Inequality, Immoral Behavior, and Responsibility: Essays in Behavioral and Experimental Economics

Felix Klimm

Dissertation

Felix Klimm

Inequality, Immoral Behavior, and Responsibility: Essays in Behavioral and Experimental Economics

March, 2018

Inequality, Immoral Behavior, and Responsibility: Essays in Behavioral and Experimental Economics

Inaugural-Dissertation

zur Erlangung des Grades

Doctor oeconomiae publicae (Dr. oec. publ.)

an der Volkswirtschaftlichen Fakultät

der Ludwig-Maximilians-Universität München

2018

vorgelegt von

Felix Klimm

Referent:	Prof. Dr. Martin Kocher
Koreferent:	Prof. Dr. Florian Englmaier
Promotionsabschlussberatung:	11. Juli 2018
Berichterstatter:	Prof. Dr. Martin Kocher
	Prof. Dr. Florian Englmaier
	Prof. Dr. Lars Hornuf
Datum der mündlichen Prüfung:	2. Juli 2018

Danksagung

Als erstes möchte ich mich bei den anderen Stipendiaten und Bürokollegen bedanken, die mir vor etwas über drei Jahren die Anfangszeit in München und der Promotion überaus angenehm gemacht haben. Insbesondere Flo Loipersberger, Mira Breckner, Marie Lechler, Stefan Grimm und Basti Horn wurden gute Freunde, zum Teil Koautoren und sogar Mitbewohner. Wir konnten uns immer über fachliche Themen austauschen, gemeinsame Interessen teilen und während der gesamten Promotion gegenseitig motivieren.

Flo stand mir immer mit Rat und Tat zur Seite. Wir verbrachten zahlreiche Stunden mit Stata, Latex und zTree, und er war permanent per Telefon für jegliches ökonometrische Problem zu erreichen (u.a. für unser gemeinsames Projekt, Kapitel 3 dieser Dissertation). Als Koautor durfte ich Flo die Vorzüge von ökonomischen Experimenten näherbringen. Im Gegenzug habe ich viel in Bezug auf Programmieren von ihm gelernt.

Außerdem möchte ich mich bei Stefan bedanken, der anderen unverzichtbaren Hälfte des "Grimm-Klimm-Büros". Gemeinsam bekleideten wir auch die Stellen der Vertreter der wissenschaftlichen Mitarbeiter an unserer Fakultät oder standen vor Kreisverwaltungsreferaten in München, um Geflüchtete für unser gemeinsames Projekt zu rekrutieren (Kapitel 4 der vorliegenden Arbeit). Jegliche Koordination — für unser Forschungsprojekt, Veranstaltungen der Fakultät oder das gemeinsame Halten von

Lehre im Rahmen von “Lehre@LMU” — war immer unkompliziert, was mein tägliches Arbeiten grundlegend erleichterte.

Erwähnen möchte ich an dieser Stelle auch Marco Schwarz, Konstantin Lucks und Daniel Gietl, mit denen ich im Rahmen des Promotionsstudiums aber auch außerhalb der Universität interessante Gespräche führen durfte.

Besonderer Dank geht an unser Graduiertenkolleg (GRK), insbesondere an Julia Zimmermann für die Koordination und Hilfe bei vielen administrativen Aufgaben. Hier sind auch weitere Fakultätsmitglieder zu nennen, die sich für unser Stipendienprogramm eingesetzt haben, allen voran der GRK-Sprecher Carsten Eckel. Von Carsten habe ich auch gelernt, dass man bei einer E-Mail, die einen dazu auffordert, per E-Mail seine Schuhgröße mitzuteilen, damit man vom bayrischen Staat Gummistiefel wegen eines angeblichen Wasserschadens gestellt bekommt, immer darauf achten sollte, ob der Absender diese E-Mail am 1. April verfasst hat. Des Weiteren bin ich dankbar für die Unterstützung der Deutschen Forschungsgemeinschaft durch das GRK, welches in den ersten drei Jahren meiner Promotion mein Stipendium sowie meine Experimente finanzierte. Das Stipendium ermöglichte mir darüber hinaus, an zahlreichen Konferenzen und Workshops teilzunehmen, u.a. in Bergen (wodurch teilweise die Ideen zu Kapitel 2 und 3 entstanden), Iseo und San Diego.

Besonders dankbar bin ich für meinen Forschungsaufenthalt an der University of Oxford, und in diesem Zusammenhang vor allem Johannes Abeler, der mich während meiner Zeit in Oxford betreute. Für Johannes durfte ich auch an einer Metastudie zu ökonomischem Betrugsverhalten mitwirken, die bald in *Econometrica* veröffentlicht wird. Vom Arbeiten an dieser Studie profitierte ich seither durch einen guten Überblick der Literatur in diesem Bereich.

Ferner möchte ich mich bei meinem Doktorvater Martin Kocher bedanken. Neben seiner Tätigkeit als Professor an der LMU München und der Leitung des

Instituts für Höhere Studien in Wien nahm er sich immer Zeit, mit mir meine Ideen zu diskutieren und gab mir hilfreiches Feedback zu jedem der Projekte dieser Dissertation — sei es im persönlichem Gespräch in München oder per Skype Gespräch zwischen Wien und Oxford. Dadurch hat die vorliegende Doktorarbeit substantiell profitiert.

Außerdem möchte ich meinen Eltern für ihre Unterstützung während meiner Promotion danken. Auch wenn ich sie in den letzten Jahren viel zu selten in meiner Heimatstadt Konstanz gesehen habe, standen sie mir immer zur Seite.

München im März 2018

Contents

List of Figures

List of Tables

Introduction and Summary 1

Traditional models of economic decision making assume that individuals act rationally, purely self-interested, and maximize their material payoffs. For the most part of the last century, economists have predicted economic behavior based on these assumptions. Meanwhile, behavioral economics has provided ample evidence about non-rational and social behavior by incorporating insights from psychology into economic research (e.g., Kahneman and Tversky, 1979; Fehr and Gächter, 2000). Laboratory experiments have played an important role in accumulating this rigorous empirical evidence and are nowadays accepted as a major source of knowledge in economics (Falk and Heckman, 2009). Similar to medical studies that employ placebos, economic experiments implement exogenous treatment variations, which offers tight control over potentially confounding factors of influence and thus allows drawing causal inferences.

This dissertation reports results from three laboratory experiments that involve "behavioral regularities", i.e., systematic deviations from behavior implied by traditional economic theory. One economically relevant domain in which these behavioral regularities matter is immoral behavior. Over decades, economists have assumed that individuals act perfectly immoral if this serves their material self-interest. However, the recent literature in economics and psychology has provided clean evidence that people often refrain from immoral behavior (e.g., Fischbacher and Föllmi-Heusi, 2013). Contributing to this literature, Chapter 2 examines inequalities that might arise due to cheating. In particular, I focus on how these inequalities affect people's preferences for redistribution. How do people's views on redistributive policies change when they suspect that the "rich" acquired their wealth by means of cheating? Chapter 3 also concerns the domain of immoral decision making. An open question in the literature is whether people justify their dishonesty by shifting responsibility to another person's choice. Are individuals more likely to lie at the expense of another person if this other person self-selected into a situation where being lied to is possible? The notion underlying this justification is that people are responsible for outcomes which result from their own choices if a different choice would have yielded a different outcome (Dworkin, 1981a,b).

In addition to immoral behavior, responsibility is also subject to behavioral regularities. Traditional economic models assume that individuals are not intrinsically concerned about who is responsible for an economic outcome. However, typically, it is considered to be fair to hold people responsible for the outcomes that result from their actions (Cappelen et al., 2016). This may be difficult as many situations entail uncertainty about people's actions, which leaves room for holding someone responsible for a certain outcome. For instance, responsibility for economic success or failure can either be attributed to an individual's action, e.g. high effort, or to external factors, e.g. demand or supply shocks. Related to this, Chapter 4 investigates non-rational responsibility attribution to refugees — a group which has become increasingly important for many developed economies — by asking: Do natives blame refugees for negative economic events?

As argued above, this thesis documents effects and determinants of inequality, immoral behavior, and responsibility by providing experimental evidence on behavioral regularities. In the remainder of Chapter 1, I summarize the following three chapters of this dissertation. Each of these chapters is self-contained and, thus, can be read independently. Each chapter's appendix follows after the chapter's main text, while the references are presented at the end of this dissertation.

Chapter 2. Different views on the necessity of redistributive policies rest upon the sources of inequality. Supporters of left-wing parties typically argue that unequal outcomes considerably emerge due to circumstances beyond individual control and thus place more emphasis on redistribution than right-wing voters (e.g., Alesina and Angeletos, 2005). I investigate whether this difference in tolerating inequality is amplified by suspicious success — achievements that may arise from cheating.

Prominent examples of fraudulent behavior revealed by the "Panama Papers" and the "Paradise Papers" have shown that a substantial fraction of global financial wealth is generated by dishonest means. According to recent estimations, tax evasion of this kind results in annually forgone tax revenues of $190 billion (Zucman, 2014). This might leave people suspicious about the wealth of the very successful. Another prominent example of cheating that leads to suspicious success is doping in professional sports. People may be skeptical about athletes' performances at the Tour de France or the Olympic Games because their achievements seem just to good to be true.

I investigate the question of how suspicious success affects redistributive preferences using a laboratory experiment. For this purpose, I exogenously vary cheating opportunities for stakeholders who work on a real effort task and earn money according to their self-reported performances. An impartial spectator may redistribute the earnings between the stakeholders. In the control condition, stakeholders are perfectly monitored and thus cannot cheat. In contrast, stakeholders are able to overstate their performances in the treatment condition. Importantly, dishonest stakeholders cannot be identified. Hence, spectators might speculate about suspicious success when observing large income differences in the treatment condition, but they do not know whether this suspicion is justified. This is why, in the presence of potential cheating, some spectators might eliminate inequalities, while others refrain from doing so.

I find that the opportunity to cheat leads to different views on whether to accept inequality. Left-wing spectators substantially reduce inequality when cheating is possible, while the treatment has no significant effect on choices of right-wing spectators. Furthermore, left-wing spectators' decisions are affected by cheating opportunities only in situations with high pre-redistribution inequality, i.e., cases when one might become suspicious about the success of a high performer. This provides evidence for the mechanism of the treatment effect. Left-wing spectators seem to redistribute more in the treatment condition because they suspect the "rich" to be cheating.

My setup enables me to distinguish between three different explanations for the polarization of redistributive preferences: (i) differences in beliefs about cheating, (ii) differences in whether spectators find cheating acceptable (i.e., norms), and (iii) mere differences in the preference for redistribution when the source of income inequality is unclear (cheating versus honest performance). Since neither beliefs nor norms about cheating are significantly different across the two political camps, my findings seem to be driven by a difference in preferences. These results suggest that redistributive preferences will diverge even more once public awareness increases that inequality may be to a certain extent created by cheating.

Chapter 3. Recent research on dishonesty suggests that people want to keep a positive image of themselves when engaging in lying behavior (e.g., Mazar et al., 2008; Abeler et al., 2016). Therefore, they must come up with excuses for dishonesty. Together with Florian Loipersberger, I set up a laboratory experiment to study whether the presence of a choice is used as such an excuse.

Choices become increasingly prevalent in most developed economies through the extension of market mechanisms to various aspects of life (Cappelen et al., 2016). For instance, nowadays, people ought to choose between different investments for their retirement savings, while a couple of years ago, many governments took full responsibility for their citizen's retirement benefits. Despite the many advantages of free choices, they might increase the chances of being lied to. Consider, for example, a bank employee who offers several financial products to a customer. The bank employee does not recommend the best fitting option but the one that leaves her with the highest commission in order to maximize her income. To convince herself of her action being morally acceptable, she brings to her mind that the customer was free to choose a different bank at any point in time.

We address this issue by conducting a laboratory experiment where a potential liar can lie at the cost of another participant (the "other participant"). The other participant faces two options: interacting with the potential liar or receiving an alternative payment. In our control condition, the other participant is randomly assigned to one of these two options. In contrast, he chooses between these alternatives in our treatment condition.

We find that the introduction of a choice leads to a positive but insignificant increase in the probability of behaving dishonestly. Following the large literature on gender differences in dishonesty (e.g., Dreber and Johannesson, 2008; Grosch and Rau, 2017), we investigate whether this results holds for both genders separately. Self-selection of the other participant has no significant effect on lying decisions of females. For men, however, we find a significant treatment effect of about 56% increased dishonesty. Thus, our results suggest that some males excuse their dishonest behavior by shifting responsibility for the outcome to the choice of the other participant.

Chapter 4. This chapter is joint work with Stefan Grimm. We investigate whether people blame refugees for negative events. The large inflow of refugees to Europe in the last couple of years has revived the heated political debate about whether and how to integrate refugees. The content of this debate is highly relevant in economic terms, as, for instance, the future of labor markets in many Western societies depends on the integration of refugees. While a large part of this discussion focuses on whether refugees can be held responsible for negative events such as rising crime and unemployment rates, surprisingly little is known about how natives attribute responsibility towards refugees.

We propose a novel experimental paradigm to measure discrimination in responsibility attribution towards Arabic refugees. In our experiment, German participants are either paired with another German or a refugee. These German participants experience a positive or negative income shock, which is with equal probability caused by a random draw or another 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. Moreover, to investigate whether our results are driven by statistical discrimination, we elicit beliefs about the partner's performance.

We find evidence for reverse discrimination. Germans attribute responsibility more favorably to refugees than to other Germans. In particular, refugees are less often held responsible for negative income shocks. Since neither actual performance differences nor beliefs about Germans' and refugees' performances can explain our finding of reverse discrimination, we rule out statistical discrimination as the driving force. Moreover, we find that Germans with negative implicit associations towards Arabic names attribute responsibility less favorably to refugees than Germans with positive associations. This indicates that implicit associations, which have predictive power for relevant field behavior such as hiring decisions (Greenwald et al., 2009), are positively related to explicit attribution behavior towards refugees.

Our findings cannot be explained by standard economic theory since German participants are willing to forgo parts of their earnings in order to attribute responsibility favorably to refugees. Instead, we suggest to interpret our findings with explanations based on theories of self-image and identity concerns. These theories assume that people want to view themselves as behaving in line with a positive self-image, which can result in self-serving beliefs about other people (Di Tella et al., 2015). Applied to our setting, assuming that our participants care about not being someone who discriminates refugees, identity concerns are likely to explain our result of reverse discrimination.

2 Suspicious Success — Cheating, Inequality Acceptance, and Political Preferences

2.1 Introduction

Whether unequal outcomes are considered to be fair primarily depends on the sources of inequality. People prefer to eliminate income disparities that have resulted from factors beyond individual control such as pure luck, physical handicap, gender, or family background, yet they tend to accept inequalities based on differences in effort, initiative, or the willingness to take risks (Konow, 2000; Fong, 2001; Cappelen et al., 2013a; Möllerström et al., 2015; Almås et al., 2016). The sources of inequality also lie at the core of the political debate about redistribution. Whereas right-wingers believe that one's fortunes are mainly the consequences of effort and choices, left-wingers place more emphasis on the notion that uncontrollable luck determines income (Piketty, 1995; Alesina and Angeletos, 2005; Cappelen et al., 2010; Cappelen et al., 2016).

In this chapter, I investigate whether the difference in redistributive preferences between the two political camps persists with regard to another source of inequality — cheating. Because everyday life is permeated with cheating opportunities, ranging from an employee tempted to overstate hours worked to potential submission of false claims by a physician, people might be suspicious of the wealth of the successful. For example, the recent leaks of the "Panama Papers" as well as the "Paradise Papers" have revealed that a large fraction of global financial wealth is held in tax havens.[1] Zucman (2014) estimates that annually foregone tax revenues due to offshore tax evasion amount to $190 billion, which suggests that a significant share of wealth is created by illegal financial activities.[2]

A particular feature of inequalities which arise from fraudulent behavior is that although cheating is within individual control, it is, as opposed to effort, unlikely to be regarded as fair (e.g., Kirchler et al., 2003). Therefore, it remains an open question whether left-wingers also demand more redistribution than right-wingers in the presence of cheating opportunities. However, with regard to the prevalence of suspicious success, this question needs to be answered in order to understand the origins of different views on the necessity of redistributive policies.

I address this question by conducting a between-subjects experiment, where some participants work on a real effort task (henceforth called stakeholders). Two stakeholders are matched with another and split a fixed amount of money according to their performances. Stakeholders can overstate their performance in the *Cheat* treatment, which does not affect the total income of the two stakeholders but shifts the distribution of income in favor of the misreporting stakeholder. This captures the impact of cheating behavior in many situations of economic relevance. For instance, tax evasion does not alter the amount of money necessary to provide public goods, but at the same time honest tax payers bear the cost of cheating in the long

[1] See, e.g., http://www.bbc.com/news/world-41880153, last accessed on March 5, 2018.

[2] Inequalities based on cheating are not limited to tax evasion, but there are various other forms of performance cheating that cause someone to be more successful than others. For instance, businessmen fabricate their curriculum vitae to get better paid jobs, athletes take performance enhancing doping substances to win prestigious competitions, and firms manipulate software to maximize profits. (see, e.g., http://www.dailymail.co.uk/news/article-2669969/CV-fake-hired-Myer-considered-companies.html, http://www.telegraph.co.uk/sport/othersports/cycling/lancearmstrong/9810199/Lance-Armstrong-tells-Oprah-Winfrey-he-doped-during-all-seven-Tour-de-France-victories.html, https://www.nytimes.com/interactive/2015/business/international/vw-diesel-emissions-scandal-explained.html, last accessed on March 5, 2018).

run. Contrary to the *Cheat* treatment, stakeholders' performances are audited in the *Monitor* treatment, which renders misreporting impossible.

Third-party participants (henceforth called spectators) are able to redistribute the earnings of the two stakeholders (following Cappelen et al., 2013a). In both treatments, they are fully aware of the rules for working on the real effort task and the (lacking) possibility to misreport own performance. Importantly, in the *Cheat* treatment, dishonest stakeholders cannot be identified, and spectators thus never know whether a stakeholder reported untruthfully. Therefore, if one of the two stakeholders earns considerably more than the other one, spectators might believe that high income results from cheating but do not know whether their suspicion is accurate. This uncertainty leaves room to justify both eliminating inequalities as well as refraining from doing so.

Using a laboratory experiment allows to provide clean evidence on the effect of cheating opportunities on inequality acceptance for two main reasons. First, it allows for exogenous manipulation of the availability to cheat, which is difficult to achieve in field settings given the nature of naturally occurring cheating opportunities. Second, eliciting redistributive preferences from impartial spectators makes it possible to exclude confounding factors such as selfishness, self-centered inequality aversion (e.g., Fehr and Schmidt, 1999), or reciprocity (e.g., Rabin, 1993).

My results show that the treatment effect depends on political preferences. Right-wing spectators hesitate to redistribute on the basis of potential cheating as they implement the same levels of inequality in *Monitor* and *Cheat*. In contrast, distributive choices of left-wing spectators reveal an increase of 74% in inequality reduction due to cheating opportunities. The analysis of the treatment effect for different levels of pre-redistribution inequality shows that left-wing spectators react to potential cheating only for high levels of inequality. This provides evidence that they believe that the "rich" stakeholder is cheating in these situations.

