Essays on prosocial and competitive behavior

Inauguraldissertation zur Erlangung des Doktorgrades
der Wirtschafts- und Sozialwissenschaftlichen Fakultät
der Universität zu Köln

2018

vorgelegt von
Uta Katharina Schier, M.Sc.
aus Berlin-Steglitz

Referent: Prof. Dr. Axel Ockenfels

Korreferenten: Prof. Dr. Wilhelm Hofmann, Prof. Dr. Felix Kölle

Tag der Promotion: 9.10.2018

ACKNOWLEDGMENTS

First, I would like to express my sincere gratitude to my advisor Axel Ockenfels for his continuous trust and support, for the chance to work in an excellent academic environment with incredible colleagues, and for his unwavering encouragement to develop and discuss my research topics!

My sincere thanks also go to Wilhelm Hofmann for his support and his insightful comments, as well as to the other members of my thesis committee, Felix Kölle and Oliver Gürtler.

I am also grateful for colleagues who encouraged me in different ways which made work much fun. Among others, these are Suparee Boonmanunt, Kevin Breuer, Lea Cassar, Katharina Diel, Christoph Feldhaus, Agne Kajackaite, Kiryl Khalmetski, Susanne Ludewig-Greiner, Andreas Pollak, Margit Schmidt, Marcin Waligora, and Peter Werner.

I would also like to thank our student assistants for their valuable support in the experiments, notably Tobias Dahint, Max Grossmann, Lea Mohnen, Yero Ndiaye, and Wolfram Uerlich.

I dedicate this thesis to my family: my parents, my siblings, my grandfather, and in particular Ali – I cherish your support, love and cheerfulness!

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1:
Introduction

For most economists, the *Homo Oeconomicus* who is rational, selfish and maximizes his expected utility is rather a useful approximation than a precise description of actual human behavior (Roth, 1996). In the recent years, behavioral and experimental economics brought forward important social utility models that incorporate selfish as well as social preferences (Fehr and Schmidt, 1999: Bolton and Ockenfels, 2000; Charness and Rabin, 2002). Social preferences result in behavior like fairness, altruism or trust that cannot be explained by selfish preferences. These models sparked a range of economic studies testing conditions under which people display social preferences. In this thesis, I contribute to this literature by investigating what drives prosocial and competitive behavior in different contexts. The thesis consists of four studies that are all based on evidence from online or laboratory experiments, in which I investigate how people trade-off social and competitive preferences compared to selfish utility-maximizing preferences. For this, I rely on economic models combined with insights and methods from psychology like moral values scales, framing and ego-depletion. The first three studies focus on prosocial behavior such as fairness, trust and cooperation that is driven by social preferences.

In specific, Chapter 2 is titled "*Moral values and increasing stakes in a dictator game*" and is published in the Journal of Economic Psychology. It is joint work with Wilhelm Hofmann and Axel Ockenfels, and all authors contributed equally to the project.[1] Chapter 2 focuses on how moral values and increasing financial incentives influence fairness behavior, and how they

[1] The research idea was developed by Wilhelm Hofmann and Axel Ockenfels. Statistical analyses were carried out by Wilhelm Hofmann and Uta Schier. Uta Schier wrote the first draft. Wilhelm Hofmann and Axel Ockenfels gave feedback on the draft.

both interact. This is tested using a so-called dictator game (Forsythe et al., 1994), in which a randomly chosen dictator is endowed with $10 and has to decide what amount he/she is willing to share with an anonymous co-player. In a typical laboratory experiment, only 36% of dictators display selfish preferences by keeping all of their money, whereas the rest displays social preferences by sharing on average 28% of their endowment (Engel, 2011).

To understand what drives prosocial behavior, we compare what people state in a hypothetical moral fairness values survey with actual fairness behavior in a dictator game with either high ($500) or low ($10) stakes, conducted by researchers from Duke University. We find that people with higher moral fairness values behave more prosocially and that prosocial behavior decreases with higher stake size. Most importantly, we find that people with high moral fairness values fail to live up to their high fairness standards when the stake size increases. The finding that many participants systematically underestimate the impact of stake size on their fairness behavior cannot be explained by standard economic social utility models nor by standard models that consider moral identity (e.g. Bolton and Ockenfels, 2000; Benabou and Tirole, 2011). However, psychological temptation theories can complement the economic social utility models by highlighting that morality should be understood as a relative concept that people might strive for but struggle to achieve in face of temptations.

In Chapter 3, we further investigate the role of self-control. The chapter is titled "Se*lf-control and social comparison information: does ego-depletion affect ingroup-outgroup discrimination?*" and is joint work with Katharina Diel, Wilhelm Hofmann and Axel Ockenfels. All authors contributed equally to the project.[2] In this chapter, we contribute to the literature seeking to understand whether prosocial behavior comes naturally to humans or whether it requires cognitive control. Several studies examined the question whether selfish behavior or generosity is the more intuitive response in the dictator game using ego-depletion but the

[2] The research idea was generated by Wilhelm Hofmann and Axel Ockenfels. Uta Schier designed and conducted the experiment, carried out the statistical analyses, and wrote the first draft. Wilhelm Hofmann, Axel Ockenfels and Katharina Diel gave feedback on the draft.

evidence is rather mixed (e.g. Rand et al., 2012; Halali, Bereby-Meyer and Ockenfels, 2013). Ego-depletion is a state in which self-control ressources are temporarily exhausted (Hagger et al., 2010). We hypothesize that ego-depletion, or a shortage of self-control resources, will make people rely more on social comparison information. To test this, we build on previous literature showing that information on group membership can affect behavior in a dictator game, by people being more generous towards a member of their own group than towards a member of a different group (Chen and Li, 2009). In a laboratory experiment, we therefore test the hypothesis that this ingroup-outgroup discrimination is more pronounced when people are ego-depleted. However, we find no evidence that people give more towards members of their ingroup, neither with nor without depletion.

In Chapter 4, we study a different aspect of prosocial behavior, namely trust and cooperation. Chapter 4 is titled "*Games as frames*" and is joint work with Axel Ockenfels. All authors contributed equally to the project.[3] We examine whether beliefs and behavior change with a trust or distrust mindset. We hypothesize that economic games, which are supposed to provide a "neutral" platform to study social behavior, can induce a trust or distrust mindset that affects beliefs and preferences in a trust game and a stag-hunt game. Whereas incentives between players in a stag-hunt game are aligned, incentives in a trust game are misaligned between players. In two laboratory experiments, we analyze whether deliberating on trust games versus stag-hunt games without feedback changes trust and cooperation behavior in a subsequent game. We find that subjects who play trust games hold more pessimistic beliefs about other players' cooperation in a subsequent game than players who played stag-hunt games. We also find that deliberating on trust games compared to stag-hunt games affects behavior in a subsequent game. We conclude that economic games with aligned versus

[3] The research idea and design was developed by Axel Ockenfels and Uta Schier. Uta Schier planned and conducted the experiments and analysed the data. Axel Ockenfels and Uta Schier wrote the paper.

misaligned interests inevitably frame the decision context in ways that systematically and significantly affect preferences and beliefs.

The last chapter focuses on competitive behavior. Chapter 5 is titled "*Female and male role models and competitiveness*" and is single-authored. Based on a growing experimental literature that suggests that women seem to shy away from competitive environments (Niederle and Vesterlund, 2007), this project examines whether successful women can serve as role models for other women. This is still an open question in the economic literature. The goal of this paper is to close this gap by providing experimental evidence regarding the importance of role models in encouraging competitiveness. In particular, how observing a woman competing and succeeding (i.e. a female role model) affects women's and men's competitiveness and how this differs from observing a similar male role model. In addition, I examine how people update their beliefs in face of a female or male role model. In a laboratory experiment, I find that a female role model increases women's self-confidence and their competitiveness. In comparison, men's competitiveness is not affected by a male or a female role model, nor do women significantly respond to male role models. These findings are important, as the effect of role models on behavior, and not only on aspirations, has hardly been studied in economics.

Together, the results from all four experimental studies presented in this thesis contribute to our understanding and knowledge about when and why people display social and competitive preferences. In particular, (i) people with high moral fairness values often seem to underestimate the temptation of high financial stakes, (ii) we do not find an ingroup-outgroup discrimination, (iii) economic games *per se* can provide contextual clues and thereby impact beliefs and preferences of players, and (iv) a female role model can increase women's self-confidence and their competitiveness. Overall, this thesis emphasizes the relevance of context for behavior and demonstrates that human behavior often diverges from selfish preferences described by the *Homo Oeconomicus*. In addition, the thesis also shows that economic social utility models benefit from incorporating insights from psychology to give a more precise picture of actual economic and social behavior.

References

Benabou, R. J., and Tirole, J. (2011). Identity, morals and taboos: Beliefs as assets. *Quarterly Journal of Economics, 162*(2), 805–855.

Bolton, G. E., and Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review, 90*(1), 166-193.

Charness, G., and Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics, 117*, 817-869.

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

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics, 14*(4), 583–610.

Fehr, E., and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics, 114*(3), 817–868.

Forsythe, R., Horowitz, J. L., Savin, N. E., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior, 6*, 347–369.

Hagger, M. S., Wood, C., Stiff, C., and Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin, 136*(4), 495-525.

Halali, E., Bereby-Meyer, Y., and Ockenfels, A. (2013). Is it all about the self? The effects of self-control depletion on ultimatum game proposers. *Frontiers in Human Neuroscience, 7*, 240.

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.

Rand, D. G., Greene, J. D., and Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature, 489*, 427–430.

Roth, A. E. (1996). Individual rationality as a useful approximation: Comments on Tversky's "Rational theory and constructive choice". In K. Arrow, E. Colombatto, M. Perlman, and C. Schmidt (Eds.), *The rational foundations of economic behavior* (pp. 198-202). New York: Macmillan.

CHAPTER 2:
Moral values and increasing stakes in a dictator game

Joint with Wilhelm Hofmann and Axel Ockenfels *

Abstract

Using data from a large representative US sample (N=1,519), we compare hypothetical moral fairness values from the Moral Foundations Sacredness Scale with actual fairness behavior in an incentivized dictator game with either low or high stakes. We find that people with high moral fairness values fail to live up to their high fairness standards, when stake size increases. This violates principles from consistency theories according to which moral values are supposedly aligned with moral behavior, but is in line with temptation theories that question the absoluteness of morality values.

* Wilhelm Hofmann, Social Cognition Center Cologne, University of Cologne, Richard-Strauss-Str. 2, 50931 Cologne, Germany, e-mail: wilhelm.hofmann@uni-koeln.de. Uta K. Schier and Axel Ockenfels, Department of Economics, University of Cologne, Albert-Magnus-Platz, 50937 Cologne, Germany, e-mail: schier@wiso.uni-koeln.de, ockenfels@uni-koeln.de. We thank Duke University for access to the "Measuring Morality Survey" data and Stephen Vaisey for helpful comments. Financial support of the German Research Foundation (DFG) through the Research Unit "Design & Behavior: Economic Engineering of Firms and Markets" (FOR 1371), as well as through the Research Unit "Relativity in Social Cognition: Antecedents and Consequences of Comparative Thinking" (FOR 2150) is gratefully acknowledged.

2.1 Introduction

One topic that economist as well as psychologist are increasingly interested in is self-control in face of everyday temptations. Exercising self-control is not only relevant for health and nutrition decisions but also for retirement plans, education, and all investment decisions in which long-term preferences may be suppressed in favor of short-term gratifications (e.g., Duckworth & Seligman, 2005; DellaVigna, 2009; Meier & Sprenger, 2010; Moffitt et al., 2011; Sutter et al., 2013; Achtziger et al., 2016). Temptation in this context describes all situational cues that might prompt decision makers to temporarily disregard their long-term preferences. To understand successful resistance to temptations, the question arises whether people with strong moral convictions are better equipped to overcome temptations and whether moral values can be a self-control device. Do they steer behavior and can they have a buffering effect against temptation?

To study these questions, we analyze data from a probability-based web panel of 1,519 US citizens, designed to be representative of the US. These data come from the "Measuring Morality Survey", collected from Duke University in 2013, and have so far been published in methodological research, in the context of voting behavior or prosocial behavior and income inequality (Johnson et al., 2014; Miles; 2014; Miles & Vaisey, 2015; Vaisey & Miles, 2014, Côté et al., 2015; Padilla-Walker & Jensen, 2015; Piff et al., 2015). The survey includes a variant of the dictator game. Specifically, dictators were endowed with 10 "tickets" and were asked how many tickets they are willing to share with an anonymous co-player. Tickets could be submitted to an online raffle with either a prize of $10 or $500, depending on the treatment (the total number of distributed tickets was unknown to participants). At the same time, the survey also includes a variety of psychological morality tests. In our analyses, we will focus on moral concerns regarding fairness and thus on the *fairness/reciprocity* subscale of the Moral

Foundations Sacredness Scale, which we will refer to as "*moral fairness*" (Graham & Haidt, 2012).[4]

To analyze fairness behavior of moral people in face of temptation, we will examine the following three research questions: (*i*) Is self-reported moral fairness reflected in more prosocial behavior in a dictator game? (*ii*) Does temptation in terms of high financial stakes decrease giving in a dictator game? (*iii*) Can high moral values buffer against the effects of temptation in high stakes situations? We find that participants harboring strong moral fairness values are on average more willing to share their endowment with an unknown co-player. However, participants with high moral fairness values fail to control themselves in face of high temptation: An increase in stake size from \$10 to \$500 reduces sharing to a greater extent for participants with high moral fairness values than for subjects with low moral fairness values. In particular, we find that people that claim they would never behave unfairly, "*not for a million dollars*", fail to live up to this claim in an actual dictator game, when the stake size increases to \$500. This suggests that moral values are not absolute, and that people seem to underestimate the power of temptations.

2.2 Background and hypotheses

2.2.1 Previous evidence on stake size effects in dictator games

Behavioral economists have used dictator games for over two decades to study prosocial behavior (Forsythe et al., 1994). In the most common version of the game, the dictator receives an initial endowment of \$10 and is asked what amount he is willing to share with an anonymous co-player. Usually, laboratory experiments reveal that only about 30% of dictators keep all of their money, whereas the rest is willing to share their money with the recipient. These results

[4] Indeed, as it will become clear in the Results section, we argue that it is a moral concept of fairness that influences social behavior and drives our results.

seem quite robust across different contexts, including experiments that were conducted with children (Fehr, Bernhard & Rockenbach, 2008) or in small-scale societies (Henrich et al., 2001).

That said, it has been found that changing the stake size can affect generosity. Several studies testing an increase in stake size from $10 to $100 (Carpenter, Verhoogen & Burks, 2005) and $20 to $100 (List & Cherry, 2008) find a slight shift towards relatively less generous offers by dictators but cannot confirm these changes to be statistically significant. However, a recent meta-study comprising 131 experimental papers on dictator games shows that increasing the stake size reduces the dictator's generosity in relative terms (Engel, 2011).[5] In a study of purely hypothetical decisions, Novakova & Flegr (2013) find the same effect such that dictators tend to reduce their relative proposed share as stake size increases. Similarly, Blake and Rand (2010) found evidence that higher stakes decrease generosity in dictator games with children, when they play with low versus high valued stickers.

