
Bonus45	0.432^{***}	0.436^{***}	0.376^{***}	0.244^{**}
	(0.088)	(0.093)	(0.092)	(0.102)
Bonus60	0.233^{*}	0.267^{**}	0.392^{***}	0.449^{***}
	(0.131)	(0.114)	(0.126)	(0.134)
Reference Point	0.002	−0.001	0.102	0.131
	(0.106)	(0.108)	(0.114)	(0.086)
Constant	5.842^{***}	4.044^{***}	4.225^{***}	3.713^{***}
	(0.081)	(0.317)	(0.310)	(0.329)
Control Variables	No	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes
Week Fixed Effects	No	No	No	Yes
Observations	722	722	722	722

Notes: Coefficients from a generalized linear model regression with a log link of the remaining time on our treatment indicators (with *Control* being the base category). Control variables, staff and week fixed effects as in Table 1.2. Robust standard errors clustered at the day level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

A.5 Additional Analyses for the Framed Field Experiment

A.5.1 Regression Analysis for Remaining Time as Dependent Variable (Framed Field Experiment)

Table A.6 shows results from GLM regressions on the remaining time. Column (1) shows a positive and statistically significant effect of the bonus treatment on remaining times. The coefficient and its standard error remain roughly unchanged with the addition of controls and fixed effects. Column (5) shows the regression on the non-pooled framing treatments. The coefficients for both frames are highly significant and equality of coefficients of *Gain45* and *Loss45* cannot be rejected (p-value = 0.88).

Table A.6: GLM regressions: Remaining time (student sample)

	GLM: Remaining time in seconds				
	(1)	(2)	(3)	(4)	(5)
Bonus45	0.658***	0.673***	0.664***	0.661***	
	(0.216)	(0.217)	(0.210)	(0.213)	
Gain45					0.676***
					(0.238)
Loss45					0.647***
					(0.226)
Constant	5.135***	3.816***	4.039***	3.684***	3.690***
	(0.195)	(0.678)	(0.723)	(0.894)	(0.889)
Control Variables	No	Yes	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes	Yes
Week Fixed Effects	No	No	No	Yes	Yes
Observations	268	268	268	268	268

Notes: Coefficients from a generalized linear model regression with a log link of the remaining time on our treatment indicators (with *Control* being the base category). Control variables, staff and week fixed effects as in Table 1.7. Robust standard errors clustered at the session level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

A.5.2 Probability of Solving the Task in 45 Minutes (Framed Field Experiment)

Table A.7: Main treatments probit regressions: Excluding weeks with no variation in the outcome variable (student sample)

	Probit (ME): Solved in less than 45 minutes				
	(1)	(2)	(3)	(4)	(5)
Bonus45	0.107*	0.097*	0.104**	0.111**	
	(0.055)	(0.054)	(0.052)	(0.051)	
Gain45					0.142**
					(0.057)
Loss45					0.072
					(0.055)
Fraction of control teams solving the task in less than 45 min	0.06	0.06	0.06	0.06	0.06
Control Variables	No	Yes	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes	Yes
Week Fixed Effects	No	No	No	Yes	Yes
Observations	191	191	191	191	191

Notes: The table reports average average marginal effects from probit regressions of whether a team solved the task within 45 minutes on our treatment indicators (with *Control* as base category). Control variables, staff and week fixed effects as in Table 1.7. All models exclude weeks that perfectly predict failure to receive the bonus. Robust standard errors clustered at the session level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Table A.7 reports the results for the regression columns (1) to (5) from Table 1.7 excluding those weeks where we do not observe variation in the outcome variable. This confirms our previous findings.

A.6 Ordered Probit Regressions for Natural and Framed Field Experiment: Hint taking

Table A.8: Ordered probit regressions: Number of hints requested

	Ordered probit: Number of hints requested							
	Field experiment (1)-(4)				Framed field experiment (5)-(8)			
	within 60 min		within 45 min		within 60 min		within 45 min	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Bonus45	0.116	0.086	0.341**	0.190	0.401***	0.395***	0.878***	0.933***
	(0.123)	(0.148)	(0.133)	(0.129)	(0.151)	(0.148)	(0.144)	(0.147)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Staff FE	No	Yes	No	Yes	No	Yes	No	Yes
Week FE	No	Yes	No	Yes	No	Yes	No	Yes
Observations	487	487	487	487	268	268	268	268

Notes: Coefficients from an ordered probit model of the number of hints requested within 60 minutes or 45 minutes regressed on our treatment indicator *Bonus45* (pooled). Controls and fixed effects (FE) identical to previous tables. Robust standard errors clustered at the day (field experiment) and at the session (framed field experiment) level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

A.7 Hint Taking at a Specific Step in the Task

We have argued that it is unlikely that hint-taking behavior alone can explain the observed performance increase of the customer teams facing incentives. In the following, we provide some additional evidence on the relationship between hint taking and performance in our experiment. When doing so, we have to deal with two opposing effects. First, from a theoretical perspective, worse teams are more likely to use hints (which is also reflected in the positive correlation between finishing times and number of hints taken). Second, faster teams are more likely to take hints earlier on, as they are likely to reach a difficult quest faster than slower teams. That is, if incentives make (worse) teams faster, these teams may also mechanically take more hints and this effect accumulates over time. In order to reduce in particular the importance of the second effect, we collected information on the time at which teams reach a specific intermediate step for a subsample of 461 out of the 487 teams and compare the number of hints taken at that specific

Table A.9: Ordered probit regressions: Number of hints taken when entering last room (field experiment)

	Ordered probit: Number of hints taken				
	(1)	(2)	(3)	(4)	(5)
Bonus45	−0.018 (0.115)	0.012 (0.113)	0.113 (0.084)	0.050 (0.110)	0.134 (0.137)
Control Variables	No	Yes	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes	Yes
Week Fixed Effects	No	No	No	Yes	Yes
Room Fixed Effects	No	No	No	No	Yes
Observations	461	461	461	461	461

Notes: Coefficients from an ordered probit model. Dependent variable: Number of hints taken at the intermediate step of entering the last room. Control variables, staff and week fixed effects as in Table 1.2. Robust standard errors clustered at the day level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

step. This allows us to control the number of quests solved and to relate fixed progress in the task to hints taken. We focus on the point in time at which teams entered the last room of their specific task (*Zombie Apocalypse*, *The Bomb*, *Madness*), as teams reach this step on average rather early in the escape game. Teams facing incentives complete this step on average after 22 minutes whereas teams in the control condition need on average 24 minutes (Mann–Whitney test, p-value= 0.018). Hence, teams facing the incentive condition outperform control teams also early in the task. In Table A.9 we report results from ordered probit models to study whether teams facing incentives take more hints before the intermediate step. All five specifications reveal that team incentives do not significantly affect the number of hints taken and also none of the marginal effects of moving from one category (e.g. from one to two hints) to another category turns out to be statistically significant.

In contrast to the customer teams, we have shown that student teams (confronted with the task by us) took on average more hints when facing incentives. Repeating the analysis on reaching the intermediate step for the student sample shows that students facing incentives reached the intermediate step significantly earlier (they entered the last room on average after 31 minutes in *Control* and after 27 minutes when facing incentives, Mann–Whitney test, p-value= 0.004) but also took significantly more hints before reaching this step (see Table A.10).

Table A.10: Ordered probit regressions: Number of hints taken when entering last room (framed field experiment)

	Ordered probit: Number of hints taken				
	(1)	(2)	(3)	(4)	(5)
Bonus45	0.244**	0.235*	0.285**	0.306***	0.361**
	(0.122)	(0.123)	(0.119)	(0.117)	(0.154)
Control Variables	No	Yes	Yes	Yes	Yes
Staff Fixed Effects	No	No	Yes	Yes	Yes
Week Fixed Effects	No	No	No	Yes	Yes
Room Fixed Effects	No	No	No	No	Yes
Observations	267	267	267	267	267

Notes: Coefficients from an ordered probit model. Dependent variable: Number of hints taken at the intermediate step of entering the last room. Control variables, staff and week fixed effects as in Table 1.7. Robust standard errors clustered at the session level reported in parentheses, and *** p<0.01, ** p<0.05, * p<0.1.

Appendix B

Sharing or Gambling? On Risk Attitudes in Social Contexts

B.1 Social Orientation Questionnaire

The design and description of the social orientation questionnaire is based on Sutter et al. (2010). The questionnaire consists of 24 choices (see Table B.1) between two own-other payoff allocations in constant and anonymous pairs of subjects. The two options in all 24 choices each assign an amount of money to the subject herself (x) and a certain amount to the matched player (y). Subjects knew that everybody received the same questionnaire, and there was no feedback given about the matched player's choices while subjects were filling in the questionnaire. For all payoff allocations $r^2 = 15^2 = x^2 + y^2$ holds, such that each option represents a vector in a Cartesian plane lying on a circle with radius $r = 15$ centered at the origin.

By adding up x and y of all 24 choices, the motivational vector M can be constructed, yielding an angle θ of vector M (x on the x-axis and y on the y-axis, see Figure B.1). With this angle subjects can be classified into one of the following eight categories based on their social motivation: individualism, altruism, cooperation, competition, martyrdom, masochism, sadomasochism, and aggression.

The classification of subjects can be seen in Figure B.1:

Table B.1: Ring Test: 24 choices for own-other payoff allocations

Question number	self (x)	other (y)	self (x)	other (y)
1	15	0	14.5	−3.9
2	13	7.5	14.5	3.9
3	7.5	−13	3.9	−14.5
4	−13	−7.5	−14.5	−3.9
5	−7.5	13	−3.9	14.5
6	−10.6	−10.6	−13	−7.5
7	3.9	14.5	7.5	13
8	−14.5	−3.9	−15	0
9	10.6	10.6	13	7.5
10	14.5	−3.9	13	−7.5
11	3.9	−14.5	0	−15
12	14.5	3.9	15	0
13	7.5	13	10.6	10.6
14	−14.5	3.9	−13	7.5
15	0	−15	−3.9	−14.5
16	−10.6	10.6	−7.5	13
17	−3.9	−14.5	−7.5	−13
18	13	−7.5	10.6	10.6
19	0	15	3.9	14.5
20	−15	0	−14.5	3.9
21	−7.5	−13	−10.6	−10.6
22	−13	7.5	−10.6	10.6
23	−3.9	14.5	0	15
24	10.6	−10.6	7.5	−13

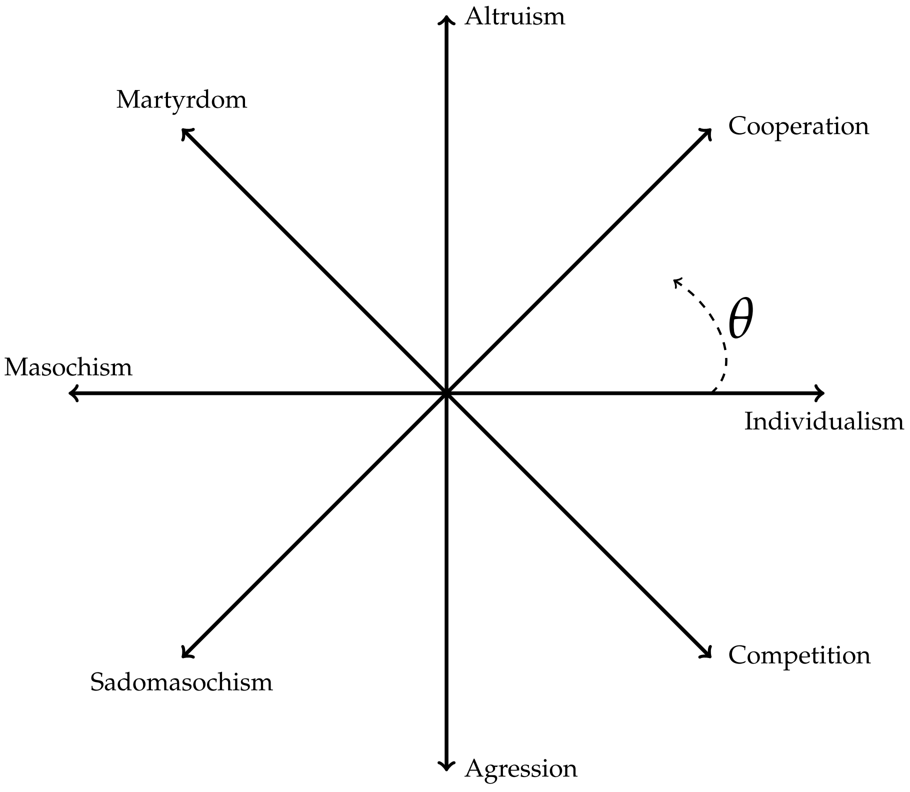


Figure B.1: Vectors defining the basic social motivation

Subjects with a θ between 0° and 22.5° or 337.5° and 0° are classified as individualistic; subjects with an angle between 22.5° and 67.5° as cooperative. More infrequent types are altruism (between 67.5° and 112.5°), martyrdom (between 112.5° and 157.5°), masochism (between 157.5° and 202.5°), sadomasochism (between 202.5° and 247.5°), aggression (between 247.5° and 292.5°), and competitive (between 292.5° and 337.5°).

Subjects' earnings in part 5 were given by the sum of choices made by the subject herself (sum of own x) and by the sum of choices made by the matched player (sum of other's y).

B.2 Instructions for all Parts

B.2.1 Instructions Before the Start of the Experiment

Please do not talk to other participants anymore and remain silent throughout the entire experiment. For simplicity we will use masculine terms in the following. These will refer to both male as well as female participants.

General information regarding procedures

The experiment aims at investigating decision making. You can earn money which will be paid out at the end of the experiment in private and in cash.

The entire experiment will last around 45 minutes. It consists of two completely independent parts in which you have to make decisions. The first part is divided into four blocks. In block 1 and 2 your earnings can depend on the decisions of another participant, who will be randomly assigned to you. In block 3 and 4 your earnings will be solely determined by your own decisions. In the second part of the experiment your earnings will again depend on your own decisions and the decisions of another participant. For this purpose, you will again be randomly assigned to another person. We will not use the same pairs as in part 1, but make new random pairs. After part 2 we will ask you to answer a general questionnaire.

While you make your decisions a clock will run down in the upper right corner of the screen. This provides guidance for how much time you can use for your decisions. If the clock is down to zero, please come to a decision. However, you can still complete your decisions with the clock down to zero.

If you still have questions after the instructions or during the experiment, please raise your hand or press the red button on your keyboard. One of the experimenters will then come to your seat and answer your question in private. If the question is of interest to all participants, we will repeat the question and answer publicly.

Anonymity

None of the other participants will be able to reconstruct your decisions in the experiment. Moreover, the data from the experiment will be analyzed anonymously. For accounting reasons you have to sign a receipt for your earnings at the end of the experiment. Your name cannot be linked to your decisions in the experiment.