There are essentially three different explanations for the polarization in redistributive preferences between the two political camps. (i) Right-wing spectators might believe to a lesser extent that stakeholders are cheating than left-wingers. (ii) Right-wing spectators' norms about cheating differ from those of left-wing supporters: They find it more acceptable to cheat when possible. (iii) Right-wingers prefer not to redistribute due to potential cheating if they do not know whether a stakeholder indeed cheated, although they know that misreporting is prevalent. In order to distinguish between these three explanations, I examine beliefs and norms about cheating and find no differences between left-wingers and right-wingers. Therefore, the political divide in how to deal with unequal outcomes that might arise from dishonest behavior seems to reflect different preferences. This suggests that different views on the importance of redistributive policies diverge even more in the light of scandals about cheating by the "rich and successful" as we know that redistributive preferences are highly elastic to information (Kuziemko et al., 2015).

This chapter contributes to several strands of the literature. First, it relates to studies on the determinants of redistributive preferences. Papers that use survey data find that personal characteristics such as gender, race, and education as well as cultural background and past experience of personal traumas (e.g., divorce, hospitalization, or death of a relative) predict redistributive preferences (Alesina and Ferrara, 2005; Alesina and Giuliano, 2011). Moreover, using data from the General Social Survey, Fong (2001) shows that people who believe that luck causes poverty and wealth support redistribution to a much larger extent than people who believe that effort causes poverty and wealth. In addition, experimental studies indicate that people also care about whether someone can be held responsible for one's own luck by choosing a risky or a safe option (Cappelen et al., 2013a; Möllerström et al., 2015). Closely related to this chapter, Bortolotti et al. (2017) experimentally investigate redistributive preferences when cheating with regard to a risky outcome, a coin flip, is possible. The authors document a shift in fairness views due to potential cheating in favor of strict egalitarianism, i.e., implementing an equal distribution of income independent of subjects' choices that affected earnings in the first place. Importantly, while Bortolotti et al. (2017) study redistributive preferences in the light of cheating in the luck domain, I focus on situations where people can cheat regarding their performance.

Second, my results show that political preferences matter for accepting inequalities that might arise from cheating behavior. Interestingly, the evidence on whether political preferences affect choices in allocation decisions is mixed. While a number of studies report significant differences across political preferences (Van Lange et al., 2012; Cappelen et al., 2013a; Cappelen et al., 2016; Bortolotti et al., 2017), there are two studies that find only weakly significant or insignificant effects of political preferences (Frohlich et al., 1984; Fehr et al., 2006). Therefore, the impact of political preferences on distributional choices seems to depend on the specific context.

Third, giving participants the opportunity to cheat relates this chapter to a growing experimental literature on dishonesty (e.g., Gneezy, 2005; Mazar et al., 2008; Fischbacher and Föllmi-Heusi, 2013; Shalvi and De Dreu, 2014; Conrads and Lotz, 2015; Houser et al., 2016).[3] While this literature is primarily concerned with the extent and the causes of cheating, my study is one of the few that deal with the consequences of cheating by showing that dishonesty affects the behavior of third parties (Pigors and Rockenbach, 2016; Bortolotti et al., 2017; Cappelen et al., 2017).

[3] See Abeler et al. (2016) for a meta-study on data from 72 experimental studies on dishonesty.

The remainder of this chapter is structured as follows. Section 2.2 describes the experimental design in detail. Section 2.3 presents the results and Section 2.4 concludes.

2.2 Experimental Design

The experiment consists of two main parts. Part 1 concerns the real effort provision and potential cheating in the matrix task, which was introduced in the literature by Mazar et al. (2008). Part 2 uses decisions from impartial spectators in order to measure redistributive preferences, similar to Cappelen et al. (2013a). Thereafter, beliefs and political preferences are elicited.

2.2.1 Part 1: The Matrix Task

All subjects receive an exercise sheet with 20 matrices, each containing a set of twelve numbers with two decimal places (see Appendix 2.5.4.1 for an example). Only two of these twelve numbers add up to exactly 10. The task is to find these two numbers and solve as many matrices as possible within 6 minutes.

Two players A (the stakeholders) are randomly matched in order to determine their preliminary income (i.e., income before redistribution). Proportionally to their performance in the matrix task, €10 are split up among these two participants. This distribution of income is rounded to 50 cents.

Treatment variation. After working on the task, the stakeholders are provided with the correct solutions on their screens and are asked to compare them with their own solutions (see Figure 2.7 in the Appendix for an example of the stakeholders' decision screen). They then report for each matrix whether they solved it correctly. Subjects in *Monitor* are informed that all exercise sheets are collected to verify their reported performance and that, if necessary, their reports will be changed to their actual performance.[4] In contrast, subjects in *Cheat* are informed that they will shred their exercise sheet at the end of the experiment, and it is thus impossible to monitor their solutions.

The design of Part 1 has at least three desirable features for the purpose of this chapter. First, the real effort task mimics a wide range of field settings where people engage in performance cheating in order to serve their self-interests. Consider, for instance, an employee who misreports his number of hours worked to receive either a higher wage or more days off. Second, exaggerating own performance implies a negative externality for the other stakeholder, which reflects the adverse consequences of cheating in many "real-world" situations. Coming back to the example of overstating hours worked, honest colleagues might be affected through a lower likelihood of being promoted due to inferior relative performance. Third, having a fixed sum of payments for the two stakeholders excludes efficiency concerns as a motivation for cheating. This is important for redistribution decisions because if cheating was efficiency enhancing, this might confound the moral assessment of such behavior and it would be difficult to account for this motive in the experiment.

2.2.2 Part 2: Redistribution Decisions

Each player B (the spectator) is matched with a pair of stakeholders. The strategy method is used for the spectators' decisions. Hence, for each of the eleven possible distributions of preliminary income (going in steps of 50 cents from maximum inequality in the case of (10,0) to full equality in the case of (5,5)), they can transfer money within a pair of stakeholders and consequently determine the two stakeholders'

[4] The fraction of stakeholders for whom performances had to be corrected downwards is 10% (upwards 5%). In 57% of these cases, the difference between reported and actual performance was one task.

final income (i.e., income after redistribution).[5] Redistribution of the preliminary income is possible in steps of 10 cents. In order to be able to unambiguously refer to the stakeholders' final income, the stakeholder with the higher or equal preliminary income is called player A1 and the other stakeholder player A2 (see Figure 2.8 in the Appendix for the spectators' decision screen).

Spectators receive a fixed income of €10 for their redistribution decisions. Giving spectators at least the sum of earnings of a pair of stakeholders assures that self-centered inequity aversion based on Fehr and Schmidt (1999) does not affect spectators' behavior.[6] Moreover, spectators are informed that either their decision or the decision of another spectator will be randomly implemented, which is designed to increase the number of decisions taken by spectators.[7]

2.2.3 Belief Elicitation

After stakeholders report their performance and spectators make their redistribution decisions, two beliefs are elicited. First, stakeholders and spectators are asked to guess the average reported performance of the stakeholders in their own session (*belief-own-treat*). Consequently, subjects in *Monitor* report their belief about how many tasks were actually solved, while subjects in *Cheat* guess how many tasks the stakeholders reported to have solved. Second, I elicit beliefs about how many correctly solved tasks the stakeholders reported in the respective other treatment (*belief-other-treat*). Therefore, subjects are informed that stakeholders worked on exactly the same task in a previously run experiment but that reported performance was monitored differently

[5] Brandts and Charness (2011) provide an analysis of 29 studies in order to compare the strategy method with the direct respond method. Since they do not find a single case in which there is a treatment effect using the strategy method that vanishes with the direct respond method, the strategy method is likely to yield a lower bound for this experiment's treatment effect.

[6] For instance, the Fehr-Schmidt model predicts the spectator to choose full equality when receiving a fixed income of €5, independent of the distribution of preliminary income.

[7] As a consequence, the number of spectators ($n = 182$) equals the number of stakeholders ($n = 182$).

(instructions can be found in Appendix 2.5.4.4). Subjects learn about the respective other treatment not before beliefs are elicited.

Beliefs are incentivized with €2 for deviations up to one task and €1 for deviations up to two tasks, whereas larger deviations are not paid. Only one of the two beliefs is randomly chosen for payment in order to prevent hedging. Since *belief-other-treat* is based on the first session of the respective other treatment, it is not elicited in the first session of each treatment.

Moreover, I elicit a third belief at the end of each session of the *Cheat* treatment. Subjects guess the fraction of stakeholders who did not report truthfully (*belief-frac-cheat*). It is impossible to incentivize these beliefs because I refrain from individual cheating detection.

2.2.4 Political Preferences

At the end of the experiment, I ask participants about their political preferences (i.e., which party they would vote for if there were federal elections next Sunday). In my analysis, being left-wing is defined as indicating to vote for the Social Democrats (SPD), the Green Party (Die Grünen), the socialist party (Die Linke), or the Pirate Party Germany (Die Piraten).[8] Subjects belonging to the remaining categories are treated as being right-wing. Following this definition, 37.36% of the spectators are classified as left-wing and 62.64% as right-wing. The distribution of spectators' votes closely resembles the results of the 2017 German national election. Thus, in terms of political preferences, the sample of the experiment is similar to the German population. In order to validate my classification of political parties, subjects are asked to indicate where they rate their general political attitudes on a scale from 1 to 10 with 1 being

[8] A coalition of the three established parties SPD, Die Grünen, and Die Linke is also called "left-wing coalition", see, e.g., `http://www.telegraph.co.uk/news/worldnews/europe/germany/10318949/Germanys-coalitions-What-happens-next.html` (last accessed on March 5, 2018). The Pirate Party Germany, which was found in 2006, is typically classified as being left-wing.

left and 10 being right. Although this question may be susceptible to the central tendency bias (64% of subjects indicate a score of 4, 5, or 6), left-wing spectators rate themselves lower on this scale than right-wing spectators (mean left-wing $= 3.94$, mean right-wing $= 5.37$). The difference in ratings across political preferences is significant ($p < 0.0001$, Mann-Whitney *U*-test, two-sided). Table 2.2 in the Appendix shows which parties spectators would vote for, the 2017 German federal election results as well as the average scores of spectators' general political attitudes.

2.2.5 Procedural Details

The experiment was conducted with 364 participants at the Munich Experimental Laboratory for Economic and Social Sciences (MELESSA) at the University of Munich in May 2016 and May 2017. In total, 184 subjects were assigned to eight sessions of the *Monitor* treatment and 180 subjects to eight sessions of the *Cheat* treatment.[9] Subjects were students from various fields of study and recruited using the online system “ORSEE" (Greiner, 2015). Each subject was randomly assigned to one of the two treatments and participated in one session only. The experiment was programmed and conducted with the software “z-Tree" (Fischbacher, 2007).

Upon arrival at the laboratory, subjects found a printed version of the instructions of Part 1 at their seats, which was read aloud by the experimenter (myself) to ensure common knowledge about the rules of the real effort task (see Appendix 2.5.4.1). In addition, these instructions informed participants that only later on they will be assigned to one of the two roles. Hence, stakeholders as well as spectators worked on the matrix task. In this way, spectators were familiar with the difficulty of the matrix task, which was important for eliciting their beliefs about the stakeholders' performances. Furthermore, subjects were told that the spectator can distribute earnings of

[9] For each of the two treatments, half of the sessions were conducted in 2016 and the other half in 2017. The time of conducting the experiment does not affect the results (see Section 2.3.2).

Part 1 and that the spectator will receive more detailed information on this at a later point in time.[10] Thereafter, subjects received the exercise sheet and were provided with pens for marking their solutions. After timeout, a second set of instructions about how to compare own with correct solutions appeared on the subjects' screens and were read aloud (see Appendix 2.5.4.2). Subsequently, subjects were displayed their role, and stakeholders self-reported their performance while spectators received detailed instructions about Part 2 (see Appendix 2.5.4.3) and made their redistribution decisions. Thus, *only* stakeholders compared their own solutions with the correct ones. After that, the exercise sheets were collected and verified in *Monitor*, and subjects in both treatments indicated on a 4 point Likert scale whether they considered it to be fair that preliminary income was proportional to self-reported performance. Next, *belief-own-treat* and *belief-other-treat* were elicited.

Finally, the participants answered a questionnaire about their political preferences, opinion on income inequality in Germany on a scale from 1 (inequality should be reduced) to 10 (inequality should be enlarged in order to provide incentives), socio-demographic characteristics, and their belief about the fraction of stakeholders who did not report truthfully (*belief-frac-cheat*, only in the *Cheat* treatment). Subjects received their payments privately after the experiment and earned €12.22 on average, including an average show-up fee of €4.5.[11] Sessions lasted on average 50 minutes.

[10] Little information about Part 2 cannot exclude the possibility of strategic effort provision. However, this would not affect spectators redistributive choices because they were elicited with the strategy method and are thus independent of actual performances.

[11] The show-up fee was €4 in 2016 and €5 in 2017 due to changes of the rules of MELESSA.

2.3 Results

2.3.1 Matrix Task Performance

Figure 2.1 shows that reported performances are significantly higher in *Cheat* than in *Monitor* and indicates that manipulation by giving stakeholders the opportunity to cheat was successful ($p < 0.001$, Mann-Whitney *U*-test, two-sided). Average reported performance is higher in *Cheat* (13.8) than in *Monitor* (11.5). Moreover, while only 5.4% of the stakeholders in *Monitor* indicate to have solved all matrices, this is the case for 16.7% of the stakeholders in *Cheat*. This is line with the finding of previous studies on dishonesty that although some people are cheating, the assumption of people always submitting payoff-maximizing reports is empirically not valid (e.g., Mazar et al., 2008; Fischbacher and Föllmi-Heusi, 2013; Jiang, 2013; Cohn et al., 2014; Abeler et al., 2016).

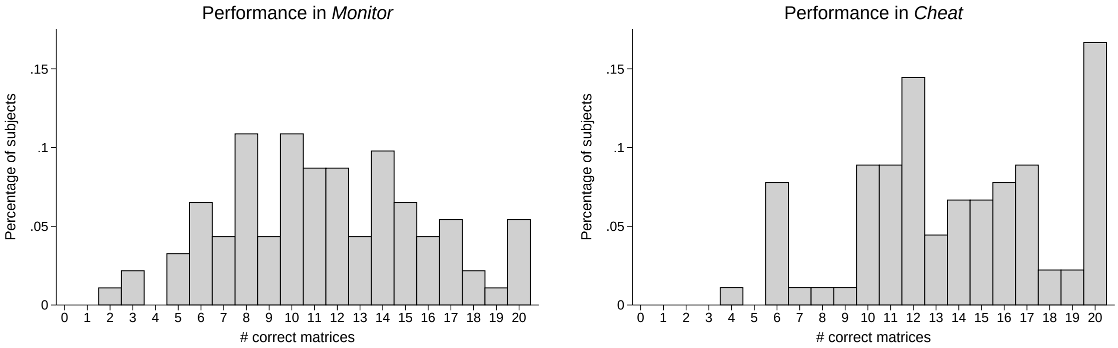


Figure 2.1: Distribution of reported performances

In addition, subjects' beliefs about average reported performance in their own session is significantly different across treatments ($p < 0.0001$, Mann-Whitney *U*-test, two-sided). On average, *belief-own-treat* is 1.9 tasks higher in *Cheat* than in *Monitor*, which is close to the actual difference of 2.3 tasks. Furthermore, subjects in *Monitor* find the income generating process of preliminary income significantly more fair than subjects in *Cheat* ($p = 0.037$, Mann-Whitney *U*-test, two-sided). Both of these results are further indications of successful treatment manipulation.

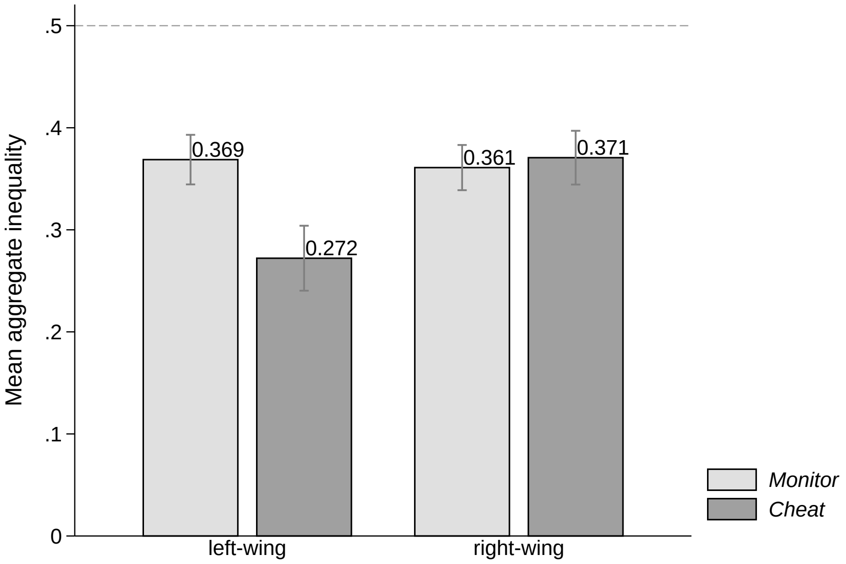
2.3.2 Inequality Acceptance and Political Preferences

The redistribution decisions of the spectators determine the final income distribution between two stakeholders. In order to quantify the extent to which spectators are willing to accept inequalities, I use the Gini coefficient as inequality measure:

$$\text{Inequality} = \frac{|\text{Income Player A1} \; - \; \text{Income Player A2}|}{\text{Income Player A1} \; + \; \text{Income Player A2}}$$

This measure of inequality relates the absolute difference in income to the total income and is zero in cases of full equality and one if one of the two stakeholders receives the entire total income. I define aggregate inequality as the average over the Gini coefficients of the eleven possible distributions of preliminary income. Thus, aggregate inequality contains the spectator's decisions for perfectly unequal preliminary incomes of (10,0), full equality in the case of (5,5) as well as all cases in between. Using this measure yields a lower bound for the treatment effect because one can expect hardly any redistribution in cases of low inequality (e.g., (5,5)) in both treatments.

Figure 2.2 shows the mean aggregate inequality across the two treatments and political preferences. If a spectator never redistributes income, aggregate inequality is 0.5, which is indicated by the horizontal dashed line. Left-wing spectators implement significantly lower inequality in *Cheat* than in *Monitor* ($p = 0.027$, Mann-Whitney *U*-test, two-sided). While they eliminate 26.2% of initial inequality in *Monitor*, they reduce inequality by 45.6% in *Cheat*, implying an increase of 74% in inequality reduction. In contrast, the treatment has no significant effect for right-wing spectators



Notes: The figure shows aggregate inequality defined as the average Gini coefficient of all eleven redistribution decisions. The dashed line indicates aggregate inequality in case of no redistribution. Error bars indicate standard errors of the mean.

Figure 2.2: Aggregate inequality

($p = 0.641$, Mann-Whitney *U*-test, two-sided). They reduce inequality by 27.8% in *Monitor* and 25.8% in *Cheat*.