In addition, the discussion about stakes in dictator games also evolved around the question of how behavior changes between hypothetical and financially incentivized decisions, for the same cake size. For instance, Forsythe et al. (1994) were the first to report that for a given cake size dictators are less generous in an incentivized context than in a hypothetical context. Similar results were confirmed by Camerer and Hogarth (1999) who reviewed 74 experimental papers and found higher generosity of dictators in hypothetical games with no incentives. Moreover, Ockenfels (1998) found in his experiment that actually paying subjects does not affect average generosity in dictator games compared to hypothetical choices, but leads to 'less round' money amounts given to recipients (utilizing an objective measure of the roundness of the data based on the prominence structure of the decimal system). The higher occurrence of less prominent outcomes seems to suggest that monetary incentives trigger more complex decision processes. Camerer and Hogarth (1999), too, argue that higher incentives might not change behavior

[5] Evidence from the ultimatum game more robustly finds a decrease in generosity with higher stake size (e.g., Andersen, Ertac, Gneezy, Hofmann & List, 2011).

substantially on average but reduce variance in responses. Overall, the effect of stake size turns out to be rather mixed, although there is a tendency of reduced relative generosity with higher stake size. In our study, we extend this line of research and study how not only social behavior in the dictator game but also moral fairness values interact with stake size. The next subsection reviews two lines of literature that guide us when formulating our hypotheses.

2.2.2 Consistency theories vs. temptation theories

There are at least two competing families of models addressing how moral values can guide social behavior: consistency theories and temptation theories. Consistency theories propose that people align values with behavior to appear more consistent to others and to themselves. This line of reasoning includes, for instance, Heider's (1946) balance theory, Osgood and Tannenbaum's (1955) congruity theory, and Ajzen & Fishbein's (1980) theory of planned behavior. These theories suggest that people aim for consistency between moral values and actual behavior. For this, they either adapt their moral values to reflect their previous behavior, or behave in a way consistent with previously stated moral values. As such, consistency theories suggest a strong relationship between people's moral values and their allocation behavior in the dictator game. Such a relationship can take the form of a buffering effect such that being reminded of or deliberating and deciding about one's moral values may act as a pre-commitment device for subsequent tempting situations (Aquino et al., 2009). In sum, consistency theories would predict that people who have expressed strong moral fairness values "on paper" should be better protected against the lure of temptation when stake size increases than those who have expressed weak moral fairness concerns.

In contrast, temptation theories like the moral hypocrisy theory by Batson et al. (1997) propose that people often have moral standards which they aim to achieve but might fail to adhere to in the face of temptation. Rustichini and Villeval (2014) present experimental evidence on moral hypocrisy by measuring fairness judgments of allocation games, and a week later actual allocation decision and the same fairness judgments again. They find that people retrospectively justify a selfish decision by adjusting their previously stated perceived norm of fairness. This adjustment of moral judgment is larger in games without strategic consequences or more bargaining power, i.e. dictator games compared to ultimatum games (Rustichini & Villeval, 2014). Similarly, Lönnqvist, Irlenbusch and Walkowitz (2014) find laboratory evidence on moral hypocrisy in dictator games which they suggest is driven rather by a desire to appear moral rather than by self-deception (see also Ockenfels & Werner, 2014a,b). Likewise, there is a concordant literature in psychology on such biases of motivated reasoning (i.e., arriving at desired conclusions) attesting to Hume's famous dictum that reason may often be the "slave" of passion (e.g., Kunda, 2000; De Witt Huberts, Evers & De Ridder, 2014). These theories of motivated reasoning as well as contemporary models of temptation and self-control (e.g., Kotabe & Hofmann, 2015) assume that the outcome of conflicting motivational dilemmas such as between selfish and prosocial motives would depend, among other things, on the strength of the temptation: As temptation increases, so does the desire to arrive at a particular conclusion—the conclusion that indulging in temptation in this moment, would, somehow, be justified, earned, or constitute an important exception (Kivetz & Simonson, 2002; De Witt Huberts, Evers & De Ridder, 2014). Moreover, when being in a "cold" state, people may seriously underestimate how they would react in the "heat" of the moment, that is, when faced with strong temptation that may affect their senses and occupy their minds more strongly than predicted, contributing to inconsistencies between values/intentions and actual behavior (Kavanagh, Andrade & May, 2005; Loewenstein, 1996; Nordgren, van Harreveld & van der Pligt, 2009). In support, a study on everyday temptations finds that, in spite of people's stated self-control goals, people often fail to resist temptation, but especially so when desire strength for the temptation is strong (Hofmann, Baumeister, Förster & Vohs, 2012). In sum, temptation theories incorporate the idea that strong temptation may lead people to temporarily revise their intentions in a way that promotes indulgence, leading to the phenomenon of weakness of will

(Holton, 2009). In light of these converging lines of evidence and our own prior findings regarding desire strength (Hofmann et al., 2012), we expected to find evidence for stronger inconsistencies between stated moral fairness judgments and actual fairness behavior in a dictator game as temptation (i.e., stake size) increases.[6]

In this paper, we examine how stated moral fairness values correlate with actual fairness behavior, and how they interact with stake size. Based on the two families of theories reviewed above, we raise the question of whether people reporting strong moral fairness values behave more consistently with those values than people with weak moral fairness values when stake size increases.

2.2.3 Measurement of moral fairness: The Moral Foundations Sacredness Scale ("MFSS")

To understand how moral principles might motivate behavior, Graham and Haidt (2012) developed the Moral Foundations Sacredness Scale (MFSS). The MFSS offers a map of the moral space, with sacredness as protected moral value. These protected moral values are in a sense absolute and should not be traded off for money (Ritov & Baron, 1999). Graham and Haidt (2012) studied this absolute morality partly to understand how morality can potentially motivate idealistic violence. For that matter, morality is defined as the sets of values, virtues, norms, practices, identities, institutions, technologies, and psychological mechanisms that regulate selfishness and guide social life (Graham & Haidt, 2012). As such, morality and in particular absolute morality are concepts that deviate from the standard economic perspective of rational, self-interested agents (see Roth, 2007, for a broader economics approach).

In this context, the MFSS aims to measure five innate, psychological foundations on which culturally different moralities can evolve: harm/care, fairness/reciprocity, in-group/loyalty, authority/respect, and purity/sanctity. In addition, attitudes towards non-moral values are

[6] We will comment on the relationship to the social preferences literature in economics in our concluding section.

included, to control for unpleasant outcomes that are not relevant to morality. Thus, compared to other morality measures, Graham and Haidt (2012) suggest that the MFSS triggers an intuitive response, as well as activates deliberative reasoning.

Each foundation is measured with three to four items, depending on the short or long version of the scale, and are presented in random order without foundation labels. For instance, to quantify the subscale fairness/reciprocity (which we will use as our "moral fairness" measure), participants have to answer for which amount of money they would be willing to "*cheat in a game of cards played for money with some people you don't know well*" (Item 1) and "*throw out a box of ballots, during an election, to help your favored candidate win*" (Item 2) and "*sign a secret-but-binding pledge to only hire people of your race in your company*" (Item 3). Responses for each item are given on an 8-point scale, from "$0 (I'd do it for free)", then "$10" and increasing numbers by a factor of 10, up to "$1 million dollars or more", and the final option of "never for any amount of money". Additionally, participants could refuse an answer. Subscales for each foundation are measured by the average score across items. Typically, the internal consistencies are relatively low, around 0.64, indicating broad constructs in line with the idea that the items grasp moral attitudes for a range of topics, from nation, to family or club identity (Graham & Haidt, 2012).

2.2.4 Hypotheses

Our research question in this paper is to examine how stake size moderates the effect of strong moral judgments on fairness behavior. In particular, we are interested in how people with strong moral values behave in a dictator game, when stake size increases. For this, we predict, based on previous research, that increasing stake size in a dictator game reduces the willingness to share an endowment with others. Thus, our first hypothesis is:

> *Hypothesis 1 (stake size main effect): Higher stake size in a dictator game decreases giving (as measured by the number of tickets, out of ten, given).*

Second, we assume that people with strong moral foundations in fairness are generally more altruistic and will be more willing to share their endowment with an unknown co-player.

> *Hypothesis 2 (moral fairness concerns main effect): Giving in a dictator game increases along the moral fairness (MFSS fairness/reciprocity) dimension.*

Hypothesis 2 is crucial for showing that hypothetical statements made on the MFSS fairness/reciprocity dimension are not cheap-talk but, in fact, do signal a type of people behaving fairer in a dictator game. Hence, Hypothesis 2 is an essential precondition for Hypothesis 3.

Based on the previous two hypotheses, Hypothesis 3 is then concerned with the interaction effect between stake size and moral fairness values. We investigate whether stake size moderates how people with different moral foundations behave in a dictator game. Are people harboring strong moral fairness values better at resisting financial temptation? Contradictory to previously discussed consistency theories of moral behavior, we expect that strong moral foundations can lead to moral inconsistencies. Therefore, we propose that people with strong moral fairness values fail to control behavior, when temptation increases:

> *Hypothesis 3 (stake size × moral fairness concerns interaction): According to temptation theories, we expect that people reporting high moral fairness values behave relatively less fair in a high stakes dictator game, compared to people with lower moral fairness values. Alternatively, consistency theories would predict no difference in behavior across varying stake size scenarios between people with high moral fairness values and people with lower moral fairness values.*

2.3 Methods

We used data from a large representative sample of 1,519 US adults, collected for the "Measuring Morality Survey" by the Kenan Institute for Ethics of the Duke University in 2012. Subjects were randomly selected from a GfK panel and invited by e-mail to participate in an online survey.[7] The study comprised several psychological tests on morality, including a reduced version of the MFSS, and a dictator game. All tests were presented in a randomized order.

The dictator game was manipulated with either a low stakes condition ($10) or a high stakes condition ($500), to which participants were randomly assigned. 772 subjects were assigned to the $10 stake size condition, 747 participants to the $500 stake size condition. They all received an initial endowment of 10 tickets and were asked how many tickets they are willing to share with an anonymous co-player. All participants were playing the dictator role but were made believe that every other, alternate participant is the receiver. Specifically, odd subjects were told that even subjects are receivers, and vice versa. Although such deception would be widely unacceptable for economic experiments, we suggest it does not impair our results, since participants had no reason to doubt that every second player (odd or even) is a receiver and had no chance of communicating with each other.[8] We also emphasize that the data were neither collected by economists nor in an economics laboratory, and thus this study does not threat the reputation for a no-deception policy that economics laboratories are committed to.

Tickets could be used as entry tickets to an online raffle that subjects automatically entered at the end of the survey. Within their assigned treatment group, subjects had a 0.13% chance of winning the prize of either $10 or $500 but this was not known to subjects. 21 out of 1,519

[7] More detailed information on the study and access to the data: http://kenan.ethics.duke.edu/attitudes/resources/measuring-morality/.

[8] In fact, if at all, the participants in the survey were on average more social than what is found in a typical laboratory experiment (Engel, 2011). E.g., only 13% of all participants decided to give nothing, and 54% chose the equal split.

participants refused to decide and their answer was coded as missing value. On the MFSS, 13 out of 1,519 subjects refused to give any answer.

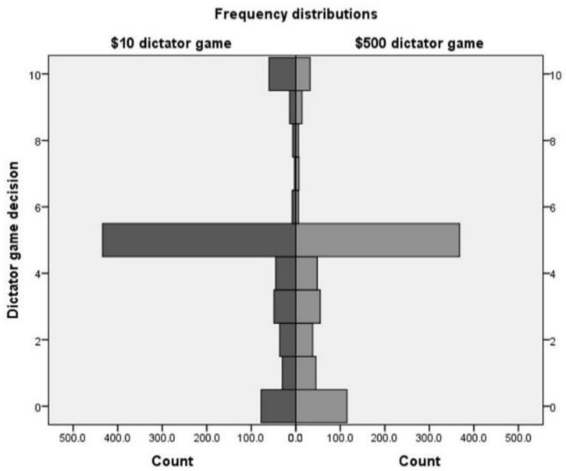
2.4 Results

In this section, we present results following our three main hypotheses. Then, we discuss our findings, and finally present additional robustness checks.

2.4.1 Stake size

Figure 1 shows the distributions of tickets given for both stake size treatments.[9] As predicted by *Hypothesis 1*, we find a strong effect of stake size on the dictator's decision. Overall, participants in the low stakes condition shared significantly more tickets (M = 4.52; SD = 2.46) than participants in the high stakes condition (M = 3.95; SD = 2.48; t = 4.46, df = 1,496, p < .001)[10]. That is, consistent with the literature, we find that an increase in stake size tends to reduce generosity in relative terms but to increase generosity in absolute terms.

Figure 1: Distribution of giving in the dictator game by stake size (N=1,498).



[9] See Table 2 in the Supplementary Material for a descriptive summary of the number of tickets given in the dictator game by reported moral fairness values and by stake size.
[10] Mann-Whitney U test: p < .001.

2.4.2 Moral fairness

As predicted by *Hypothesis 2*, results from the Morality Foundation Sacredness Scale (MFSS) demonstrate that people with high moral fairness values contribute more in a dictator game. In general, participants in the survey stated relatively strong moral fairness foundations, with a mean of 7.26 (SD = 1.26) on an 8pt Likert scale with a score of 8 representing the highest moral fairness (Cronbach's alpha = .71).

As for the moral fairness (fairness/reciprocity) dimension, we also tested a base regression model with *giving* in the dictator game as outcome variable and the MFSS *moral fairness* values as predictor variable. *Moral fairness* scores had a significant, positive influence on *giving* in the dictator game (B = 0.36, p < .001)[11]. This confirms that, on average, people stating high *moral fairness* values on the MFSS do behave more prosocially in a dictator game.

2.4.3 Interaction between stake size and moral fairness

To test Hypothesis 3 and the interaction effect between moral fairness concerns and stake size on fairness behavior, we conducted a moderated regression analysis using centered continuous predictors. In the overall model, we included stake size and moral fairness as predictors for the outcome variable giving, including the interaction effect between stake size and moral fairness following the procedure recommended by Aiken and West (1991). Stake size was modelled as a dummy variable for the low and high stakes conditions, with low stake size as baseline (i.e., coded 0). The remaining four moral foundations harm, loyalty, authority and purity were included as covariates, centered on their mean, as statistical control.

The outcome of this moderated regression model is depicted in Table 1. The estimated allocation decisions for the low and high stakes condition are plotted in Figure 2 across a

[11] Kruskal-Wallis test: χ^2 = 54.28, p < .001.

meaningful range of moral fairness scores ranging from 6 to 8 on the scale for presentational purposes (these values encompass 87% of the range of participants' moral fairness scores).[12]

The central interaction was significant (*B* = -0.24, *p* = .017), indicating that the impact of moral fairness concerns on more prosocial allocation decisions declines with increasing stake size. Accordingly, simple slope analyses showed that whereas there was a highly reliable effect of moral fairness concerns in the low stakes condition (*B* = 0.32, *p* < .001), the effect was not significantly different from zero in the high stakes condition (*B* = 0.08, *p* = .345, see Figure 2).

Table 1: Moderated regression with giving as outcome variable, moral fairness (MFSS) as predictor and stake size scenario as moderator.