The Experiment — Part 1

Block 1 and 2

In block 1 and 2 you will be randomly assigned a role: active or passive participant. Your decisions will only be relevant for your earnings and the earnings of your matched participant if you are the active participant. Decisions of the passive participant have no impact on earnings. However, your role will only be revealed at the end of the experiment. For that reason please assume for these decisions that you are the active participant. Otherwise decisions might be implemented that you want to avoid. In block 1 you will make decisions for 9 scenarios. In block 2 there is one scenario.

Block 3 and 4

Assigned roles are irrelevant in block 3 and 4. Your potential earnings only depend on your own decisions. In block 3 you will make decisions for 6 scenarios. In block 4 there will be 9 scenarios.

Payment

For all decisions in part 1 all potential earnings will be stated in Euro. Since you will make many different decisions in these blocks, the computer will randomly draw one single decision at the end of the experiment. This decision will be relevant for your earnings. The procedure is as follows: Only one out of the 4 blocks is relevant. This relevant block will be randomly determined by the computer. Within this block, one specific decision (scenario) will again be determined randomly to be payoff relevant. If the chosen decision involves uncertain payments (probabilities) the computer will again determine randomly which probabilistic event will be realized. Further, the computer will randomly assign roles of active and passive participants for all randomly matched pairs of participants. This role will only be relevant for your earnings if block 1 or 2 is relevant for your earnings. Let us assume you are assigned the active role and block 1 was determined to be payoff relevant. Based on the randomly chosen scenario you and your matched participant will receive earnings based on your decision in this scenario. If you were assigned the passive role and block 1 was determined relevant, you will receive earnings based on the decision of the matched participant in the respective scenario. Every decision in blocks 1 to 4 can be relevant for your earnings. Choose your answers carefully.

The Experiment — Part 2

Upon finishing part 1 of the experiment you will start with part 2. This part is completely independent of part 1. Here, you will again be randomly assigned to one other participant. The pairs, however, will be randomly drawn anew. After part 1 you will be provided with more information on part 2.

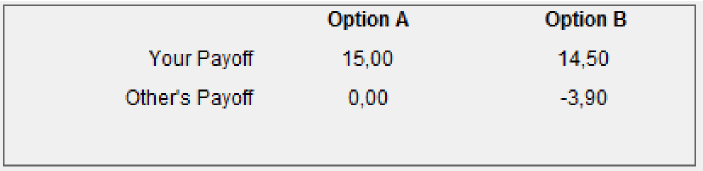
Your total earnings in today's experiment hence will consist of the described earning from part 1 and the earnings from part 2. In addition, you receive €4 for showing up on time.

B.2.2 Instructions Before Part 2 of the Experiment (Distributed and Read out After Part 1)

In part 2 you will again be randomly assigned to one other participant. You will make multiple decisions which affect your own payoff as well as the payoff of your matched participant. There will be no roles in this part of the experiment. That is, both your decisions as well as the decisions of your matched participant will be implemented. Both you and the other participant will remain anonymous.

You will make 24 decisions with two options each (Option A and Option B). Each option assigns a certain amount of the experimental currency 'Taler' to your account ('Your Payoff') and a certain amount to the account of your matched participant ('Other's Payoff').

An example:



	Option A	Option B
Your Payoff	15,00	14,50
Other's Payoff	0,00	-3,90

If you choose option A, 15 Taler will be transferred to your account and zero Taler to the account of the other participant. If you choose option B you receive 14.50 Taler and the other participant will receive -3.90 Taler (3.90 Taler will be deducted from his account).

Your total earnings of part 2 will be the sum of 'Your Payoff' of your 24 decisions. The payoff for the other participant based on your decisions is the sum of 'Other's Payoff'. That is, every single decision in this part of the experiment will affect your own and the other's earnings.

Your matched participant makes decisions for exactly the same choices. Hence, in addition to the sum of 'Your Payoff' of your own decisions you will receive the sum of 'Other's Payoff' of the decisions of your matched participant. Similarly, next to the earnings from your decisions, the other participant receives a payment based on his own decisions, too.

The resulting total earnings in 'Taler' will then be converted to Euros and represent your earnings from part 2 of the experiment. The exchange rate is: 10 Taler = 1.50 Euro.

During the experiment you will not receive feedback on any decision of your matched participant. Only at the end of the experiment will you see the sums of 'Your Payoff', 'Other's Payoff' and 'Other's Payoff' of your matched participant, as well as your total earnings from part 2.

Potential negative earnings in single parts of the experiment will be offset by earnings from the other part and the €4 received for showing up on time such that total earnings of the experiment will always be positive.

If you have any questions please raise your hand now. We will then come to your seat and answer your questions in private.

B.3 Additional Results for the Cluster Analysis

The categorization of subjects can help in explaining the aggregate pattern. Remember that for type-1 subjects the increase in risk taking in the unfavorable range when going from the individual to the social context is very pronounced, while there is a reduction in risk taking in the favorable range. Type-3 subjects increase overall risk-seeking behavior in the favorable range and reduce risk taking in the unfavorable range, when they are in the social context. Type 2 individuals are characterized by their leap towards indifference in the social lotteries – especially in the unfavorable domain. Statistical tests confirm this first impression from the cluster analysis.

For type-1 individuals, risk taking very strongly and significantly increases in the unfavorable range ($p < 0.01$, Stuart-Maxwell test) for all lotteries when going from the individual context to the social context. The reverse is true for the favorable situations. Here, type-1 individuals even reduce risk taking in the social context. For lotteries T7(i) to T9(i) this difference is significant (at the 5%-level using the McNemar test).[1] For type-3 subjects most comparisons do not result in significant differences, most probably due to a lack of statistical power, given the much smaller number of decision makers than in the type-1 cluster. There is at least some tentative evidence that risk taking increases from T9i to T9 ($p < 0.1$, Stuart-Maxwell test) and that risk taking decreases in the unfavorable range at least from T3i to T3 ($p < 0.01$ for McNemar's test), in contrast to the aggregate pattern. As indicated before, the change in pure risk-taking behavior for type-2 individuals is less clear cut due to the large number of indifference choices. This trend towards more indifference however is clearly significant (mostly at the 1%-level) in a McNemar test, grouping indifference against risky and safe choices for all lotteries except T5 vs. T5i and T9 vs. T9i.

B.4 Evidence From a Classroom Experiment

B.4.1 Experimental Design

The experiment was divided into three parts: a series of risky choices in the social context (similar to part 1 of our lab experiment), two dictator games (see our part 2), and a series of individual decisions under risk (see our part 4). The first part, as before, is the core of the study and aims at measuring how risk attitude is affected by social contexts, whereas the latter two again provide a control for social concerns and risk attitude in a purely individual context.

In the first part of the experiment, subjects faced tasks where fifty Euros had to be allocated (either deterministically or randomly) between the decision maker and the receiver. This is equivalent to our design described in the chapter. However, in the classroom experiment, €50 instead of €10 had to be divided with expected payoffs for the decision maker ranging from €5 ("T5") to €45 ("T45") in steps of €5. Order

[1] Since there are no indifference choices for type-1 individuals, we can only look at McNemar's test. This also holds for type-3 individuals in T3 and T3i.

effects were controlled for by presenting the choices in ascending orders to half the subjects and in descending order to the other half.

Parts 2 and 3 aim at measuring social preferences in a risk-free environment and risk preferences in an individual setting (without social context). Part 2 was equivalent to part 2 of our lab experiment in that subjects had to play the two dictator games. Here again, €50 were to be distributed. Part 3 consisted of nine binary decisions under risk. The first three were a truncated and adapted Holt and Laury (2002) procedure to estimate subjects' risk attitudes with stakes comparable to the one used in the main part of our experiment (see first half of part 3 in the lab experiment). The next three tasks were aimed at measuring loss aversion (second half of part 3 above), and the last three tasks were risky binary choices that were exactly equivalent to three of the tasks in part 1, but without any social component (part 4 above). Hence, in contrast to the lab experiment, we only have three equivalent individual tasks to compare to risk taking in the social context. One of these tasks was in the unfavorable range (expected payoff for the decision maker of €15, "T15i"), one was in the favorable range (expected payoff of €35, "T35i") and one was the equal spilt task (expected payoff of €25, "T25i"). Finally, as in the lab experiment, subjects were asked to provide some socio-demographic characteristics.

B.4.2 Experimental Procedures

The design described above was implemented as a classroom experiment with 82 undergraduates in economics at the University of Munich. Their role – either decision maker or receiver - was only determined after the experiment. Decision sheets and instructions were first distributed for parts 1 and 2 together, and upon finishing, also for part 3. Subjects knew that there were three parts of the experiments already at the beginning. For payment, four randomly selected decision makers were matched with four randomly selected receivers. For each pair one of the 'social' tasks (parts one and two, including the question regarding their preference for the regular dictator game or the probabilistic one) was randomly selected for payment. In addition, four participants were randomly picked for payment in the individual lottery part, where one task was once again randomly picked to be implemented. Payments were provided individually and confidentially. All design details and the procedural details were common knowledge among participants.

These design and procedure details result in three major differences between the classroom and lab experiment, apart from the obvious differences in the setting. First and most importantly, in the classroom experiment we did not collect data on all choices in the individual context. Second, we only paid a small fraction of subjects while the amounts to be shared were much higher. Third, in the lab experiment we included the ring test to measure social value orientation to have a better individual control for social preferences.

B.4.3 Results

In the social decisions under risk (part 1), supporting the results from our lab experiment, subjects clearly become more risk taking the more unfavorable the tasks become in the unfavorable range. In contrast to the results in Chapter 2 (Figure 2.1), however, risk taking in the social context is more U-shaped. That is, risk taking also increases in the favorable range towards the extremely favorable decision T45. The proportion of subjects taking the risky option in T30 is significantly lower than in T45 ($p < 0.01$, Stuart-Maxwell marginal homogeneity test), although conventional levels of significance are not reached with T40 and T35 vs. T45. Nevertheless also here, the U-shaped pattern seems to be asymmetric. The number of risky choices appears higher in the case of unfavorable inequity for the decision maker than in the case of favorable inequity. Leaving the case of the equal split aside, all comparisons between tasks corresponding to sure payoffs adding up to 50 (T5 vs. T45, T10 vs. T40, T15 vs. T35, and T20 vs. T30) suggest that the risky option is relatively more appealing when the sure option implies unfavorable inequity: The differences are significant according to Stuart-Maxwell tests at the 5%-level. By the same token, comparing the number of times decision makers have chosen the risky option in the four favorable situations against the same number in the four unfavorable situations yields a significant difference (using a two-sided Wilcoxon signed ranks test for matched observations; $p = 0.02$).

Core to our analysis, however, is the comparison between otherwise identical decisions in the social and individual context. Comparing the three individual tasks from part 3 of the experiment (T15i, T25i, T35i) with the social context counterparts T15, T25 and T35 yields very much similar results as in the lab experiment (see Figure B.2). In T15 vs. T15i, where the social situation is unfavorable to the decision maker, individuals take significantly more risk than in the equivalent individual lottery and this difference is strongly significant ($p < 0.01$, Stuart-Maxwell test). However, in case of a favorable social context (T35 vs. T35i) the difference is not significant at conventional significance levels. The effect also points in the opposite direction. In the favorable range of the social context, decision makers, if anything, seem to reduce risk taking compared to the equivalent individual decision. That is, as in the lab experiment, decision makers seem to be affected by social context when

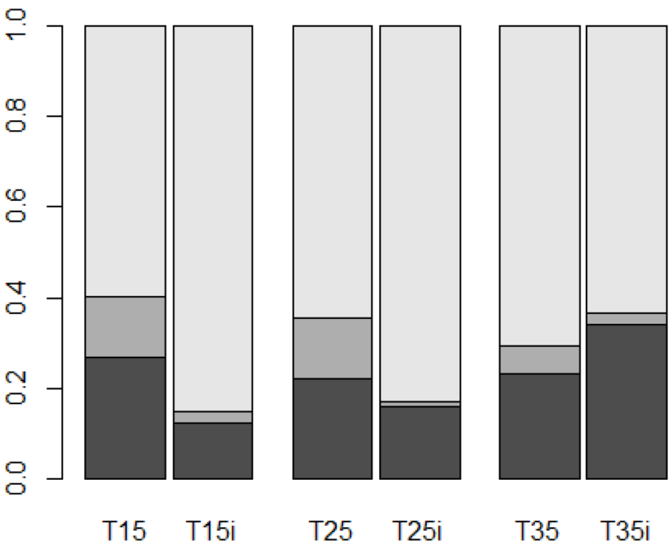


Figure B.2: Choices in the social vs. individual context (risky: black; safe: light grey)

making a risky decision, but not in a homogeneous way: They take more risk when the situation is unfavorable or equal, but less when it is favorable to them. Different to the results in Chapter 2, we can also observe a difference between the contexts for the equal situation (T25 vs. T25i), where subjects take more risk in the social lottery than in the individual one ($p < 0.01$, Stuart-Maxwell test).

As in Chapter 2, we can also look at individual patterns and heterogeneity between participants. If we split the sample based on the median offer in the dictator game in part 2 of the classroom experiment, we can see that the difference between the contexts in the unfavorable range seems to be driven by selfish subjects only ($p < 0.01$, Stuart-Maxwell test, and not significant difference for pro-social subjects). We also ran a k-medians cluster analysis on all social lotteries dividing the subjects into three clusters. This leads to the following characterization here: 20 individuals (type 1) exhibit a very dichotomous pattern of risk attitude in the social context (strongly risk-seeking in the unfavorable case, and risk-averse in the favorable one), 41 subjects (type 2) are rather overall risk-averse, and 21 individuals (type 3) show a relatively stable attitude towards risk, except in the case of high probabilities of winning (T35, T40, T45), where they strongly increase risk taking. Comparing the effect of the decision context for the different types, we can draw similar conclusions as in Chapter 2: Type 1 subjects seem to be most strongly affected by social context ($p<0.01$, Stuart-Maxwell test for T15 vs. T15i and $p < 0.05$ for T35 vs. T35i), increasing risk taking in the unfavorable range and decreasing risk taking in the favorable range. The same pattern also holds for type 2 subjects, even though the differences are less pronounced (not significant in the favorable range). There is no significant effect for type 3 subjects. That means that also for these subjects in this experimental setting, a clear majority of subjects (roughly three quarters) exhibit the pattern observed in the laboratory. They tend to take more risks in unfavorable situations than in equivalent individual contexts, and they – if anything – seem to be more risk-averse in socially favorable situations.

B.5 Theoretical Model

The decision makers have to choose (for each possible question) between a safe lottery $L_s = \big(1; (x\pi, (1-x)\pi)\big)$ and $L_r = \big(x, (\pi, 0); 1-x, (0, \pi)\big)$ with π the pie and x the probability (or the share of the pie). The function V represents the individual's preferences over lotteries.