Table 2.1 contains a series of Tobit regressions to account for using the censored dependent variable aggregate inequality. Column (1) suggests that aggregate inequality is not affected by the treatment for the pooled sample, which is due to the fact that the majority of spectators are right-wing. Column (2) confirms the result depicted in Figure 2.2. Inequality is significantly reduced through cheating opportunities when being left-wing ($p = 0.004$). However, the treatment does not affect implemented inequality of right-wing spectators. The sum of the treatment dummy and the interaction term "Cheat × Right-wing" in column (2) indicates that the treatment effect for right-wing spectators is insignificant ($p = 0.791$, *F*-Test).[12] Furthermore, being

[12] Interpreting the coefficient of an interaction term can be misleading in Tobit models (Ai and Norton, 2003). To examine this problem, I perform an alternative calculation of the interaction effect by computing the predicted values of aggregate inequality separately for left-wing and right-wing spectators in *Monitor* and *Cheat*. The respective difference in differences of these four groups' predicted values are of the same size as the marginal effect of the interaction terms in the models of column (2)

Table 2.1: Aggregate inequality

Dependent variable	Aggregate inequality					
	(1)	(2)	(3)	(4)	(5)	(6)
Cheat	−0.036	−0.112***	−0.110***	−0.110***	−0.089**	−0.118***
	(0.025)	(0.038)	(0.037)	(0.040)	(0.041)	(0.039)
Right-wing		−0.010	−0.028	−0.032	−0.006	−0.021
		(0.042)	(0.041)	(0.040)	(0.048)	(0.048)
Cheat × Right-wing		0.120**	0.116**	0.115**	0.128**	0.164***
		(0.048)	(0.045)	(0.047)	(0.062)	(0.058)
We need inequality			0.014*	0.013*	0.010	0.012
(1 = no, 10 = yes)			(0.007)	(0.007)	(0.007)	(0.009)
Constant	0.358***	0.364***	0.316***	0.334***	0.306***	0.215**
	(0.010)	(0.030)	(0.040)	(0.077)	(0.075)	(0.101)
Additional controls	No	No	No	Yes	Yes	Yes
Observations	182	182	182	182	182	144
Log likelihood	1.433	4.589	6.502	7.931	9.596	21.228

Notes: Two-limit Tobit regressions regressions on aggregate inequality. Columns (4) to (6) include a binary variable for whether the experiment was conducted in 2016 or 2017 (insignificant in all specifications) and additional covariates from the questionnaire: age, gender, semester, and number of experiments so far (all insignificant in all specifications). Column (5) includes the categories “other party” and nonvoters in the definition of being left-wing. Column (6) excludes spectators who would vote for “other party” as well as nonvoters. 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.

right-wing has no impact on aggregate inequality in the *Monitor* treatment ($p = 0.805$). In column (3), I add as regressor the answer to whether the spectators are of the opinion that income inequality should be reduced, or enlarged in order to provide incentives for individual performance. Believing that “we need inequality” increases implemented income inequality, while the treatment effect remains significant for left-wing spectators ($p = 0.003$). Hence, differences in the opinion on inequality[13] cannot explain that the effect of cheating opportunities depends on political preferences. The results are robust to adding a time dummy (whether the experiment was conducted in 2016 or 2017) as well as personal background characteristics in column (4). In column (5), I include nonvoters and spectators who indicate to vote

to (6) in Table 2.1. Thus, the bias induced by using interaction terms in a nonlinear model is negligible in my estimations.

[13] Right-wing spectators (mean $= 4.95$) favor inequality significantly more than left-wing spectators (mean $= 3.49$, $p < 0.0001$, Mann-Whitney *U*-test, two-sided).

for “other party” in the definition of left-wing instead of right-wing. The results are robust to this specification (with the exception that the significance of the treatment dummy decreases to $p = 0.031$). Furthermore, the results are also robust to excluding nonvoters and spectators who would vote for “other party” in column (6).[14]

In order to provide further evidence that left-wing spectators’ decisions are affected by the treatment *because* they suspect player A1 to be cheating, I analyze each redistribution decision separately. A spectator should doubt the performance of a stakeholder if it is far above the other stakeholder’s performance, which results in high income inequality.[15] In contrast, there is little reason to attribute cheating to a stakeholder when income is evenly distributed and thus reported performances of both stakeholders are similar.[16]

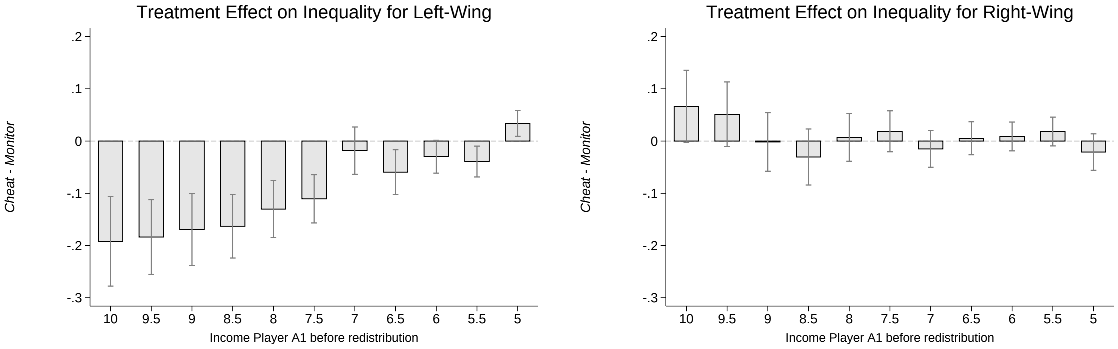
Figure 2.3 shows the treatment effect depending on preliminary income distribution for left-wing (left panel) and right-wing spectators (right panel). The very left and the very right bar represent the treatment effect when the income distribution before redistribution is (10,0) and (5,5) respectively.[17] The treatment has a significant negative effect on the implemented inequality of left-wing spectators for high pre-redistribution inequality. Two-sided Mann-Whitney *U*-tests indicate a

[14] In addition to investigating the Gini coefficient, I use the amount of money redistributed by spectators to study treatment differences. This is important because spectators might redistribute away from player A1 — who receives the higher preliminary income of the two stakeholders except for the case of full equality — in such a way that after redistribution player A2 has more income than player A1. For instance, a spectator in *Cheat* might determine the final income distribution to be (2,8) when preliminary incomes were (8,2), which results in the same inequality before and after redistribution. However, it is important to capture this incidence of redistribution as it might reflect punishing player A1 for potential cheating. Results on redistribution are reported in Appendix 2.5.1 and support the results of analyzing the Gini coefficient.

[15] This holds true even more when only one of the two stakeholders can cheat. To analyze the impact of asymmetric cheating opportunities within a pair of stakeholders, I ran an additional treatment where only one of the two stakeholders could misreport the own performance. The results of this treatment are reported in Appendix 2.5.2.

[16] In this case, both stakeholders could be cheating. However, there is no possibility to redistribute away from a suspicious potential cheater to another presumably more honest stakeholder because it is not possible to detect cheating.

[17] Levels of inequality for both treatments are shown in Figure 2.6 in the Appendix.

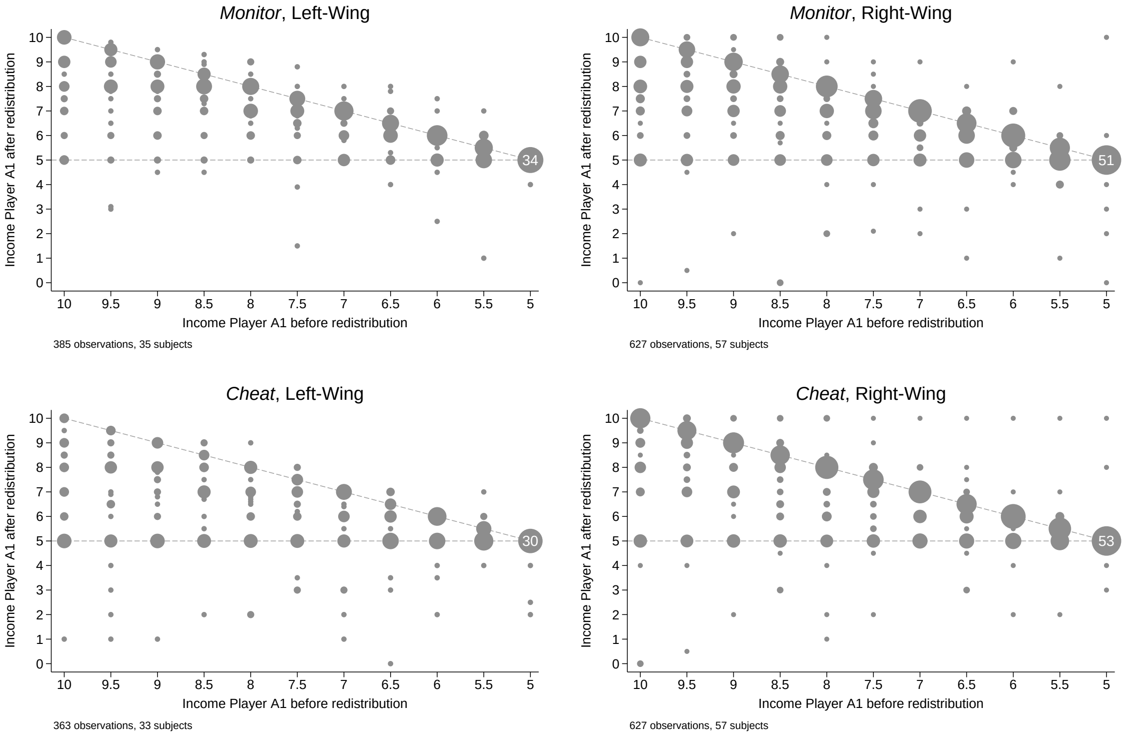


Notes: The figure shows the treatment effect on implemented inequality by subtracting inequality in the *Monitor* treatment from inequality in the *Cheat* treatment. The effect is shown separately for each preliminary level of income of player A1. The left panel displays the effect for left-wing spectators and the right panel for right-wing spectators. Error bars indicate standard errors of the mean.

Figure 2.3: Treatment effect on inequality

reduction in inequality if player A1 has a preliminary income of at least 7.5 (p-values do not exceed 0.033), while there is no significant treatment effect otherwise (with the exception of player A1 having €6.5 before redistribution, $p = 0.072$). These results suggest that left-wing spectators suspect player A1 to cheat when initial inequality is high and therefore reduce inequality in these cases. There is no significant effect of cheating opportunities for any of the preliminary income distributions when being right-wing.

Despite analyzing implemented inequality, the question remains which kind of redistribution decisions drive the treatment effect for left-wing spectators. Therefore, Figure 2.4 depicts player A1's income after redistribution contingent on his income before redistribution. Circles on the downward-sloping line indicate cases of no redistribution, circles on the horizontal line cases of redistribution resulting in full equality, and circles between the two lines are associated with redistribution away from player A1 such that the ranking of incomes is maintained. The few circles above the downward-sloping line represent "negative redistribution", which leaves the pair of stakeholders with higher inequality after redistribution at the expense



Notes: The figure shows the spectators' redistribution decisions depending on treatment (upper panels show *Monitor* and lower panels *Cheat*) and political preferences (left panels show left-wing and right panels right-wing spectators). Numbers in circles in case of no redistribution when the preliminary income of player A1 is €5 indicate the number of observations and serve as a benchmark for the remaining circles.

Figure 2.4: Overview redistribution decisions

of player A2. Circles below the horizontal line indicate "overredistribution", where player A1 receives a lower income than player A2.

Supporting the previous findings, systematic treatment differences can be inferred from Figure 2.4 only for left-wing spectators. Two-sided Fisher's exact tests reveal that the fraction of full equality is significantly higher in *Cheat* (lower left panel) than in *Monitor* (upper left panel) if player A1 has a preliminary income between 10 and 7.5, or 6.5 (five comparisons are significant at the 5% and two at the 10% level). This is in line with the result depicted in Figure 2.3 that left-wing spectators only react to cheating opportunities by implementing a lower inequality when the preliminary income distribution is unequal. In addition, there is a higher fraction of left-wing spectators in *Cheat* than in *Monitor* that always implement full equality independent of preliminary incomes (21.2% vs. 5.7%; $p = 0.079$, Fisher's exact test, two-sided). These two results suggest that the treatment difference for left-wing spectators is driven by an increase in redistribution decisions that result in full equality when cheating opportunities are present.

The analysis of this section shows that left-wing spectators are less willing to accept inequalities when cheating is possible. Looking at their decisions contingent on the income distribution before spectators can redistribute suggests that this is the case because they suspect the stakeholder with higher initial income of cheating. As a consequence, left-wing spectators implement more often a perfectly equal income distribution between the two stakeholders. The interaction between cheating opportunities and being left-wing cannot be explained by differences in their opinion on income inequality. In the next section, I therefore investigate whether different beliefs or norms about cheating across left-wing and right-wing spectators can account for this finding.

2.3.3 Beliefs and Norms about Cheating

Apart from preferences, there are two alternative explanations for why treatment differences depend on political color. (i) Beliefs about cheating might interact with being left-wing. If in the *Cheat* treatment left-wing spectators believe to a larger extent that stakeholders are cheating than right-wing spectators, this might account for the treatment effect. (ii) Norms about cheating might differ between left-wing and right-wing spectators. If right-wing spectators find it more acceptable to cheat when there is an opportunity to do so than left-wing spectators, this could also explain the results.

Beliefs about cheating can be inferred from the three different measures of beliefs in the *Cheat* treatment. First, subjects were asked to guess the average reported performance in their own session (*belief-own-treat*) and the average reported performance in the *Monitor* treatment (*belief-other-treat*). Subtracting the latter from the former one indicates the spectator's belief to which extent stakeholders cheated on average. This difference is not significantly different between left-wing (mean $= 2.65$) and right-wing (mean $= 2.02$) spectators ($p = 0.389$, Mann-Whitney *U*-test, two-sided). Second, I also compare *belief-own-treat* between the two groups because spectators' answers to the belief about the *Monitor* treatment (*belief-other-treat*) could suffer from self-serving ex-post rationalization of their redistribution decisions. *Belief-own-treat* is not significantly different between left-wing (mean $= 10.88$) and right-wing spectators (mean $= 11.21$; $p = 0.666$, Mann-Whitney *U*-test, two-sided). Third, subjects in *Cheat* were asked to guess the fraction of dishonest stakeholders (*belief-frac-cheat*). Again, beliefs do not differ between left-wing (58%) and right-wing spectators (57%; $p = 0.807$, Mann-Whitney *U*-test, two-sided). Thus, although right-wing spectators believe to the same extent as left-wing spectators that stakeholders are cheating, they are not willing to redistribute more in *Cheat* than in *Monitor*.

It has been shown that norms (behavior that people perceive as appropriate) have predictive power for subjects' actual behavior (e.g., Krupka and Weber, 2013). In addition, several studies in the economics literature use actual behavior to identify norms (e.g., Camerer and Fehr, 2004; Fehr and Fischbacher, 2004). Therefore, I use actual cheating behavior of the stakeholders as a proxy for the spectators' norms about cheating. Performances in the *Monitor* treatment show that left-wing and right-wing stakeholders are equally able to work on the matrix task ($p = 0.977$, Mann-Whitney *U*-test, two-sided). While their average performance in *Monitor* is 11.53 and 11.5 tasks respectively, in *Cheat*, left-wing stakeholders report to have solved 14.35 tasks and right-wing stakeholders 13.08 tasks. This difference between the two groups is not significant ($p = 0.283$, Mann-Whitney *U*-test, two-sided). If anything, left-wing stakeholders tend to cheat more than right-wing stakeholders. However, there is no significant evidence that the norm about cheating depends on political preferences. In particular, looking at actual behavior suggests that right-wing stakeholders do not find it more acceptable to cheat than left-wing stakeholders and I assume that the same holds true for right-wing and left-wing spectators.[18]

Finding no differences in beliefs and norms about cheating between left-wing and right-wing spectators suggests that the difference in their choices reflects a difference in preferences. Right-wing spectators are reluctant to take away money from a stakeholder due to potential cheating if they do not know whether this stakeholder actually cheated. In contrast, left-wing spectators are willing to redistribute more if they believe that someone has cheated — even without being able to detect cheating.

2.4 Concluding Remarks

The sources of inequality largely influence what people consider to be a fair distribution of income and wealth within a society. Assuming that fiscal imbalances will rise in most western countries due to demographic trends, these redistributive preferences will be particularly relevant for designing welfare policies in the future and thus constitute an important issue in public economics (Kuziemko et al., 2015). In this chapter, I focus on cheating as a potential source of unequal outcomes. I find large differences in how to deal with these inequalities depending on political preferences. Supporters of left-wing parties substantially redistribute incomes when cheating is possible, while supporters of right-wing parties refrain from redistribution. As a consequence, cheating opportunities — which receive increasing public attention

[18] Since roles were randomly assigned in the experiment, norms of stakeholders and spectators should not systematically differ from each other.

through recent revelations about fraudulent behavior — amplify the disagreement over redistributive policies between the political left and right.

A deeper look into my data reveals that left-wing spectators' redistribution decisions are only affected by potential misreporting in cases of large income differences. This provides strong evidence that left-wing spectators suspect a "rich" stakeholder to be cheating. Furthermore, my results suggest that both beliefs and norms about cheating do not depend on political color. Hence, right-wingers refrain from redistribution although they believe that stakeholders are cheating and although they are themselves reluctant to cheat to the full extent. This shows that right-wing spectators hesitate to redistribute if they do not know for sure that high relative income was acquired by dishonest means, while left-wingers are less concerned about this.

These findings might help to understand the political debate about how to tackle tax evasion and can inform politicians about the consequences of preventing fraudulent behavior. An example for fighting tax evasion are a couple of German federal states which bought tax CDs that contain information about German tax dodgers' Swiss bank accounts. While the Social Democrats advocate the potentially illegal purchases from whistleblowers, the Conservatives object such measures.[19] One reason for these different strategies might be that tax CDs purchases raises attention to potential cheating, which is, according to my findings, beneficial for left-wing parties with regard to justifying redistributive policies.

Further implications might be drawn concerning the different extent of redistributive policies between Europe and the United States. Europeans prefer substantially more redistribution than U.S. Americans (Almås et al., 2016), which can be partly explained by differences in beliefs about whether luck or effort determine

[19] See, e.g., http://www.spiegel.de/international/germany/german-authorities-investigate-ubs-in-relation-to-tax-evasion-a-849366.html, last accessed on March 5, 2018.

inequalities (Alesina and Angeletos, 2005). As U.S. Americans seem to be far less skeptical towards the very rich, perceived cheating opportunities might also contribute to explain cross-country evidence on redistributive policies. Therefore, exploring how redistributive preferences are affected by potential cheating in the United States as opposed to Europe is a fruitful avenue for further research.