Regression Summary

R-sq	F	df1	df2	p	n
0.04	8.87	7	1480	0.000	1488

	b	se	t	p
Constant	4.05	0.67	6.04	0.000
Stake size (baseline=low)	-0.55	0.13	-4.35	0.000
Moral fairness	0.32	0.09	3.78	0.000
Interaction (moral fairness*stake size)	-0.24	0.10	-2.39	0.017
Harm	-0.07	0.09	-0.75	0.453
Loyalty	0.01	0.07	0.17	0.866
Authority	0.10	0.05	1.94	0.052
Purity	0.03	0.05	0.55	0.585

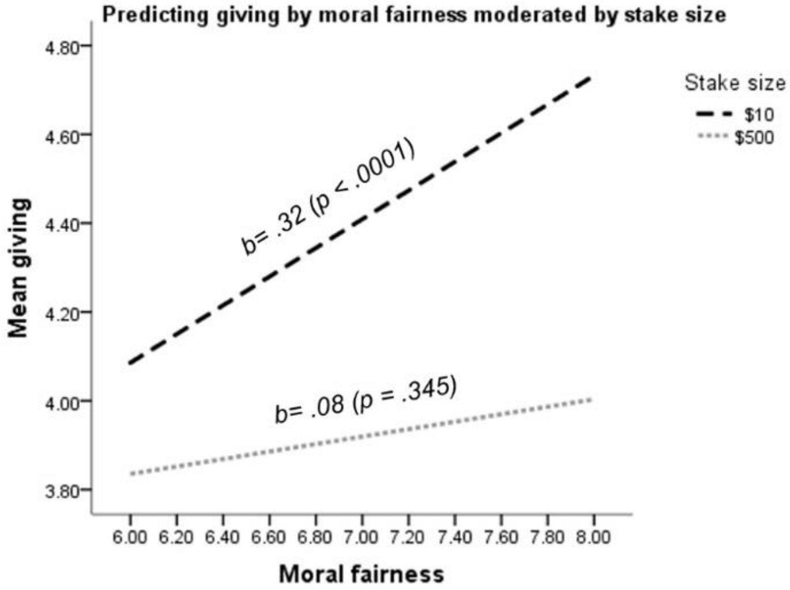
Conditional Effect of Focal Predictor at Values of the Moderator Variable

Stake size	b	se	t	p	LLCI(b)	ULCI(b)
High ($500)	.08	.09	.95	.345	-.09	.26
Low ($10)	.32	.09	3.78	.000	.16	.49

[12] This is a meaningful range because the distribution of moral fairness is left-skewed.

Further analysis showed that this significant interaction effect between moral fairness and stake size was independent of the covariates. Even in a model excluding the covariates *harm*, *authority*, *loyalty* and *purity*, the interaction effect remained significant (R^2 = .04, p = .022). Moreover, none of these variables interacted with stake size (all p > .128). Similarly, the interaction effect survived in a model including *non-morality* as covariate. *Non-morality* is the sixth subscale of the MFSS, suggested to measure extreme attitudes of participants that are independent of moral concerns (Graham & Haidt, 2012). However, even with *non-morality* as covariate in the model, the interaction effect remained significant.[13]

Figure 2: Interaction effect between moral fairness values (MFSS) and stake size on mean giving in a dictator game, for moral fairness score between 6 and 8 (87% of data).



2.4.4 Discussion

We found that the interaction effect is mostly driven by responses on Item 1 (MFSS1) of the moral fairness dimension. Item 1 asked whether respondents would cheat in a game of cards played with some people they do not know well. Running the moderated regression on

[13] This holds for a model with only *non-morality* as covariate (R^2 = .04; *non-morality p* = .67; *interaction p* = .03), as well as for a model including all other MFSS subscales and non-morality as covariate.

predicting the dictator's decision only with *item1*, *stake size* and their interaction effect resulted in a significant interaction effect (B = -0.176, p = .01). In contrast, models with Item 2 or Item 3 yielded interaction effects with stake size that were not significant at the conventional level of p = .05 (Item 2: B = -0.089, p = .324; Item 3: B = -0.164, p = .071). Item 2 described moral fairness behavior with respect to cheating in an election to help the favored candidate win, whereas Item 3 described moral fairness with respect to discriminating people of different race.

In addition, the interaction effect is driven by absolute moralists, defined as participants that reported the highest possible value on each of the three *moral fairness* items MFSS1, MFSS2, and MFSS3. This applied to 57% of subjects in the survey (867 out of 1,519). These subjects believed they would never behave unfairly, "not for a million dollars" but failed to behave consistently when stake size increases from a \$10 to a \$500 pie (mean tickets shared = 4.94 in low stakes scenario, mean = 4.13 in high stakes scenario). More specifically, the stake size effect observed in Hypothesis 1 also holds, when we only consider absolute moralists (t = 4.98, p < .001; Mann-Whitney U test: p < .001). In contrast, for relative moralists stake size did not significantly influence their decision in the dictator game (t = 0.88, p = .378; Mann-Whitney U test: p = .272). In other words, participants with high moral fairness scores were comparatively more sensitive to financial temptation. Similarly, in line with intuition, moral fairness scores drive generosity in the dictator game more strongly for absolute moralists than for relative moralists.[14] These findings strongly suggest that the interaction effect is mostly driven by absolute moralists. And indeed, considering only relative moralists for the moderated regression, the interaction effect disappears.[15] This implies that our results are driven by absolute moralists who fail to behave consistently when temptation increases. In line with temptation theories (but inconsistent with consistency theories), our results therefore suggest

[14] In addition, there is a significant relationship between absolute moralists vs. everybody else and the equal split in the dictator game (χ^2 = 37.84, p < .001; see Table 3 in the Supplementary Material).
[15] See Table 4 in the Supplementary Material. A moderated regression model separately for absolute moralists is not feasible because the focal predictor (moral fairness) could only take one value.

that participants with higher moral fairness values behave relatively less fair in a high stakes scenario than subjects with lower moral fairness values.

In addition, we considered a double-hurdle model, assuming that the question to share or not to share in a dictator game is different from the question of how much to share, conditional on having decided to share something (McDowell, 2003).[16] For this, the decision to give in a dictator game was modelled as a two-step process: The decision to give or not to give was constructed as a binary probability model using a probit regression. Whereas the decision of how much to give, conditional on a positive contribution, was modelled as a truncated-at-zero OLS model following a Poisson process. The probit model (*LR Chi-Square test* = 27.74, $p < .0001$) suggested that moral fairness values predict the dictator's decision to give something or keep everything ($B = 0.16$, $p < .001$), but stake size and the interaction effect between moral fairness and stake size were not significant predictors (stake size: $B = 0.34$, $p = .424$; interaction effect: $B = -0.09$, $p = .149$). In addition, the truncated Poisson model revealed that the size of the dictator's contribution seemed to be driven by moral fairness values ($B = 0.06$, $p < 0.001$), as well as weakly significantly by the interaction effect between moral fairness and stake size ($B = -0.04$, $p = .077$). This suggests that the interaction effect observed in the main results can be mostly attributed to dictators who have decided to share something but are unsure about how much to share. However, selfish dictators who are not willing to share anything are not influenced in their decision by the stake size. As such, it seems that the interaction effect between moral fairness and stake size applies less to selfish dictators but rather drives the size of contribution of prosocial dictators.

Finally, we found that it is indeed valuing fairness that drives the results. A comparison of the fairness concept (MFSS) with related measures assessed in the survey, like benevolence (Schwartz values), community concerns (Ethics Value Assessment) and engagement (Triune

[16] See Table 5 in the Supplementary Material.

Ethics Theory), revealed no significant corrected item-total correlations (below .40) and low internal validity (Cronbach's alpha of .48) between the variables. That is, participants that scored extremely high on fairness only scored average on benevolence, community concerns and engagement. In addition, neither benevolence, nor community, nor engagement had a significant direct or indirect effect on giving. Thus, only moral fairness concerns distinctively influenced giving in a dictator game and significantly affected giving under different stake size scenarios.[17]

2.5 Discussion and conclusions

We find that moral fairness concerns influence prosocial giving in a dictator game, and that this effect is moderated by stake size. Increasing stake size from $10 to $500 in a dictator game reduces the average willingness to share to a greater extent for people harboring high moral fairness values than for people with lower moral fairness concerns. This pattern of findings can be accounted for by temptation models assuming that stronger temptations may compromise people's ability to live up to the moral standards they maintain otherwise (i.e., when stakes are low), but cannot be easily explained through consistency theories. In particular, we find no evidence that high moral values may better shield people from temptation. This finding is particularly pronounced for *absolute moralists* who claim they would never behave unfairly for any amount of money. To the extent that *absolute moralists* fail to behave fairly in a dictator game, the present results also provide evidence for moral hypocrisy (Batson et al., 1997; Rustichini & Villeval, 2014).

Economic social preference models typically do not explicitly address the role of moral judgments, and how stake size interacts with such moral judgments. Some work (e.g., Ariely, Bracha & Meier, 2009; Bénabou & Tirole, 2011) assumes that social behavior is affected by

[17] Our results are robust when controlling for potential demographic confounds (age, education, household income, etc.), as well as when considering dictator game data to be left-censored. See "Robustness Checks" in the Supplementary Material.

(self-) image concerns or moral identity, which might be interpreted as modeling a consistent relationship between social behavior and moral judgments. Other economic models often do take into account the trade-off between social parameters and the size of monetary gains, although the predictions can depend on the exact model. For instance, the model of inequity aversion by Fehr & Schmidt (1999) is invariant to any change in stake size for given social utility parameters, and thus would suggest that social behavior in dictator games should only depend on fairness parameters but be independent of stake size. Bolton & Ockenfels' (2000) model, on the other hand, is consistent with different stake size effects, including the one we observe (which is also predicted by their example motivation function, as well as a concave version of the Fehr-Schmidt model). That is, if the self-reported *moral fairness* in our data is interpreted as a fairness parameter in the social utility function, different social preference models correspond to the hypotheses with respect to stake size effect that we derived from models in psychology. At the same time, we note that none of the economic models predicts that many participants grossly understate the impact of stake size on their fairness behavior. From this perspective, the temptation theories in psychology complement the economic social utility approach by emphasizing that morality should be seen as a relative concept that people might aim for but struggle to achieve in face of temptations. That is, financial temptation can sometimes be more difficult to resist than estimated. We thus conclude that even though moral fairness values do guide prosocial behavior, they fail to function as an effective self-control device in high stakes situations.

While we study a large representative subject pool, our dictator game differs from standard dictator games. In particular, dictators were not endowed with money, but rather with ten tickets to participate in an online raffle with a given prize. The number of other tickets that could possibly enter the raffle and thus impact the chances to win the raffle was not disclosed to the participants. Such uncertainty is known to decrease the willingness to behave pro-socially (see Engel, 2011; and Ockenfels, Sliwka & Werner, 2015), and is often found to interact with social preferences over certain outcomes in non-trivial ways (e.g., Bolton, Ockenfels & Stauf, 2015). There is no obvious reason why the dictator game variant in our study would affect results in specific ways and thus confound our conclusions. Still, it would be useful to replicate our study with more standard dictator games, as well as with a broader set of games of social behavior, to see how robust our findings are.

Also, from a psychology perspective, it would be interesting to study the consequences of gaps between moral claims and actual fairness behavior. Some researchers suggest that people with strong moral foundations might suffer from a greater distress from not behaving consistently with their moral values (Aquino et al., 2009). However, Rustichini and Villeval (2014) suggest that people adjust their moral norms to be in line with their previous behavior. Alternatively, people may conjure up certain justifications that would help frame the specific behavior as somehow deserved or a defendable exception from the rule, in line with theories of motivated reasoning (Kunda, 2000). Although our data are consistent with this view, especially for absolute moralists, the study of such consequences was beyond the scope of the present research. Nevertheless, this could be a promising avenue for further research.

References

Achtziger, A., Alós-Ferrer, C., & Wagner, A. K. (2016). The impact of self-control depletion on social preferences in the ultimatum game. *Journal of Economic Psychology, 53*, 1-16.

Aiken, L. S. & West, S. G. (1991). *Multiple Regression: Testing and Interpreting Interactions*. Newbury Park, CA: Sage Publications.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice-Hall.

Andersen, S., Ertaç, S., Gneezy, U., Hoffman, M., & List, J. A. (2011). Stakes Matter in Ultimatum Games. *American Economic Review, 101*(7), 3427-3439.

Aquino, K., Freeman, D., Reed II., A., & Lim, V. K. G. (2009). Testing a social-cognitive model of moral behavior: The interactive influence of situations and moral identity centrality. *Journal of Personality and Social Psychology, 97*(1), 123-141.

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review, 99*(1), 544-555.

Batson, C. D., Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C., & Wilson, A. D. (1997). In a very different voice: Unmasking moral hypocrisy. *Journal of Personality and Social Psychology, 72*(6), 1335-1348.

Benabou, R. J., & Tirole, J. (2011). Identity, morals and taboos: beliefs as assets. *Quarterly Journal of Economics, 162*(2), 805-855.

Blake, P. R., & Rand, D. G. (2010). Currency value moderates equity preference among young children. *Evolution and Human Behavior, 31*(3), 210-218.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory on equity, reciprocity, and competition. *American Economic Review, 90*(1), 166-193.

Bolton, G. E., Ockenfels, A., & Stauf, J. (2015). Social responsibility promotes conservative risk behavior. *European Economic Review, 74*, 109-127.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: a review and capital-labor-production framework. *Journal of Risk and Uncertainty, 19*(1-3), 7-42.

Carpenter, J., Verhoogen, E., & Burks, S. (2005). The effect of stakes in distribution experiments. *Economic Letters, 86*(3), 393-398.

Côté, S., House, J., & Willer, R. (2015). High economic inequality leads to higher-income individuals to be less generous. *Proceedings of the National Academy of Sciences of the Unites States of America, 112*(52), 15838-15843.

DellaVigna, S. (2009). Psychology and Economics: Evidence from the Field. *Journal of Economic Literature, 47*(2), 315-372.

De Witt Huberts, J. C., Evers, C., & De Ridder, D. T. (2014). "Because I am worth it": a theoretical framework and empirical review of a justification-based account of self-regulation failure. *Personality and Social Psychology Review, 18*(2), 119-138.

Duckworth, A. L., & Seligman, M. E. P. (2005). Self-Discipline Outdoes IQ in Predicting Academic Performance of Adolescents. *Psychological Science, 16*(12), 939-944.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics, 14*(4), 583-610.

Fehr, E., Bernhard, H. & Rockenbach, B. (2008). Egalitarianism in young children. *Nature, 454*, 1079-1083.

Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics, 114*(3), 817-868.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior, 6*, 347-369.

Graham, J., & Haidt, J. (2012). Sacred values and evil adversaries: A moral foundations approach. In P. Shaver & M. Mikulincer (Eds.), *The Social Psychology of Morality: Exploring the Causes of Good and Evil* (pp. 11-31). New York: APA Books.

Heider, F. (1946). Attitudes and cognitive organization. *The Journal of Psychology, 21*, 107-112.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review, 91*(2), 73-78.

Hofmann, W., Baumeister, R. F., Foerster, G., & Vohs, K. D. (2012). Everyday Temptations: An Experience Sampling Study of Desire, Conflict, and Self-Control. *Journal of Personality and Social Psychology, 102*(6), 1318-1335.