Ex post social preferences

We assume here that individuals have social concerns that apply only to final allocations (i.e. the probabilistic distribution of the outcomes does not play a role as for ex ante social preferences, see below).

For a given individual, we set that her preferences over final allocations are represented by $u : (x_1, x_2) \mapsto u(x_1, x_2)$ with x_1 the payoff of the decision maker and x_2 the payoff of the recipient.

Under the usual assumption of expected utility maximization, we have for any x:

$$V(L_s) = u(x\pi, (1-x)\pi)$$

and

$$V(L_r) = xu(\pi, 0) + (1-x)u(0, \pi)$$

First, note that $(x\pi, (1-x)\pi)$ is a convex combination of $(\pi, 0)$ and $(0, \pi)$:

$$\begin{aligned} (x\pi, (1-x)\pi) &= (x\pi + (1-x) \times 0, (1-x)\pi + x \times 0) \\ &= x(\pi, 0) + (1-x)(0, \pi) \end{aligned}$$

We observe experimentally that a significant share of individuals have $V(L_s) > V(L_r)$ for large x (greater than or equal to $\frac{1}{2}$), but that for small x, the opposite holds: $V(L_s) < V(L_r)$.

This implies that u cannot be concave on $(0, 1)$ since for x small:

$$u(x(\pi, 0) + (1 - x)(0, \pi)) < xu(\pi, 0) + (1 - x)u(0, \pi)$$

Nor can it be convex, since for $x \geq \frac{1}{2}$:

$$u(x(\pi, 0) + (1 - x)(0, \pi)) > xu(\pi, 0) + (1 - x)u(0, \pi)$$

And for the same reasons it cannot be linear either.

Consider the function $f(x) = u(x(\pi, 0) + (1 - x)(0, \pi))$. It is continuous (if u is) and hence by the intermediate value theorem $\bar{x} \in (0, 1)$ exists such that

$$f(\bar{x}) = u(\bar{x}(\pi, 0) + (1 - \bar{x})(0, \pi)) = \bar{x}u(\pi, 0) + (1 - \bar{x})u(0, \pi)$$

Assume for the sake of simplicity that $\bar{x}$ is unique.[2]

For $0 < x < \bar{x}$,

$$u(x(\pi, 0) + (1 - x)(0, \pi)) < xu(\pi, 0) + (1 - x)u(0, \pi)$$

Or denoting $\pi_1 = (\pi, 0)$ and $\pi_2 = (0, \pi)$ for the sake of conciseness:
For $0 < x < \bar{x}$,

$$u(x\pi_1 + (1 - x)\pi_2) < xu(\pi_1) + (1 - x)u(\pi_2)$$

Given that any point on the line $[\bar{x}\pi_1 + (1 - \bar{x})\pi_2; \pi_1]$ is such that $0 < x < \bar{x}$, it ensues that for any convex combination $\alpha\big(\bar{x}\pi_1 + (1 - \bar{x})\pi_2\big) + (1 - \alpha)\pi_1$:

$$u[\alpha\big(\bar{x}\pi_1 + (1 - \bar{x})\pi_2\big) + (1 - \alpha)\pi_1] < \alpha u(x\pi_1 + (1 - x)\pi_2)) + (1 - \alpha)u(\pi_1)$$

Hence u is convex on $[\bar{x}\pi_1 + (1 - \bar{x})\pi_2, \pi_1]$.

For $\bar{x} < x < 1$,

$$u(x(\pi, 0) + (1 - x)(0, \pi)) > xu(\pi, 0) + (1 - x)u(0, \pi)$$

And the same reasoning as for $x < \bar{x}$ yields:

$$u[\alpha\big(\bar{x}\pi_1 + (1 - \bar{x})\pi_2\big) + (1 - \alpha)\pi_2] < \alpha u(x\pi_1 + (1 - x)\pi_2)) + (1 - \alpha)u(\pi_2)$$

[2] The multiplicity of $\bar{x}$ does not change the argumentation. It suffices to take the minimum and maximum of those multiple $\bar{x}$ to obtain the same result.

Hence u is concave on $[\pi_2, \bar{x}\pi_1 + (1-\bar{x})\pi_2]$.

To conclude, for *any ex post social preference*, there exists a share $\bar{x}$ such that the utility is concave for $\mathrm{x} > \bar{x}$ and is convex for $\mathrm{x} < \bar{x}$. Said differently, independently of the model of social preference under consideration (altruism, inequity aversion, maximin/efficiency, spitefulness, selfishness), a change in the curvature (from convexity to concavity) is required to observe a change of choice between the safe and the risky social lottery.

The observation that a significant share of individuals choose the risky social lottery for x small (when they chose the safe individual one for the same x) implies a change in the curvature of their utility function. This change of behavior is independent of their pro-social or anti-social motivations.

This also applies to a self-interested utility maximizer: A risk-averse individual ($u'' < 0$) always chooses the safe lottery, and a risk-seeking individual ($u'' > 0$) always chooses the risky one. Note here that the variance of the risky lottery in terms of individual payoff (which increases as x gets further away from $\frac{1}{2}$) does not play any role.

Ex ante social preferences

Under the assumption that the ex ante term of social preferences is based on expected payoff, the ex ante fairness of the safe social lottery and the risky social lottery is constant for each decision. Hence it cannot play a role.

When considering a mix of ex ante and ex post social consideration, the same is true and the only relevant driver for the switch from a risky to a safe social lottery can be, as in the case of ex post social preferences, the curvature of the utility function. Once again, independently of the motives under consideration.

Appendix C

Show What You Risk — Norms for Risk Taking

C.1 Seating Arrangement

The separating wooden walls between opposing seats in the laboratory were taken out before the start of the experiment to allow participants to identify their matched partner. Separating walls to the left and right remained (see Figure C.1).

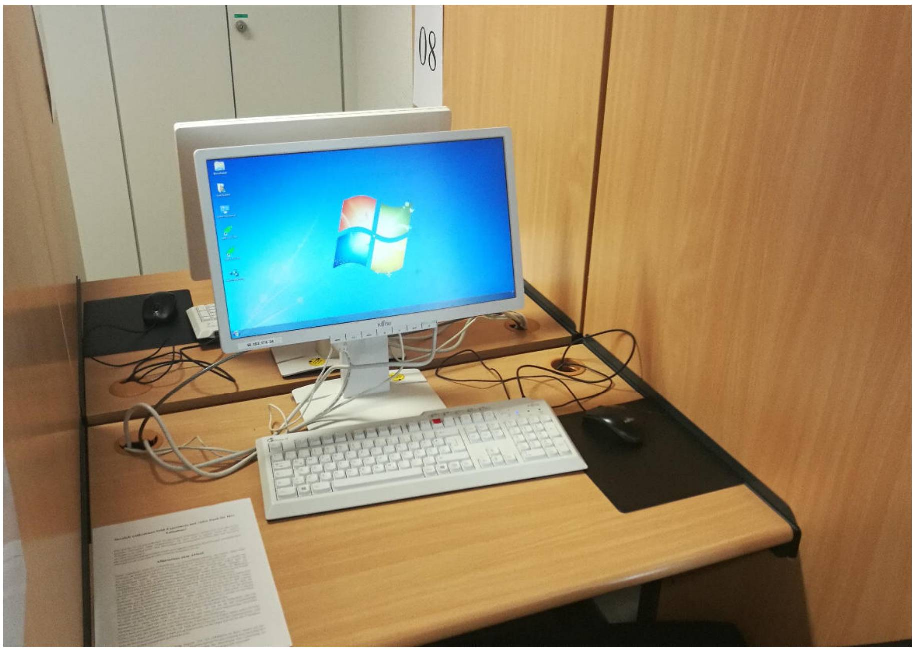


Figure C.1: Seat arrangement for matched participants

C.2 Picture Consent Form

At the beginning of the following experiment, a picture will be taken of all participants. The anonymous picture will be saved on servers of MELESSA, can only be accessed by the research team and will be deleted after the end of the study.

The picture taken of you might be shown to other participants during the experiment.

I hereby state that I read and understood above-standing information and that I agree to the generation and use of my personal data as stated in this form.

C.3 Instructions

C.3.1 Instructions Before the Start of the Experiment

Welcome to the experiment and thank you for your participation!

Please do not talk to other participants during the experiment and stay quiet during the entire time. Also please do not communicate in any other form with other participants in this experiment. Violations of these rules will lead to the exclusion of all experimental payments. Even though, for simplicity, we will only use masculine terms in these instructions to describe participants, these remarks refer to both male and female participants.

General information

This experiment is meant to study economic decision making. You can earn money during the experiment. This money will be paid to you in private after the experiment.

The entire experiment will last about 45 minutes. It consists of two parts in which you will make decisions. These decisions will affect your payment. In addition, your payment might depend on other participants' decisions as well as on chance. The specific decisions and exact payment rules will be explained to you right before each decision onscreen. Today's payment consists of the money earned through your decisions plus €5 for showing up on time.

In part 1 you will make one decision that is relevant for payment. The outcome of this decision will be paid out with certainty. Additionally, you will answer a few questions. In part 2, there will be 5 subparts in which you will make decisions. Only one of these subparts, however, will be relevant for payment. The computer will randomly choose one of the subparts to be paid. You will be informed in more detail about this procedure and payment rules at the beginning of part 2. After finishing part 2 we will ask you to answer some general questions about you and the experiment.

While you make your decisions there will be a clock running down in the upper right corner of your screen. This time will provide some indication for how much time you can take to make your decision. When this clock runs down, please come to a decision. However, even with the time at zero you can still complete your answers.

If you have any questions after the instructions or during the experiment, please simply raise your hand or press the red button on your keyboard. We will then come to your seat and answer your question in private. In case the question is relevant to all participants, we will publicly repeat the question and answer in public.

Anonymity

The analysis of the data generated in this experiment will be anonymous. That is, we will never link your name to data from the experiments. At the end of the experiment you will have to sign a receipt. This is only for accounting purposes. Also the photos taken will only be used anonymously.

C.3.2 Instructions for all Experimental Tasks

The following passages are the instructions that participants read on-screen. Text in italics denotes treatment manipulations and text in brackets denotes self-explaining comments. These accentuations are added for illustrative reasons and were not part of the original instructions.

[first screen]
Before the start of the experiment
In this experiment, you are matched with another participant. This participant sits in the seat vis-à-vis your own. Below you see the picture of your matched participant.

During the experiment, the matching will become relevant. We will point out once it is relevant. Independent of the relevance of the matching, the picture of your matched participant will be shown on all decision screens.

[Picture of matched participant here]

[new screen]
Part 1
You will make one decision in this part. This decision will definitely be payoff relevant for you. Your payoff depends on your decision and chance.

[new screen]
Part 1
You receive 100 Taler for this decision. 100 Taler corresponds to €5. The exchange rate is 1 Taler=€0.05. You can now invest any integer amount between 0 and 100 Taler in a risky option. You will keep the amount that you do not invest.

With a probability of 50% the investment in the risky option will be successful. If it is successful, you receive 2.5 times the invested amount. If it is not successful, you lose the invested amount.

Your earnings from this part of the experiment are made up of the amount not invested, and (potentially) the earnings from your investment in the risky option. Your earnings will be converted to Euros at the end.

[for participants in Control:] *Your decision is anonymous.*

[for participants in Treatment:] *Your decision is not anonymous. Your matched participant will be shown your choice (not your earnings) on-screen at the end of the experiment.*

Please indicate now the amount in Taler (0-100) that you want to invest in the risky option.

[Small picture of matched participant here]

[new screen]
Part 1
Please answer a couple of questions on a scale from 1 to 10.

How would you rate yourself: Are you in general a risk-tolerant person, or do you try to avoid risks? Please indicate a value on the following scale from 1 to 10, in which 0 translates to "not at all willing to take risks" and 10 translates to "very risk tolerant".

[for participants in Control:] *Your decision is anonymous.*

[for participants in Treatment:] *Your decision is still not anonymous. Your matched participant will be shown your choice on-screen at the end of the experiment.*

[Small picture of matched participant here]

[new screen]
Part 1
People's decisions often depend on the context in which these decisions take place. How would you rate your risk tolerance with respect to the following domains?

Please indicate a value on the following scale from 1 to 10, in which 0 translates to "not at all willing to take risks" and 10 translates to "very risk tolerant".

[for participants in Control:] *Your decisions are still anonymous.*

[for participants in Treatment:] *Your decisions are still not anonymous. Your matched participant will be shown your choices on-screen at the end of the experiment.*

How about the domain... [each domain and its scale were presented in a different row]

Car driving? Investing money? Sports and leisure activities? Professional career? Health? Trusting other people?

[Small picture of matched participant here]

[new screen]
Part 1
You finished part 1 of the experiment. You will next start part 2 of the experiment.

[new screen]
Part 2 — General information
We now start part 2 of the experiment. This part consists of five blocks. One of these five blocks will be chosen for payment at the end of the experiment. Your payment for part 2 will be based on your earnings in the respective block. Please make all decisions conscientiously since each block can be payoff relevant. The probability to be payoff relevant is the same for all blocks.

The specific rules for payments in each block will be explained once you start the respective block. Your earnings depend on your own decisions and the decisions of other participants.

[new screen]
Part 2 — General information
In all following decision tasks in part 2 you will have to provide assessments and estimates regarding the behavior of other participants in part 1.

[for participants in Control:] *In part 1, you and in total 50% of the participants made anonymous choices. Your matched participant and the other half of the participants made non-anonymous choices. That is, these participants were told in part 1 that their choices in part 1 would be shown to their matched participants at the end of the experiment. That means that you will be shown the decisions of your matched participant at the end of the experiment (you will not see the earnings of your matched participant). Your decisions remain anonymous.*

[for participants in Treatment:] *In part 1, you and in total 50% of the participants made choices that will be shown to the matched participants at the end of the experiment. This was pointed out to you before making your decisions. Your matched participant and the other half of the participants made anonymous choices. Since their choices remain anonymous, they also were not told that their choices were non-anonymous. Hence, you will not be shown the decision of your matched participant at the end of the experiment.*

In part 2, however, all participants make the same decisions. All these decisions will be anonymous for all participants.

[new screen]
Part 2 — Block 1
In block 1 of part 2 you are supposed to provide an estimate for the choice of your matched participant in part 1. In part 1 participants could invest any integer amount between 0 and 100 Taler (corresponds to between €0 and €5) in a risky option.

Please indicate now in Taler your estimate regarding your matched participant's invested amount.

In case this block will be payoff relevant, your earnings for this block will depend on your answer and the invested amount of your matched participant in part 1. If you deviate by 10 Taler or less from the actual invested amount of your matched participant, you will receive €5 for this block. If you deviate by more than 10 Taler, you will not receive any payment for this block.

What do you think your matched participant invested in the risky option (amount in Taler between 0 and 100)?