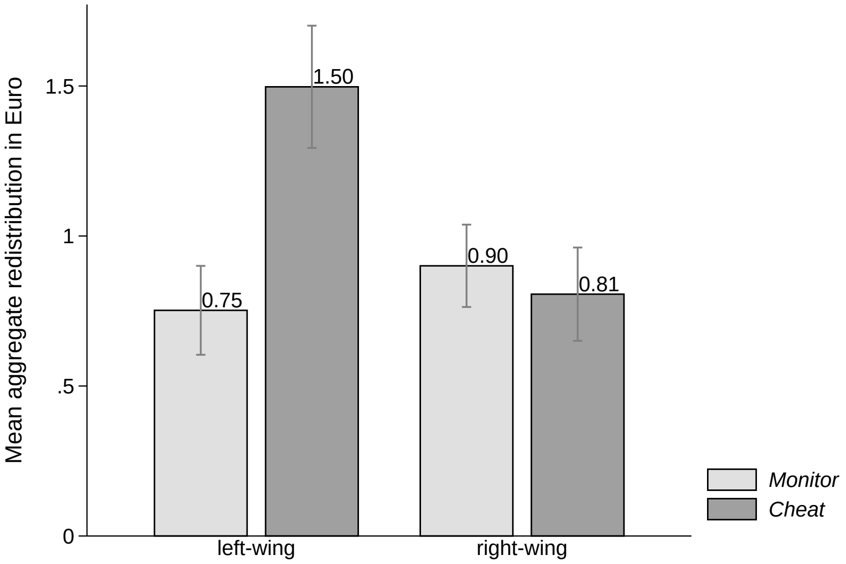
2.5 Appendix

2.5.1 Supplementary Results

Table 2.2: Voting behavior and political left-right score of spectators & 2017 German federal election results

Political Party	Experiment (in %)	2017 federal election (in %)	left-right scale
CDU/CSU	28.57	25.02	5.73
SPD	14.84	15.58	4.04
Die Grünen	14.29	7.02	4.08
Die Linke	6.04	6.79	3.45
AFD	3.30	8.17	7.67
FDP	9.89	9.60	5.56
Die Piraten	2.20	0.28	3.75
Other party	6.04	3.51	4.73
Would not go to the election	14.84	24.03	4.30

Notes: The parties are the Christian Democrats (CDU/CSU), the Social Democrats (SPD), the Green Party (Die Grünen), the socialist party (Die Linke), the right-wing populist Alternative for Germany (AFD), the libertarian party (FDP), and the Pirate Party Germany (Die Piraten). Results of the 2017 German federal election are based on the “second vote” and calculated without rejected votes. The scale for political attitudes ranges from 1 (left) to 10 (right).



Notes: The figure shows aggregate redistribution in € defined as the average redistribution of all eleven redistribution decisions. Error bars indicate standard errors of the mean.

Figure 2.5: Aggregate redistribution

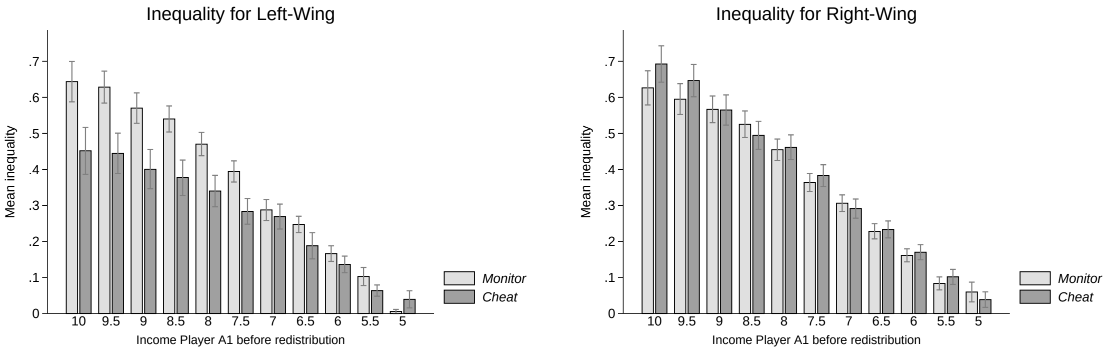
Figure 2.5 is equivalent to Figure 2.2 for redistribution instead of inequality. As for inequality, aggregate redistribution is defined as the average over all eleven decisions of the spectator.[20] Left-wing spectators react to the treatment even stronger when looking at redistribution instead of inequality. They redistribute twice as much in the *Cheat* treatment than in the *Monitor* treatment, which is highly significant ($p < 0.01$, Mann-Whitney *U*-test, two-sided). There is no significant effect of cheating opportunities for right-wing spectators ($p = 0.433$, Mann-Whitney *U*-test, two-sided). Regression results for aggregate redistribution can be found in Table 2.3 and are in line with the findings on inequality in Table 2.1.

[20] Decisions in which the spectator redistributes away from player A2 are also taken into account. Consequently, I use the absolute values of money redistributed to calculate the aggregate redistribution of one spectator.

Table 2.3: Aggregate redistribution

Dependent variable	Aggregate redistribution in €					
	(1)	(2)	(3)	(4)	(5)	(6)
Cheat	0.215	0.745***	0.737***	0.758***	0.478*	0.795***
	(0.142)	(0.246)	(0.240)	(0.252)	(0.265)	(0.260)
Right-wing		0.148	0.233	0.247	0.010	0.210
		(0.236)	(0.237)	(0.231)	(0.320)	(0.300)
Cheat × Right-wing		−0.840***	−0.818***	−0.820***	−0.593	−0.930***
		(0.307)	(0.284)	(0.280)	(0.374)	(0.341)
We need inequality			−0.065**	−0.064**	−0.052	−0.077*
(1 = no, 10 = yes)			(0.029)	(0.028)	(0.034)	(0.045)
Constant	0.844***	0.752***	0.983***	0.800*	0.959**	1.268**
	(0.078)	(0.165)	(0.176)	(0.478)	(0.457)	(0.520)
Additional controls	No	No	No	Yes	Yes	Yes
Observations	182	182	182	182	182	144
Log likelihood	-274.643	-270.166	-268.661	-264.546	-265.249	-202.152

Notes: Two-limit Tobit regressions regressions on aggregate redistribution (lower limit: –82.5, upper limit: 82.5). Columns (4) to (6) include a binary variable for whether the experiment was conducted in 2016 or 2017 (insignificant in all specifications) and additional covariates from the questionnaire: age, gender, semester, and number of experiments so far. Column (5) includes the categories “other party” and nonvoters in the definition of being left-wing. Column (6) excludes spectators who would vote for “other party” as well as nonvoters. 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.



Notes: The figure shows inequality separately for each level of preliminary income of player A1. Error bars indicate standard errors of the mean.

Figure 2.6: Inequality by preliminary income

Figure 2.6 shows inequality depending on preliminary income distribution and treatment for left-wing (left panel) and right-wing spectators (right panel). The two very left and very right bars represent inequality when the income distribution before redistribution is (10,0) and (5,5) respectively. Inequality is not significantly different between left-wing and right-wing spectators in any of the redistribution decisions in the *Monitor* treatment (Mann-Whitney *U*-tests, two-sided).

2.5.2 The *Mixed* Treatment

I conducted an additional treatment to find out how unequal cheating opportunities affect inequality acceptance. This is motivated by “real-world” examples where only some people are able to cheat, while it is much harder or even impossible for others to report untruthfully. Consider, for instance, the cases of tax evasion, doping in sports, and faking educational achievements. Only individuals who are subject to tax can evade taxes, and only the “rich” might have the means to do this large-scale. Only professional athletes have access to well known doping doctors. And only people belonging to the higher education system have the opportunity to plagiarize a PhD thesis. Therefore, I implement a treatment called *Mixed*, where only one of the two stakeholders is able to cheat. While in the *Cheat* treatment it was unclear whether both or only one of the two stakeholders cheated, even in cases with high preliminary inequality, it is clearer that the stakeholder with a much higher income is presumably cheating in the *Mixed* treatment. Furthermore, spectators might perceive it as unfair that only one of the two stakeholders has the opportunity to cheat. For these reasons, inequality should be even more reduced in *Mixed* than in *Cheat*.

Procedural Details. Eight sessions of the *Mixed* treatment were conducted in May and June 2017 with a total of 172 subjects. Subjects earned €12.94 on average, including a €5 show-up fee. The exercise sheet was collected only from one of the two stakeholders. After working on the matrix task, it was publicly announced that subjects will next be informed about their role and whether their exercise sheet will be collected (see Appendix 2.5.4.2). Hence, all participants, including the spectators, knew that the stakeholders were aware of their cheating opportunities before stating their performance and that only one of the two stakeholders could cheat.

Results. In the *Mixed* treatment, reported performance of subjects that were monitored (mean of 11 tasks) do not significantly differ from those that were able to cheat (12.63; $p = 0.158$, Mann-Whitney *U*-test, two-sided). Since reported performance of the monitored stakeholders in *Mixed* is generated under the exactly same conditions as of stakeholders in *Monitor*, these observations can be pooled and compared with potential cheaters in *Mixed*. Again, performances do not significantly differ ($p = 0.176$, Mann-Whitney *U*-test, two-sided).[21] In line with stakeholders' actual performances, beliefs of spectators about reported performance in their own session (*belief-own-treat*) do not significantly differ between *Monitor* (9.35 on average) and *Mixed* (9.31; $p = 0.832$, Mann-Whitney *U*-test, two-sided). In addition, asking specta-

[21] Based on these observations (135 subjects that cannot cheat with a mean of 11.35 tasks solved as baseline and 43 potential cheaters), I calculate the minimal detectable difference of a two-sided Mann-Whitney *U*-test for the 5% significance level. A treatment difference of at least 2.25 tasks is detected with a statistical power of 80%. Thus, the sample is large enough to detect treatment differences in performance of a similar size as implied by the difference between *Monitor* and *Cheat* (2.29 tasks).

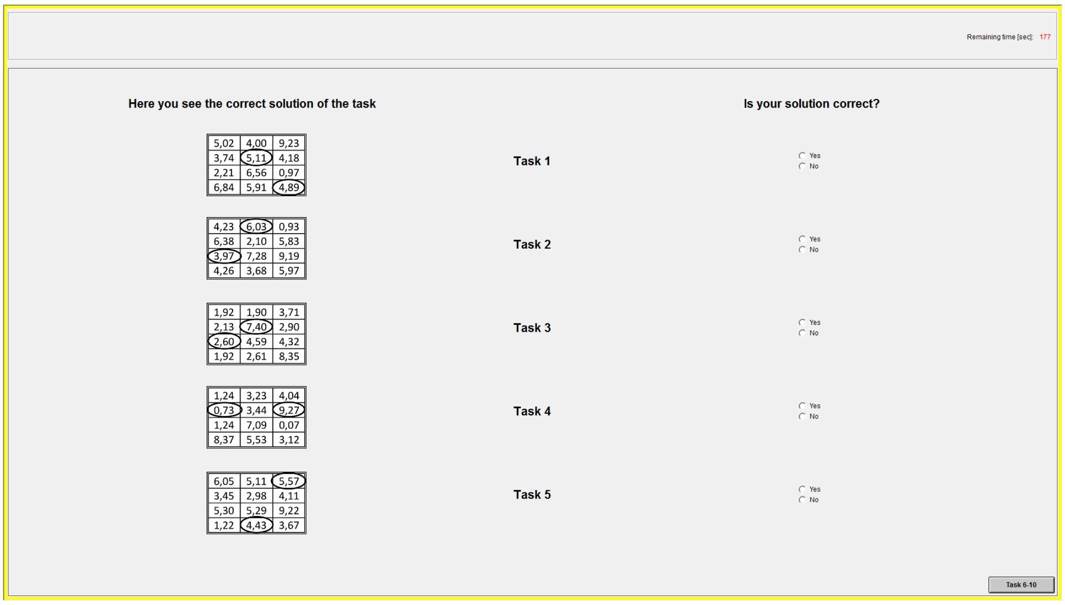
tors in *Mixed* about the average performance of stakeholders in the *Cheat* treatment (*belief-other-treat* in the *Mixed* treatment) reveals that spectators in *Mixed* believe reported performance to be higher in *Cheat* (12.87) than in *Mixed* ($p < 0.0001$, Wilcoxon signed rank test, two-sided). Moreover, spectators' beliefs about the ratio of cheaters to those stakeholders who are able to cheat (*belief-frac-cheat*) are higher in *Cheat* (57%) than in *Mixed* (50%; $p = 0.036$, Mann-Whitney *U*-test, two-sided).[22] Hence, spectators seem to anticipate stakeholders (missing) dishonest behavior as there is no clear evidence that spectators in *Mixed* believe stakeholders be cheating.

As a consequence of failed treatment manipulation in the *Mixed* treatment, implemented aggregate inequality does neither differ for right-wing ($p = 0.467$) nor left-wing spectators ($p = 0.670$) between *Monitor* and *Mixed* (Mann-Whitney *U*-tests, two-sided). In summary, the null hypothesis of equal performance between stakeholders that are monitored and those who are not cannot be rejected. The low or nonexistent occurrence of cheating might be a result of stakeholders finding it unfair that only one of them can cheat. When designing this treatment, it was difficult to predict this finding since this is, to the best of my knowledge, the first treatment where only one of two subjects who are otherwise in exactly the same position can cheat.[23] In line with this, spectators in *Mixed* seem to anticipate stakeholders' behavior as I find no evidence that they believe stakeholders to be cheating. Given this, it is not surprising that I do not find a treatment effect in *Mixed*.

[22] *Belief-frac-cheat* is not incentivized and likely to be overstated due to demand effects by explicitly asking for the fraction of cheaters. Therefore, I refrain from interpreting the size of this belief but only compare the difference across treatments.

[23] This conclusion is drawn from comparing the *Mixed* treatment to treatments of 72 papers analyzed in a meta-study by Abeler et al. (2016).

2.5.3 Screenshots of Decision Screens



Remaining time [sec]: 177

Here you see the correct solution of the task

Is your solution correct?

5,02	4,00	9,23
3,74	5,11	4,18
2,21	6,56	0,97
6,84	5,91	4,89

Task 1 ○ Yes ○ No

4,23	6,03	0,93
6,38	2,10	5,83
3,97	7,28	9,19
4,26	3,68	5,97

Task 2 ○ Yes ○ No

1,92	1,90	3,71
2,13	7,40	2,90
2,60	4,59	4,32
1,92	2,61	8,35

Task 3 ○ Yes ○ No

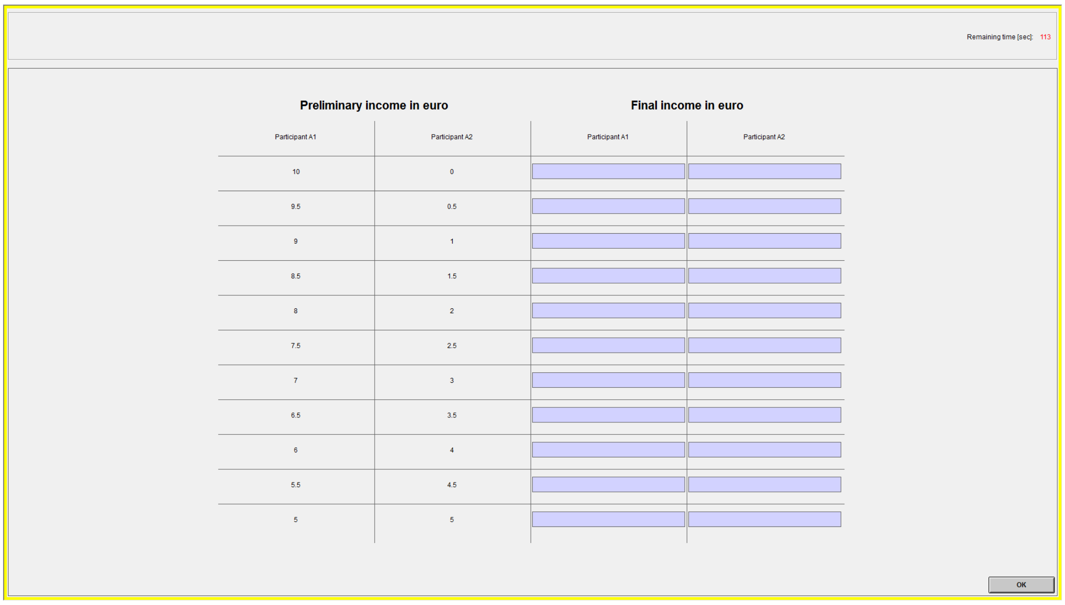
1,24	3,23	4,04
0,73	3,44	9,27
1,24	7,09	0,07
8,37	5,53	3,12

Task 4 ○ Yes ○ No

6,05	5,11	5,57
3,45	2,98	4,11
5,30	5,29	9,22
1,22	4,43	3,67

Task 5 ○ Yes ○ No

Figure 2.7: Screenshot of one of the stakeholder's decision screens



Remaining time [sec]: 113

Preliminary income in euro		Final income in euro	
Participant A1	Participant A2	Participant A1	Participant A2
10	0		
9.5	0.5		
9	1		
8.5	1.5		
8	2		
7.5	2.5		
7	3		
6.5	3.5		
6	4		
5.5	4.5		
5	5		

Figure 2.8: Screenshot of the spectator's decision screen

2.5.4 Instructions

2.5.4.1 General Instructions at the Beginning of the Experiment

[In paper form][24]

Welcome to the experiment and thank you for your participation!

Please do not talk to other participants of the experiment from now on.

General information on the procedure

This experiment serves to investigate economic decision making behavior. You can earn money, which will be paid to you individually and in cash after the experiment has ended.

If you have any questions after reading these instructions or during the experiment, please raise your hand or press the red button on your keyboard. We will then come to you and answer your question in private.

During the experiment, you and the other participants will make decisions. Your own decisions as well as the decisions of other participants can determine your payoffs. These payoffs are determined according to the rules which are explained in the following.

Payment

At the end of the experiment, you will receive in cash the money that you have earned during the experiment and additional 5 euro for showing up in time. Therefore, we will call every participant based on his seat number, i.e., none of the other participants gets to know your payment, and also you will not get to know the payments of other participants.

[24] The instructions were translated from German. The original version is available upon request.

Anonymity

Data from this experiment will be analyzed anonymously, i.e., we will never link your name to the data of the experiment. At the end of the experiment, you have to sign a receipt confirming that you received your payment. This receipt serves accounting purposes only.

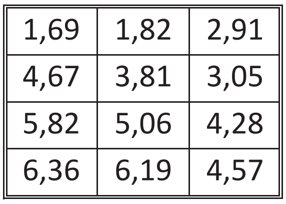
Assignment of Roles

There are two different roles in this experiment: A and B. The role of each participant is determined randomly. Whether you are participant A or B will be communicated to you at a later point in time on your screen.

Your Task

At the beginning of the experiment, we will hand out an exercise sheet, which we will place upside down on your desk. Please turn the sheet over only when you are asked to do so. There are 20 tasks on the exercise, which are numbered top down. It does not matter on which task you work first.

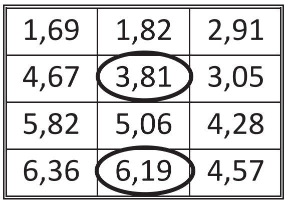
Each task consists of a box containing 12 numbers. Here is an example:



1,69	1,82	2,91
4,67	3,81	3,05
5,82	5,06	4,28
6,36	6,19	4,57

Example

Only two numbers in the box add up to 10.00. It is your task to find these two numbers and to circle them. In the following, you see the correct solution for the example.



1,69	1,82	2,91
4,67	[✓ 3,81]	3,05
5,82	5,06	4,28
6,36	[✓ 6,19]	4,57

Correct solution of the example

You have 6 minutes to work on the tasks. For your guidance, a clock will display the remaining time on your screen. After the 6 minutes expired, please put down your pen. Subsequently, we will collect your pens.

Income of Participant A

After timeout, participant A compares the numbers which he marked to the solution. You will receive more information on this after working on the task. Participant A receives 1 point for every correct solution.

Two participants A are randomly assigned to each other in order to determine their incomes. 10 euro will be split up among these two participants. Income is proportional to the number of points achieved in the preceding task and rounded to 50 cents.