Johnson, K. M., Iyer, R., Wojcik, S. P., Vaisey, S., Miles, A., Chu, V., & Graham, J. (2014). Ideology-Specific Patterns of Moral Indifference Predict Intentions Not to Vote. *Analyses of Social Issues and Public Policy. Special Issue: The Social Psychology of the 2012 US Presidential Election*, 1-7. DOI: 10.1111/asap.12039

Holton, R. (2009). *Willing, wanting, waiting.* New York: Oxford University Press.

Kavanagh, D. J., Andrade, J., & May, J. (2005). Imaginary relish and exquisite torture: The Elaborated Intrusion theory of desire. *Psychological Review, 112*(2), 446-467.

Kivetz, R., & Simonson, I. (2002). Earning the right to indulge: Effort as a determinant of customer preferences toward frequency program rewards. *Journal of Marketing Research, 39*(2), 155-170.

Kotabe, H. P., & Hofmann, W. (2015). On integrating the components of self-control. *Perspectives on Psychological Science, 10,* 618–638.

Kunda, Z. (2000). The case for motivated reasoning. In E. T. Higgins & A. W. Kruglanski (Eds.), *Motivational science: Social and personality perspectives* (pp. 313-335). New York, NY: Psychology Press.

List, J. A., & Cherry, T. L. (2008). Examining the role of fairness in high stakes allocation decisions. *Journal of Economic Behavior and Organization, 65*(1), 1-8.

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes, 65*(3), 272-292.

Lönnqvist, J.-E., Irlenbusch, B., & Walkowitz, G. (2014). Moral hypocrisy: impression management or self-deception? *Journal of Experimental Social Psychology, 55*, 53-62.

McDowell, A. (2003). From the help desk: hurdle models. *The Stata Journal, 3*(2), 178-184.

Meier, S., & Sprenger, C. (2010). Present-Biased Preferences and Credit Card Borrowing. *American Economic Journal: Applied Economics, 2*(1), 193-210.

Miles, A. (2014). Demographic Correlates of Moral Differences in the Contemporary United States. *Poetics, 46,* 75-88.

Miles, A., & Vaisey, A. (2015). Morality and Politics: Comparing Alternate Theories. *Social Science Research, 53*, 252-269.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., Houts, R., Poulton, R., Roberts, B. W., Ross, S., Sears, M. R., Thomson, W., M., Caspi, A., & Heckman, J. J. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences of the United States of America, 108*(7), 2693-2698.

Nordgren, L. F., van Harreveld, F., & van der Pligt, J. (2009). The Restraint Bias: How the Illusion of Self-Restraint Promotes Impulsive Behavior. *Psychological Science, 20*(12), 1523-1528.

Novakova, J., & Flegr, J. (2013). How much is our fairness worth? The effect of raising stakes on offers by proposers and minimum acceptable offers in dictator and ultimatum games. *PLOS ONE, 8*(4), e60966.

Ockenfels, A. (1998). *Fairness, Reziprozität und Eigennutz: Ökonomische Theorie und experimentelle Evidenz. Die Einheit der Gesellschaftswissenschaften, 108*. Tübingen: Mohr Siebeck.

Ockenfels, A., & Werner, P. (2014a). Beliefs and Ingroup Favoritism. *Journal of Economic Behavior and Organization, 108,* 453-462.

Ockenfels, A., & Werner, P. (2014b). Scale Manipulation in Dictator Games. *Journal of Economic Behavior and Organization, 97*, 138-142.

Ockenfels, A., Sliwka, D., & Werner, P. (2015). Bonus Payments and Reference Point Violations. *Management Science, 61*(7), 1496-1513.

Osgood, C. E., & Tannenbaum, P. H. (1955). The principle of congruity in the prediction of attitude change. *Psychological Review, 62*(1), 42-55.

Padilla-Walker, L. M., & Jensen, L. A. (2015). Validation of the long- and short-form of the Ethical Values Assessment (EVA): A questionnaire measuring the three ethics approach to moral psychology. *International Journal of Behavioral Development, 40*(2), 181-192.

Piff, P. K., Dietzke, P., Feinberg, M., Stancato, D. M., & Keltner, D. (2015). Awe, the small self, and prosocial behavior. *Journal of Personality and Social Psychology, 108*(6), 883-899.

Ritov, I., & Baron, J. (1999). Protected Values and Omission Bias. *Organizational Behavior and Human Decision Processes, 79*(2), 79-94.

Roth, A. E. (2007). Repugnance as a Constraint on Markets. *Journal of Economic Perspectives, 21*(3), 37-58.

Rustichini, A., & Villeval, M. C. (2014). Moral hypocrisy, power and social preferences. *Journal of Economic Behavior and Organization*, 107, 10-24.

Sutter, M., Kocher, M. G., Glaetzle-Ruetzler, D., & Trautmann, S. T. (2013). Impatience and uncertainty: experimental decisions predict adolescents' field behavior. *American Economic Review, 103*(1), 510-531.

Vaisey, S., & Miles, A. (2014). Tools from moral psychology for measuring personal moral culture. *Theory and Society, 43*(3), 311-332.

Supplementary Material

Table 2. Summary statistics: giving in the dictator game by moral fairness scores and stake size.

Giving in the Dictator Game				
	\$10 Treatment		*\$500 Treatment*	
Moral fairness	Mean	N	Mean	N
1.00	3.00	3	0.33	3
1.33	2.00	1	-	-
1.67	3.00	2	5.00	1
2.00	0.00	1	10.00	1
2.33	3.00	2	3.00	2
2.67	5.00	1	-	-
3.00	3.40	5	4.67	6
3.33	3.50	8	1.67	3
3.67	3.50	6	5.00	2
4.00	4.29	7	6.40	5
4.33	3.00	9	2.89	9
4.67	5.17	6	3.91	11
5.00	3.60	10	2.92	13
5.33	3.56	9	2.91	22
5.67	3.32	22	4.08	26
6.00	3.75	12	3.71	21
6.33	3.24	29	3.57	30
6.67	3.81	27	4.00	35
7.00	4.17	48	3.57	37
7.33	4.55	40	3.40	47
7.50	5.00	1	-	-
7.67	4.07	59	4.10	51
8.00	4.94	453	4.13	403
Total	*4.50*	*761*	*3.94*	*728*

Table 3: Do absolute moralists share equitably more often? (observed frequencies).

		Equal split in dictator game		
		Yes (decision = 5)	No (decision <> 5)	**Row total**
OVERALL	**Relative moralists** (Moral fairness score < 8)	281	352	633
		44%	*56%*	*100%*
	Absolute moralists (Moral fairness score = 8)	517	339	856
		60%	*40%*	*100%*
LOW STAKE ($10)	**Relative moralists** (Moral fairness score < 8)	148	160	308
		48%	*52%*	*100%*
	Absolute moralists (Moral fairness score = 8)	286	167	453
		63%	*37%*	*100%*
HIGH STAKE ($500)	**Relative moralists** (Moral fairness score < 8)	133	192	325
		41%	*59%*	*100%*
	Absolute moralists (Moral fairness score = 8)	231	172	403
		57%	*43%*	*100%*

Overall: Pearson chi2(1) = 37.48 Pr < .001, Fisher's exact < .001
Low stake scenario: Pearson chi2(1) = 17.02 Pr < .001, Fisher's exact < .001
High stake scenario: Pearson chi2(1) = 19.35 Pr < .001, Fisher's exact < .001

Table 4: Moderated regression considering only relative moralists: with giving as outcome variable, moral fairness as predictor and stake size scenario as moderator.

Regression Summary

R-sq	F	df1	df2	p	n
0.02	1.34	7	625	0.229	633

	b	se	t	p
Constant	3.21	0.75	4.26	0.000
Stake size (baseline=low)	-0.16	0.20	-0.80	0.426
Moral fairness	0.14	0.11	1.27	0.203
Interaction (moral fairness*stake size)	-0.15	0.14	-1.02	0.310
Harm	-0.05	0.11	-0.48	0.632
Loyalty	-0.01	0.09	-0.11	0.910
Authority	0.08	0.08	1.04	0.300
Purity	0.10	0.08	1.26	0.208

Conditional Effect of Focal Predictor at Values of the Moderator Variable

Stake size	b	se	t	p	LLCI(b)	ULCI(b)
High ($500)	.00	.12	-0.01	.992	-.24	.24
Low ($10)	.14	.11	1.27	.203	-.08	.37

Table 5: Double-hurdle model, assuming that the dictator game decision is a two-step process.

I. Binary decision to give something or nothing

I. Probit Regression: Positive Decision						
Log likelihood			LR chi2(3)	Prob > chi2	Pseudo R2	n
-562.31			27.74	0.0000	0.0241	1503
	b	se	z	p	[95% CI]	
Constant	0.12	0.30	0.41	0.680	-0.46	0.71
Stake size (baseline = $10)	0.34	0.43	0.80	0.424	-0.50	1.18
Moral fairness	0.16	0.04	3.90	0.000	0.08	0.24
Interaction (moral fairness * stake size)	-0.09	0.06	-1.44	0.149	-0.20	0.03

II. Conditional on giving something, deciding how much to give

II. Zero-Truncated Poisson Estimates: Size of Positive Decision						
Log likelihood			Model chi2(3)	Prob > chi2	Pseudo R2	n
-2697.36			23.83	0.0000	0.0044	1296
	b	se	z	p	[95% CI]	
Constant	1.18	0.12	9.88	0.000	0.94	1.41
Stake size (baseline = $10)	0.23	0.17	1.33	0.182	-0.11	0.56
Moral fairness	0.06	0.02	3.67	0.000	0.03	0.09
Interaction (moral fairness * stake size)	-0.04	0.02	-1.77	0.077	-0.08	0.00

Robustness checks

We controlled for potential demographic confounds such as gender, age, education, or household income. However, the interaction effect between stake size and moral fairness values driving giving in a dictator game qualitatively remained identical when controlling for these demographic covariates. Also, none of these variables significantly interacted with stake size. On a side note, in line with Graham and Haidt's findings (2012), we observed a gender effect with women indicating stronger values than men on each of the MFSS foundations (*Mann-Whitney U tests: each p* < .001).

Considering the argument that dictator game data are left censored (Engel, 2011), we ran a Tobit model to account for the lower censor of the variable *giving* at zero (see Table 6). The Tobit model showed that harboring higher *moral fairness* values also increased expected *giving* in the dictator game. Although *stake size* did not have a significant direct effect on *giving*, the interaction effect between *moral fairness* and *stake size* remained significant.

Table 6: Tobit model, assuming that the dictator game decision is censored at zero.

Tobit Regression						
Log likelihood			LR chi2(7)	Prob > chi2	Pseudo R2	n
-3379.47			60.95	0.0000	0.0089	1488
	b	se	t	p	[95% CI]	
Constant	1.10	0.70	1.56	0.118	-0.28	2.48
Stake size (baseline = $10)	1.30	0.85	1.53	0.127	-0.37	2.96
Moral fairness	0.37	0.10	3.76	0.000	0.18	0.56
Interaction (moral fairness * stake size)	-0.26	0.12	-2.30	0.022	-0.49	-0.04
Harm	-0.06	0.10	-0.62	0.535	-0.26	0.13
Loyalty	0.02	0.08	0.22	0.824	-0.14	0.18
Authority	0.10	0.06	1.75	0.080	-0.14	0.18
Purity	0.04	0.06	0.61	0.540	-0.08	0.16
/Sigma	2.47	0.06			2.63	2.85

Obs. Summary: 193 left-censored observations at decision <= 0
1295 uncensored observations
0 right-censored observations

CHAPTER 3:
Self-control and social comparison information: Does ego-depletion affect ingroup-outgroup discrimination?

Joint with Katharina Diel, Wilhelm Hofmann and Axel Ockenfels *

Abstract

We investigate the interaction between ego-depletion and group identity in a laboratory dictator game experiment. Our hypothesis is that people discriminate more between ingroup and outgroup members when they are depleted, as we expect that people that are deprived of self-control resources will rely more heavily on social comparison information available. However, we find no evidence for discrimination conditional on group identity, neither with nor without depletion.

* Katharina Diel and Wilhelm Hofmann, Social Cognition Center Cologne, University of Cologne, Richard-Strauss-Str. 2, 50931 Cologne, Germany, e-mail: katharina.diel@uni-koeln.de, wilhelm.hofmann@uni-koeln.de. Uta K. Schier and Axel Ockenfels, Department of Economics, University of Cologne, Albert-Magnus-Platz, 50937 Cologne, Germany, e-mail: schier@wiso.uni-koeln.de, ockenfels@uni-koeln.de. This work benefitted from financial support of the German Research Foundation (DFG) through the Research Unit "Relativity in Social Cognition: Antecedents and Consequences of Comparative Thinking" (FOR 2150), as well as through the Research Unit "Design & Behavior: Economic Engineering of Firms and Markets" (FOR 1371). We thank Yoella Bereby-Meyer, Yan Chen, and participants of the Research Unit seminars in Cologne for helpful comments and feedback. We also thank Wolfram Uerlich for assistance in the experiment.

3.1 Introduction

Social behavior often depends on the group membership of the person one is interacting with. Studies from both social psychology and economics suggest, for example, that people mostly favor ingroup members and discriminate against outgroup members (Tajfel & Turner, 1979; Chen & Li, 2009). This has been studied with natural identities such as football fans, party affiliation or gender (Weisel & Böhm, 2015; Akerlof & Kranton, 2010), but also with artificial groups created in the lab (Chen & Li, 2009). Identity economics research (Akerlof & Kranton, 2010) suggests that this discrimination might be driven by different social preferences for ingroup and outgroup members, when group identity is incorporated in people's utility function (Chen and Li, 2009).

In this paper, we propose that ego-depletion, a state in which self-control efforts are temporarily exhausted (Hagger et al., 2010), might lead to a more pronounced reliance on social comparison information such as ingroup/outgroup membership. Specifically, we expect that a shortage of cognitive control (ego-depletion) will lead people to rely more on categorical information about the person they interact with. Hence, we expect that people discriminate more between ingroup and outgroup members when they are depleted. We test this notion in a lab study comparing prosocial behavior towards ingroup and outgroup members in a dictator game while subjects are either depleted or not. We find no evidence that people give more towards members of their ingroup, neither with nor without depletion. Mostly, this is because we fail to replicate ingroup favoritism in the lab using a group identity manipulation previously shown to be effective in inducing group identities in the lab (Chen & Li, 2009). In addition, our manipulation of self-control resources using depletion methods might be too short-lived to be reflected in subsequent dictator game behavior, as consistent with recent failures to find substantial ego depletion effects (Hagger et al., 2016).

3.2 Background and hypotheses

3.2.1 Previous evidence on prosocial behavior and ego-depletion

Economic studies on fairness behavior robustly show that people are not only selfish, as standard economic theory assumes, but that they also care about others' payoffs. In a meta-study on dictator game behavior, Engel (2011) shows that in fact only 36% of all subjects behave selfishly while the average shares 28% of the endowment.