[Small picture of matched participant here]

[new screen]
Part 2 — Block 1
How confident are you in your answer from the screen before?

[Scale from 1 (“Not at all confident. I basically guessed randomly.”) to 5 (“I am completely convinced that I gave the correct answer.”). Other options were 2 (“I did not guess randomly, but I am still very uncertain.”), 3 (“I am somewhat uncertain, but I had some idea of the correct answer.”), and 4 (“I am rather certain that I gave the correct answer.”).]

[This question on confidence was used after all following decision screens in exactly the same way. Therefore, hereinafter, these screens will not be shown again.]

[new screen]
Part 2 — Block 2
In block 2 of part 2 you are supposed to provide an estimate for the choices of all participants in part 1. In part 1 participants could invest any integer amount between 0 and 100 Taler (corresponds to between €0 and €5) in a risky option.

Please indicate now in Taler your estimate regarding the average invested amount of all participants.

In case this block will be payoff relevant, your earnings for this block will depend on your answer and the invested amounts of all participants in part 1. If you deviate by 10 Taler or less from the actual average invested amount of all participants in part 1, you will receive €5 for this block. If you deviate by more than 10 Taler, you will not receive any payment for this block.

What do you think did participants on average invest in the risky option (amount in Taler between 0 and 100)?

[Small picture of matched participant here]

[new screen]
Part 2 — Block 3
In block 3 of part 2 you are again supposed to provide an estimate regarding the choices of other participants. These will be different assessments though.

Later, four other participants will be shown your picture (they will also be told whether your choice in part 1 was anonymous or not). These participants will then, based on your picture, indicate what answer in part 1 would have been appropriate for you to make. They will not know your actual investment when making that assessment. Each of these four participants will be paid for his/her assessment if it does not deviate by more than 10 Taler from the average assessment of the other three participants seeing your picture.

You will receive €5 for this block, if you do not deviate by more than 10 Taler from the average assessment of the four other participants. If you deviate by more than 10 Taler, you will not receive any payment for this block.

What investment in Taler (between 0 and 100) do you think the other four participants deem appropriate for you; i.e., what do they think you should have invested?

[Small picture of matched participant here]

[new screen]
Part 2 — Block 4
One of the four other participants that will be shown your picture and that will indicate the appropriate investment for you will be your matched participant from part 1.

In block 4 of part 2 you are supposed to indicate the following: What do think your matched participant thinks would have been the appropriate investment that you should have made?

You will receive €5 for this block, if your answer does not deviate by more than 10 Taler from the actual answer of your matched participant. If your answer deviates by more than 10 Taler from the actual answer of your matched participant, you will not receive any payment for this block.

What does your matched participant think would have been the appropriate investment that you should have made (in Taler between 0 and 100)?

[Small picture of matched participant here]

[new screen]
Part 2 — Block 5
In block 5 of part 2 you will make four decisions. In case block 5 is payoff relevant, one of those four decisions will be chosen for payoff (with equal probabilities). The rules determining your payoff in that case will be explained to you when making the respective decision.

[new screen]
Part 2 — Block 5 — Decision 1
In the first decision you see the picture of another participant ([Here, the treatment condition of the participant in the picture was indicated. If “picture 1 participant” was in *Control* it said: *anonymous decision in part 1*; if “picture 1 participant” was in *Treatment* it said: *non-anonymous decision in part 1*]). You are supposed to

indicate, what you think, what would have been the appropriate investment (in Taler between 0 and 100) that the person in the picture should have invested in part 1.

Three other participants see the same picture and answer the same question that you will answer.

In case this decision becomes payoff relevant, you will earn €5 if your answer does not deviate by more than 10 Taler from the average answer of the other three participants. If your answer deviates by more than 10 Taler from the average answer of the three other participants, you will not receive any payment for this decision.

What investment in Taler (0 to 100) would have been the appropriate investment for the person in the picture; i.e., what amount should that person have invested?

[Small picture of picture 1 participant here]

[Screens for decision 2, 3, and 4 in block 5 of part 2 were equivalent to the screen for decision 1. They only referred to and showed the picture of picture 2 participant, picture 3 participant, and picture 4 participant (note that picture 4 participant always was the matched participant). Decision 4 in block 5 of part 2 concluded the main part of the experiment.]

C.4 Randomization

Table C.1 reports results from a randomization test between *Control* and *Treatment*. It lists average values by treatment for the subject characteristics. These include individual observables (age, gender, nationality, mother tongue, math grade, relationship status, as well as decision — i.e. reading — time for the first screen which was the same for both treatments) and information obtained from the pictures (whether they made eye contact, looked friendly and how attractive they were rated).

Table C.1: Balance table by treatment conditions

Variable	Control	Treatment	p-value of difference
Age	22.935	23.521	0.354
Female	0.474	0.521	0.334
Geman nationality	0.795	0.822	0.490
German mother tongue	0.791	0.779	0.775
Math grade	2.101	2.072	0.621
Relationship status	0.460	0.484	0.632
Time left after first screen	21.181	21.329	0.962
Eye contact with picture	0.926	0.944	0.449
Friendly face in picture	0.474	0.469	0.919
Attractive (0-10)	4.836	4.891	0.554
Total earnings	12.429	12.942	0.222

Notes: Math grade refers to the final (last) math grade in high school. *Time left after first screen* is seconds left upon reading the instructions onscreen and can serve as proxy for reading and comprehension speed. Measures regarding the photo taken were rated independently by 4 RAs and the average was taken for *attractive*, while for the binary measures regarding eye contact and facial expression the dummy is coded as one if at least three out of four RAs indicated one. The test of difference between the treatments used is either Chi^2 or Mann-Whitney, depending on whether variables only have categorical values or distributions.

C.5 Details on Investment Choices

C.5.1 Investment Patterns for Round Numbers

Partitioning investment choices in specific focal and non-focal investment amounts gives some insights into behavior depending on treatment. While there is neither an overall treatment effect nor a general change in the distribution, it seems that focal (salient) investment amounts become somewhat more important in *Treatment*. Table C.2 displays these patterns. Directionally, more *Treatment* subjects choose to invest all, nothing as well as exactly half of their endowment. Combining these investment amounts into one group results even in a weakly significantly higher share investing either 0, 50 or 100 in *Treatment* (p-value $= 0.05$, Chi^2-test). While not more subjects in *Treatment* invest 25 or 75 (one fourth or three fourth), the difference between the treatments when only considering subjects investing multiples of ten (or zero) becomes even more apparent: More than 90% in *Treatment* invest in such a manner, while only 78% do so in *Control* (p-value < 0.01, Chi^2-test).

While the effects are not very large in magnitude and hence should be interpreted with caution (i.e. multiple testing), they are in line with evidence from accountability studies mentioned in Section 3.1. These studies indicate that subjects who need to justify their choices to the experimenter after the experiment, choose more easily

Table C.2: Fraction of subjects investing focal amounts

	Control	Treatment	p-value
Fraction of subjects investing 100:	0.12	0.14	0.45
Fraction of subjects investing 0:	0.04	0.06	0.49
Fraction of subjects investing 50:	0.24	0.29	0.21
Fraction of subjects investing 0, 50, or 100:	0.40	0.49	0.05*
Fraction of subjects investing 0, 25, 50, 75 or 100:	0.47	0.54	0.18
Fraction of subjects investing multiples of 10:	0.78	0.91	0.00***

Notes: Fractions of subjects, by treatment condition, making investment choices based on certain patterns: Investing the entire endowment, investing nothing, investing exactly have of the endowment, investing one of either of these focal points, investing any amount represented by 25 point increments or investing any amount represented by 10 point increments. The p-value of the test for difference between the treatments is based on Chi^2-tests (expected cells size $>$ 50, results robust to using Fisher exact test). Stars indicate significance, with *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

justifiable options. Investing round numbers can be interpreted in a similar way. Even though subjects do not have to explicitly justify behavior in front of their matched participant, they might still expect uneven investments to be a very specific signal. This indicates that if researchers or marketing departments are indeed interested in exact (and sometimes "weird") values, they should consider these accountability and observability effects in study designs.

C.5.2 Distribution of Investment Choices by Gender

Figure C.2 displays cumulative distribution functions by treatment for males and females separately. This clearly shows that considering average investments in Figure 3.4 does not obscure any more subtle treatment effects on the distribution of choices. Kolmogorov-Smirnov tests confirm this assessment (p-value = 0.98 for males, p-value = 0.66 for females).

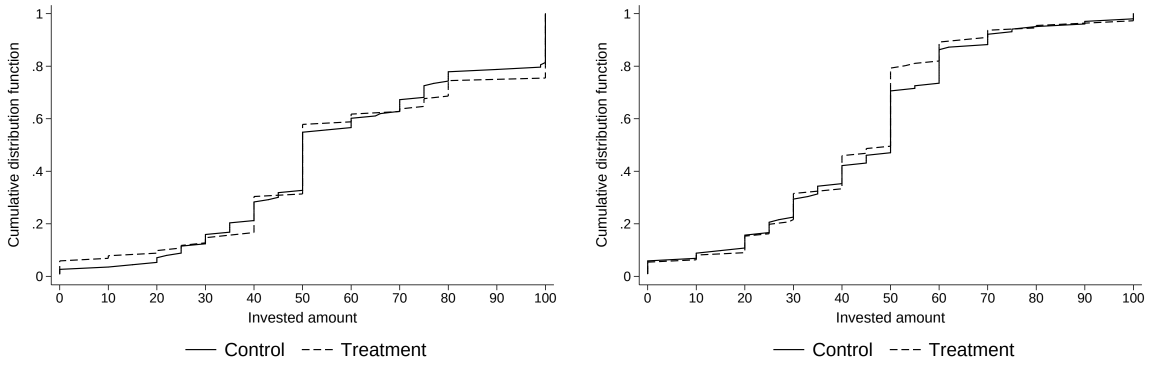
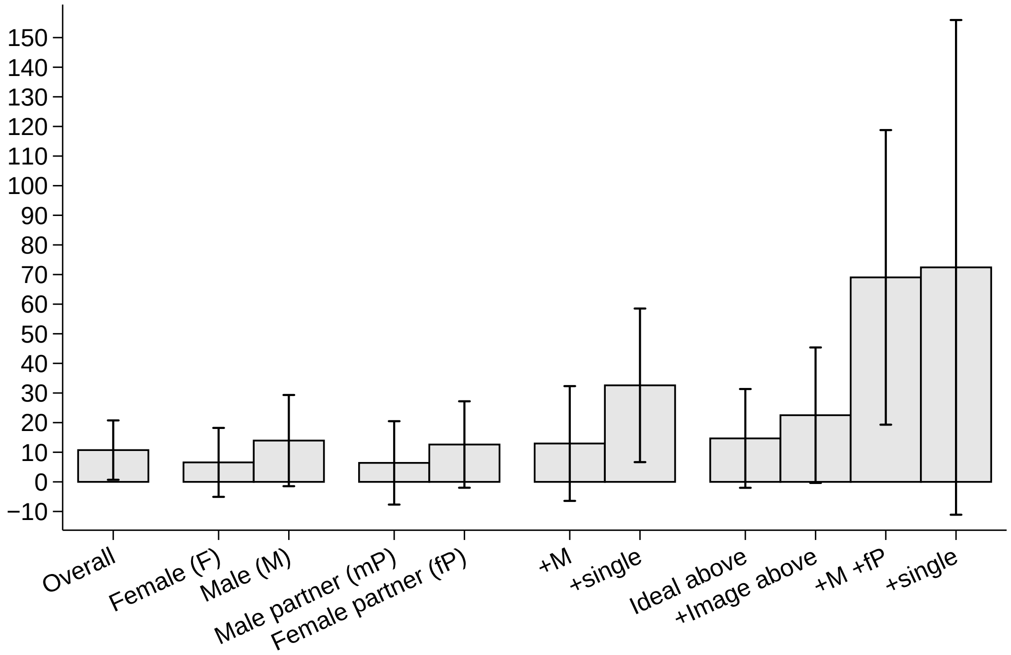


Figure C.2: Cumulative distribution function of investment by treatment for males (left) and females (right) separately

C.5.3 Heterogeneity in Treatment Effects

Figure C.3 displays the difference in the treatment effects between subjects matched with an attractive partner and those matched with an unattractive partner.

Every bar relates to a specific sample considered and the positive values almost always arise from both (directionally) negative treatment effects for those matched with an unattractive participant and (directionally) positive treatment effects for those matched with an attractive participant. Confidence intervals are based on t-tests. For some subsamples, these differences become very large. Interstingly, the further I move to the right in the figure and the higher the intuitively expected treatment effect differences should get, indeed the more pronounced effects I observe.

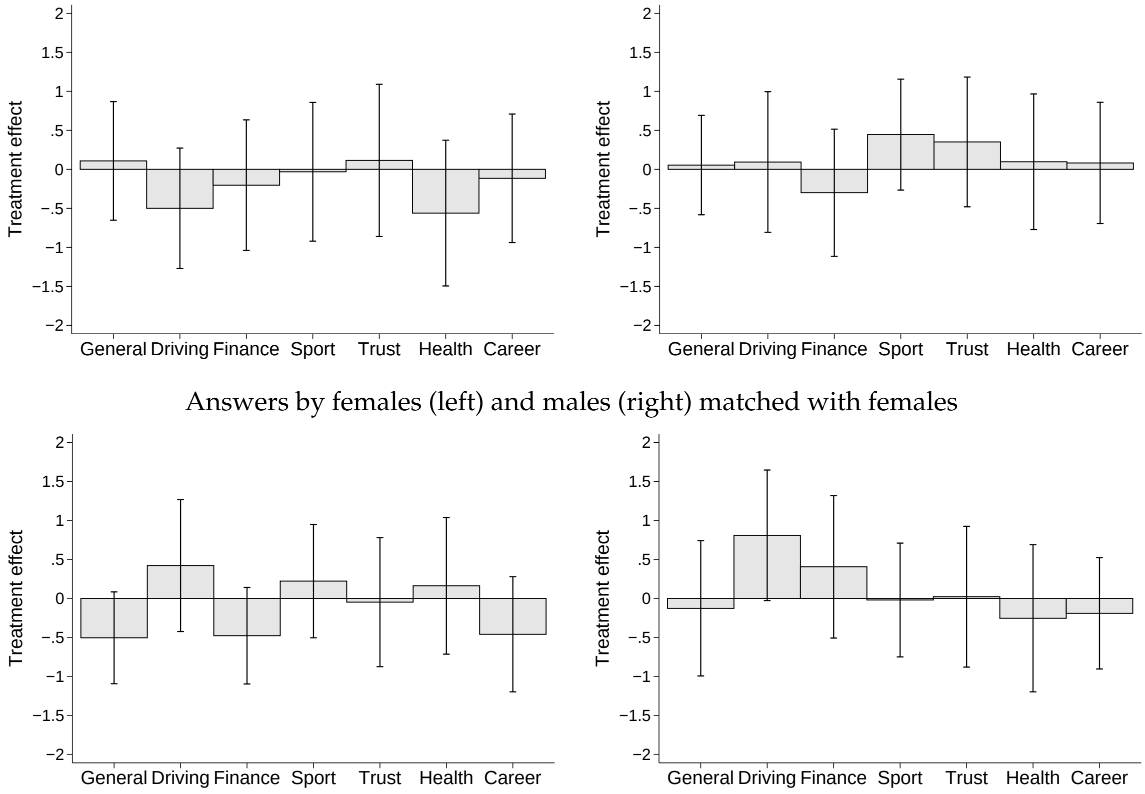


Notes: X-axis labels with a "+" indicate that the respective subsample constraint is put on top of the subsample definition of the bar to the left.