Example 1: One participant A achieved 6 points and the other participant A 4 points. Thus, both participants A have achieved 10 points in total. The participant A with 6 points therefore receives an income of 6.00 euro ((6 points / 10 points) x 10 euro = 6 euro). The participant A with 4 points receives an income of 4 Euro ((4 points / 10 points) x 10 euro = 4.00 euro).

Example 2: One participant A achieved 12 points and the other participant A 7 points. Thus, both participants A have achieved 19 points in total. The participant A with 12 points therefore receives an income of 6.50 euro ((12 points / 19 points) x 10 euro = 6.32, rounded to 50 cents). The participant A with 7 points receives an income of 3.50 euro ((7 points / 19 points) x 10 euro = 3.68 euro, rounded to 50 cents).

Income of Participant B

Participant B can distribute the incomes which participants A earned. Participants B will receive detailed information about this on their screen later on.

Further Procedures

We will soon hand out an exercise sheet to each of you. Please leave this sheet upside down until we announce the beginning of the task. After you have finished the task, the computer determines whether you are participant A or B. You will be informed about this on your screen. The assignment of roles is random. During the experiment, you will receive further information on your screen.

2.5.4.2 Treatment Variation after the Matrix Task

[On the screens of the participants]

Entering of Solutions

All participants have now worked on the task.

Participants A will soon have 3 minutes to compare the numbers which he marked to the solution as follows:

On the left hand side of the screen, participant A sees the correct solution of the tasks. On the right hand side of the screen, participant A should indicate whether the correct solution corresponds to the numbers that he marked on his exercise sheet.

After timeout, participants A automatically proceed to the next screen and can no longer compare solutions.

On the next screen, you will see an example of the screen on which participant A compares his solutions with the correct solutions. This screen will be displayed to you for 15 seconds. You do neither have to indicate something nor click on OK. The tasks on the following screen are examples. Please click now on OK.

[new screen]

[*Monitor* treatment only]

Important: After all participants A have compared their solutions to the correct solutions, we collect the exercise sheets of all participants. We then verify that participant A did not make any mistake when comparing his solutions to the correct ones. If participant A made a mistake, we will correct the number of points of participant A. You will soon be informed whether you are participant A or B. Please click now on OK.

[*Cheat* treatment only]

Important: You will receive your payment in the room next door. There is also a shredder in this room. At the end of the experiment, you will shred your exercise sheet and afterwards receive your payment. This ensures that we cannot trace back your solutions. You will soon be informed whether you are participant A or B. Please click now on OK.

[*Mixed* treatment only, see Appendix 2.5.2]

Important: After all participants A have compared their solutions to the correct solutions, we collect the exercise sheets of **one** of the two matched participants A. We then verify that this participant A did not make any mistake when comparing his solutions to the correct ones. If this participant A made a mistake, we will correct the number of points of this participant A. We **do not** collect the exercise sheet of the other participant A. You will receive your payment in the room next door. There is also a shredder in this room. If we do not collect your exercise sheet, please put it in the envelope which you find on your desk. **All** participants have to seal their envelopes and shred it before receiving their payoff. This ensures that we cannot trace back the solutions of the participants whose exercise sheets we do not collect. You will soon be informed whether you are participant A or B, and whether we will collect your exercise sheet. Please click now on OK.

2.5.4.3 Instructions for Participant B

[On screens of participants B while participants A compare their solutions to correct solutions; not read aloud by experimenter]

Your Decisions

As participant B you will be randomly assigned to 2 participants A among whom 10 euro will be split up as described in the instructions. We call this income, which they receive for working on the task, *preliminary income* of participants A. You will now determine the *final income* of participants A for *every possible* distribution of the preliminary incomes.

You see the table where you will enter the final incomes further below. [table without possibility to enter something is shown at the bottom of the screen] Final incomes of participants A, which you determine, must add up to 10 euro. You may enter final incomes in 10 cents steps. In order to assign final incomes unambiguously, we call the two participants A “participant A1” and “participant A2”.

After the two participants A compared their solutions to the correct solutions, preliminary incomes will be determined. This income then corresponds to one row in the table: e.g., 7 euro for participant A1 and 3 euro for participant A2. Participants A will receive the final incomes that you enter in the same row on the right side of the table (e.g., next to 7 euro for participant A1 and 3 euro for participant A2). Hence, each of your decisions can be decisive for the payoffs of the participants A!

[new screen]

Implementation of Your Decision

At the end of the experiment, the computer will randomly determine whether your decision or the decision of another participant B will be implemented. The probability for your decision to be implemented is 50%.

Your Income

As participant B you receive a fixed income of 10 euro independent of your decision.

You will make your decisions on the next screen.

2.5.4.4 Incentivized Belief Elicitation

[On screens of participants; not read aloud by experimenter; text in *Monitor*]

Assessment

You will provide two assessments in the following. You will receive details hereto on the next two screens. You are paid for the accuracy of your assessments.
However, only one of the two assessments is paid. At the end of the experiment, the computer will randomly determine which of the two assessments is paid. The probability that assessment 1 or assessment 2 is paid is 50% respectively.

[new screen]

Assessment 1

All participants A now have compared their solutions to the correct ones. Please assess how many points participants A achieved on average in the task at the beginning of the experiment.
You are paid for the accuracy of your assessment. If your assessment deviates less than 1 point from the actual average, you will receive **2 euro** additional to your remaining income from the experiment. If your assessment deviates at least 1 point but less than

2 points from the actual average, you will receive **1 euro**. Larger deviations are not paid.
How many points did the participants A achieve on average? (integers only)

[new screen]

Assessment 2
We now ask you to provide an assessment for a similar experiment. In the other experiment, participants worked on exactly the same task on which you have worked in the current experiment.

Important
In contrast to the current experiment, exercise sheets were not collected and it was not verified that the solutions had been compared without making a mistake. Participants were aware of this before comparing the solutions.

Please assess how many points participants A achieved on average in the other experiment.
As before, you are paid for the accuracy of your assessment. If your assessment deviates less than 1 point from the actual average, you will receive **2 euro** additional to your remaining income from the experiment. If your assessment deviates at least 1 point but less than 2 points from the actual average, you will receive **1 euro**. Larger deviations are not paid.
How many points did the participants A achieve on average in the other experiment? (integers only)

Choice as Justification for Dishonesty

3

Joint with Florian Loipersberger

3.1 Introduction

The expanding literature on dishonesty indicates that some people want to benefit from the gains of lying but simultaneously prefer to appear honest in front of others and towards themselves (e.g., Mazar et al., 2008; Fischbacher and Föllmi-Heusi, 2013; Abeler et al., 2016). Thus, these people care about the norm of truth-telling and therefore do not bluntly lie to the full extent. As a consequence, they create excuses for dishonest behavior to cope with the conflict between profits from lying and moral considerations.

In this chapter, we examine whether people justify dishonest behavior towards another person with the idea that this other person self-selected into a situation where being lied to is possible. Consider, for example, the case of an employer who promises his employees that their workload will not be increased in the future, even though this is exactly what he plans to implement. To justify this lie, he tells himself that his employees were free to choose to work for a different company. As another example, consider a bank employee who recommends an unfit financial product to a customer in order to receive a high commission. In light of her guilty conscience, she brings to her mind that every customer chooses her bank and financial products herself. In these examples, individuals apply the following principle to justify immoral behavior. People should be held personally responsible for their outcomes in life, in particular if they could have chosen differently (Dworkin, 1981a,b).

We address this issue by conducting a laboratory experiment. Two participants are randomly matched with each other. One of them (the "potential liar") is able to lie to the other (the "other participant"). In our design, the potential liar benefits in monetary terms from behaving dishonestly by reducing the other participant's payoff.[1] However, the other participant does not necessarily engage in an interaction with the potential liar but might receive an alternative payment instead. In the *Random* treatment, chance determines whether both participants interact with each other. Yet in the *Choice* treatment, the other participant can choose to interact with the potential liar or to take the alternative payment.

Our experimental setup enables us to cleanly identify the effect of making a choice on the probability of being lied to. This is difficult to achieve in field settings as most choices in the field involve reasonable alternatives. For instance, consider an employee who can choose between two comparable employers. Deciding for one of the two employers signals trust that the chosen employer will keep his promises. In this situation, trust and choice coincide. Therefore, their effects on lying behavior cannot be disentangled. In order to exclude trust as a confounding factor, we follow Cappelen et al. (2016) and implement a meaningless choice. In particular, we set the alternative payment merely €0.05 higher than the minimum that can be obtained in the interaction. By not offering an acceptable alternative, we de facto force the other participant into the interaction. Hence, both treatments differ in one aspect only. In contrast to the *Random* treatment, the other participant made a "forced choice", i.e., a choice without acceptable alternatives, in the *Choice* treatment.

Our results suggest that there is no overall effect of the possibility to self-select into a situation on lying behavior. Lying is only slightly more prevalent in *Choice* compared to *Random*. Supporting a large body of the literature (e.g., Dreber and

[1] Thus, we study selfish black lies. Several other types of lying are studied in the literature such as lies that benefit other people. For an overview of different types of lying, see Erat and Gneezy (2012).

Johannesson, 2008; Conrads et al., 2013; Rosenbaum et al., 2014; Abeler et al., 2016; Grosch and Rau, 2017), we find that men are much more likely to lie than females. Interestingly, the treatment has a strong and significant effect on males. They are 56% more likely to lie to the other participant in the *Choice* treatment compared to in the *Random* treatment. In contrast, we find no significant treatment effect for females. This suggests that forced choices can induce males to justify dishonest behavior. Females, however, do not seem to consider forced choices as a legitimate excuse for lying. In addition, we find that economics and business students lie significantly more than students from other fields of study. Moreover, participants who generally trust other people lie substantially less and thus can be interpreted to be more trustworthy themselves.

Our study offers several contributions to the literature. First, various studies have employed choice as treatment manipulation, for instance, through exogenous and endogenous group formation in team production (Herbst et al., 2015), exogenous or voting based rules in public good games (Sutter et al., 2010), or to find out whether intentions matter in interactions where people are able to reciprocate (Falk et al., 2008).[2] However, we are aware of only one other study that uses choices without acceptable alternatives as treatment variation in order to study the mere effect of choice. Cappelen et al. (2016) study how a forced choice, similar to the choice we implement, and a “nominal choice”, a choice between two ex ante identical lotteries, affect the willingness to accept inequalities. The authors find that both of these choices increases inequality acceptance. In this sense, their finding is in line with our result (for male potential liars) since in both studies a forced choice serves as justification for certain behavior. In particular, it seems as if people are held to a

[2] There is also a large strand of literature dealing with the determinants of individual choices, for instance, in the domains of financial decision making, education, or health. However, we do not study which economic variables affect choices but the *consequences* of the possibility to make a choice on lying behavior.

certain extent responsible for their own forced choices and it is thus morally acceptable to disadvantage them.

Second, we contribute to the growing experimental literature on dishonesty (e.g., Mazar et al., 2008; Cappelen et al., 2013b; Cohn et al., 2015). Our design is related to studies that build on the die roll paradigm introduced in the literature by Fischbacher and Föllmi-Heusi (2013): Subjects report the outcome of a random variable and are paid according to their report. Since the outcome is observed in private, subjects may lie about the observed outcome to increase their payoff. These studies detect lying at the group level by comparing the empirical distribution of reported outcomes to the underlying theoretical distribution. In contrast to that, we employ individual lying detection. Hence, our participants do not observe the outcome in private but are observed by the experimenter.[3] This may affect the overall extent of lying in our experiment. Therefore, we focus on comparing the two different treatments rather than interpreting the absolute levels of dishonesty.

Finally, our results contribute to the strand of literature that examines gender differences in economic behavior. There are, for instance, pronounced differences between males and females with regard to risk preferences or preferences for competition. Males seem to be less risk averse than females (e.g., Eckel and Grossman, 2008; Charness and Gneezy, 2012) and more competitive (e.g., Gneezy et al., 2003; Niederle and Vesterlund, 2007). More specifically, our findings contribute to the evidence about gender differences in dishonesty. We stress that we did not have a hypothesis on gender effects since the literature provides no consistent evidence on the relationship between gender and lying. As mentioned above, there are many studies that find that males are significantly more likely to lie than females. However, there are other papers that find no significant gender difference (e.g., Childs, 2012;

[3] See Kocher et al. (2017) and Gneezy et al. (2018) for two recent studies that also implement individual lying detection.

Gylfason et al., 2013; Abeler et al., 2014). This mixed evidence is reflected by our results. While we find an overall substantial and significant gender effect, this is solely driven by differences in the *Choice* treatment. We do not observe any differences between males and females in the *Random* treatment. Kajackaite and Gneezy (2017) also provide evidence that gender differences in lying are not stable across different experimental designs. While the authors show that males lie more often in a version of the Fischbacher-Föllmi-Heusi paradigm, they find no gender differences in a version of the mind game, where participants imagine to throw a die and have to report the respective number (Jiang, 2013). It can therefore be concluded that the gender effect in lying decisions depends on the specific context.

The remainder of this chapter is structured as follows. In Section 3.2, we describe the experimental design as well as the procedural details and a power calculation. We present our results in Section 3.3. In Section 3.4, we conclude.

3.2 Experiment

3.2.1 Experimental Design

Basic Setup and Interaction. At the beginning of the experiment, each subject is randomly assigned one of two different roles, role A or role B. Every participant is informed about his own role on the screen. Thereafter, each player A is randomly matched with one of the players B. Player B either engages in an interaction with player A (left half of the game tree, see Figure 3.1) or receives an outside option (right half of Figure 3.1).

The interaction gives player A the opportunity to lie at the expense of player B.[4] In order to implement this feature, player A takes part in a lottery. The computer randomly displays either of two colors, green (the "high state") or orange (the "low state"), on player A's screen.[5] Chances are $q = 0.1$ for the high state and $1 - q = 0.9$ for the low state to occur. We then ask player A to report the color that he has seen to player B, who does not know the displayed color.[6] The payoff of both players depends on the reported color. If player A reports the high state, player A receives €5 and player B €1. Payoffs reverse if player A reports the low state. In this way, we obtain a zero-sum game. Thus, lying is neither efficiency enhancing nor decreasing, which excludes efficiency concerns as a motive for lying behavior. Furthermore, when the displayed color is orange, there is an incentive for player A to lie to player B, as player A then receives the high payoff.

When there is no interaction, player A takes part in the same lottery. The difference here is that the color which player A reports does not affect player B's payoff. Instead, player B receives an outside option of €1.05 and does not observe the reported color in this case.

In order to study the mere effect of choice, we implement two treatments. In our control setting, which we denote by *Random*, the computer randomly determines whether player B faces the interaction or the outside option. In contrast, player B actively takes this decision in our treatment group, which we therefore denote by *Choice*. The outside option of €1.05 is only marginally higher than the lowest payoff that can be obtained in the interaction (€1). As a consequence, we implement a choice without acceptable alternatives and isolate the effect of having a choice from

[4] Player A is also able to lie in favor of player B (and at the expense of himself). However, there is no monetary incentive to do so. Moreover, none of our participants did engage in such downward lying. We also discuss this issue in Section 3.3.

[5] Figure 3.4 depicts how the color is displayed on the screen.

[6] See Figure 3.5 for player A's decision screen.

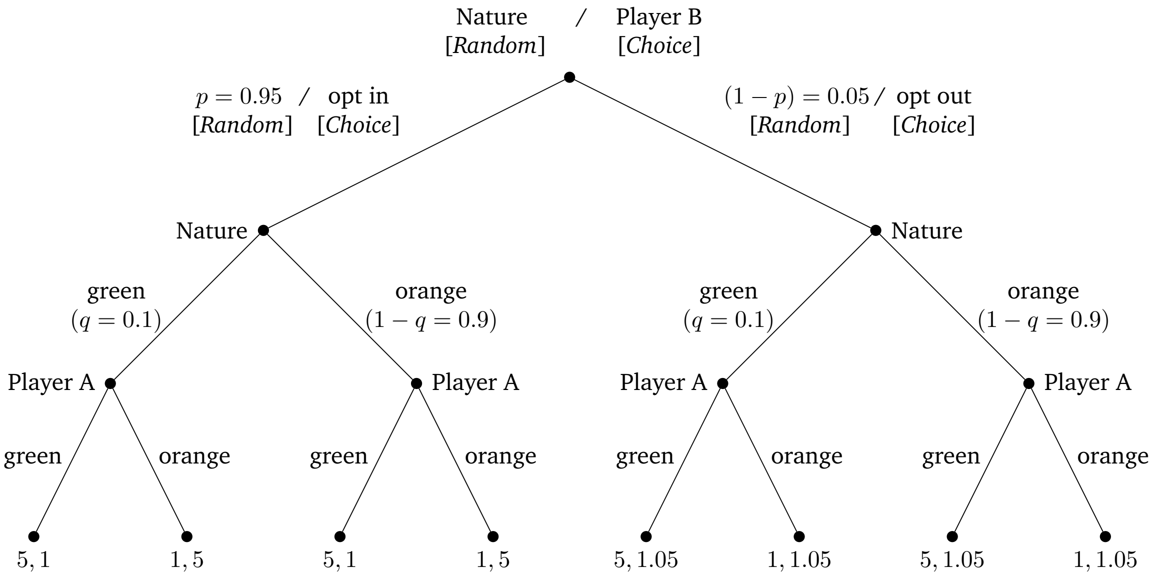


Figure 3.1: Game tree

other confounding factors (similar to Cappelen et al., 2016). For instance, if we gave player B a "real" outside option, say €2, opting in would signal that player B trusts player A to report the true color. In this case, player A might feel compelled to tell the truth because player B signals trust. This motive would be indistinguishable from acting upon player B's choice. In addition, player B's expected value of opting in is €4.60 in case of truth-telling and thus amounts to a mark-up of €3.60 above the lowest possible payoff, while the mark-up is €0.05 in case of taking the outside option. Hence, there is a 72 times higher mark-up under the assumption of truth-telling for opting in, which leaves the outside option to be a barely acceptable alternative. Therefore, we de facto force player B to decide in favor of the interaction.

Predictions. A model with purely self-interested agents yields a subgame-perfect Nash equilibrium for the *Choice* treatment in which player B opts out of the interaction because he knows that player A always reports green. Furthermore, standard economic theory predicts that player A always reports green in the *Random* treatment. Deviations from these predictions can be obtained by assuming a certain extent of lying aversion for player A. As a consequence, player A must report the true color with a probability of at least roughly 1.4% such that player B opts in, assuming that player B is a risk-neutral expected utility maximizer and player A does not engage in downward lying. Therefore, we hypothesize that a positive number of players B selects into the interaction with player A. This is necessary to examine the effect of self-selection into the interaction on the probability of being lied to.

Probabilities in the *Random* Treatment. In the *Random* treatment, we need to define the probability for engaging in the interaction. To obtain similarity to the *Choice* treatment, this probability should be close to the fraction of players B choosing the interaction in *Choice*. In this way, player A faces the same probability of interacting with player B across treatments. As it can be expected that a low fraction of players B decides in favor of the outside option in the *Choice* treatment, we set the probability of receiving the outside option in the *Random* treatment to the low but non-zero value of $1 - p = 0.05$.[7]

Matching. Our design implies that only player A has the possibility to lie. As the focus of our analysis is this lying decision, we need to generate a high number of role A observations. Therefore, we assign role A more often than role B. Next, we form groups by matching several players A to one player B. At the end of the experiment, the computer randomly selects one player A in every group. This individual's decision is implemented and therefore also determines the payoff of player B in that group. The remaining players A receive a flat payment of €3. We notify participants that it is still optimal to choose as if their decision was implemented.