To understand whether prosocial behavior comes naturally to humans or whether it requires cognitive control, several studies examined the question whether selfish behavior or generosity is the more intuitive response in the dictator game. By inducing a shortage of cognitive control (e.g. ego-depletion, cognitive load or time constraints), these studies tested what the default response is when cognitive control is temporarily exhausted. However, the evidence is rather mixed. On the one hand, some studies find that a shortage of cognitive control increases prosocial behavior, suggesting that generosity might be the intuitive response (Schulz et al., 2014; Rand et al., 2012; Lotz, 2015; Cappelen et al., 2016). On the other hand, other studies show that people become more selfish under ego-depletion (Halali, Bereby-Meyer and Ockenfels, 2013; Achtziger, Alós-Ferrer and Wagner, 2015).[1]

We suspect that one reason why the evidence is so mixed is that the reaction towards ego-depletion might depend on which social comparison context is rendered salient. In those studies that found generosity to be the more intuitive response (Schulz et al., 2014; Cappelen et al., 2016), subjects did not know their payoff-relevant role in the dictator game before they had to make an allocation decision. For instance, in two consecutive dictator games subjects played

[1] Similar evidence is found by Benjamin, Brown and Shapiro (2013) who find that with cognitive load selfish behavior in children slightly increases in hypothetical dictator game. However, this effect is not significant. Ben-Ner et al. (2015) find a correlation between generosity of parents and their 2-5 year old children in a field experiment which suggests that generosity might be a learned response. Roch et al. (2000) find that under cognitive load the amount people take from a common resource increases. Finally, Balafoutas et al. (2017) suggest that ego-depletion has no effect on social preferences, and Banker et al. (2017) propose that depletion does not directly influence selfishness but rather increases people's susceptibility to social clues like anchors.

the role of the dictator, as well as the role of the recipient. In contrast, in studies that found selfish behavior to be more intuitive (Halali et al., 2013; Achtziger et al., 2015), the payoff-relevant role of subjects was determined before dictators made their decision. We suspect that these differences in role uncertainty constitute different social comparison contexts, in which subjects sympathize more or less with recipients, depending on whether they themselves could be in the role of a recipient. As such, we argue that depletion may sometimes result in more selfish and sometimes in more cooperative behavior, depending on how people socially compare with others. To study this, we use group identity to induce a social environment in which people have been shown to behave differently depending on whether they face ingroup or outgroup members.[2]

3.2.2 Group identity and dictator games

Social identity theory suggests that a person's sense of self partly derives from memberships in social groups (Tajfel & Turner, 1979; Akerlof & Kranton, 2000; Chen & Li, 2009). People categorize others and themselves for example as being female, a banker or a Republican and identify with and favor the ingroup and discriminate against the outgroup (Tajfel & Turner, 1979).

Chen and Li (2009) show that people actually discriminate between ingroup and outgroup members in incentivized experiments. Specifically, they find that with artificial, near-minimal groups created in the lab, subjects are more altruistic towards ingroup members across various games, including the dictator game. Chen and Chen (2011) further find that interacting with ingroup members increases efficiency in coordination games, although this could not be replicated in a recent study (Camerer et al., 2016). Yamagishi & Kiyonari (2000) find in a

[2] The social comparison theory of Mussweiler (2003) suggests that people either assimilate with or contrast away from others which could provide a psychological understanding of ingroup-outgroup discriminations in behavior. Fehr (2009) argues that if somatic similarity is associated with ingroup members, people will tend to behave more altruistic towards others who are more similar.

modified prisoner's dilemma game with simultaneous moves that players cooperate more with ingroup than with outgroup members.

Using natural groups, Ockenfels and Werner (2014) investigate the role of beliefs on the effects of group identity and find that, when the shared identity is not known to the recipient, dictators do not give more to ingroup members.[3] Also using natural groups (football club fans or political party affiliation), Weisel and Böhm (2015) show in a modified prisoner's dilemma that subjects like to help ingroup members but refrain from harming outgroup members. Goette, Huffman, and Meier (2006) find that Swiss army officers cooperate more in a prisoner's dilemma with officers from their own platoon and are also less likely to punish norm violators if the victim of a norm is from their own platoon. This ingroup bias in third-party-punishment has also been shown in other field studies (e.g., Bernhard, Fehr, & Fischbacher, 2006).[4] However, Buchan, Johnson, and Croson (2006) show that ingroup favoritism in the context of trust games is not universal but can depend on cultural orientation.

Taken together, the literature shows that prosocial behavior might depend on whether subjects compare to ingroup or outgroup members. Although group identity might be stronger among members of natural groups, we will rely on artificial groups created in the lab by a near-minimal paradigm which has been shown to be sufficient to induce ingroup-outgroup discriminations by still maximizing experimental control of variables.

3.2.3 Social comparison processes with ego-depletion

Previous research suggests that social comparison is relevant for economic behavior (e.g. Bernheim, 1994; Cason & Mui, 1998; Bolton & Ockenfels, 2000). However, little is known about the downstream effects of self-control on social comparison processes. We argue that

[3] This suggests that ingroup-outgroup discrimination might critically hinge on mutual knowledge about the common identity.
[4] Experimental evidence also shows that social identities affect risk and time preferences (Benjamin et al., 2010) and can influence economic outcomes (Bertrand et al., 2015).

people in a depleted state have fewer resources to deliberate about the pros and cons of their behavioral choices; thus, we propose that depleted individuals are more sensitive to social comparison information in their environment. This idea is indirectly supported by research showing that comparative thinking allows for more efficient information processing under suboptimal conditions (Mussweiler & Epstude, 2009). On this note, two recent papers have specifically looked at how a lack of self-control resources might influence the effects of group identity.

First, De Dreu et al. (2015) investigate cooperation with ingroup versus outgroup members in a Prisoner's Dilemma-Maximizing Differences Game, when subjects' cognitive deliberation is either restricted or not. The authors find that decisions to contribute are made faster than decisions to behave selfishly. In addition, subjects who are not depleted with a Stroop task (as explained below) keep on average more for themselves than depleted subjects. Furthermore, depleted subjects punish outgroup members more than non-depleted subjects, suggesting that outgroup discrimination might rather be an intuitive than a deliberative process. Thus, this study suggests that self-control depletion enhances the gap between prosocial behavior towards ingroup and outgroup members. Therefore, the study provides indicative evidence concerning our hypothesis that social comparison processes might intensify when subjects are depleted.

Second, similar evidence is found by Yudkin et al. (2016) in third-party punishment in two online experiments with natural groups. Restricting cognitive resources with a cognitive load task, they observe that subjects punish outgroup members significantly more than ingroup members. In contrast, subjects in the low cognitive load condition do not differentiate between groups. However, in this study punishment behavior was not actually payoff-relevant. Therefore, the respective results might be questionable from an economic perspective but they still provide some evidence that intergroup biases might be exacerbated when cognitive resources are restricted.

In the present study, we aim at testing this prediction while incentivizing subjects' behavior. We do so in a dictator game and argue that depletion sometimes results in more prosocial than selfish behavior, depending on which social comparison standard is made salient. Consequently, the following hypotheses guided us in our research.

3.2.4 Hypotheses

Hypothesis 1 (group effect): Subjects facing an ingroup member in the dictator game give more to the recipient than subjects facing an outgroup member.

Based on the group identity literature discussed above, we predict that dictators will show some ingroup favoritism in their allocation decisions.

Hypothesis 2 (depletion effect): Ego-depletion increases the reliance on social comparison information (i.e. group membership) in social interactions, and thus the ingroup-outgroup discrimination of Hypothesis 1 is exacerbated with ego-depletion.

The second hypothesis suggests that the ingroup-outgroup discrimination is amplified when subjects are ego-depleted. In a state of ego-depletion, subjects only have limited cognitive control and thus, should rely more strongly on available social comparison information when making allocation decisions.

3.3 Methods

3.3.1 Sample

To study our hypotheses, a laboratory experiment was conducted in May and June 2016 in the Cologne Laboratory for Economic Research (CLER). 202 participants were recruited via ORSEE (Greiner, 2015) and participated in an experiment that was implemented using the z-Tree software (Fischbacher, 2007). Subjects were students from various disciplines, with a mean age of 23 (range: 18-41) and 57% being female. They earned on average 14.41 EUR for an hour of participation, including a 4.00 EUR show-up fee. Two subjects reported having a color deficiency which might have limited the effect of the depletion task. However, both subjects were recipients in the dictator game and as such not relevant for our analyses.

3.3.2 Experimental procedure

We make use of a between-subject, 2 × 2 factorial design. Subjects were randomly assigned either to a depletion or a control group and matched either with an ingroup or an outgroup partner.

The experiment consisted of three parts.[5] In the first part, group identities were induced based on preferences for paintings by Paul Klee or Wassily Kandinsky as suggested by Chen & Li (2009). Participants saw five pairs of paintings with no information about the painters, and had to make five decisions which painting they preferred. Thereby, subjects were assigned either to a KLEE group or to a KANDINSKY group, depending on their preferred artist. Subsequently, their group membership, the group size in the experimental session and information about the artists and paintings were revealed to participants. Their individual group membership was prominently displayed on their screen for the rest of the experiment. In the next step, subjects saw two additional paintings from the same artists, and were asked to guess who of them painted which picture. To do so, they could chat for ten minutes with the other members of their group. After the chat, subjects individually made guesses about which artist painted which painting, and could earn 1 EUR per correct guess. The chat stage was added to enhance group identity (Chen and Chen, 2011). Only after completing part I, participants received instructions on part II and III.

In part II, subjects were either depleted or not depleted using a Stroop task (Halali et al., 2013). In this task, subjects are asked to determine the color that a word is displayed in, by hitting a corresponding key on the keyboard. In addition, the displayed words are also colors, e.g. “Blue”, “Yellow”, “Green”, or “Red”. In the control group, subjects had to identify only

[5] See Appendix A for the instructions given to participants, and Appendix B for exemplary experimental screens.

congruent trials, in which the content of the word also matched its color. For instance, the word "Blue" was always displayed in blue font.[6] In contrast, in the depletion group subjects faced only incongruent trials, meaning that the content of the word did not match its color. For example, subjects in the depletion group saw the word "Blue" displayed in yellow font. As such, subjects in the depletion group had to suppress the first response of reading the word "Blue" and instead determine the correct answer "Yellow" by hitting a corresponding key. Subjects played the Stroop task for 80 rounds. They could earn 0.5 tokens (5ct) per correct key hit within 5 seconds (Achtziger et al., 2015). Before the 80 rounds started, subjects faced two non-incentivized test rounds, where they received feedback on their responses being correct or incorrect. To avoid income effects, feedback on subjects' performance from the 80 incentivized rounds was only revealed to them at the end of the experiment.

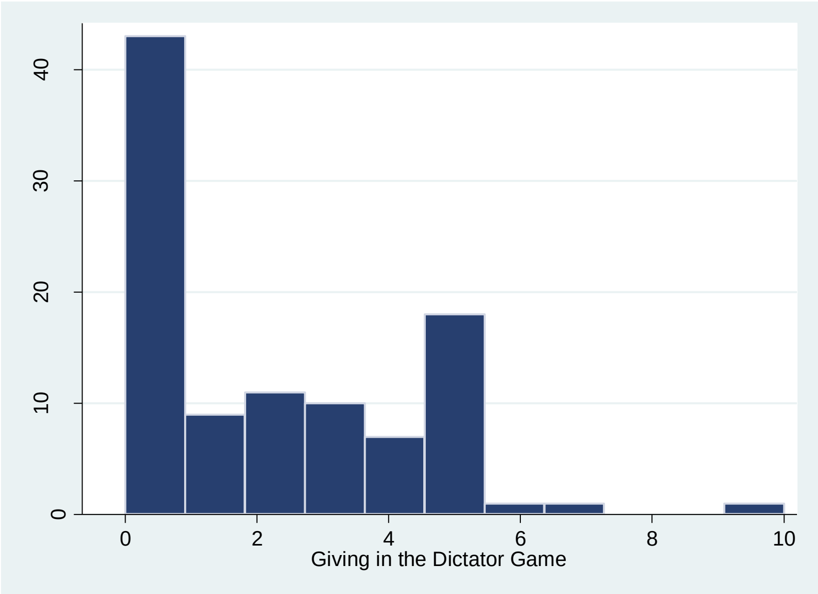
Finally, in part III, subjects played a dictator game. They were randomly matched with another participant, and randomly assigned either the role of Player A (dictator) or Player B (recipient). Dictators were told whether their recipient was part of their ingroup or of the outgroup. They were endowed with 100 tokens (10 EUR) and had to choose how much they wanted to transfer to their recipient. Recipients were informed about the group membership of their assigned dictator and about how much she transferred. After the dictator game, subjects answered a post-hoc questionnaire, received information about their earnings and were paid out privately in cash.

[6] See Appendix B for experimental screens of the Stroop task.

3.4 Results

All statistical tests are performed at a 5% significance level. For our analyses, we consider the 101 observations from the dictators. Dictators shared on average 2.06 EUR (*sd* = 2.19 EUR). 39 out of 101 dictators shared nothing, 17 chose the equal split, and only 3 shared more than 50%. Hence, our results look fairly standard (see Figure 3; Engel, 2011).

Figure 3: Distribution of giving in the dictator game (N=101).



3.4.1 Treatment differences in generosity in the dictator game

To test our hypotheses, we examine the amount dictators share conditional on the treatment (see Table 7). Overall, we do not find evidence for ingroup-outgroup discrimination. Dictators facing an ingroup recipient share on average 1.87 EUR (*SD* = 2.02) while dictators matched with an outgroup recipient give on average 2.23 EUR (*SD* = 2.34). This difference is not statistically significant ($t = 0.81$, $df = 99$, $p = .419$)[7]. As such we cannot confirm our *Hypothesis 1* that dictators favor ingroup recipients in a dictator game. To the contrary, we rather find a

[7] Mann-Whitney U test: $p = .543$.

tendency that dictators give more to outgroup recipients but this is rather driven by two observations in the outgroup/no-depletion treatment that shared 70% and 100% of the pie.[8]

Table 7: Amount given by dictators out of 10€.

N = 101	Ingroup Recipient (n = 47)	Outgroup Recipient (n = 54)	Total (N = 101)
Depletion Group (n = 50)	2.06€ (n = 22)	2.50€ (n = 28)	2.31€
Control Group (n = 51)	1.70€ (n = 25)	1.93€ (n = 26)	1.82€
Total (N = 101)	1.87€	2.23€	

With regard to the effect of depletion on prosocial behavior, we find that depleted dictators shared on average more than non-depleted dictators ($m_{depleted}$ = 2.31 EUR, $m_{non\text{-}depleted}$ = 1.82 EUR). Although this difference seems to be relatively large, it is not statistically significant (t = 1.13, df = 99, p = .263)[9]. Hence, we cannot identify a statistically robust effect of depletion on prosocial behavior, but again our evidence might be distorted by the two unusual observations in the outgroup/no-depletion treatment[10].