Figure C.3: Treatment effect difference between matched with an attractive vs. unattractive participant by subsamples

C.5.4 Non-Incentivized Domain-Specific Stated Willingness to Take Risks

Just as for overall treatment effects on domain-specific non-incentivized risk taking, I do not observe strong patterns when considering gender pairs separately. See Figure C.4 for the equivalents of Figure 3.5 by gender pairs.



Answers by females (left) and males (right) matched with females

Answers by females (left) and males (right) matched with males

Figure C.4: Treatment effect on non-incentivized domain-specific risk taking by gender pairs

The largest effect observed is the treatment effect on willingness to take risks in driving for males matched with males. When being observed, these subjects state higher risk attitudes than in *Control*. The effect size, however, is still below one and only weakly significant (p-value = 0.08, Mann-Whitney test; p-value = 0.06, two-sided t-test as displayed in Figure C.4). Apart from this difference, based on t-tests, only females' general risk assessment is weakly higher in *Treatment* compared to *Control* when matched with males (p-value = 0.09; p-value = 0.11, Mann-Whitney test). One statistic that does not show up in the figure (and t-tests) is the difference in sports for males matched with females. Based on a non-parametric test response behavior is clearly different in *Treatment* (p-value = 0.02, Mann-Whitney test). This is insignificant with a t-test since the treatment means are not too far apart. However, answers are much more dispersed in *Treatment* such that the Mann-Whitney test results in a significant statistic. For no other gender pairing or domain a similar effect could be observed.

While for these subsamples I do not have sufficient power to detect small effects, I'd still be able to detect economically important differences. Importantly, the statistically small effects would become even less meaningful once corrected for multiple testing.

C.6 More Details on Norms

C.6.1 Norm Choices

For completeness, Table C.3 displays all elicited beliefs (i.e. norms) by gender, matched gender and treatment cells. To allow for a comparison to actual investment choices, average invested amounts by cell are included.

Table C.3: All norms by gender, matched gender and treatment cells

	MMC	MMT	MFC	MFT	FMC	FMT	FFC	FFT
Guess partner	49.60	45.53	44.44	43.84	52.22	48.88	40.23	41.44
Guess all	46.19	42.28	50.74	48.55	49.87	44.68	45.40	47.44
Perceived norm	52.19	44.77	49.79	51.65	46.13	39.32	41.83	45.60
Perceived norm partner	54.32	47.32	49.45	50.04	50.47	42.03	46.00	46.29
Stated norm (1-3)	47.99	44.13	47.23	49.89	50.11	44.94	45.82	49.81
Stated norm (4)	50.30	44.96	43.70	45.11	52.82	47.55	42.77	45.87
Investment	56.02	53.94	61.48	64.36	49.95	41.55	38.40	46.11

Notes: All choices in the belief elicitation part of the experiment (and investment) by gender, matched partner gender and treatment with MMC (males in *Control* matched with males), MMT (males in *Treatment* matched with males), MFC (males in *Control* matched with females), MFT (males in *Treatment* matched with females), FMC, FMT, FFC and FFT (the four groups equivalent to before only for female decision makers) denoting the different treatment combination cells. *Guess partner* refers to the average guess for the investment of the matched partner for subjects in the respective cell. *Guess all* equivalently refers to the guessed overall investment in the given session. *Perceived norm* is the subject's belief regarding the average stated norm of the four people being shown the subject's picture. *Perceived norm partner* denotes the beliefs about that stated norm by the matched partner. *Stated norm (1-3)* and *stated norm (4)* refer to average stated norms by subjects in the respective cell (for the first three pictures seen, and for the fourth picture — the picture of the matched partner).

Figure C.5 shows kernel densities for *perceived norm* and *investment*. The distributions show that a strong focal point at half the investment amount can neither per se explain the null effect of the treatment on investment, nor the average perceived norm level of 50. For both outcomes there is sufficient variation in participant answers.

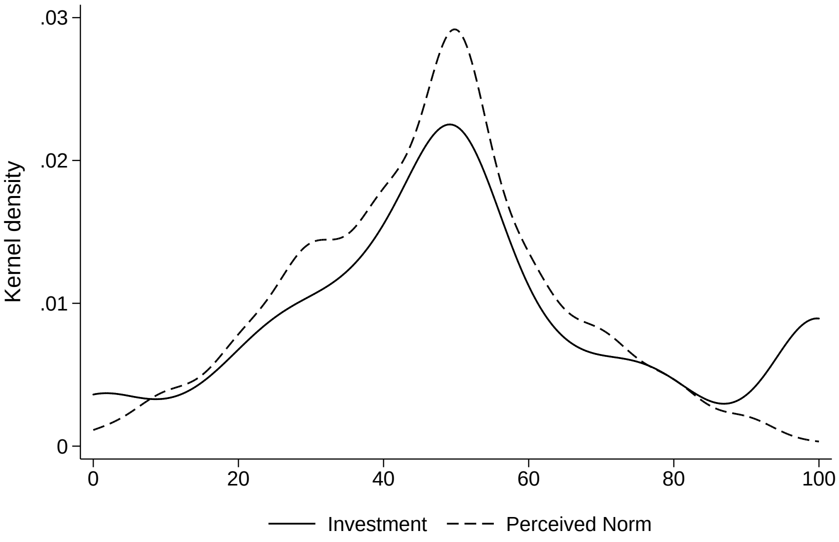
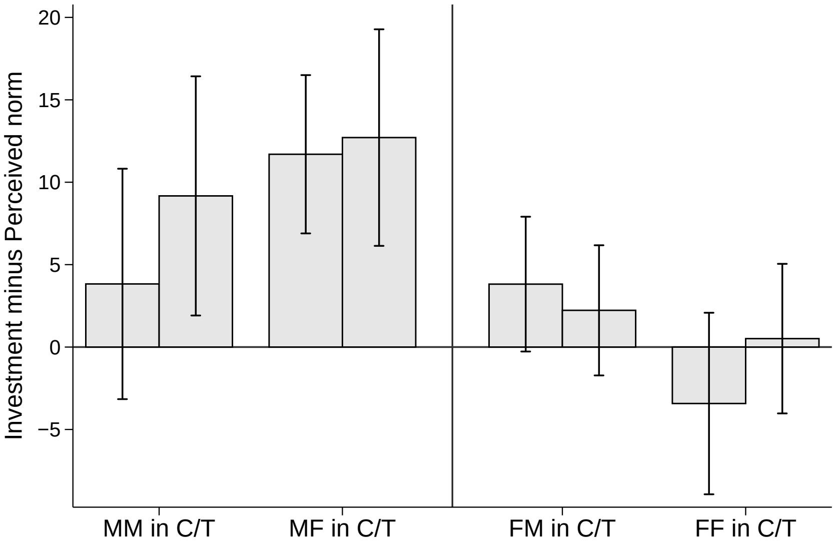


Figure C.5: Kernel density functions for *perceived norm* and *investment*

C.6.2 Norm Following



Notes: MM (males matched with males), MF (males matched with females), FM (females matched with males) and FF (females matched with females) refer to the four gender pairing cells. Within a gender pair, norm following behavior is split by treatment condition with behavior in *Control* (C) displayed always on the left and *Treatment* always shown on the right.

Figure C.6: Norm following by treatment, gender and partner gender cells

As indicated in Section 3.4.2 norm following overall does not depend on treatment condition. Figure C.6 shows norm following not only by gender pairs, but also by treatment condition. The first letter of the x-axis labels refers to the gender of the decision maker and the second letter to the gender of the matched participant. "C" and "T" denote *Control* and *Treatment*, respectively.

All differences, including the differences for males matched with males and females matched with females, by treatment are not significant. Hence, norm following does neither overall nor for the different gender pairs separately depend on the treatment.

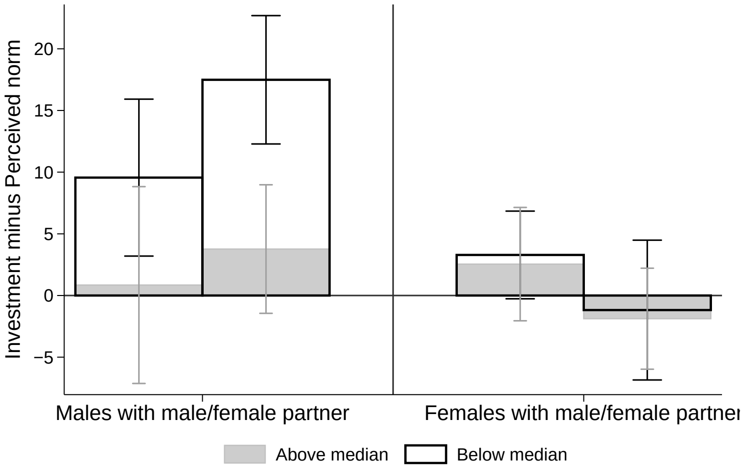


Figure C.7: Norm following by gender and matched gender for above and below median norm conformists

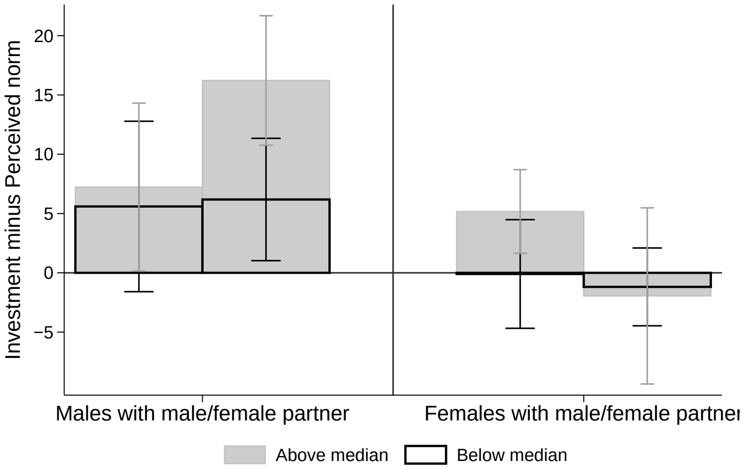


Figure C.8: Norm following by gender and matched gender for above and below rule breakers

Figure C.7 and C.8 are equivalent to Figure 3.7 in the main text, but split the sample by median norm conformity and rule breaking preferences, respectively. This clearly indicates that the "overshooting" observed for males is almost entirely driven by and very strong for males with below median *norm conformity*. For subjects with high norm conformity preferences, investment does not significantly deviate from perceived norms.

I expect a similar relationship between rule-breaking preferences and norm following, just in the reverse direction. The concept measures — to a large extent — very similar aspects as *norm conformity*. This is what I find. Again, "overshooting" by males is mainly driven by above median "rule breakers". However, for *rule breaking*, also below median males do "overshoot" perceived norms when making investment choices. Further, also for females matched with males "rule breakers" significantly "overshoot" (p-value = 0.01, Wicoxon signrank test).

Not relying on median splits, but rather using the entire distribution of *rule breaking* and *norm conformity* in a linear regression model leads to the same inference. Overall, *norm conformity* is negatively (p-value = 0.02) and *rule breaking* positively (p-value < 0.01) linked to individual norm following. Also here, there is no significant interaction with *Treatment*.

The arguments above and in the main text relate to norm following defined as invested amount minus individual perceived norm. This is adequate to describing and assessing the norm "overshooting" apparent in my data. However, I can also look at absolute norm deviations only instead of differentiating between positive and negative deviations. This results in very similar conclusions. Males do significantly more strongly deviate (ignoring the direction of the deviation) from norms than females. Further, also this absolute norm deviation is related to *norm conformity* and *rule breaking*. Both, below median norm conformists and above median rule breakers do significantly more strongly deviate from norms than their counterparts. Lastly and importantly, absolute deviations from norms are again independent of the treatment condition.

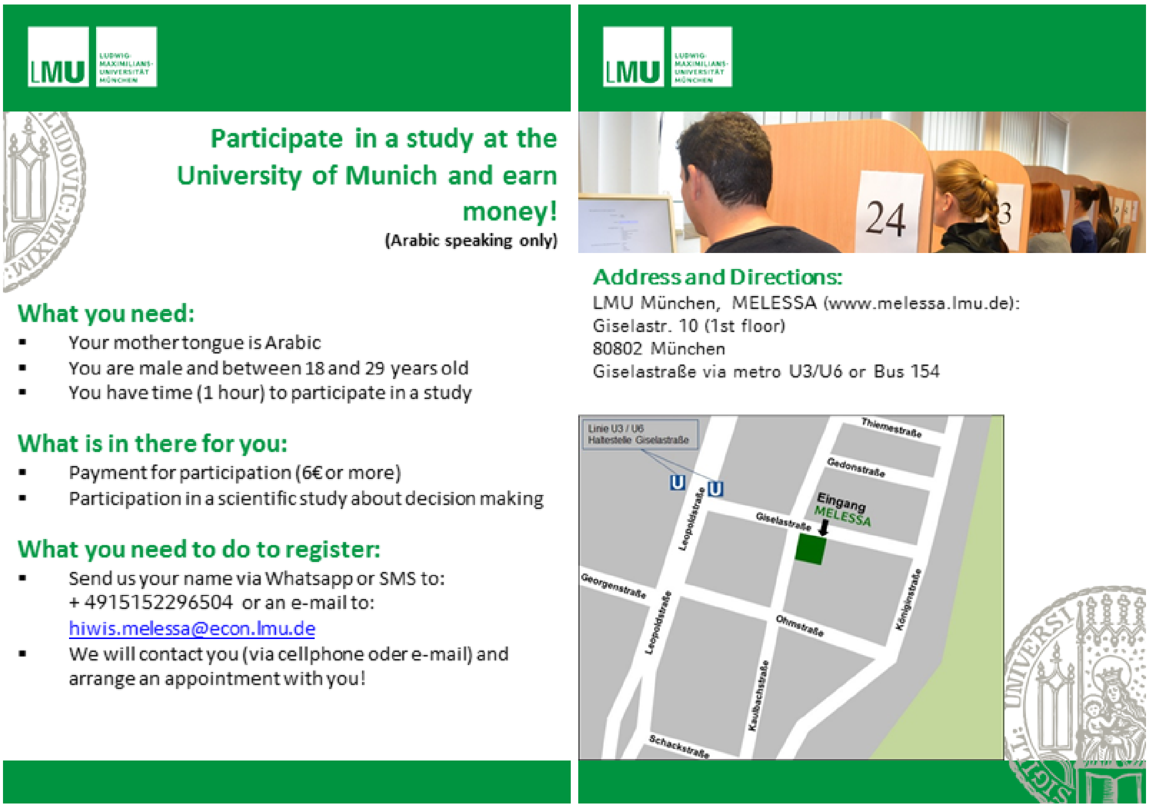
Appendix D

Blaming the Refugees? Experimental Evidence on Responsibility Attribution

D.1 Refugee Recruiting Details

Refugees were recruited by distributing the leaflet shown in Figure D.1. The actual first names of the refugees taking part in the experiment and which were visible to the matched partner were: Abdo, Abduh, Abdullah (2x), Adnan, Ahmad (3x), Alaa, Ali, Alkhder, Almhklf, Amjad, Anas, Bshr, Firas, Ghassan, Ghiath, Giwan, Hafez, Hasan, Khaled (2x), Louay, Mazen (2x), Mohamad, Mohamd, Mohammad, Mohammed (3x), Mounir, Nizar, Obaida, Odai, Omar, Sabri, Saleem, Schindar, Wissam, Yazan, Youssef.