3.2.2 Procedural Details

In total, 263 subjects participated in our experiment at the Munich Experimental Laboratory for Economic and Social Sciences (MELESSA) in June and July 2017.

[7] As it turns out, not a single participant decided to opt out in the *Choice* treatment.

We conducted nine sessions of the *Random* treatment with 121 subjects and nine session of the *Choice* treatment with 142 subjects. Our participants were university students from various fields of study and randomly assigned to one treatment. We used the online recruiting system "ORSEE" (Greiner, 2015) to recruit our subjects and programmed and conducted the experiment with the software "z-Tree" (Fischbacher, 2007). Each subjects participated in one session only.

At the beginning of the experiment, subjects received a printed version of the instructions, which can be found in Appendix 3.5.3. The instructions were read aloud such that it was common knowledge that all subjects received the same information. Next, subjects had to answer a couple of control questions, and only after every subject correctly answered all questions, the experiment proceeded.

At the end of the experiment, subjects anonymously answered a questionnaire about whether one can trust people in general[8] as well as socio-demographic characteristics such as field of study, age, or sex. Our subjects are on average 23.11 years old and 37.3% of them are male. After the experiment, subjects received their earnings, which amounted to €8 on average, including a €5 show-up fee. An average session lasted about 20 minutes.

3.2.3 Power Calculation

Given the number of relevant observations in the two treatments,[9] we compute the minimal detectable treatment effect size for lying behavior. We base our calculations on a two-sided χ^2-test of proportions and assume that the fraction of players A who lie in *Random* is 41% (the actual fraction of liars). With a statistical power of 80% and a

[8] The question is taken from the World Value Survey and asks: "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?" The answer to the question is to either support the former or the latter part of the question.

[9] We are only interested in players A who are able to lie to player B (83 in *Random* and 107 in *Choice*). At the beginning of Section 3.3, we explain how we obtain these numbers.

significance level of 5%, the minimal detectable treatment difference is 20 percentage points. As this is certainly not a small treatment difference, one should be cautious when interpreting insignificant treatment effects in our experiment.

3.3 Results

The aim of this chapter is to investigate whether the choice to interact with someone increases the probability of being lied to. This implies that we only analyze type A individuals who had the possibility to lie to player B (i.e., they interacted with player B *and* the computer displayed the low state on their screen).[10] That leaves us with 190 relevant observations, 107 in the *Choice* treatment and 83 in the *Random* treatment.[11] Based on these observations, Figure 3.2 depicts the fraction of liars, i.e., players A who report the high state when actually having seen the low state, in both treatments. We observe that subjects in *Choice* lie slightly more than in *Random*. However, the treatment effect of 4.8 percentage points is statistically insignificant, which suggests that subjects did not react to our treatment ($p = 0.506$, χ^2-test, two-sided).[12]

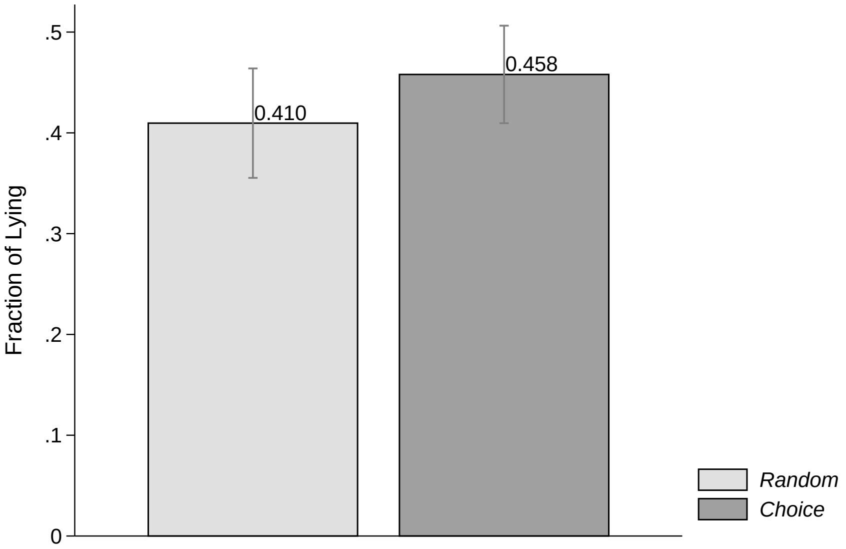
Table 3.1 reports marginal effects of probit regressions with lying as dependent variable.[13] Column (1) is the parametric equivalent of Figure 3.2 as it only includes the treatment dummy as explanatory variable. Supporting our non-parametric result, we

[10] Downward lying, i.e., seeing the high state but reporting the low state, did not occur. This is in line with the literature. We are not aware of any study with a unilateral lying decision — thus, excluding sender-receiver games, where strategic lying is possible (e.g., Gneezy, 2005; Sutter, 2009) — and direct lying observability that provides evidence for downward lying (see Kocher et al., 2017; Gneezy et al., 2018). Concerning studies that infer lying behavior from an underlying probability distribution, such as in Fischbacher and Föllmi-Heusi (2013), we are only aware of one study that reports downward lying (Utikal and Fischbacher, 2013). However, this finding can be attributed to the specific subject pool of this experiment, namely nuns.

[11] In total, 219 of our participants were assigned to role A and 44 to role B.

[12] We also conduct a second power calculation based on the observed treatment difference and a two-sided χ^2-test of proportions. Given a power of 80%, this calculation reveals that we would require 3136 players A who are able to lie to player B in order to obtain a treatment effect which is significant at the 5% level.

[13] Coefficients of the probit regression can be found in Table 3.2 in the Appendix.



Notes: The figure shows the share of liars in the *Random* and the *Choice* treatment. Error bars indicate standard errors of the mean.

Figure 3.2: Share of liars

find a positive but insignificant effect of having a choice on dishonesty. In a meta-study which investigates 72 experimental studies on dishonesty, Abeler et al. (2016) find gender to be the only socio-demographic background variable that affects dishonesty. They report a positive and significant effect of being male on lying.[14] Therefore, we include a gender dummy in column (2) and find that also in our experiment, males are significantly more likely to lie.[15] Being female decreases the likelihood of lying by 19 percentage points ($p = 0.004$). Interestingly, including an interaction term between our treatment and gender in column (3) reveals that there is a highly significant and sizeable treatment effect for males. They are 24.7 percentage points more likely to lie at the expense of player B when player B chooses to interact with player A as compared to when player B is randomly allocated to the interaction with player A ($p = 0.006$). The sum of the treatment dummy "Choice" and the interaction term "Choice × Female" yields the treatment effect for females, which is insignificant

[14] Comparing 63 economic and psychological experiments on dishonesty, Rosenbaum et al. (2014) also find that lying is more prevalent among males than females.

[15] Again, we emphasize that this step of the analysis is exploratory.

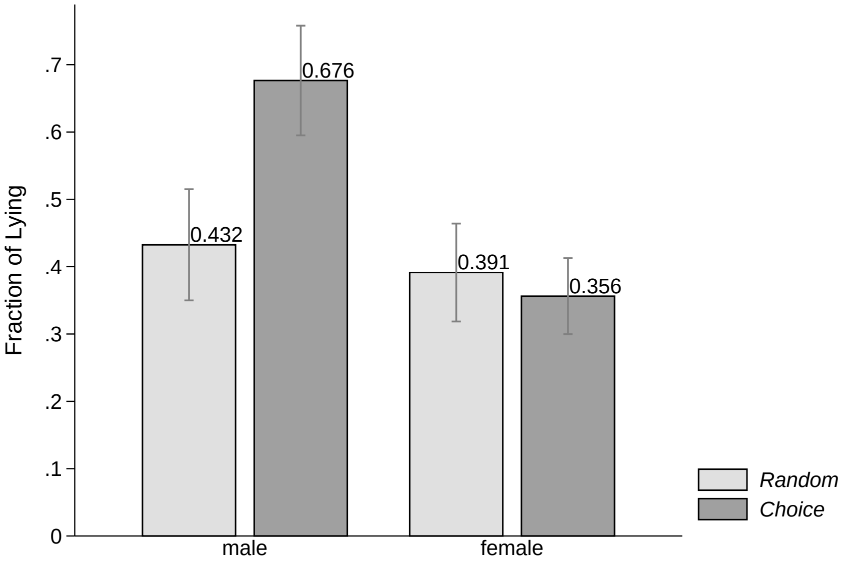
Table 3.1: Lying behavior. Probit marginal effects

Dependent variable	Lying			
	(1)	(2)	(3)	(4)
Choice	0.048	0.074	0.247***	0.372***
	(0.049)	(0.057)	(0.091)	(0.124)
Female		−0.190***	−0.042	0.057
		(0.066)	(0.061)	(0.090)
Choice × Female			−0.284**	−0.423**
			(0.119)	(0.190)
Economics/Business				0.233***
				(0.088)
Trust				−0.233**
				(0.091)
Age				−0.007
				(0.009)
Math Grade				−0.041
				(0.040)
#Experiments				0.008
				(0.008)
Observations	190	190	190	190
Pseudo-R^2	0.002	0.026	0.040	0.142

Notes: This table presents marginal effects at means from probit regressions with lying as independent variable. Column (4) includes a dummy for whether being an economics or business student, a dummy for whether one trusts people in general, age, "math grade", which is the last grade in mathematics during high school ranging from 1 (very good) to 6 (insufficient), as well as number of experiments participated in 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.

($p = 0.621$, *F*-test).[16] In addition, the female dummy shows that gender does not play a role in the *Random* treatment ($p = 0.493$). These results are robust to including additional variables from the questionnaire in column (4). By doing so, we find that economics and business students are 23.3 percentage more likely to lie than students of other fields of study ($p = 0.009$). Furthermore, subjects who find other people

[16] One has to be cautious when using interaction terms in probit models since in this case, the marginal effect of the interaction term is not the same as the interaction effect (Ai and Norton, 2003; Greene, 2010). We therefore also manually calculate the interaction effect by taking the difference in differences of the predicted values of lying separately for males and females in the *Random* and the *Choice* treatment. As the difference in differences for the models in column (3) and (4) of Table 3.1 is very similar to the marginal effect of the interaction term in the respective column, our estimates appear to be only marginally biased by using interaction terms in a probit model. Moreover, our results are robust to using a linear probability model. These results are reported in Table 3.3 in the Appendix.



Notes: The figure shows the share of liars separately for males and females in the *Random* and the *Choice* treatment. Error bars indicate standard errors of the mean.

Figure 3.3: Share of liars separately for males and females

generally trustworthy are 23.3 percentage points less likely to engage in dishonest behavior themselves ($p = 0.011$).[17]

As our regression analysis indicates that the treatment effect depends on whether a player A is male, we replicate Figure 3.2 for males and females separately in Figure 3.3. We observe a different reaction to the treatment depending on gender, which is in line with our parametric results. While 43.2% of the males in the *Random* treatment lie, this fraction significantly increases to 67.6% in the *Choice* treatment ($p = 0.039$, χ^2-test, two-sided). This amounts to a 56% increase in lying behavior. In contrast, the fraction of liars does not differ across treatments among females ($p = 0.699$, χ^2-test, two-sided). In addition, there is no significant gender difference in lying behavior in the *Random* treatment ($p = 0.705$, χ^2-test, two-sided). These results support the findings of our parametric analysis.

[17] Replicating column (4) with either interacting "Trust" or "Economics/Business" with "Choice" instead of including the interaction term "Choice × Female" yields insignificant interaction effects.

3.4 Conclusion

People who engage in dishonest behavior can take advantage of their private information. At the same time, however, some individuals want to avoid a guilty conscience. Therefore, they might justify their behavior with excuses in order to minimize the moral costs of lying. In this chapter, we study whether choice can serve as such a justification. We find a positive but insignificant average effect of the possibility to self-select into a situation on the probability of being lied to.

Furthermore, our analysis reveals that females do not excuse dishonesty with choice. In contrast, males are significantly more likely to lie to another person when this person self-selected into an interaction with them. This increase is substantial since it amounts to more than 50% of the baseline proportion of lying. Moreover, the gender effect occurs in the *Choice* treatment only. The difference in dishonesty between males and females is insignificant in the *Random* treatment. This is in line with the mixed evidence on gender differences in lying behavior found in previous studies. In addition, we find that economics and business students as well as subjects who do not trust people in general are significantly more likely to lie.

There are several implications of our findings with regard to dishonesty of males. First, the possibility to choose is a core value of Western societies. However, our results suggest that this possibility to choose may come at the cost of inducing dishonest behavior. In order to prevent fraudulent behavior, organizations might want to frame interactions that entail the temptation to lie to another person in a way such that the other person had no choice.

Second, regulatory measures should aim to avoid any redundant decision making of the entity which should be protected. For instance, regulation in industries that face inherent information asymmetries often includes the following feature. By law, firms need to retrieve certain characteristics from each customer. Depending on these characteristics, a firm can offer a limited set of alternatives to the customer. As an example, banks in Germany need to ask each client (among other things) about risk preferences and investment horizon before offering them financial products. Our results suggest that alternatives which are rather similar and only exist to give customers the illusion of choice should be discouraged by regulation. In the previous example, the bank may not offer several investment funds per risk-investment horizon combination that pursue a similar strategy. In this way, it is harder for the bank employee to keep up a positive self-image when selling unfit products with high mark-ups, due to the reduced choice set of the client. This may lead the bank to restructure its products in a more customer-oriented way.

Finally, our finding that only males excuse dishonest behavior with choice raises the question whether this also holds true for other types of justification. It thus would be interesting to explore whether females hide behind other excuses for immoral behavior, such as not being pivotal (Falk and Szech, 2017) or that others also engage in immoral behavior (Falk and Szech, 2013). In addition, it remains an open question how our findings translate to non-student subject pools, and whether they depend on cultural determinants.

3.5 Appendix

3.5.1 Tables

Table 3.2: Lying behavior. Probit coefficients

Dependent variable	Lying			
	(1)	(2)	(3)	(4)
Choice	0.123	0.187	0.628***	0.948***
	(0.124)	(0.146)	(0.231)	(0.316)
Female		−0.483***	−0.106	0.146
		(0.168)	(0.154)	(0.230)
Choice × Female			−0.721**	−1.078**
			(0.299)	(0.486)
Economics/Business				0.593***
				(0.225)
Trust				−0.592**
				(0.234)
Age				−0.018
				(0.022)
Math Grade				−0.105
				(0.102)
#Experiments				0.020
				(0.020)
Constant	−0.228*	0.036	−0.170	0.214
	(0.121)	(0.153)	(0.123)	(0.505)
Observations	190	190	190	190
Pseudo-R^2	0.002	0.026	0.040	0.142

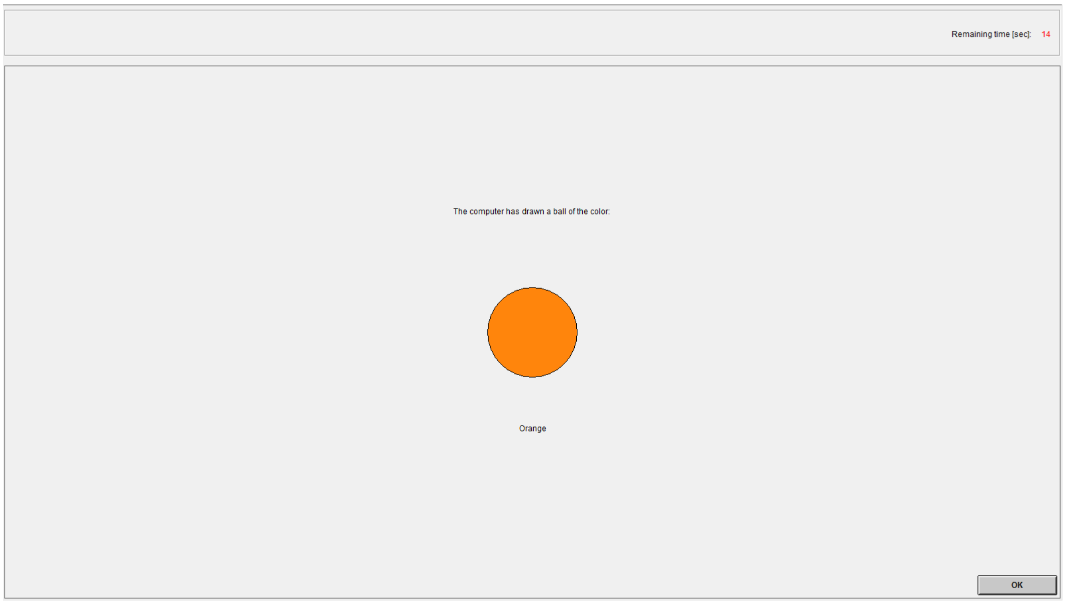
Notes: This table presents coefficients from probit regressions with lying as independent variable. Column (4) includes a dummy for whether being an economics or business student, a dummy for whether one trusts people in general, age, “math grade”, which is the last grade in mathematics during high school ranging from 1 (very good) to 6 (insufficient), as well as number of experiments participated in 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.

Table 3.3: Lying behavior. Linear probability model

Dependent variable	Lying			
	(1)	(2)	(3)	(4)
Choice	0.048	0.073	0.244**	0.328***
	(0.049)	(0.056)	(0.086)	(0.099)
Female		−0.189***	−0.041	0.050
		(0.065)	(0.060)	(0.081)
Choice × Female			−0.279**	−0.371**
			(0.111)	(0.150)
Economics/Business				0.214**
				(0.080)
Trust				−0.204**
				(0.076)
Age				−0.004
				(0.006)
Math Grade				−0.037
				(0.034)
#Experiments				0.007
				(0.007)
Constant	0.410***	0.515***	0.432***	0.524***
	(0.047)	(0.060)	(0.049)	(0.133)
Observations	190	190	190	190
R^2	0.002	0.036	0.054	0.182

Notes: This table presents coefficients from regressions of a linear probability model with lying as independent variable. Column (4) includes a dummy for whether being an economics or business student, a dummy for whether one trusts people in general, age, "math grade", which is the last grade in mathematics during high school ranging from 1 (very good) to 6 (insufficient), as well as number of experiments participated in 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.