Furthermore, we find no evidence for *Hypothesis 2* that subjects rely more on social comparison information when they are depleted. We would expect this to be mostly reflected in exacerbated ingroup-outgroup discrimination when subjects are depleted. Comparing the rows of Table 7, it in fact seems like depleted dictators discriminate more between ingroup and outgroup recipients (difference in means: 0.44€) than non-depleted dictators from the control group (difference in means: 0.23€). However, in both rows the discrimination goes in the opposite direction as expected, as we would have expected ingroup members to be favored. In

[8] The reasons for this unlikely high prosocial behavior are unclear to us. However, these two observations might be driving the observed effect. The difference in giving to outgroup members versus ingroup members becomes smaller without these two observations (outgroup: m = 1.98€, ingroup: m = 1.87; t = 0.28, df = 97, p=.778).
[9] Mann-Whitney U test: p = .170.
[10] Without the two observations that shared 70% and 100% of the pie, the effect of depletion on giving in the dictator game increases but remains non-significant (t-test: p < .057, Mann-Whitney U test: p < .076).

addition, the difference in ingroup-outgroup discrimination between the depletion and the control group is not statistically significant when tested as an interaction effect in a factorial ANOVA. For this, we test the main effects of depletion and group identity, as well as their interaction on giving in the dictator game. However, all effects were statistically non-significant (depletion: $p = .339$, ingroup: $p = .719$, depletion × ingroup: $p = .802$). Even when we only look at dictators from the depletion group which seem to discriminate most, the difference in average giving is not significant (t-test: $p = .464$)[11].

3.4.2 Manipulation checks

These results provide no evidence for our hypotheses. To check whether our treatment manipulations were actually effective, we next analyze the results of a post-hoc questionnaire. In this questionnaire, subjects had to make judgments on scales from 1 ("not at all") to 7 ("completely").

Regarding perceived group identification, we asked participants how much they identify with the KLEE-group and how much they identify with the KANDINSKY-group. The 39 subjects assigned to the KLEE group identified significantly more with the KLEE group ($m = 3.51$, $sd = 1.92$) than with the KANDINSKY group ($m = 2.15$, $sd = 1.37$). This difference is statistically significant (paired t-test: $p < .0002$)[12]. Similarly, the 62 participants preferring paintings by Kandinsky identified more with their assigned KANDINSKY group ($m = 3.58$, $sd = 1.90$) than with the KLEE group ($m = 1.66$, $sd = 0.96$) which is also statistically significant (paired t-test: $p < .0001$)[13]. This suggests that the group identity manipulation was indeed successful.

We further test whether participants in the depletion condition actually perceived the Stroop task as more difficult than participants in the control condition. In fact, subjects from

[11] Mann-Whitney U test: $p = .3701$.
[12] Wilcoxon signed-rank test: $p < .0001$.
[13] Wilcoxon signed-rank test: $p < .0001$.

the depletion condition judge the Stroop task as more difficult (m = 1.94, sd = 1.10) than subjects from the control group (m = 1.39, sd = 0.60). This difference (t-test: $p < .0025$)[14] suggests that the depletion manipulation was effective. This is also supported by the performance of subjects across treatment groups: in the depletion condition, subjects on average solved 78.88 out of 80 trials correctly, whereas in the control condition 79.55 out of 80 trials were completed correctly (t-test: $p < .0030$)[15]. With regard to the decision times during the Stroop task, we further observe a weak treatment difference across the 80 rounds but this is statistically not significant (Mann-Whitney U test: $p = .0564$)[16].

Finally, we also see no irregularities in the chat protocols. Across all sessions, participants in the KLEE chat, as well as in the KANDINSKY chat mostly discussed their assigned task, the paintings and rarely instructions for part I. No subject had to be excluded due to anonymity concerns or other forbidden content during the chat communication.

3.4.3 Robustness checks

To get a clearer picture of whether and how our experimental treatments might affect behavior, we analyze our data irrespective of selfish dictators who give nothing, regardless of contextual factors like group identity. Therefore, we ran some additional tests, considering only dictators that shared a positive amount in the dictator game (n =62). With regard to ingroup-outgroup discrimination, we find that those dictators do not significantly differentiate between ingroup and outgroup recipients (Mann-Whitney U test: $p = .532$), regardless of whether dictators were depleted (MWU test: $p = .804$) or non-depleted (MWU test: $p = .509$). Similarly, depletion seems to have no effect on giving of dictators that share a positive amount (MWU test: $p = .258$), regardless of whether they share with an ingroup member (MWU test: $p = .350$)

[14] Mann-Whitney U test: $p < .003$.
[15] Mann-Whitney U test: $p < .013$.
[16] See also Figure 6 in the Supplementary Material.

or an outgroup member (MWU test: p = .569). Hence, neither group identity nor depletion seems to affect prosocial giving in our data.

To understand the underlying mechanism driving behavior in our data, we collected additional evidence in the post-hoc questionnaire. First, we elicited a short version of the Social Comparison Orientation scale to examine whether how much subjects care about social comparison affects giving in our setting (Gibbons & Buunk, 1999; Schneider & Schupp, 2014). For this, participants had to report on scales from 1 ("not at all") to 7 ("completely") how much they compare with others with regard to abilities and opinions. According to our second hypothesis[17], subjects scoring high on Social Comparison Orientation (SCO) might rely stronger on group membership information than subjects with a low SCO score. In a regression predicting giving in the dictator game based on depletion, group identity, SCO score, as well as their interaction effects, we find that subjects with a higher SCO score react stronger to group information, especially when they are depleted, but none of these effects are statistically significant ($F(7,93)$ = 1.39, R^2 = 0.09; depletion: p = .709, ingroup: p = .124, SCO: p = .259, depletion × ingroup: p = .387, depletion × SCO: p = .997, SCO × ingroup: p = .134, SCO × depletion × ingroup: p = .449).

Second, we measured performance in the Cognitive Reflection Test (Frederick, 2005), to assess whether depleted subjects are more prone to intuitive answers. The Cognitive Reflection Test consists of three brainteaser questions, in which subjects are supposed to suppress an intuitive, wrong answer in favor of a reflective, correct answer. We find that depleted subjects on average solve 1.67 out of 3 questions correctly, whereas non-depleted subjects answer 1.73 out of 3 questions correctly which suggests no effect of ego-depletion on the Cognitive Reflection Test (t-test: p = .715).[18]

[17] See Hypothesis 2, p. 7.
[18] In addition, the Cognitive Reflection Test does not show significant differences between the depletion and no-depletion group when we only consider those subjects that decided to give a positive amount. However, it might be that our depletion manipulation does not affect the Cognitive Reflection Test because ego-depletion has been

3.5 Discussion and conclusion

In this study, we tested the hypothesis that people rely more on social comparison information when their cognitive resources are depleted. We find no significant ingroup-outgroup discrimination in a dictator game, neither with nor without depletion. Mostly, this is due to the fact that we fail to replicate ingroup favoritism in dictator games which has been shown by previous research (Chen & Li, 2009). It might be that the group identities we induced in the lab were not strong or salient enough for participants. As such, it would be interesting to rerun the study with naturally existing groups, such as students from rivalling cities (Ockenfels & Werner, 2014). This might not only increase identification with and favoring of one's own group but also has the advantage that group sizes are more controllable than in the Klee-Kandinsky paradigm. As such, it could be insightful to test whether ingroup favoritism with natural groups increases when subjects are depleted.

Another extension would be to use cognitive load manipulations instead of self-depletion. Previous research has suggested that ego-depletion is a temporary effect (Hagger et al., 2010) and that self-control resources get "recharged" over time. Cognitive load is a similar manipulation to ego-depletion but has been suggested to cognitively limit subjects for a longer time period and to decrease self-control (Shiv & Fedorikhin, 1999). Also, more recent attempts have cast doubt on the robustness and strength of the ego-depletion effect (Hagger et al., 2016). This multi-lab endeavor found a very low mean depletion effect for an often-used paradigm, and substantial heterogeneity in effect sizes, indicating moderators that have yet to be clearly identified.

With regard to the effect of depletion on prosocial behavior, we find some suggestive but statistically not significant evidence for our hypothesis. Looking at the previous literature,

suggested to be a temporary effect (Hagger et al., 2010). As such, depleted resources of the participants of the depletion treatment might have been "recharged" when the post-hoc questionnaire is conducted.

this might be cautiously interpreted as slightly favoring the argument that fairness is a more intuitive response, and selfishness a reflective process (Rand et al., 2012).

In conclusion, our study did not reveal the predicted interplay of social comparison information such as group membership and available cognitive resources in a standard ego-depletion paradigm. Before drawing strong conclusions about the absence of such an effect, it seems important to build a larger database and to go beyond the methods and paradigms used in this study, such as by using natural groups in combination with cognitive load tasks.

References

Achtziger, A., Alós-Ferrer, C., & Wagner, A. K. (2015). Money, depletion, and prosociality in the dictator game. *Journal of Neuroscience, Psychology, and Economics, 8*(1), 1-14.

Akerlof, G. A., & Kranton, R. E. (2000). Economics and Identity. *Quarterly Journal of Economics, 115*(3), 715-753.

Akerlof, G. A., & Kranton, R. E. (2010). *Identity economics: How our identities shape our work, wages, and well-being.* Princeton, NJ: Princeton University Press.

Balafoutas, L., Kerschbamer, R., & Oexl, R., (2017). Distributional preferences and ego-depletion. *Journal of Neuroscience, Psychology and Economics, forthcoming*.

Banker, S., Ainsworth, S. E., Baumeister, R. F., Ariely, D., & Vohs, K. (2017). The sticky anchor hypothesis: ego depletion increases susceptibility to situational cues. *Journal of Behavioral Decision Making, 30*(5), 1027-1040.

Ben-Ner, A., List, J. A., Putterman, L., & Samek, A. (2015). Learned generosity: A field experiment with parents and their children. Working Paper.

Benjamin, D. J., Brown, S. A., & Shapiro, J. M. (2013). Who is 'behavioral'? Cognitive abilities and anomalous preferences. *Journal of the European Economic Association, 11*(6), 1231-1255.

Benjamin, D. J., Choi, J. J., & Strickland, A. J. (2010). Social identity and preferences. *American Economic Review, 100*(4), 1913-1928.

Bernhard, H., Fehr, E., & Fischbacher, U. (2006). Group affiliation and altruistic norm enforcement. *American Economic Review, 96*(2), 217-221.

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy, 102*(5), 841-877.

Bertrand, M., Kamenica, E., & Pan, J. (2015). Gender identity and relative income within households. *Quarterly Journal of Economics, 130*(2), 571-614.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review, 90*(1), 166-193.

Buchan, N. R., Johnson, E. J., & Croson, R. T. A. (2006). Let's get personal: An international examination of the influence of communication, culture and social distance on other regarding preference. *Journal of Economic Behavior & Organization, 60,* 373-398.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433-1436.

Cappelen, A. W., Nielsen, U. H., Tungodden, B., Tyran, J.-R., & Wengström, E. (2016). Fairness is intuitive. *Experimental Economics, 19,* 727-740.

Cason, T. N., & Mui, V.-L. (1998). Social influence in the sequential dictator game. *Journal of Mathematical Psychology, 42*(2-3), 248-265.

Chen, R., & Chen, Y. (2011). The potential of social identity for equilibrium selection. *American Economic Review, 101*(6), 2562-2589.

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

De Dreu, C. K. W., Dussel, D. B., & Ten Velden, F. S. (2015). In intergroup conflict, self-sacrifice is stronger among pro-social individuals, and parochial altruism emerges especially among cognitively taxed individuals. *Frontiers in Psychology, 6*(572), 1-9.

Engel, C. (2011). Dictator games: A meta-study. *Experimental Economics, 14*(4), 583-610.

Fehr, E. (2009). On the economics and biology of trust. *Journal of the European Economic Association, 7*(2), 235-266.

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

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25-42.

Gibbons, F. X., & Buunk, B.P. (1999). Individual differences in social comparison: The development of a scale of social comparison orientation. *Journal of Personality and Social psychology*, *76,* 129-142.

Goette, L., Huffman, D., & Meier, S. (2006). The impact of group membership on cooperation and norm enforcement: evidence using random assignment to real social groups. *American Economic Review, 96*(2), 212-216.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114-125.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., . . . Zwienenberg, M., (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*(4), 546–573.

Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. D. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin, 136*(4), 495-525.

Halali, E., Bereby-Meyer, Y., & Ockenfels, A. (2013). Is it all about the self? The effects of self-control depletion on ultimatum game proposers. *Frontiers in Human Neuroscience, 7*, 240.

Lotz, S. (2015). Spontaneous giving under structural inequality: Intuition promotes cooperation in asymmetric social dilemmas. *PLOS one, 10*(7), e0131562.

Mussweiler, T. (2003). Comparison process in social judgment: Mechanisms and consequences. *Psychological Review, 110*(3), 472-489.

Mussweiler, T., & Epstude, K. (2009). Relatively fast! Efficiency advantages of comparative thinking. *Journal of Experimental Psychology: General, 138*, 1–21.

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

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature, 489*, 427–430.

Roch, S. G., Lane, J. A. S., Samuelson, C. D., Allison, S. T., & Dent, J. L. (2000). Cognitive load and the equality heuristic: a two-stage model of resource over-consumption in small groups. *Organizational Behavior and Human Decision Processes, 83,* 185–212.

Schneider, S., & Schupp, J. (2014). The social comparison scale. *Social Indicators Research, 115*(2), 767-789

Schulz, J. F., Fischbacher, U., Thöni, C., & Utikal, V. (2014). Affect and fairness: dictator games under cognitive load. *Journal of Economic Psychology, 41*, 77–87.

Shiv, B., & Fedorikhin, A. (1999). Heart and mind in conflict: The interplay of affect and cognition in consumer decision making. *Journal of Consumer Research*, *26*, 278–292.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In S. Worchel & W. Austin (Eds.), *The social psychology of intergroup relations* (33-47). Monterey, CA: Brooks/Cole.

Weisel, O., & Böhm, R. (2015). "Ingroup love" and "outgroup hate" in intergroup conflict between natural groups. *Journal of Experimental Social Psychology, 60,* 110-120.

Yamagishi, T., & Kiyonari, T. (2000). The group as the container of generalized reciprocity. *Social Psychology Quarterly, 63*(2), 116-132.

Yudkin, D. A., Rothmund, T., Twardawski, M., Thalla, N, & Van Bavel, J. J. (2016). Reflexive intergroup bias in third-party punishment. *Journal of Experimental Psychology General, 145*(11), 1448-1459.

Appendix A. Experimental Instructions

GENERAL INSTRUCTIONS

Welcome to the experiment!

Please read the following instructions carefully and from now on, please do not communicate with other participants. If you have any questions during the experiment, please raise your hand and we will come to you and answer your questions.

In this experiment, you can earn money. How much you can earn, depends on your decisions and the decisions of one other participant during the experiment. Regardless of this, you will receive 4 EUR for showing-up to the experiment. Your earnings will be displayed in points during the experiment. The experiment comprises several parts. At the end of the experiments, your points from all parts will be added, converted into EUR and paid out to you in cash. The exchange rate is:

10 points = 1 EUR.

Your decisions during the experiment are anonymous. In addition, your earnings will be kept in confidence.

During the experiment, we ask you to turn off your cell phone and to store it out of reach. All documents not related to the experiment (lecture notes, books, etc.) must not be used. Breach of the rules can lead to exclusion from the experiment and all payments.

On the next page, you are given instructions on the experimental procedure. When you are finished reading them, please remain at your seat and wait for the experiment to start.

--- *next page* ---

INSTRUCTIONS ON THE EXPERIMENTAL PROCEDURE

These are instructions to the first part of the experiment. After completing the first part of the experiment, you will be given instructions on the second and third part of the experiment.

<u>Part I</u>

In the first part of the experiment, you will be presented with five pairs of paintings by two different painters. For each pair of paintings, please decide which painting you prefer by clicking the preferred option on-screen. Based on which painter you prefer, you will be then assigned to one of two groups. In addition, you will receive information on the painters and the paintings you have previously seen.