The names of the German participants were: Aleksandar, Alex, Alexander (3x), Aljoscha, Andi, Andreas (2x), Axel, Ben, Benedikt, Benjamin, Benno, Bernhard, Caspar, Chris, Christian (3x), Christoph, Christopher, Daniel (4x), David (4x), Dominic, Dominik (2x), Eric, Fabian (7x), Felix (3x), Fiete, Florian (2x), Franz, Franziskus, Fridtjof, Gregor, Ion, Jan, Jan Fedor, Jens, Joel, Johannes (4x), Jonas (3x), Jonathan (2x), Josaphat, Julian (3x), Kevin, Konstantin (2x), Korbinian (2x), Laurian, Lennart, Leon, Leonard, Lion, Louis, Lukas (2x), Manuel, Marcus (3x), Marian, Marius (4x), Markus (3x), Martin (2x), Matthias (5x), Maurus, Max (5x), Maximilian (3x), Michael (4x), Moritz, Niclas, Niklas, Niko, Oswald, Pascal, Patrick, Paul, Philipp (4x), Raffael, Richie, Roman, Sebastian (3x), Simon, Stefan (3x), Steffen, Stephan (2x), Thomas (3x), Tilman, Tim, Timo, Tobi, Tobias (3x), Tom, Valentin, Vincent.



Participate in a study at the University of Munich and earn money!
(Arabic speaking only)

What you need:

- Your mother tongue is Arabic
- You are male and between 18 and 29 years old
- You have time (1 hour) to participate in a study

What is in there for you:

- Payment for participation (6€ or more)
- Participation in a scientific study about decision making

What you need to do to register:

- Send us your name via Whatsapp or SMS to: + 4915152296504 or an e-mail to: hiwis.melessa@econ.lmu.de
- We will contact you (via cellphone oder e-mail) and arrange an appointment with you!

Address and Directions:

LMU München, MELESSA (www.melessa.lmu.de):
Giselastr. 10 (1st floor)
80802 München
Giselastraße via metro U3/U6 or Bus 154

Figure D.1: Leaflet for recruiting refugees (translated from Arabic)

D.2 Instructions

The following passages are the instructions for *Cond* translated from German. Text in italics refers to instructions read out aloud by the experimenter (alternating one of the two authors), which were repeated in Arabic. Text in brackets indicates self-explaining comments. Text in normal letters refers to instruction that the subjects read on screen (either in German or Arabic).

[upon arrival at the laboratory]
Hello everybody. We provide refugees with the possibility to take part in a series of experiments. This is why there are refugees among the participants today. In order to assign you to the seat with the correct language [experimenter points at the two bags labeled with "German" or "Arabic"] *Arabic-speaking participants draw a card with a seat number from the bag with the label Arabic and German-speaking participants a card from the bag with the label German.*

[in the laboratory after seating took place]
Welcome to MELESSA. Thank you very much for showing up to this experiment on time. My name is Felix Klimm/Stefan Grimm, and I will conduct this experiment today.

Please do not talk to other participants during the experiment.

For the sake of simplicity, you find the instructions on your screen. The instructions are the same for all participants. Please follow the instructions. If you have any questions, please raise your hand or press the red button on your keyboard. We will then come to you and answer your question in private.

[first screen]
General Procedures I
This experiment is meant to study economic decision making. It will last about 1 hour. You can earn money during the experiment. This money will be paid to you in private after the experiment. You will make decisions in this study. These decisions will affect your payment. In addition, your payment might depend on other participant's decisions as well as on chance. Further rules will be explained to you right before each decision. Hence, today's payment is the sum of money earned with your decisions plus €6 for showing up on time.

[new screen]
General Procedures II
The experiment consists of 2 parts. You will see the instructions for each part right before the respective part starts. Data from this experiment will be analyzed anonymously. At the end of the experiment, you will have to sign a receipt. This is only for accounting purposes.

[new screen]
Part 1
In part 1 of the experiment, you need to perform a task. You receive €3 for performing this task. Your task is to correctly solve as many puzzles as possible. This task is suited for everybody as puzzles are well known in most parts of the world. For this purpose, there are 8 puzzles next to your keyboard. You are allowed to start as soon as we tell you to do so. After 10 minutes, you need to stop, and we will count the number of correct puzzles. There will be a clock on your screen displaying the remaining time. Click on OK if you understand the procedure. Please still wait with solving a puzzle until we tell you to start.

[Subjects perform real effort and the experimenter and student research assistants checks the number of correctly solved puzzles.]

[new screen]
Part 2
You are now matched with another participant. Please enter your first name for this purpose. Thereafter, the first name of your matched participant will be shown to you. Your matched participant will see your first name.

Your first name: ≪own name≫

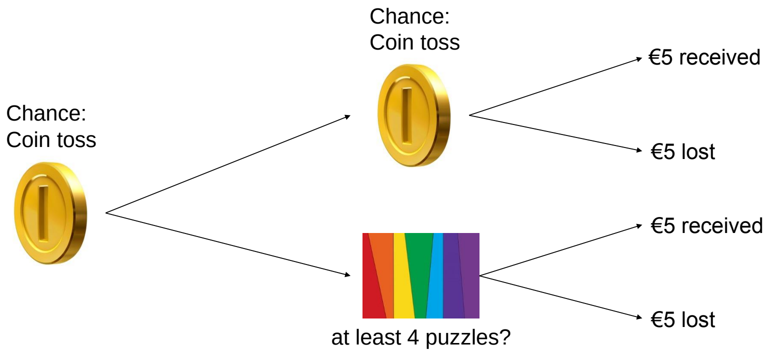
[new screen]
Your matched participant is: ≪name partner≫

[new screen]
Your payoff might depend on your matched participant's decisions. Reminder: Your matched participant is ≪name partner≫. In the following, you can receive additional €5 or lose €5. Whether you are receiving or losing €5 depends on chance or the other participant. First, the computer will determine via a virtual coin flip whether chance or the other participant is responsible for your payment. Both cases are equally likely (50/50). Hence, there are 2 possibilities:

1. If chance is responsible, you will receive €5 with 50% probability. Hence, a coin will be flipped again.

2. If ≪name partner≫ is responsible, the number of puzzles that ≪name partner≫ solved correctly in part 1 will determine whether you receive or lose €5. If ≪name partner≫ solved at least 4 puzzles, you will receive €5. If ≪name partner≫ solved fewer than 4 puzzles, you will lose €5.

The graph below illustrates the procedure.



[new screen]
You will know about your payment in a second. However, you will not know whether chance or ≪name partner≫ is responsible for this payment.

Please answer four test questions in order to be sure that you understand the procedure.

[new screen]
1. If ≪name partner≫ solved at least 4 puzzles, will you receive €5 in any case?

2. If ≪name partner≫ solved 3 or fewer puzzles and chance was selected to be responsible for your payment, how likely is it that you will receive €5?

3. If chance was selected to be relevant for your payment, does your payment depend on the number of correctly solved puzzles by ≪name partner≫ in this case?

4. How much lower will your payment be if you lose €5 compared to the case in which you receive €5?

[new screen]
You have answered all the questions correctly. On the next screen you will see whether you receive or lose €5.

[new screen]
Your income:

Reminder: The computer randomly determined whether chance or ≪name partner≫ is relevant for your payment. According to these rules:

You receive/lose €5.

[new screen]
We now ask you to answer 4 questions. One of the questions will be randomly selected at the end of the experiment. You will then receive payment according to your answer to this question.

[new screen]
Question 1
Do you believe that chance or ≪name partner≫ was responsible for your payment?

If your answer is correct and this questions will be selected to be payoff relevant, you receive €5.

[new screen]
Question 2
You will now make a sequence of decisions. Each of the decisions contains 2 options — A and B. Both options give you once more the chance to receive another €5.

One of the 9 rows will be randomly chosen for payment if question 2 will be payoff relevant.

If you choose option A in one of the 9 rows, you will receive €5 if ≪name partner / chance≫ [name of partner or chance displayed depending on the answer to Question 1 — name of the partner displayed if subject indicated that the partner is responsible] was responsible for your payment.

If you choose option B, you will receive €5 with a certain probability. This probability varies from 10 to 90 percent and is shown to you next to every decision.

If question 2 is payoff relevant, one of your 9 decisions will be implemented. The computer will randomly select which decision will be implemented in this case.

Please consider now from which probability on (which row) you want to choose option B. If you took your decision, click on OK.

Option A You receive €5 if ≪name partner / chance≫ [here, again, name of partner or chance displayed depending on the answer to Question 1] was responsible for your payment.

Option B You receive €5 with a probability of 10% ... 90%.

[new screen]
Question 3
Do you believe that ≪name partner≫ solved at least 4 puzzles? Hence, did he solve 4, 5, 6, 7, or 8 puzzles?

If your answer is correct and this questions will be selected to be payoff relevant, you receive additional €5.

[new screen]
Question 4
In question 4 — like in question 2 — you will make a sequence of decisions. Each

of the decisions contains 2 options — A and B. Both options give you the chance to receive another €5.

One of the 9 rows will be randomly chosen for payment if question 4 will be payoff relevant.

If you choose option A in one of the 9 rows, you will receive €5 if ≪name partner≫ solved at least 4 puzzles.

If you choose option B, you will receive €5 with a certain probability. This probability varies from 10 to 90 percent and is shown to you next to every decision.

If question 4 is payoff relevant, one of your 9 decisions will be implemented. The computer will randomly select which decision will be implemented in this case.

Please consider now from which probability on (which row) you want to choose option B. If you took your decision, click on OK.

Option A You receive €5 if ≪name partner≫ solved at least 4 puzzles.

Option B You receive €5 with a probability of 10% ... 90%.

D.3 Puzzle Motives

The selected motives for the puzzles are pictures of a range of colors, a bird, a beach, a lamb, a tree in a desert, a sunset over the ocean, a water drop, and a box of bananas. They are displayed in Figure D.2.

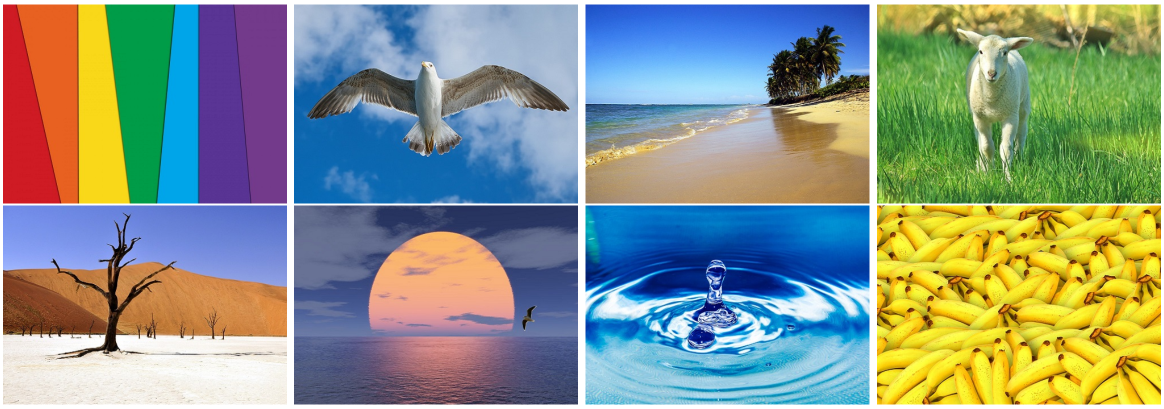
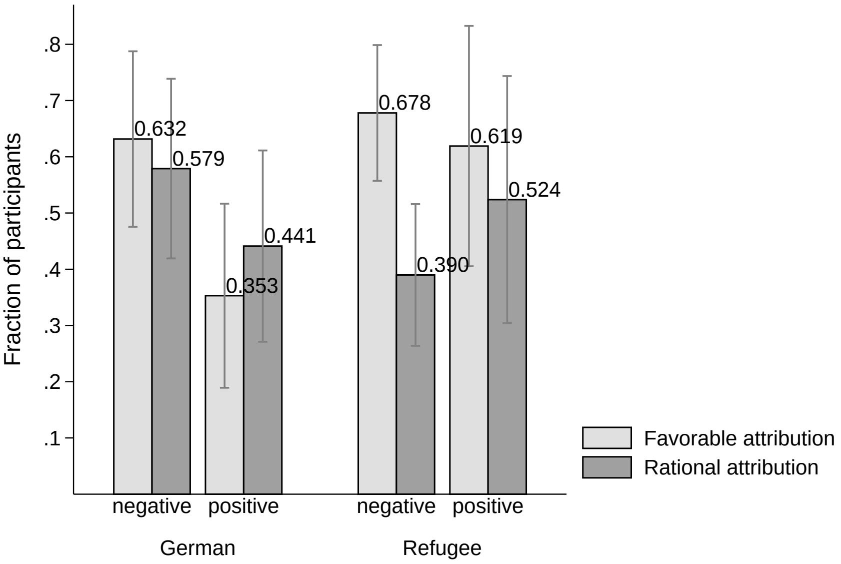


Figure D.2: Puzzle motives for real effort task

D.4 Supplementary Results

D.4.1 Responsibility Attribution by Shock



Notes: The figure shows *favorable attribution* and *rational attribution* for both treatments divided by shock direction. Error bars indicate 95% confidence intervals.