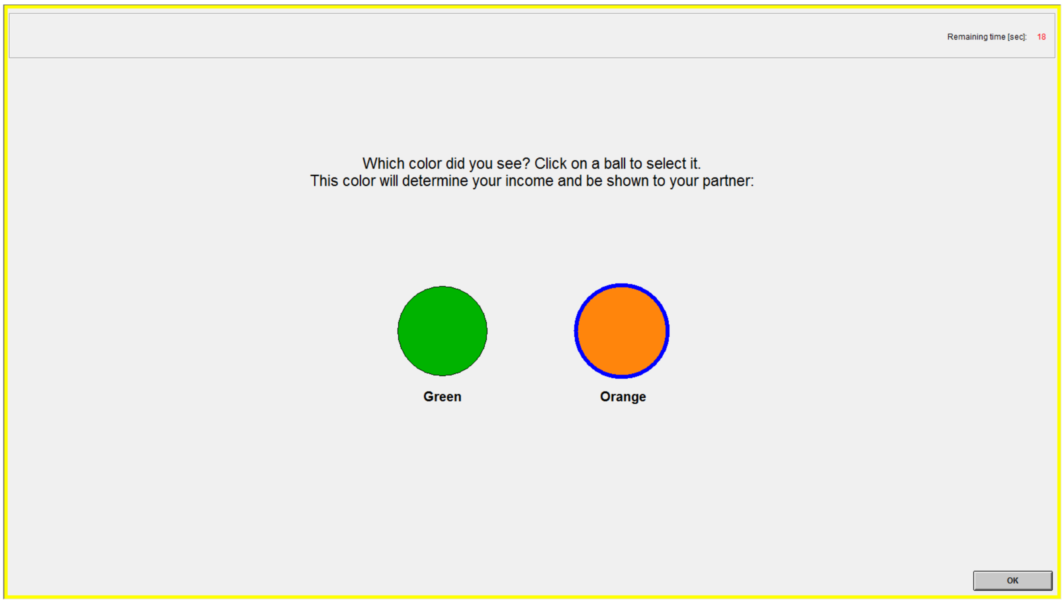
3.5.2 Screenshots



Remaining time [sec]: 14

The computer has drawn a ball of the color:

Figure 3.4: Screenshot of how the color is displayed to player A



Remaining time [sec]: 18

Which color did you see? Click on a ball to select it.
This color will determine your income and be shown to your partner:

Figure 3.5: Screenshot of player A's decision screen with orange being selected

3.5.3 Instructions

[In paper form][18]

Welcome to the experiment and thank you for your participation!

Please do not talk to other participants of the experiment from now on.

General information on the procedure

This experiment serves to investigate economic decision making behavior. You can earn money, which will be paid to you individually and in cash after the experiment has ended.

If you have any questions after reading these instructions or during the experiment, please raise your hand or press the red button on your keyboard. We will then come to you and answer your question in private.

During the experiment, you and the other participants will make decisions. Your own decisions as well as the decisions of other participants can determine your payoffs. These payoffs are determined according to the rules which are explained in the following.

Payment

At the end of the experiment, you will receive in cash the money that you have earned during the experiment. Additionally, you will receive 5 euro for showing up on time. Therefore, we will call every participant based on his seat number, i.e., none of the other participants gets to know your payment, and also you will not get to know the payments of other participants.

[18] The instructions were translated from German. The original version is available upon request.

Anonymity

Data from this experiment will be analyzed anonymously, i.e., we will never link your name to the data of the experiment. At the end of the experiment, you have to sign a receipt confirming that you received your payment. This receipt serves accounting purposes only.

Assignment of Roles

There are two different roles in this experiment: A and B. The role of each participant is determined randomly. Whether you are participant A or B will be communicated to you at a later point in time on your screen.

Matching of the Participants

After all participants were assigned to their role, the computer will randomly assign one or several participants A to one participant B.

In case you are participant A, the computer will randomly determine whether your decision or the decision of another participant A will be implemented.

Participants A whose decisions are not implemented receive 3 euros. Participants A whose decisions are implemented can determine their payment and possibly the payment of participant B.

Since each participant A is informed whether his decision is implemented only at the end of the experiment, it is optimal for participant A to decide as if their decision will in fact be implemented.

Your Task

Participant B's Decision

[*Random* treatment only]

At the beginning of the experiment, participant B is **randomly** assigned to either interacting with participant A or to receiving an alternative payment. If participant B is **not** assigned to the interaction, he receives an alternative payment of 1.05 euros. The probability of participant B **not** being assigned to the interaction is 5%.

If participant B is assigned to the interaction, his payment depends on participant A's decision in the following way:

[*Choice* treatment only]

At the beginning of the experiment, participant B decides **in favor of or against** an interaction with participant A. If participant B does **not** interact, he receives an alternative payment of €1.05.

If participant B decides **in favor of** the interaction, his payment depends on participant A's decision in the following way:

[both treatments]

Participant A's Decision

The computer randomly draws a ball from a urn which contains 90 balls of the color orange and 10 balls of the color green. Therefore, the probability that a ball of the color orange is drawn is 90%, whereas it is 10% for the color green.

The color of the drawn ball is shown to participant A but not to participant B. It is the task of participant A to remember this color and report it on the screen later in the experiment.

If participant A reports to have seen orange, participant A receives €1 and participant B €5. If participant A reports to have seen green, participant A receives €5 and participant B €1.

Reported color →	Orange	Green
Payment Participant A	1 Euro	5 Euro
Payment Participant B	5 Euro	1 Euro

[*Random* treatment only]

If participant B has **not** been assigned to the interaction with participant A, participant A determines **his** earnings with his decision in the same way. Participant B, however, always receives €1.05 in this case.

[*Choice* treatment only]

If participant B decides **against** interacting with participant A, participant A determines **his** earnings with his decision in the same way. Participant B, however, always receives €1.05 in this case.

[both treatments]

Reported color →	Orange	Green
Payment Participant A	1 Euro	5 Euro
Payment Participant B	1.05 Euro	1.05 Euro

Comprehension Questions

In case you are **participant A** and your decision is implemented:

What happens if you report to have seen green?

1. I receive €1.05.
2. I receive €5.
3. I receive €1.

What happens if you report to have seen orange and the participant B assigned to you has been assigned to interacting with you? [*Random* treatment]

What happens if you report to have seen orange and the participant B assigned to you decides to interact with you? [*Choice* treatment]

1. I receive €5 and participant B €1.
2. I receive €1 and participant B €5.
3. I receive €5 and participant B €1.05.

In case you are **participant B**:

What happens if you have been assigned to receiving the alternative payment? [*Random* treatment]

What happens if you decide against interacting with participant A? [*Choice* treatment]

1. I receive €1.05 and participant A receives €5.
2. I receive €1.05 and participant A receives €5 or €1 euro, depending on which color he reports.
3. I receive €5 and participant B €1.

What happens if you have been assigned to interacting with participant A? [*Random* treatment]

What happens if you decide in favor of interacting with participant A? [*Choice* treatment]

1. I receive €1.05.
2. I receive €5 euros if participant A reports orange.
3. I receive €5 euros if participant A reports green.

4 Blaming the Refugees? Experimental Evidence on Responsibility Attribution

Joint with Stefan Grimm

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

[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).

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

[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.

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 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 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",

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

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 4.7 in the Appendix shows an English version of the leaflet.

[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).

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 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 4.7.5. Refugees were invited to the experiment half an hour earlier than Germans to make sure they knew what to expect and to

[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).

[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 4.7.1 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.

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 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.

[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.

[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.

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 4.8 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 matched participant) and otherwise due to nature. If player B's performance was

[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.

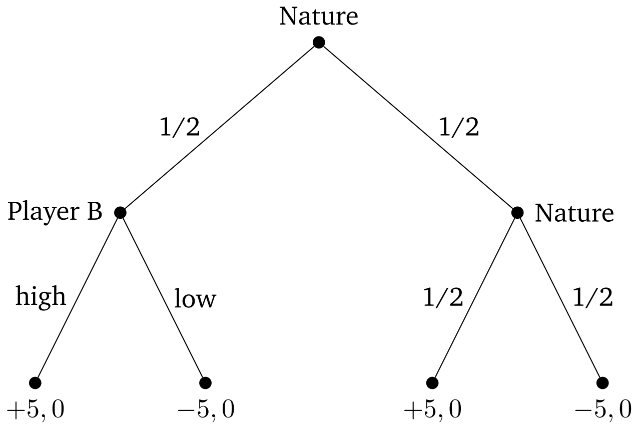


Figure 4.1: Income generating process

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

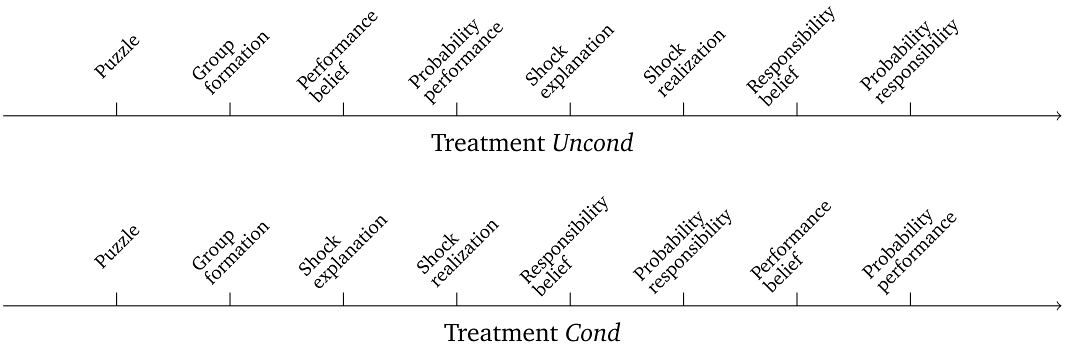


Figure 4.2: Timeline of the experiment

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 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

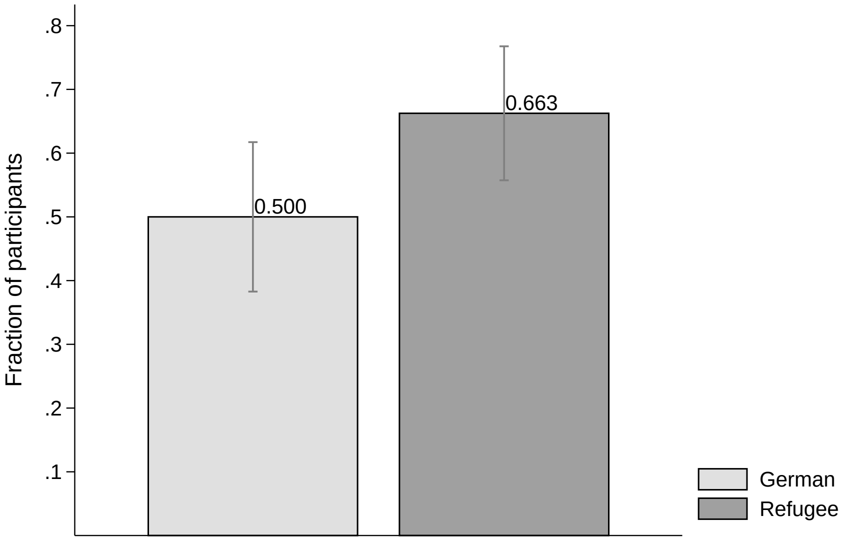
Our main results on the comparison of responsibility attribution by group assignment over all sessions combined are reported in Section 4.3.1. This abstracts from potential systematic differences between *Uncond* and *Cond*, which we analyze in 4.3.2 separately. Section 4.3.3 presents evidence for heterogeneity 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 a positive shock implies unfavorable attribution.[16] This simplification ignores potential asymmetries in

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

[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



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

Figure 4.3: *Favorable attribution* depending on group affiliation

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

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.

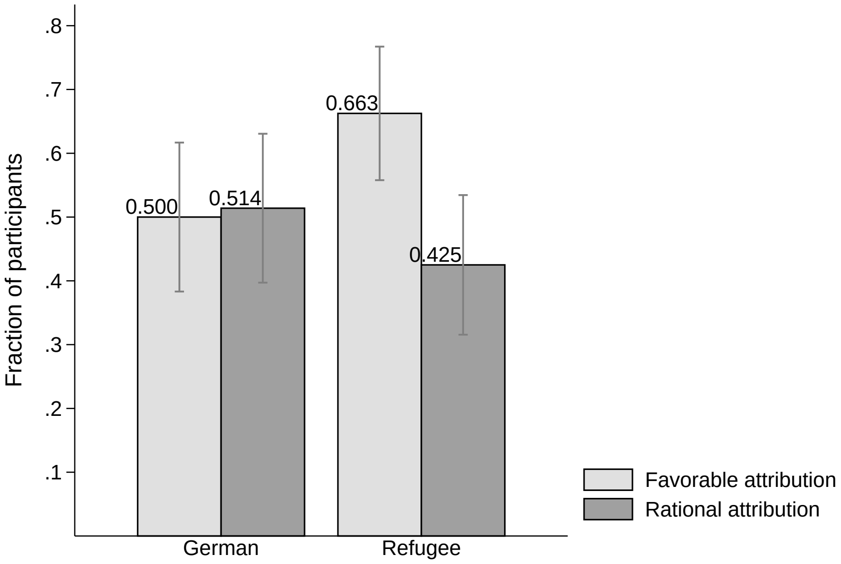
are better than Germans in solving puzzles. However, comparing performance beliefs reveals no significant difference. If anything, Germans expect refugees to perform slightly worse, which renders reverse discrimination even more pronounced. While 43% of Germans matched with a refugee expect the refugee 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]

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,

[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 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.



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

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 4.9 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

Table 4.1: Favorable responsibility attribution

Dependent variable	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.

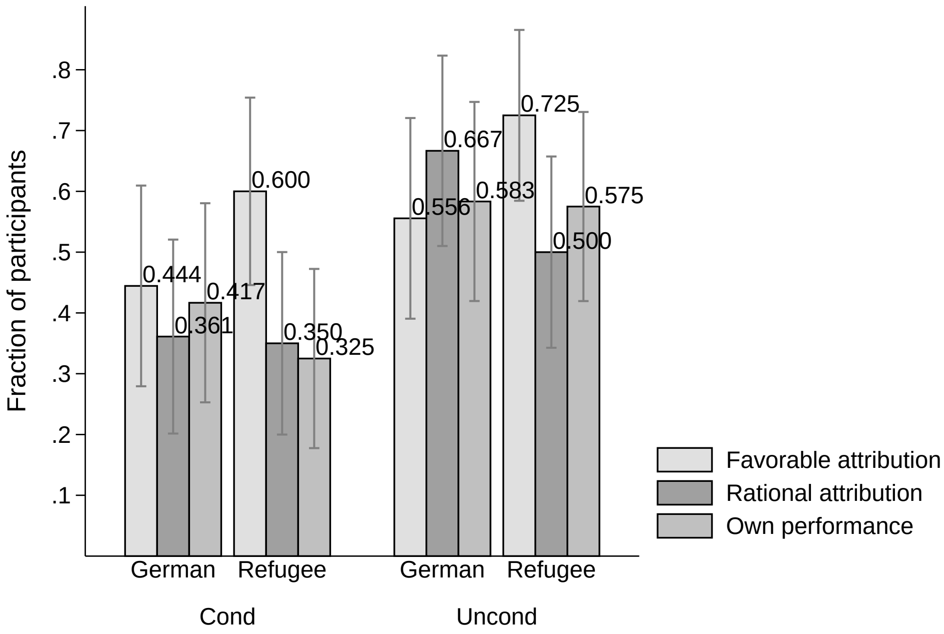
indicated by the binary variable *Refugee,* which is equal to one if a subject is matched with a refugee and zero otherwise. Column (2), equivalent to Figure 4.4, controls for performance beliefs with *belief high* as binary variable. *Belief high* is equal to one if a subject believes that the partner 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 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*), and actual



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*

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

[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 4.6 in the Appendix shows the sample balance.

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*, 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

[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.

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 4.7 in the Appendix reports results from regressions replicating Table 4.1 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

[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,

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.

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

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).

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 *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

[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.

[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 ($p = 0.143$).

Table 4.2: Favorable responsibility attribution depending on IAT

Dependent variable	Favorable attribution		
	Refugee (1)	*German* (2)	pooled (3)
IATneg	−0.272** (0.114)	0.089 (0.159)	0.084 (0.162)
# correct puzzles	0.077** (0.036)	0.104*** (0.024)	0.092*** (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.

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.

[27] Estimation of the difference in differences in predicted values can be found in Appendix 4.7.4.

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 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

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

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

[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

Dependent variable	Favorable 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) includes *performance switchpoint*. Column (2) reports average marginal effects of from a probit regression explaining *favorable attribution* with *performance switchpoint*. Subjects that clearly misunderstood the BDM elicitations 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.

of their binary counterparts in column (3), we again observe no significant reverse discrimination.

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.

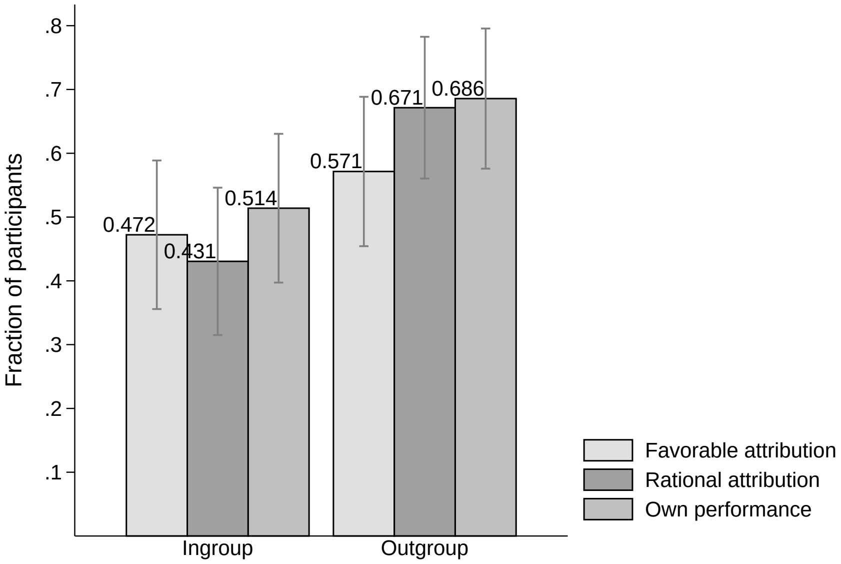
In summary, we get directionally very similar results with the non-binary belief elicitations. However, these results are weaker. Increased noise, indifference, systematic inconsistencies, and possibly increased opagueness of the normative prescription can help explaining this difference. While this 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 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,



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*

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,

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

Dependent Variable	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 binary variable *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.

given rational attribution (dark gray bars), subjects in *Outgroup* should attribute responsibility even more favorably and subjects in *Ingroup* even less 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 dis-

[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 4.8 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.

tributed 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.

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]

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

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

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 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.

[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).

[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.

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, 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

[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.

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 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.