Afterwards, you will be shown two additional paintings and have to guess which of the two previous painters painted which of the two paintings (or both). For each correct answer, you can earn 10 points. To help you decide, you can chat with your group members using a chat program. You will only see messages exchanged within your own group. You only have 10 minutes to discuss the paintings with your group members. Then, the chat will be automatically closed and you have to decide on your own, which of the two artist each painting belongs to. All points earned in part I are added to your account balance.

Please consider the following chat rules: You are not allowed to share any information on your identity or any information compromising your anonymity during the experiment (including your name, age, student ID, telephone number, email address, or your cubicle number). Any agreements on payments outside of the experiment are forbidden. Breach of these rules can lead to an exclusion from all payments and possible from participating in future experiments.

-- *next page* --

INSTRUCTIONS ON THE EXPERIMENTAL PROCEDURE

Part II

The second part of the experiment is a perception and reaction test. The test comprises 80 rounds. In each round, you have to determine the color of a word displayed to you, as fast as possible.

If the word is displayed in the color **Red**, please hit the key **"D"** on your keyboard.
If the word is displayed in the color **Yellow**, please hit the key **"F"** on your keyboard.
If the word is displayed in the color **Blue**, please hit the key **"J"** on your keyboard.
If the word is displayed in the color **Green**, please hit the key **"K"** on your keyboard.

Please note that the color in which the word is displayed counts, and not the content of the word.

For each word, you have at most 5 seconds to respond. If you do not hit any of the four keys (D, F, J or K) within 5 seconds, you earn 0 points for that round.

For each correct key hit (within 5 seconds), you earn 0.5 points for that round. For an incorrect key hit (within five seconds), you earn 0 points for that round. Once you hit one of the four keys (D, F, J or K), you automatically enter the next round. This means, per round you can only attempt once.

In total, there are 80 round. Points earned across all rounds are added to your account balance.

Hint: Ideally, you position your left hand on the keys "D" and "F", and your right hand on the keys "J" and "K".

At the beginning, there will be two test rounds, in which we will give you feedback on whether your response was correct or incorrect. You do not earn any points for these two test rounds.

-- *next page* --

Part II

For the third part of the experiment, you will be randomly matched with another participants from today's experiment. One of you will be Player A, the other one Player B. Whether you are Player A or Player B will be determined randomly.

Player A receives an endowment of 100 points, and has to decide how many points he/she wants to give to Player B. The rest of the points, will be added to Player A's account balance.

Player B has an initial endowment of 0 points. Player B earns in part III only points, given by Player A.

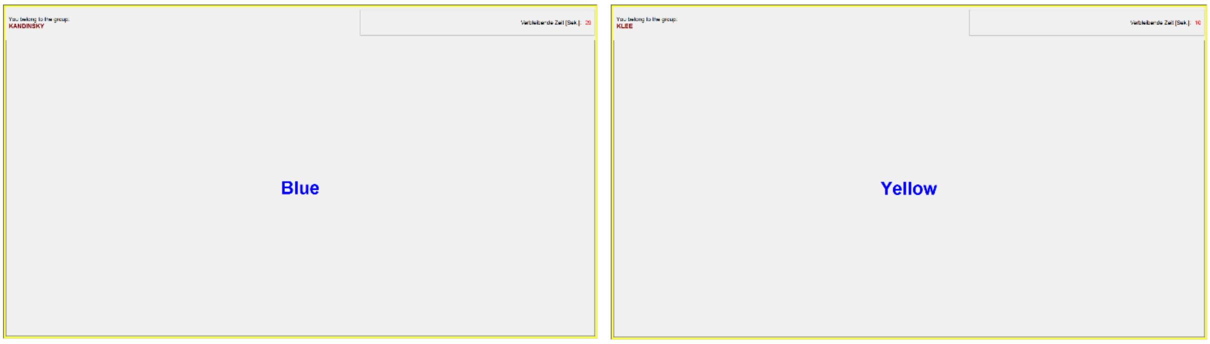
Player A's decision is anonymous, and Player B has no say in it. Player B will only receive information on how many points Player A sent him.

Appendix B. Experimental Screens

Depletion task (Stroop task):

The control group saw 80 congruent trials, where the meaning of the word matched the color the word was displayed in (see Figure 4: on the left). The depletion group faced 80 incongruent trials, whether the meaning of the word did not match the color (see Figure 4: on the right). During the Stroop task, subjects were reminded of their group membership in the upper left corner.

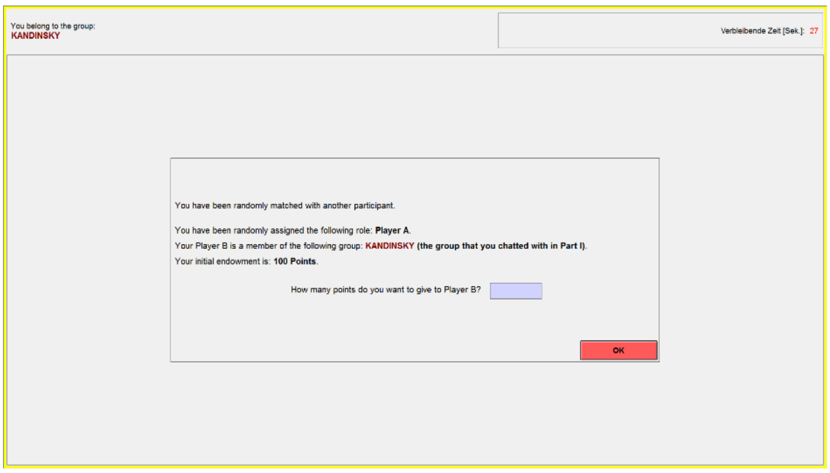
Figure 4: Stroop task for the control group (on the left) and the depletion group (on the right).



Dictator game

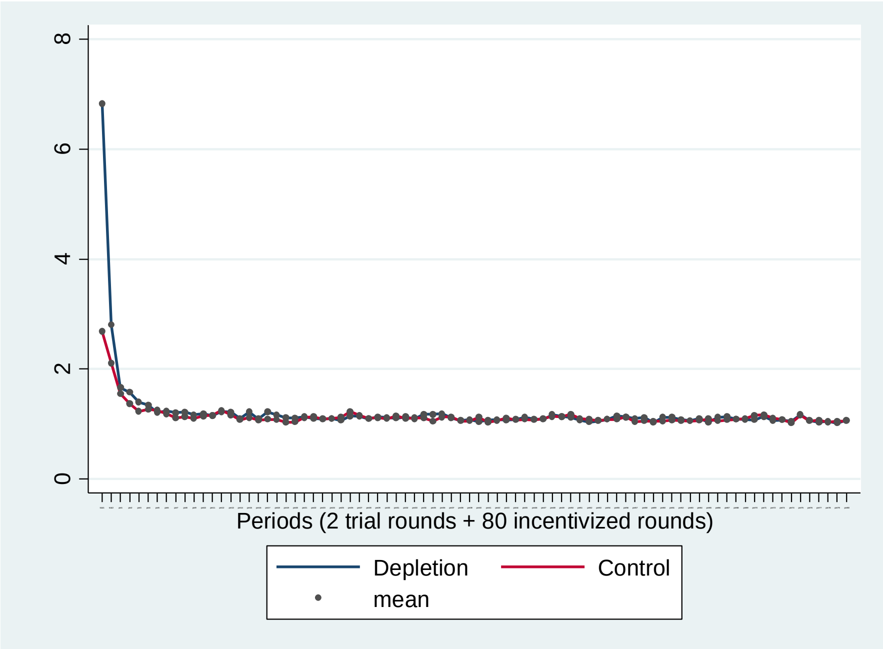
Dictators, when they reached the dictator game stage, were told that they have been randomly matched with another participant and were randomly assigned the role of "Player A". In addition, the group membership of the assigned Player B was displayed (see Figure 5). It also specified whether the group Player B belonged to was the group subjects "have chatted with in part I of the experiment" (ingroup treatment), or the group subjects "have not chatted with in part I" (outgroup treatment). Dictators had to specify, how many of 100 points they want to give to Player B. In the upper left corner, dictators were reminded of their own group membership.

Figure 5: Dictators' screen for the ingroup treatment.



Supplementary Material

Figure 6: Decision times for Stroop task between treatments.



CHAPTER 4:
Games as frames

Joint with Axel Ockenfels [*]

Abstract

We show that economic games *per se* can provide contextual clues and thereby impact behavior. In two laboratory experiments, we examine whether deliberating on trust games versus stag-hunt games without feedback changes cooperation behavior in a subsequent game. First, we find that subjects who play trust games without feedback hold more pessimistic beliefs about other players' cooperation in a subsequent game than subjects who played stag-hunt games without feedback. We also observe that deliberation on trust games versus stag-hunt games accordingly affects behavior in a subsequent, unrelated game. We argue that this is because stag-hunt games align interests between players, whereas trust games pose a conflict of interest between players. Such (mis-)alignments induced by the game *per se* offer clues that affect beliefs and behavior in subsequent games.

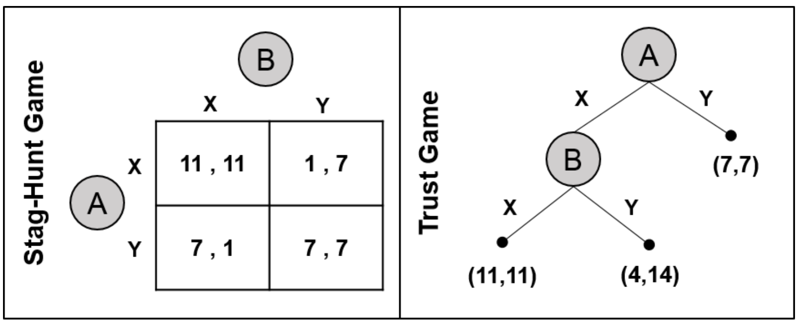
[*] Uta K. Schier and Axel Ockenfels, Department of Economics, University of Cologne, Albert-Magnus-Platz, 50937 Cologne, Germany, e-mail: schier@wiso.uni-koeln.de, ockenfels@uni-koeln.de. We thank participants of the Research Unit seminar in Cologne for helpful comments and feedback, in particular Gary Bolton, Christoph Feldhaus, Wilhelm Hofmann and Thomas Mussweiler. We also thank Tobias Dahint and Lea Mohnen for assistance in the experiment, gratefully acknowledge financial support for the experiments from the German Research Foundation (DFG) through the Research Unit "Design & Behavior: Economic Engineering of Firms and Markets" (FOR 1371), and funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 741409). This paper reflects only the authors' view and the funding agencies are not responsible for any use that may be made of the information it contains.

4.1 Introduction

People's choice depends on how it is presented. This is called framing effect (Deutsch, 1958; Tversky and Kahneman, 1981). One important line of research shows the importance of framing in social decision making. For instance, whether choice is in terms of cooperation or competition, or in a give- or take-frame, can substantially change behavior (Eiser and Bhavnani, 1974; Andreoni, 1995; Hoffman et al., 1996; Larrick and Blount, 1997; Levin et al., 1998; Burnham et al., 2000; Kay and Ross, 2003; Abbink and Hennig-Schmidt, 2006; Branas-Garza, 2007; Gächter et al., 2009; Ellingsen et al., 2012; Engel and Rand, 2014; Cartwright, 2016). For instance, framing a Prisoner's dilemma as a cooperation game, highlighting team decision making, or as a competition game, emphasizing conflicting interests amongst players, affects players' willingness to cooperate (Engel and Rand, 2014). In this and all other cases, the game and its strategy spaces and payoff functions remain unchanged, yet the language that the choices are presented in affects game outcomes. One reason why framing matters is that language provides contextual clues about the applicable social norms and about expected behavior (Dufwenberg et al., 2011; List, 2007).

In this paper, we take a reverse approach. While keeping the language constant, we show that a game with its strategies and payoffs can *per se* provide contextual clues, even when presented in neutral terms, which subsequently affect beliefs and behavior. To illustrate our point, take the trust and the stag-hunt game (as shown in Figure 7), two widely-used games in

Figure 7: Example of a stag-hunt game (left) and a trust game (right).



behavioral economics and, increasingly so, in psychology (e.g. Balliet, Wu and De Dreu, 2014; Martinez and Zeelenberg, 2015).[1] In both games, two players (A and B) each decide between two options (X and Y). In the stag-hunt game, players choose simultaneously and may coordinate either on the payoff-dominant but riskier Nash equilibrium in pure strategies (both players choose X and receive each a payoff of 11), or on the risk-dominant Nash equilibrium in pure strategies, which pays less for both players (both players choose Y and earn 7 each). In the trust game, the first-mover (A) has to decide whether to trust the second-mover (X) or to end the game (Y). In case of trust, the second-mover (B) decides whether to reciprocate the trust (X, in which case both earn 11), or to behave selfishly (Y, in which case B receives 14 and A receives 4). Hence, for payoff-maximizing players, the trust game only knows one Nash equilibrium, in that the first-mover ends the game (and both players receive a payoff of 7). However, in both games, behavior is typically heterogeneous, and all options are chosen with positive probability (e.g., Cooper et al., 1992; Bohnet and Zeckhauser, 2004; Fetchenhauer and Dunning, 2009).

For our purpose, the important difference between these two games is that the stag-hunt game perfectly aligns interests between players, while the trust game creates a conflict of interest in that the first-mover has good reason to distrust the second-mover. We hypothesize that such (mis-)alignments induced by the game *per se* can serve as clues that affect the 'mindset' of players, and thus affect beliefs and behavior in subsequent choice tasks.[2] Games

[1] Schier, Ockenfels and Hofmann (2016) study a related game, the dictator game, which too is increasingly popular both in psychology and economics.

[2] This is related to a literature in cognition research that has shown that human information processing differs under a 'trust state of mind' compared to a 'distrust state of mind'. In a seminal paper, Schul, Mayo and Burnstein (2004, p. 668) conclude from their research that "the cognitive system reacts to distrust by automatically inducing the consideration of incongruent associations — it seems designed to ask 'and what if the information were false?'". Further research shows that distrust increases cognitive flexibility, creativity (Schul et al., 2008; Mayer and Mussweiler, 2011), and more elaborate information processing (Schul et al., 1996), and reduces stereotyping (Posten and Mussweiler, 2013). To our knowledge, the only paper in economics that is related to this topic is Bolton, Feldhaus and Ockenfels (2016) who investigated whether betrayal aversion, typically ascribed to behavior in trust games, can also be found in stag-hunt games. For this, the authors disentangled strategic risk from natural risk in a stag-hunt game experiment, in which participants either play against a computer (natural risk) or a human (strategic risk). To their surprise, the authors found the opposite of betrayal aversion in that cooperation is higher with strategic risk, and they hypothesized (but did not provide

that pose a conflict of interest (such as the trust game, where there is reason to distrust the second-mover) induce more pessimistic beliefs and more cautious behavior than games that align interests (such as the stag-hunt game, where there is no reason to distrust the other player). Specifically, we hypothesize that trustors in trust games form more pessimistic beliefs about others' cooperation in a subsequent game than subjects playing stag-hunt games (which is our Hypothesis 1).[3] Similarly, subjects who deliberate on others' decision to trust subsequently behave more cautiously and cooperate less than subjects who deliberated on others' stag-hunt game behavior (Hypothesis 2). Carefully controlling for potentially confounding learning and spill-over effects, we find statistically and economically strong support for both hypotheses. We conclude that there is no such thing as a 'neutral' game description. Games with aligned vs. misaligned incentives create *per se* social clues that in turn affect how people approach social interaction.