Figure D.3: *Favorable attribution* and *rational attribution* by shock direction

Figure D.3 shows actual attribution behavior and counterfactual rational attribution based on performance beliefs for both group affiliations by shock direction. Even though, at first glance, it looks as if behavior in *Refugee* after a negative shock drives reverse discrimination, comparing behavior across the two group affiliation shows that the difference in difference is rather similar for both shocks. After a negative shock, participants in *Refugees* deviate by 0.288 from rational attribution, while those in *German* attribute responsibility more favorably by 0.053. This is a difference in difference of 0.235. After a positive shock, the deviation for participants in *Refugees* is 0.095 and -0.088 in *German*. Hence, the difference in difference sums up to 0.183, and is therefore close to 0.235 after a negative shock.

D.4.2 Balance Table *Cond* vs. *Uncond*

Table D.1: Balance table *Refugee Experiment* (*Cond* vs. *Uncond*)

	Cond (1)	*Uncond* (2)	(1) vs. (2) p-value
Own performance	0.368	0.579	0.009
Age	22.474	23.303	0.160
Semester	4.224	4.553	0.534
Number of experiments so far	5.461	8.250	0.021

Notes: Own performance indicates whether a subject solved four or more puzzles.

D.4.3 Regression Analysis Controlling for Own Performance

Table D.2 reports results from regressions equivalent to our main regressions in Table 4.1 (Section 4.3.1) only using the number of correctly solved puzzles as control variable instead of performance beliefs directly.

Table D.2: Favorable responsibility attribution (controlling for own performance)

	Favorable attribution			
	(1)	(2)	(3)	(4)
Refugee	0.160***	0.181***	0.144***	0.139***
	(0.056)	(0.055)	(0.047)	(0.044)
# correct puzzles		0.089***	0.086***	0.091***
		(0.022)	(0.023)	(0.022)
Neg shock			0.159**	0.148**
			(0.063)	(0.064)
Additional controls	No	No	No	Yes
Observations	152	152	152	152
Pseudo R^2	0.020	0.062	0.081	0.090

Notes: Probit regressions on *favorable attribution* reporting average marginal effects. Column (4) includes additional covariates from the questionnaire: age, semester, and number of experiments so far (all insignificant). Robust and clustered (on session level) standard errors in parentheses. Stars indicate significance on the levels: *** p<0.01, ** p<0.05, * p<0.1.

D.4.4 Balance Table for the *KleeKandinsky Experiment*

Table D.3: Balance table *KleeKandinsky Experiment*

	Outgroup (1)	*Ingroup* (2)	(1) vs. (2) p-value
Own performance	0.686	0.514	0.037
Age	24.729	24.875	0.842
Semester	5.129	5.736	0.220
Number of experiments so far	11.700	10.542	0.401

Notes: Own performance indicates whether a subject solved four or more puzzles.

D.5 Interaction Effect of IAT Score and Being Matched with a Refugee

For estimating the interaction effect between having a negative IAT score and our treatment, we compute predictive values for *favorable attribution* by using probit regression estimates from model (3) used in Table 4.2 for the following four groups:

- Subjects in *Refugee* with a negative IAT score:
  $\widehat{P(Y = 1|Refugee = 1, IAT < 0, X)} = 0.5862$
- Subjects in *Refugee* with a positive IAT score:
  $\widehat{(Y = 1|Refugee = 1, IAT > 0, X)} = 0.8375$
- Subjects in *German* with a negative IAT score:
  $\widehat{P(Y = 1|Refugee = 0, IAT < 0, X)} = 0.5295$
- Subjects in *German* with a positive IAT score:
  $\widehat{P(Y = 1|Refugee = 0, IAT > 0, X)} = 0.4189$

This leaves us with a difference in differences of –0.3619 ([0.5862 – 0.8375] – [0.5295 – 0.4189]). Thus, the effect of having a negative IAT score on *favorable attribution* is 36.19 percentage points lower in *Refugee* than in *German*.

Bibliography

Adam, M. T., Kroll, E. B., and Teubner, T. (2014). A note on coupled lotteries. *Economics Letters*, 124(1):96–99.

Ai, C. and Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters*, 80(1):123–129.

Akerlof, G. A. (1980). A theory of social custom, of which unemployment may be one consequence. *Quarterly Journal of Economics*, 94(4):749–775.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *Quarterly Journal of Economics*, 115(3):715–753.

Amabile, T. M. (1996). *Creativity in context: Update to the social psychology of creativity*. Westview Press, Boulder, Colorado.

Andreoni, J. and Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5):1607–1636.

Ariely, D., Bracha, A., and Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review*, 99(1):544–55.

Ariely, D. and Levav, J. (2000). Sequential choice in group settings: Taking the road less traveled and less enjoyed. *Journal of Consumer Research*, 27(3):279–290.

Auster, C. J. and Ohm, S. C. (2000). Masculinity and femininity in contemporary american society: A reevaluation using the Bem Sex-Role Inventory. *Sex Roles*, 43(7):499–528.

Autor, D. H. and Handel, M. J. (2013). Putting tasks to the test: Human capital, job tasks, and wages. *Journal of Labor Economics*, 31(S1):S59–S96.

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4):1279–1333.

Autor, D. H. and Price, B. (2013). The changing task composition of the US labor market: An update of Autor, Levy, and Murnane (2003). *Working Paper*.

Azoulay, P., Graff Zivin, J. S., and Manso, G. (2011). Incentives and creativity: Evidence from the academic life sciences. *RAND Journal of Economics*, 42(3):527–554.

Backhouse, R. E. and Medema, S. G. (2009). Defining economics: the long road to acceptance of the robbins definition. *Economica*, 76(s1):805–820.

Baker, M. D. and Maner, J. K. (2008). Risk-taking as a situationally sensitive male mating strategy. *Evolution and Human Behavior*, 29(6):391–395.

Baker, M. D. and Maner, J. K. (2009). Male risk-taking as a context-sensitive signaling device. *Journal of Experimental Social Psychology*, 45(5):1136–1139.

Bandiera, O., Barankay, I., and Rasul, I. (2005). Social preferences and the response to incentives: Evidence from personnel data. *Quarterly Journal of Economics*, 120(3):917–962.

Bandiera, O., Barankay, I., and Rasul, I. (2013). Team incentives: Evidence from a firm level experiment. *Journal of the European Economic Association*, 11(5):1079–1114.

Bartling, B. and Fischbacher, U. (2011). Shifting the blame: On delegation and responsibility. *Review of Economic Studies*, 79(1):67–87.

Bartling, B., Fischbacher, U., and Schudy, S. (2015). Pivotality and responsibility attribution in sequential voting. *Journal of Public Economics*, 128:133–139.

Bault, N., Coricelli, G., and Rustichini, A. (2008). Interdependent utilities: How social ranking affects choice behavior. *PLoS ONE*, 3(10):e3477.

Becker, G. M., DeGroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3):226–232.

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 42(2):155–162.

Bohnet, I. and Frey, B. S. (1999). Social distance and other-regarding behavior in dictator games: Comment. *American Economic Review*, 89(1):335–339.

Bolton, G. E., Brandts, J., and Ockenfels, A. (2005). Fair procedures: Evidence from games involving lotteries. *The Economic Journal*, 115(506):1054–1076.

Bolton, G. E., Katok, E., and Zwick, R. (1998). Dictator game giving: Rules of fairness versus acts of kindness. *International Journal of Game Theory*, 27(2):269–299.

Bolton, G. E. and Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193.

Bolton, G. E. and Ockenfels, A. (2010). Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States: Comment. *American Economic Review*, 100(1):628–633.

Bolton, G. E., Ockenfels, A., and Stauf, J. (2015). Social responsibility promotes conservative risk behavior. *European Economic Review*, 74:109–127.

Bradler, C., Neckermann, S., and Warnke, A. J. (2014). Rewards and performance: A comparison across a creative and a routine task. *Working Paper*.

Brandts, J. and Charness, G. (2011). The strategy versus the direct-response method: A first survey of experimental comparisons. *Experimental Economics*, 14(3):375–398.

Brennan, G., González, L. G., Güth, W., and Levati, M. V. (2008). Attitudes toward private and collective risk in individual and strategic choice situations. *Journal of Economic Behavior & Organization*, 67(1):253–262.

Brennan, G. and Pettit, P. (2004). *The economy of esteem: An essay on civil and political society*. Oxford University Press, Oxford.

Brosig, J. (2002). Identifying cooperative behavior: Some experimental results in a prisoner's dilemma game. *Journal of Economic Behavior & Organization*, 47(3):275–290.

Bursztyn, L., Ederer, F., Ferman, B., and Yuchtman, N. (2014). Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions. *Econometrica*, 82(4):1273–1301.

Byrd, D. T., Hall, D. L., Roberts, N. A., and Soto, J. A. (2015). Do politically non-conservative whites "bend over backwards" to show preferences for black politicians? *Race and Social Problems*, 7(3):227–241.

Cai, J., De Janvry, A., and Sadoulet, E. (2015). Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108.

Carneiro, P., Heckman, J. J., and Masterov, D. V. (2005). Labor market discrimination and racial differences in premarket factors. *The Journal of Law and Economics*, 48(1):1–39.

Carr, P. B. and Steele, C. M. (2010). Stereotype threat affects financial decision making. *Psychological Science*, 21(10):1411–1416.

Chakravarty, S., Harrison, G. W., Haruvy, E. E., and Rutström, E. E. (2011). Are you risk averse over other people's money? *Southern Economic Journal*, 77(4):901–913.

Charness, G. and Gneezy, U. (2012). Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization*, 83(1):50–58.

Charness, G. and Grieco, D. (2018). Creativity and incentives. *Journal of the European Economic Association*, forthcoming.

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

Chen, L.-H., Baker, S. P., Braver, E. R., and Li, G. (2000). Carrying passengers as a risk factor for crashes fatal to 16-and 17-year-old drivers. *Journal of the American Medical Association*, 283(12):1578–1582.

Chen, Y. and Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, 99(1):431–457.

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6):1015–1026.

Cooper, D. J. and Rege, M. (2011). Misery loves company: Social regret and social interaction effects in choices under risk and uncertainty. *Games and Economic Behavior*, 73(1):91–110.

Crosetto, P. and Filippin, A. (2017). Safe options induce gender differences in risk attitudes. *Working Paper*.

Croson, R. and Gächter, S. (2010). The science of experimental economics. *Journal of Economic Behavior & Organization*, 73(1):122–131.

Curley, S. P., Yates, J. F., and Abrams, R. A. (1986). Psychological sources of ambiguity avoidance. *Organizational Behavior and Human Decision Processes*, 38(2):230–256.

D'Acunto, F. (2015). Identity, overconfidence, and investment decisions. *Working Paper*.

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

Deci, E. L., Koestner, R., and Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6):627–668.

Delfgaauw, J. and Dur, R. (2010). Managerial talent, motivation, and self-selection into public management. *Journal of Public Economics*, 94(9):654 – 660.

Delfgaauw, J., Dur, R., Non, A., and Verbeke, W. (2015). The effects of prize spread and noise in elimination tournaments: A natural field experiment. *Journal of Labor Economics*, 33(3):521–569.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature*, 47(2):315–372.

DellaVigna, S. and Pope, D. (2017). What motivates effort? Evidence and expert forecasts. *Review of Economic Studies*, forthcoming.

Di Tella, R., Perez-Truglia, R., Babino, A., and Sigman, M. (2015). Conveniently upset: Avoiding altruism by distorting beliefs about others' altruism. *American Economic Review*, 105(11):3416–3442.

Diecidue, E. and Van De Ven, J. (2008). Aspiration level, probability of success and failure, and expected utility. *International Economic Review*, 49(2):683–700.

Dijk, O., Holmen, M., and Kirchler, M. (2014). Rank matters – The impact of social competition on portfolio choice. *European Economic Review*, 66:97–110.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550.

Dufwenberg, M. and Muren, A. (2006). Generosity, anonymity, gender. *Journal of Economic Behavior & Organization*, 61(1):42–49.

Duncker, K. (1945). On problem-solving. *Psychological Monographs*, 58(5):i–113.

Dutton, D. G. (1973). Reverse discrimination: The relationship of amount of perceived discrimination toward a minority group on the behaviour of majority group members. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 5(1):34–45.

Ebbesen, E. B. and Haney, M. (1973). Flirting with death: Variables affecting risk taking at intersections. *Journal of Applied Social Psychology*, 3(4):303–324.

Eckartz, K., Kirchkamp, O., and Schunk, D. (2012). How do incentives affect creativity? *Working Paper*.

Eckel, C. C. and Grossman, P. J. (2008). Men, women and risk aversion: Experimental evidence. *Handbook of Experimental Economics Results*, 1:1061–1073.

Ederer, F. and Manso, G. (2013). Is pay for performance detrimental to innovation? *Management Science*, 59(7):1496–1513.

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, 3(4):99–117.

Englmaier, F., Roider, A., and Sunde, U. (2017). The role of communication of performance schemes: Evidence from a field experiment. *Management Science*, 63(12):4061–4080.

Erat, S. and Gneezy, U. (2016). Incentives for creativity. *Experimental Economics*, 19(2):269–280.

Erev, I., Bornstein, G., and Galili, R. (1993). Constructive intergroup competition as a solution to the free rider problem: A field experiment. *Journal of Experimental Social Psychology*, 29(6):463–478.

Fafchamps, M., Kebede, B., and Zizzo, D. J. (2015). Keep up with the winners: Experimental evidence on risk taking, asset integration, and peer effects. *European Economic Review*, 79:59–79.

Farthing, G. W. (2005). Attitudes toward heroic and nonheroic physical risk takers as mates and as friends. *Evolution and Human Behavior*, 26(2):171–185.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3):817–868.

Fershtman, C. and Gneezy, U. (2001). Discrimination in a segmented society: An experimental approach. *Quarterly Journal of Economics*, 116(1):351–377.

Filippin, A. and Crosetto, P. (2016). A reconsideration of gender differences in risk attitudes. *Management Science*, 62(11):3138–3160.

Filiz-Ozbay, E. and Ozbay, E. Y. (2014). Effect of an audience in public goods provision. *Experimental Economics*, 17(2):200–214.

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

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science*, 47(1):117–132.

Forsythe, R., Horowitz, J. L., Savin, N. E., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3):347–369.

Frank, R. H. (2005). Positional externalities cause large and preventable welfare losses. *American Economic Review*, 95(2):137–141.

Frankenhuis, W. E., Dotsch, R., Karremans, J. C., and Wigboldus, D. H. (2010). Male physical risk taking in a virtual environment. *Journal of Evolutionary Psychology*, 8(1):75–86.

Frankenhuis, W. E. and Karremans, J. C. (2012). Uncommitted men match their risk taking to female preferences, while committed men do the opposite. *Journal of Experimental Social Psychology*, 48(1):428–431.

Friebel, G. and Giannetti, M. (2009). Fighting for talent: Risk-taking, corporate volatility and organisation change. *The Economic Journal*, 119(540):1344–1373.