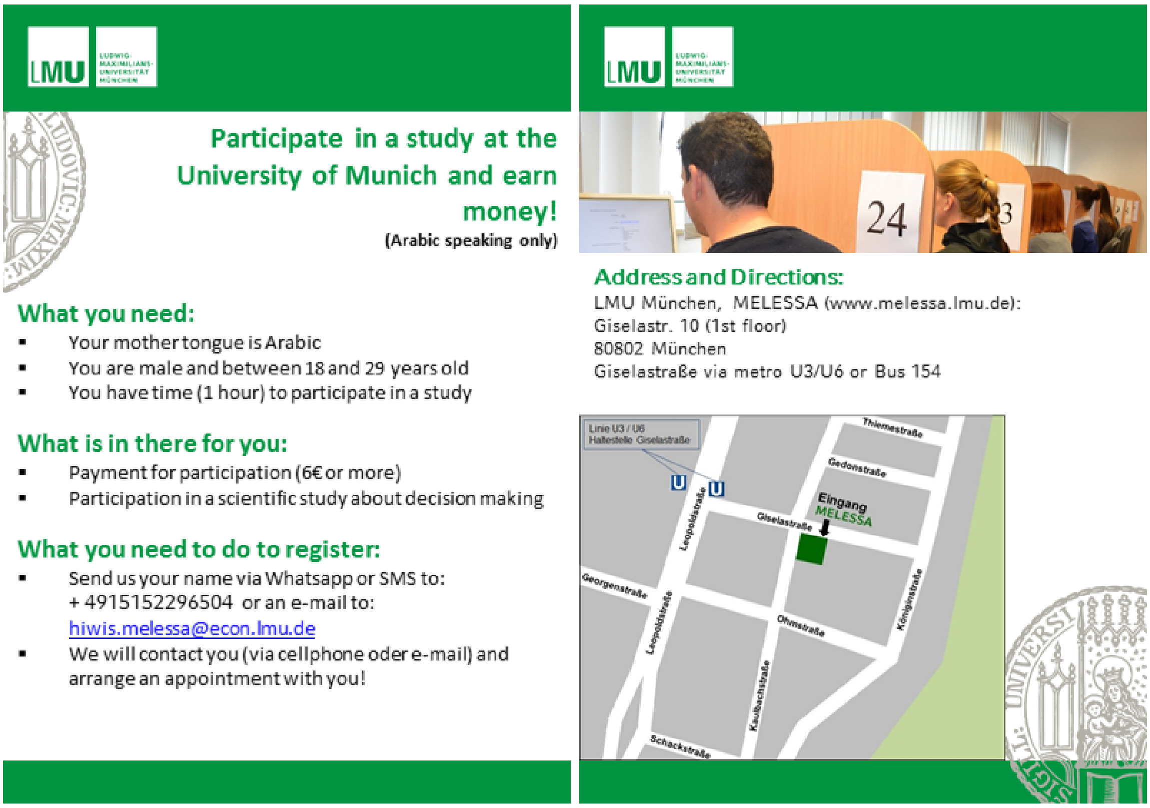
The evidence for reverse discrimination towards refugees together with our results on potential mechanisms provide fruitful avenues for future research. First, while we find strong evidence in the domain of responsibility attribution, our study cannot draw conclusions about whether our finding for the 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.

4.7 Appendix

4.7.1 Refugee Recruiting Details

Refugees were recruited by distributing the leaflet shown in Figure 4.7. 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.



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 4.7: Leaflet for recruiting refugees (translated from Arabic)

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.

4.7.2 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 4.8.

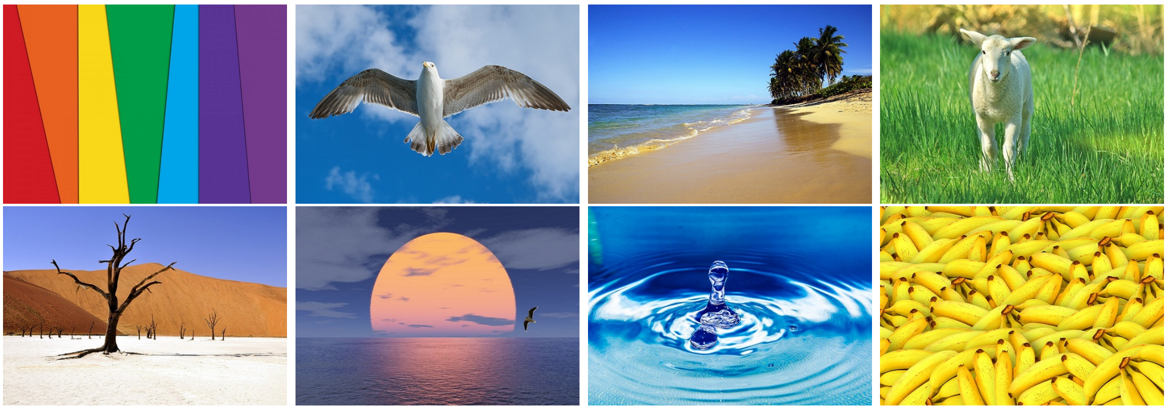
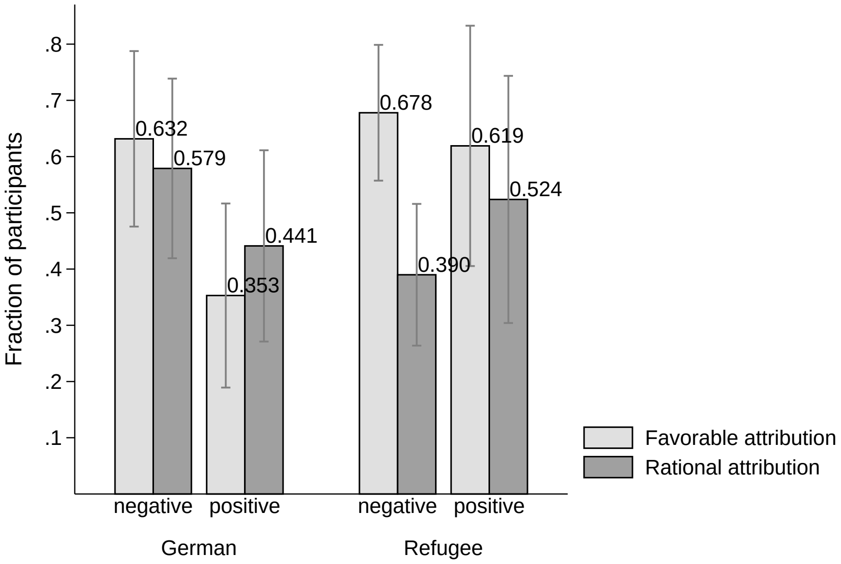


Figure 4.8: Puzzle motives for real effort task

4.7.3 Supplementary Results

4.7.3.1 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 4.9: *Favorable attribution* and *rational attribution* by shock direction

Figure 4.9 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.

4.7.3.2 Balance Table *Cond* vs. *Uncond*

Table 4.6: 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.

4.7.3.3 Regression Analysis Controlling for Own Performance

Table 4.7 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 4.7: Favorable responsibility attribution (controlling for own performance)

Dependent variable	Favorable attribution			
	(1)	(2)	(3)	(4)
Refugee	0.160*** (0.056)	0.181*** (0.055)	0.144*** (0.047)	0.139*** (0.044)
# correct puzzles		0.089*** (0.022)	0.086*** (0.023)	0.091*** (0.022)
Neg shock			0.159** (0.063)	0.148** (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 binary variable *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.

4.7.3.4 Balance Table for the *KleeKandinsky Experiment*

Table 4.8: Balance table *KleeKandinsky Experiment*

	Ingroup (1)	*Outgroup* (2)	(1) vs. (2) p-value
Own performance	0.514	0.686	0.037
Age	24.875	24.729	0.842
Semester	5.736	5.129	0.220
Number of experiments so far	10.542	11.700	0.401

Notes: Own performance indicates whether a subject solved four or more puzzles.

4.7.4 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*.

4.7.5 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≫

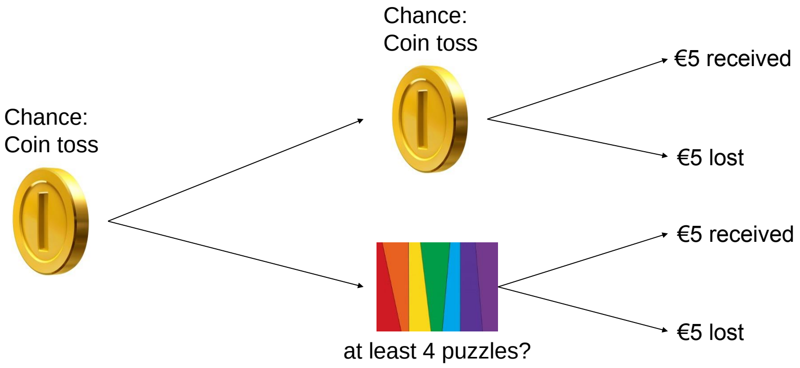
[new screen]

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%.

Bibliography

Abeler, J., Becker, A. and Falk, A. (2014). “Representative evidence on lying costs.” *Journal of Public Economics*, 113, 96–104.

Abeler, J., Nosenzo, D. and Raymond, C. (2016). “Preferences for truth-telling.” *IZA Discussion Paper*, No. 10188.

Ai, C. and Norton, E. C. (2003). “Interaction terms in logit and probit models.” *Economics Letters*, 80(1), 123–129.

Akerlof, G. A. and Kranton, R. E. (2000). “Economics and identity.” *Quarterly Journal of Economics*, 115(3), 715–753.

Alesina, A. and Angeletos, G.-M. (2005). “Fairness and redistribution.” *American Economic Review*, 95 (4), 960–980.

Alesina, A. and La Ferrara, E. (2005). “Preferences for redistribution in the land of opportunities.” *Journal of Public Economics*, 89(5), 897–931.

Alesina, A. and Giuliano, P. (2011). “Preferences for redistribution.” In *Handbook of Social Economics*, edited by J. Benhabib, A. Bisin and M. O. Jackson, 1, 93–132. Amsterdam: North-Holland.

Almås, I., A. W. Cappelen and Tungodden, B. (2016). “Cutthroat capitalism versus cuddly socialism: Are Americans more meritocratic and efficiency-seeking than

Scandinavians?” *Discussion Paper* 18/2016, NHH Dept. of Economics.

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.

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.

Becker, G. M., DeGroot, M. H. and Marschak, J. (1964). “Measuring utility by a single-response sequential method.” *Behavioral Science*, 9(3), 226–232.

Bortolotti, S., Soraperra, I., Sutter, M. and Zoller, C. (2017). “Too lucky to be true: Fairness views under the shadow of cheating.” *CESifo Working Paper Series*, No. 6563.

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.

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.

Camerer, C. F. and Fehr, E. (2004). “Measuring social norms and preferences using experimental games: A guide for social scientists.” In *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*, edited by J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr and H. Gintis, 97, 55–95. Oxford University Press.

Cappelen, A. W., Cappelen, C. and Tungodden, B. (2017). “False positives and false negatives in distributive choices.” *Working Paper*, mimeo.

Cappelen, A. W., Fest, S., Sørensen, E. Ø. and Tungodden, B. (2016). “Choice and personal responsibility: What is a morally relevant choice.” *Discussion Paper* 27/2014, NHH Dept. of Economics.

Cappelen, A. W., Konow, J., Sørensen, E. Ø. and Tungodden, B. (2013a). “Just luck: An experimental study of risk-taking and fairness.” *American Economic Review*, 103(4), 1398–1413.

Cappelen, A. W., Sørensen, E. Ø. and Tungodden, B. (2010). “Responsibility for what? Fairness and individual responsibility.” *European Economic Review*, 54(3), 429–441.

Cappelen, A. W., Sørensen, E. Ø. and Tungodden, B. (2013b). “When do we lie?” *Journal of Economic Behavior & Organization*, 93, 258–265.

Carneiro, P., Heckman, J. J. and Masterov, D. V. (2005). “Labor market discrimination and racial differences in premarket factors.” *Journal of Law and Economics*, 48(1), 1–39.

Charness, G. and Gneezy, U. (2012). “Strong evidence for gender differences in risk taking.” *Journal of Economic Behavior & Organization*, 83(1), 50–58.

Chen, Y. and Li, S. X. (2009). “Group identity and social preferences.” *American Economic Review*, 99(1), 431–457.

Childs, J. (2012). “Gender differences in lying.” *Economics Letters*, 114(2), 147–149.

Cohn, A., Fehr, E. and Maréchal, M. A. (2014). “Business culture and dishonesty in the banking industry.” *Nature*, 516(7529), 86–89.

Cohn, A., Maréchal, M. A. and Noll, T. (2015). "Bad boys: How criminal identity salience affects rule violation." *Review of Economic Studies*, 82(4), 1289–1308.

Conrads, J., Irlenbusch, B., Rilke, R. M. and Walkowitz, G. (2013). "Lying and team incentives." *Journal of Economic Psychology*, 34, 1–7.

Conrads, J. and Lotz, S. (2015). "The effect of communication channels on dishonest behavior." *Journal of Behavioral and Experimental Economics*, 58, 88–93.

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.

Dreber, A. and Johannesson, M. (2008). "Gender differences in deception." *Economics Letters*, 99(1), 197–199.

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.

Dworkin, R. (1981a). "What is equality? Part 1: Equality of welfare." *Philosophy and Public Affairs*, 10(3), 185–246.

Dworkin, R. (1981b). "What is equality? Part 2: Equality of resources." *Philosophy and Public Affairs*, 10(4), 283–345.

Eckel, C. C. and Grossman, P. J. (2008). "Forecasting risk attitudes: An experimental study using actual and forecast gamble choices." *Journal of Economic Behavior & Organization*, 68(1), 1–17.

Erat, S. and Gneezy, U. (2012). "White lies." *Management Science*, 58(4), 723–733.

Falk, A., Fehr, A. and Fischbacher, U. (2008). “Testing theories of fairness — Intentions matter.” *Games and Economic Behavior*, 62(1), 287–303.

Falk, A. and Heckman, J. J. (2009). “Lab experiments are a major source of knowledge in the social sciences.” *Science*, 326(5952), 535–538.

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

Falk, A. and Szech, N. (2017). “Diffusion of being pivotal and immoral outcomes.” *Working Paper Series in Economics, Karlsruher Institut für Technologie (KIT)*, No. 111.

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

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

Fehr, E., Naef, M. and Schmidt, K. M. (2006). “Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Comment.” *American Economic Review*, 96(5), 1912–1917.

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.

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

Fischbacher, U. and Föllmi-Heusi, F. (2013). “Lies in disguise — an experimental study on cheating.” *Journal of the European Economic Association*, 11(3), 525–547.

Fong, C. (2001). “Social preferences, self-interest, and the demand for redistribution.” *Journal of Public Economics*, 82(2), 225–246.

Frohlich, N., Oppenheimer, J., Bond, P. and Boschman, I. (1984). "Beyond economic man: Altruism, egalitarianism, and difference maximizing." *Journal of Conflict Resolution*, 28(1), 3–24.

Gneezy, U. (2005). "Deception: The role of consequences." *American Economic Review*, 95(1), 384–394.

Gneezy, U., Kajackaite, A. and Sobel, J. (2018). "Lying aversion and the size of the lie." *American Economic Review* (forthcoming).

Gneezy, U., Niederle, M. and Rustichini, A. (2003). "Performance in competitive environments: Gender differences." *Quarterly Journal of Economics*, 118(3), 1049–1074.

Gneezy, U., Saccardo, S., Serra-Garcia, M. and van Veldhuizen, R. (2016). "Motivated self-deception, identity and unethical behavior." *Working Paper*, mimeo.

Goldin, C. and Rouse, C. (2000). "Orchestrating impartiality: The impact of "blind" auditions on female musicians." *American Economic Review*, 90(4), 715–741.

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.

Grosch, K. and Rau, H. (2017). "Gender differences in honesty: The role of social value orientation." *Journal of Economic Psychology*, 62, 258–267.

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.

Gylfason, H. F., Arnardottir, A. A. and Kristinsson, K. (2013). “More on gender differences in lying.” *Economics Letters*, 119(1), 94–96.

Heckman, J. J. (1998). “Detecting discrimination.” *Journal of Economic Perspectives*, 12(2), 101–116.

Herbst, L., Konrad, K. A. and Morath, F. (2015). “Endogenous group formation in experimental contests.” *European Economic Review*, 74, 163–189.

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.

Houser, D., List, J. A., Piovesan, M., Samek, A. and Winter, J. (2016). “Dishonesty: From parents to children.” *European Economic Review*, 82, 242–254.

Jiang, Ting (2013). “Cheating in mind games: The subtlety of rules matters.” *Journal of Economic Behavior & Organization*, 93, 328–336.

Jones, E. E. and Harris, V. A. (1967). “The attribution of attitudes.” *Journal of Experimental Social Psychology*, 3(1), 1–24.

Kahneman, D. and Tversky, A. (1979). “Prospect theory: An analysis of decision under risk.” *Econometrica*, 47(2), 263–292.

Kajackaite, A. and Gneezy, U. (2017). “Incentives and cheating.” *Games and Economic Behavior*, 102, 433–444.

Kirchler, E., Maciejovsky, B. and Schneider, F. (2003). “Everyday representations of tax avoidance, tax evasion, and tax flight: Do legal differences matter?” *Journal of Economic Psychology*, 24(4), 535–553.

Kocher, M. G., Schudy, S. and Spantig, L. (2017). “I Lie? We Lie! Why? Experimental Evidence on a Dishonesty Shift in Groups.” *Management Science* (forthcoming).

Konow, J. (2000). “Fair shares: Accountability and cognitive dissonance in allocation decisions.” *American Economic Review*, 90(4), 1072–1091.

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.

Kuziemko, I., M. I. Norton, E. Saez and Stantcheva, S. (2015). “How elastic are preferences for redistribution? Evidence from randomized survey experiments.” *American Economic Review*, 105(4), 1478–1508.

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.

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.

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.

Möllerström, J., Reme, B.-A. and Sørensen, E. Ø. (2015). "Luck, choice and responsibility — an experimental study of fairness views." *Journal of Public Economics*, 131, 33–40.

Niederle, M. and Vesterlund, L. (2007). "Do women shy away from competition? Do men compete too much?" *Quarterly Journal of Economics*, 122(3), 1067–1101.

Ockenfels, A. and Werner, P. (2014). "Beliefs and ingroup favoritism." *Journal of Economic Behavior & Organization*, 108, 453–462.

Pettigrew, T. F. (1979). "The ultimate attribution error: Extending Allport's cognitive analysis of prejudice." *Personality and Social Psychology Bulletin*, 5(4), 461–476.

Pigors, M. and Rockenbach, B. (2016). "The competitive advantage of honesty." *European Economic Review*, 89, 407–424.

Piketty, T. (1995). "Social mobility and redistributive politics." *Quarterly Journal of Economics*, 110(3), 551–584.

Rabin, M. (1993). "Incorporating fairness into game theory and economics." *American Economic Review*, 83(5), 1281–1302.

Rosenbaum, S. M., Billinger, S. and Stieglitz, N. (2014). "Let's be honest: A review of experimental evidence of honesty and truth-telling." *Journal of Economic Psychology*, 45, 181–196.

Ross, L. (1977). "The intuitive psychologist and his shortcomings: Distortions in the attribution process." *Advances in Experimental Social Psychology*, 10, 173–220.

Shalvi, S. and De Dreu, C. K. (2014). "Oxytocin promotes group-serving dishonesty." *Proceedings of the National Academy of Sciences*, 111(15), 5503–5507.

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*.

Sutter, M. (2009). “Deception through telling the truth?! Experimental evidence from individuals and teams.” *Economic Journal*, 119(534), 47–60.

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.

Utikal, V. and Fischbacher, U. (2013). “Disadvantageous lies in individual decisions.” *Journal of Economic Behavior & Organization*, 85, 108–111.

Van Lange, P. A., Bekkers, R., Chirumbolo, A. and Leone, L. (2012). “Are conservatives less likely to be prosocial than liberals? From games to ideology, political preferences and voting.” *European Journal of Personality*, 26(5), 461–473.

Zizzo, D. J. (2010). “Experimenter demand effects in economic experiments.” *Experimental Economics*, 13(1), 75–98.

Zucman, G. (2014). “Taxing across borders: Tracking personal wealth and corporate profits.” *Journal of Economic Perspectives*, 28(4), 121–148.