Friebel, G., Heinz, M., Krüger, M., and Zubanov, N. (2017). Team incentives and performance: Evidence from a retail chain. *American Economic Review*, 107(8):2168–2203.

Friedl, A., De Miranda, K. L., and Schmidt, U. (2014). Insurance demand and social comparison: An experimental analysis. *Journal of Risk and Uncertainty*, 48(2):97–109.

Fryer, R., Levitt, S., List, J., and Sadoff, S. (2012). Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. *Working Paper*.

Gächter, S. and Fehr, E. (1999). Collective action as a social exchange. *Journal of Economic Behavior & Organization*, 39(4):341–369.

Gächter, S., Johnson, E. J., and Herrmann, A. (2007). Individual-level loss aversion in riskless and risky choices. *IZA Discussion Paper*.

Gerhart, B. and Fang, M. (2015). Pay, intrinsic motivation, extrinsic motivation, performance, and creativity in the workplace: Revisiting long-held beliefs. *Annual Review of Organizational Psychology and Organizational Behavior*, 2(1):489–521.

Gibbs, M., Neckermann, S., and Siemroth, C. (2017). A field experiment in motivating employee ideas. *Review of Economics and Statistics*, 99(4):577–590.

Gill, M. J. (2004). When information does not deter stereotyping: Prescriptive stereotyping can foster bias under conditions that deter descriptive stereotyping. *Journal of Experimental Social Psychology*, 40(5):619–632.

Gittelman, M. and Kogut, B. (2003). Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns. *Management Science*, 49(4):366–382.

Gneezy, U. and Potters, J. (1997). An experiment on risk taking and evaluation periods. *Quarterly Journal of Economics*, 112(2):631–645.

Gneezy, U., Saccardo, S., Serra-Garcia, M., and van Veldhuizen, R. (2016). Motivated self-deception, identity and unethical behavior. *Working Paper*.

Goldin, C. and Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, 90(4):715–741.

Gosling, S. D., Rentfrow, P. J., and Swann, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528.

Gough, H. G. (1979). A creative personality scale for the adjective check list. *Journal of Personality and Social Psychology*, 37(8):1398.

Greene, W. (2010). Testing hypotheses about interaction terms in nonlinear models. *Economics Letters*, 107(2):291–296.

Greenwald, A. G., Uhlmann, E. L., Poehlman, T. A., and Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97(1):17–41.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1):114–125.

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

Haisley, E., Mostafa, R., and Loewenstein, G. (2008). Subjective relative income and lottery ticket purchases. *Journal of Behavioral Decision Making*, 21(3):283–295.

Hall, B., Jaffe, A., and Trajtenberg, M. (2001). The NBER patent citation data file: Lessons, insights, and methodological tools. *Working Paper*.

Hamed, M. M. (2001). Analysis of pedestrians' behavior at pedestrian crossings. *Safety Science*, 38(1):63–82.

Harris, A. C. (1994). Ethnicity as a determinant of sex role identity: A replication study of item selection for the Bem Sex Role Inventory. *Sex Roles*, 31(3-4):241–273.

Harrison, G. W. and List, J. A. (2004). Field experiments. *Journal of Economic Literature*, 42(4):1009–1055.

Heckman, J. J. (1998). Detecting discrimination. *Journal of Economic Perspectives*, 12(2):101–116.

Helmreich, R. L. and Spence, J. T. (1978). The work and family orientation questionnaire: An objective instrument to assess components of achievement motivation and attitudes toward family and career. *JSAS Catalog of Selected Documents in Psychology*, 8:35.

Hennessey, B. A. and Amabile, T. M. (2010). Creativity. *Annual Review of Psychology*, 61(1):569–598.

Hewstone, M. (1990). The 'ultimate attribution error'? A review of the literature on intergroup causal attribution. *European Journal of Social Psychology*, 20(4):311–335.

Hewstone, M., Rubin, M., and Willis, H. (2002). Intergroup bias. *Annual Review of Psychology*, 53(1):575–604.

Himanen, V. and Kulmala, R. (1988). An application of logit models in analysing the behaviour of pedestrians and car drivers on pedestrian crossings. *Accident Analysis & Prevention*, 20(3):187–197.

Hoegl, M. and Gemuenden, H. G. (2001). Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence. *Organization Science*, 12(4):435–449.

Holländer, H. (1990). A social exchange approach to voluntary cooperation. *American Economic Review*, 80(5):1157–1167.

Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5):1644.

Holt, C. L. and Ellis, J. B. (1998). Assessing the current validity of the Bem Sex-Role Inventory. *Sex Roles*, 39(11):929–941.

Hossain, T. and List, J. A. (2012). The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science*, 58(12):2151–2167.

Jackson, T. T. and Gray, M. C. (1976). Field study of risk-taking behavior of automobile drivers. *Perceptual and Motor Skills*, 43(2):471–474.

Jayaraman, R., Ray, D., and de Véricourt, F. (2016). Anatomy of a contract change. *American Economic Review*, 106(2):316–358.

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

Kachelmaier, S. J., Reichert, B. E., and Williamson, M. G. (2008). Measuring and motivating quantity, creativity, or both. *Journal of Accounting Research*, 46(2):341–373.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291.

Kelly, S. and Dunbar, R. I. (2001). Who dares, wins. *Human Nature*, 12(2):89–105.

Kosfeld, M., Neckermann, S., and Yang, X. (2017). The effects of financial and recognition incentives across work contexts: The role of meaning. *Economic Inquiry*, 55(1):237–247.

Krawczyk, M. and Le Lec, F. (2010). 'Give me a chance!' An experiment in social decision under risk. *Experimental Economics*, 13(4):500–511.

Krawczyk, M. W., Trautmann, S. T., and van de Kuilen, G. (2017). Catastrophic risk: Social influences on insurance decisions. *Theory and Decision*, 82(3):309–326.

Krieger, N. (2014). Discrimination and health inequities. *International Journal of Health Services*, 44(4):643–710.

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

Kugler, T., Kausel, E. E., and Kocher, M. G. (2012). Are groups more rational than individuals? A review of interactive decision making in groups. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(4):471–482.

Lacetera, N. and Macis, M. (2010). Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme. *Journal of Economic Behavior & Organization*, 76(2):225–237.

Lahno, A. M. and Serra-Garcia, M. (2015). Peer effects in risk taking: Envy or conformity? *Journal of Risk and Uncertainty*, 50(1):73–95.

Lamiraud, K. and Vranceanu, R. (2017). Group gender composition and economic decision-making. *Journal of Economic Behavior & Organization*, forthcoming.

Lane, T. (2016). Discrimination in the laboratory: A meta-analysis of economics experiments. *European Economic Review*, 90:375–402.

Lang, K. and Manove, M. (2011). Education and labor market discrimination. *American Economic Review*, 101(4):1467–1496.

Laske, K. and Schroeder, M. (2016). Quantity, quality, and originality: The effects of incentives on creativity. *Working Paper*.

Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review*, 90(5):1346–1361.

Levitt, S. D. and Neckermann, S. (2014). What field experiments have and have not taught us about managing workers. *Oxford Review of Economic Policy*, 30(4):639–657.

Lima de Miranda, K., Detlefsen, L., and Schmidt, U. (2017). Can gender quotas prevent excessive risk taking? The effect of gender composition on group decisions under risk. *mimeo*.

Linde, J. and Sonnemans, J. (2012). Social comparison and risky choices. *Journal of Risk and Uncertainty*, 44(1):45–72.

López-Pérez, R. (2008). Aversion to norm-breaking: A model. *Games and Economic Behavior*, 64(1):237–267.

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45(6):633–644.

McCullers, J. C. (1978). Issues in learning and motivation. In Lepper, M. R. and Greene, D., editors, *The hidden costs of reward: New perspectives on the psychology of human motivation*, pages 5–18. Psychology Press, New York.

McGraw, K. O. (1978). The detrimental effects of reward on performance: A literature review and a prediction model. In Lepper, M. R. and Green, D., editors, *The hidden costs of reward: New perspectives on the psychology of human motivation*, pages 33–60. Psychology Press, New York.

Mujcic, R. and Frijters, P. (2013). Economic choices and status: Measuring preferences for income rank. *Oxford Economic Papers*, 65(1):47–73.

Müller, S. and Rau, H. A. (2016). The relation of risk attitudes and other-regarding preferences: A within-subjects analysis. *European Economic Review*, 85:1–7.

Muralidharan, K. and Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy*, 119(1):39–77.

Nelson, R. R. and Winter, S. G. (1982). *An evolutionary theory of economic change*. Harvard University Press, Cambridge.

Nuyts, E. and Vesentini, L. (2005). The relation between seat belt use of drivers and passengers. In de Waard, D., Brookhuis, K., van Egmond, R., and Boersema, T., editors, *Human Factors in Design, Safety, and Management*, pages 81–92. Shaker Publishing, Maastricht.

Ockenfels, A. and Werner, P. (2014). Beliefs and ingroup favoritism. *Journal of Economic Behavior & Organization*, 108:453–462.

Offerman, T., Sonnemans, J., and Schram, A. (1996). Value orientations, expectations and voluntary contributions in public goods. *The Economic Journal*, 106(437):817–845.

Patil, S. V., Vieider, F., and Tetlock, P. E. (2014). Process versus outcome accountability. *Oxford Handbook of Public Accountability*, pages 69–89.

Pawlowski, B., Atwal, R., and Dunbar, R. (2008). Sex differences in everyday risk-taking behavior in humans. *Evolutionary Psychology*, 6(1):29–42.

Pettigrew, T. F. (1979). The ultimate attribution error: Extending allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin*, 5(4):461–476.

Pink, D. (2009). Dan pink: The puzzle of motivation. `https://www.ted.com/talks/dan_pink_on_motivation`. Accessed: 2018-03-05.

Pink, D. H. (2011). *Drive: The surprising truth about what motivates us*. Riverhead Books, New York.

Pomeroy, S. (1996). *Oeconomicus: A Social and Historical Commentary, with a New English Translation*. Clarendon Press, Oxford.

Prentice, D. A. and Carranza, E. (2002). What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of Women Quarterly*, 26(4):269–281.

Rabin, M. (2000). Risk aversion and expected-utility theory: A calibration theorem. *Econometrica*, 68(5):1281–1292.

Ramm, J., Tjotta, S., and Torsvik, G. (2013). Incentives and creativity in groups. *Working Paper*.

Ratner, R. K. and Kahn, B. E. (2002). The impact of private versus public consumption on variety-seeking behavior. *Journal of Consumer Research*, 29(2):246–257.

Rege, M. and Telle, K. (2004). The impact of social approval and framing on cooperation in public good situations. *Journal of Public Economics*, 88(7):1625–1644.

Ronay, R. and Hippel, W. v. (2010). The presence of an attractive woman elevates testosterone and physical risk taking in young men. *Social Psychological and Personality Science*, 1(1):57–64.

Rosenkopf, L. and Nerkar, A. (2001). Beyond local search: Boundary-spanning, exploration and impact in the optical disc industry. *Strategic Management Journal*, 22:287–306.

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

Rudman, L. A. and Phelan, J. E. (2008). Backlash effects for disconfirming gender stereotypes in organizations. *Research in Organizational Behavior*, 28:61–79.

Saito, K. (2013). Social preferences under risk: Equality of opportunity versus equality of outcome. *American Economic Review*, 103(7):3084–3101.

Sandroni, A., Ludwig, S., and Kircher, P. (2013). On the difference between social and private goods. *The BE Journal of Theoretical Economics*, 13(1):151–177.

Schumpeter, J. (1934). *The Theory of Economic Development*. Harvard University Press, Cambridge.

Schwerter, F. (2013). Social reference points and risk taking. *Working Paper*.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In Sauerman, H., editor, *Beiträge zur Experimentellen Wirtschaftsforschung*, pages 136–168. Mohr, Tübingen.

Shapiro, T., Meschede, T., and Osoro, S. (2013). The roots of the widening racial wealth gap: Explaining the black-white economic divide. *Research and Policy Brief*.

Shearer, B. (2004). Piece rates, fixed wages and incentives: Evidence from a field experiment. *Review of Economic Studies*, 71(2):513–534.

Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research*, 16(2):158–174.

Simonson, I. and Nye, P. (1992). The effect of accountability on susceptibility to decision errors. *Organizational Behavior and Human Decision Processes*, 51(3):416–446.

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

Takahashi, H., Shen, J., and Ogawa, K. (2016). An experimental examination of compensation schemes and level of effort in differentiated tasks. *Journal of Behavioral and Experimental Economics*, 61:12–19.

Trautmann, S. T. (2009). A tractable model of process fairness under risk. *Journal of Economic Psychology*, 30(5):803–813.

Trautmann, S. T. and Vieider, F. M. (2012). Social influences on risk attitudes: Applications in economics. In Roeser, S., Hillerbrand, R., Sandin, P., and Petersen, M., editors, *Handbook of Risk Theory. Epistemology, Decision Theory, Ethics, and Social Implications of Risk*, pages 575–600. Springer, Amsterdam.

Trautmann, S. T. and Wakker, P. P. (2010). Process fairness and dynamic consistency. *Economics Letters*, 109(3):187–189.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.

Tversky, A. and Simonson, I. (1993). Context-dependent preferences. *Management Science*, 39(10):1179–1189.

Tymula, A. and Whitehair, J. (2018). Young adults gamble less when observed by peers. *Journal of Economic Psychology*, 68:1–15.

Van Dijk, F., Sonnemans, J., and van Winden, F. (2002). Social ties in a public good experiment. *Journal of Public Economics*, 85(2):275–299.

Van Lange, P. A., De Bruin, E., Otten, W., and Joireman, J. A. (1997). Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. *Journal of Personality and Social Psychology*, 73(4):733.

Vieider, F. M. (2009). The effect of accountability on loss aversion. *Acta Psychologica*, 132(1):96–101.

Vieider, F. M., Villegas-Palacio, C., Martinsson, P., and Mejía, M. (2016). Risk taking for oneself and others: A structural model approach. *Economic Inquiry*, 54(2):879–894.

Wakker, P. P. (2010). *Prospect theory: For risk and ambiguity*. Cambridge University Press, Cambridge.

Weigold, M. F. and Schlenker, B. R. (1991). Accountability and risk taking. *Personality and Social Psychology Bulletin*, 17(1):25–29.

Weitzman, M. (1998). Recombinant growth. *Quarterly Journal of Economics*, 113(2):331–360.

Wilke, A., Hutchinson, J. M., Todd, P. M., and Kruger, D. J. (2006). Is risk taking used as a cue in mate choice? *Evolutionary Psychology*, 4(1):367–393.

Yechiam, E., Druyan, M., and Ert, E. (2008). Observing others' behavior and risk taking in decisions from experience. *Judgment and Decision Making*, 3(7):493.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, 13(1):75–98.