4.2 Methods

Study 1: 258 university students were recruited via ORSEE (Greiner, 2015) and invited to participate in a laboratory experiment in May 2017, programmed with zTree (Fischbacher, 2007). Students came from various disciplines, with a mean age of 23 (*sd* = 2.89) and 57% being female. They earned on average 13.16 EUR for 40 minutes of participation (including a 4.00 EUR show-up fee).

Subjects were randomly allocated to one of four treatments, which specified the content of each of two parts. In the first part of treatments "TT" and "TS", subjects played five different trust games, whereas in the first part of treatments "SS" and "ST", subjects played five different stag-hunt games. In the second part, we elicited subjects' beliefs about other subjects' behavior in a previously conducted experiment. In treatments "TT" and "ST", subjects predicted

evidence) that this might be due to different mindsets that are triggered by the different game forms. Our study's hypotheses were developed behind this background, before we ran our experiments.

[3] While we will focus on first-movers in trust games, because there is no trust involved in second-movers' choices, we will also analyze second-movers' beliefs and behavior.

behavior in a trust game, whereas in treatments "SS" and "TS" subjects predicted stag-hunt game behavior.

For the first part of the experiment, subjects were randomly matched in pairs and assigned either the role of player A or player B. The structure of the games in the first part was similar to the games displayed in Figure 7 but the payoffs varied.[4] Before each game, players were informed of the exact payoff structure and then had to choose between option X and option Y. Player B in the trust game, the second-mover, decided for the case that player A chose X. The outcome of the five games played in the first part was only revealed to subjects at the end of the experiment.

In the second part of the experiment, subjects were informed that another experiment had been conducted some months earlier in the same laboratory with different participants.[5] We told subjects that participants from the previous experiment were randomly matched in pairs and assigned a role of player A or B, and then played the game illustrated in Figure 7. Then subjects had to predict how many players A in the previous experiment chose option X, on a scale from 0 to 100%. Estimations of subjects were incentivized by a quadratic scoring rule (Selten, 1998; Palfrey and Wang, 2009), so that subjects could earn an amount between 0.20€ and 1.00€ depending on the accuracy of their estimation.[6] At the end of the experiment, subjects answered a short questionnaire about their demographic background, before we revealed the outcome of the five games played in the first part and the accuracy of their estimation in the second part. For subjects' final payoff, one of the five games from the first part was randomly selected and paid out, in addition to earnings from the second part.

Study 2: We recruited 294 university students via ORSEE (Greiner, 2015) to participate in a laboratory experiment in September 2017, programmed with zTree (Fischbacher, 2007).

[4] See Figure 12 in Appendix B, for more details on the games.
[5] This data was collected with 134 students in April 2017.
[6] Subjects were incentivized according to the following rule: $\pi = 1 - 0.8 * (estimation - actual\ behavior)^2$.

Students were on average 24 years old (*sd* = 4.11) and 64% were female. They earned on average 13.02 EUR for 40 minutes of participation, including a 4.00 EUR show-up fee.[7] As in Study 1, the experiment consisted of two parts and subjects were randomly allocated to one of four treatments that we again refer to as "TT", "TS", "SS", and "ST". However, in contrast to Study 1, we reversed the experimental design for Study 2: In the first part, subjects estimated behavior in five previously played trust games (treatments "TT" and "TS") or stag-hunt games (treatments "SS" and "ST"). In the second part, subjects now played a game themselves, which was either a trust game (treatments "TT" and "ST") or a stag-hunt game (treatments "SS" and "TS"). For the estimations in the first part, subjects were given instructions analogue to Study 1, and estimated previous behavior in the five games played in Study 1. For the second part, subjects were matched in pairs and assigned the role of player A or B, and then played one of the games from Figure 7. Finally, subjects answered a short questionnaire on demographics, and then received information about the accuracy of their five estimations in the first part, the outcome of the game in the second part, and their final payoffs. For subjects' payoffs, one of the five estimations from the first part was randomly selected and paid out, as well as the game from the second part.

4.3 Results: Study 1

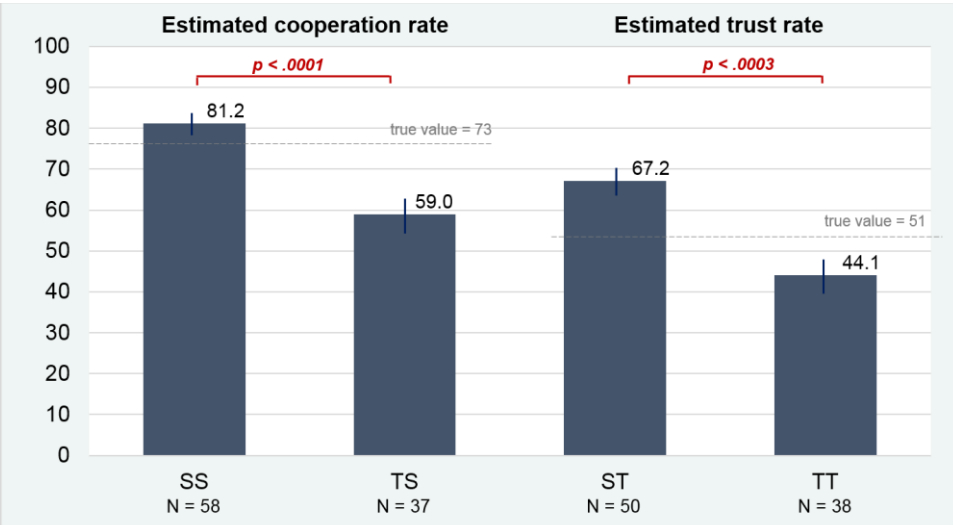
All statistical tests were performed at a 5% significance level, and differences in beliefs between subjects were tested using two-sided Mann-Whitney U tests. For our analyses, we considered all observations from the treatments SS and ST, but only decisions by the first-movers (player A) from treatments TT and TS (in sum: N = 183, with 58 observations in SS, 37 in TS, 50 in ST, and 38 in TT).

[7] Due to an error in the calculations of profits, 26% of subjects (76 out of 294) were paid 10-30 cents too much or too little. However, this does not affect our data, as payoffs were revealed only at the end of the experiment.

Our Hypothesis 1 is strongly confirmed, both with respect to effect size and statistical significance.[8] Beliefs about cooperation in the stag-hunt game (SS vs. TS) are much more optimistic when stag-hunt games were played before compared to trust games (81% vs. 59%, MWU: $p < .0001$; see Figure 8). Similarly, beliefs about trust rates in the trust game (ST vs. TT) are significantly higher when subjects previously played stag-hunt games compared to trust games (67% vs. 44%, MWU: $p < .0003$).[9]

One potential explanation for our games-as-frames effect on beliefs is learning: Subjects in the first part might learn something from being exposed to the five phase 1-games, even when

Figure 8: Games-as-frames effect on beliefs: estimating others' behavior is driven by previous game experience.



Notes. Estimating others' cooperation rate (SS/TS) or trust rate (ST/TT) from 0-100%, depending on own previous game experience with stag-hunt games (SS/ST) or trust games (TS/TT). P-values are from Mann-Whitney U tests, with standard errors bars.

no feedback is given (Rick and Weber, 2010). Since learning (if any) would seem more useful when the game being played in both phases is the same, this might potentially cause our

[8] Here, we only focus on first movers in the trust games in part I. However, we find similar results for second-movers, in that they have more pessimistic beliefs than stag-hunt game players about cooperation in a stag-hunt game (60% vs. 81%, MWU: $p < .0002$), as well as in a trust game (55% vs. 67%, MWU: $p < .03$).

[9] OLS regressions confirm the effect. Playing stag-hunt games compared to trust games in the first part significantly increases beliefs in the second part (ß_0 = 51.5, $p < .001$, ß_1 = 23.2, $p < .001$), and the effect remains significant if we control for which game subjects estimate in the second part. The treatment effect is robust to controlling for variations in the session size, as well as average profit made in the games in the first part (not known to subjects until the experiment is over), and if we control for age, gender, prior game theory knowledge and prior experience with stag-hunt and trust games. See Table 8 in the Supplementary Material for details.

treatment effect.[10] However, we find no evidence that beliefs in the treatments in which subjects predict the same game as they played before (treatments SS and TT) are more accurate. Specifically, we look at how much beliefs deviate from the true value, defined by actual behavior in the pretest. For beliefs about stag-hunt game behavior, we do not find that subjects' beliefs in the treatment SS deviate less from the true value than in TS ($m_{SS} = 20$ vs. $m_{TS} = 24$, MWU: $p < 0.563$). Also, subjects' beliefs in TT do not deviate less from the true value than beliefs in ST ($m_{TT} = 25$ vs. $m_{ST} = 27$, MWU: $p < 0.625$).[11]

We conclude that there is a strong games-as-frames-effect on beliefs, driven by the different incentive structures – aligned vs. misaligned interests – of the games *per se*.

4.4 Results: Study 2

In Study 2, subjects first make five estimations about behavior in previous experiments, and then play a trust or stag-hunt game themselves. All statistical differences in behavior in the games between subjects are tested using two-sided Fisher's exact tests. For our analyses, we consider all observations from the treatments SS and TS, but only decisions by first-movers (player A) in the trust game in treatments ST and TT (in total: N = 205, with 58 observations in SS, 58 in TS, 46 in ST, and 43 in TT).[12]

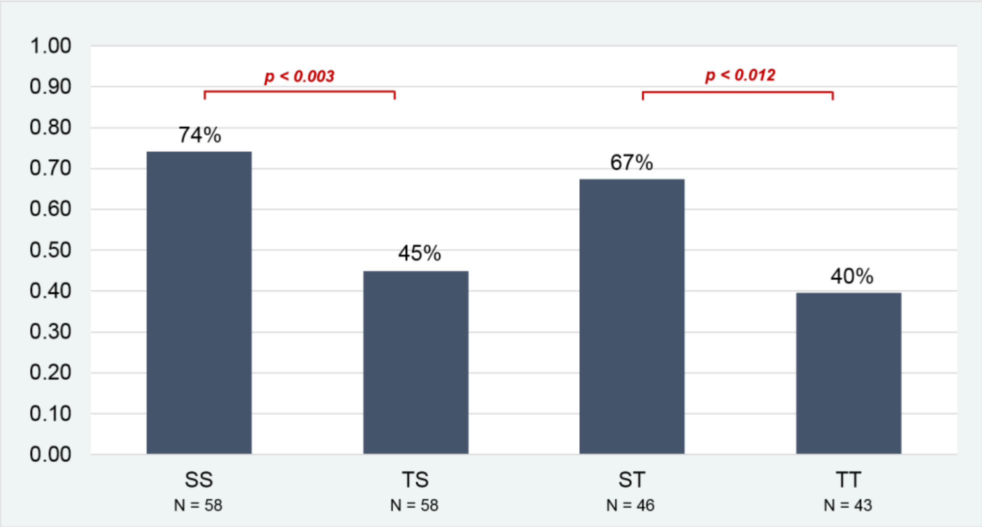
[10] In principle, there might also be spillover effects across different games. Bolton and Ockenfels (2005) find that proposers in repeated ultimatum games learn not only from playing the same game but also from playing other games. Bohnet and Huck (2004) report evidence that experiencing trustworthiness in repeated trust games with a partners matching seems to spill over to trust behavior in a subsequent trust game with strangers matching. Similarly, Schwerter and Zimmermann (2016) show that trust behavior might not be driven by stable preferences but can be affected by experiencing fair und unfair treatment in a previous, unrelated game. Moreover, Peysakhovich and Rand (2016) show that cooperation in infinitely repeated Prisoners' dilemmas can spill over to more prosocial behavior in subsequent ultimatum games, dictator games, trust games and public good games. However, in all these studies, spillover effects in behavior or beliefs rely on feedback about other players' behavior, which in our case is ruled out by design.

[11] Neither is the compound difference significant ($m_{SS+TT} = 22$ vs. $m_{TS+ST} = 26$, MWU: $p < 0.223$).

[12] For second-movers in the trust game, we find that second-movers behave more trustworthy when they previously deliberated on stag-hunt games than on trust games (61% vs. 37%, FET: $p < .022$).

In line with our second hypothesis, we observe that the kind of game subjects estimate in the first part of the experiment strongly affects game behavior in the second part. As illustrated in Figure 9, subjects substantially cooperate more in the stag-hunt game, when they previously deliberated on stag-hunt games than on trust games (74% vs. 45%, FET: $p < .003$). Similarly, first-movers in the trust game trust more, when they previously estimated stag-hunt games compared to trust games (67% vs. 40%, FET: $p < .012$).[13]

Figure 9: Games-as-frames effect on behavior: cooperation rates by players driven by previous game experience.



Notes. Cooperation rate (SS/TS) and trust rate (ST/TT), depending on previous deliberation on stag-hunt games (SS/ST) or trust games (TS/TT). P-values are from Fisher's exact test.

Similar to Study 1, we examined whether there is a learning effect in the sense that subjects in the treatments SS and TT have an advantage over subjects in the treatments ST and TS. However, we find no evidence that the beliefs within the first part become more accurate in TT/SS compared to ST/TS from the first to the fifth game (i.e. accuracy improvement from

[13] Probit regression support the results. Estimating stag-hunt games compared to trust games in the first part significantly increases cooperation in the second part ($ß\_0 = 0.19$, $p < .137$, $ß\_1 = 0.75$, $p < .001$). Moreover, this effect is robust to controlling for the kind of game subjects play in the second part, and to controlling for variations in session sizes, as well as for belief (in-)accuracy in the first part ($ß\_1= -0.55$, $p < .002$), which might be interpreted as how rational subjects' expectations about others are. Lastly, we added age, gender, prior game theory knowledge and prior experience with stag-hunt and trust games as control variables, but, none of these variables significantly determine behavior, nor do they eliminate the games-as-frames effect on behavior. See Table 9 in the Supplementary Material for details.

game 1 to game 5: $m_{SS/TT}$ = -0.04, $m_{ST/TS}$ = -0.03, MWU: p = 0.571).[14] Furthermore, in the second part subjects do not seem to make “better” (i.e. selfish utility maximizing) decisions in treatments SS/TT than in treatments ST/TS. For this, we analyzed whether subjects maximized their expected utility if they had rational expectations about the second-movers’ behavior. In the trust game treatments, 37% of second-movers in TT chose option X, compared to 61% of second-movers in ST. Hence, first-movers in TT with rational expectations about second-movers should prefer option Y ($EU_A(Y) = 7 > EU_A(X) = 6.6 = 0.37*11+0.63*4$), whereas first-movers in ST should prefer option X ($EU_A(Y) = 7 < EU_A(X) = 8.3 = 0.61*11+0.39*4$). If we look at the share of player A that actually choose the preferred option, we do not find that first-movers in TT make more payoff maximizing decisions than in ST (60.5% in TT vs. 67.4% in ST, FET: p = .516). Similarly, in the stag-hunt game treatments players in SS chose on average option X with 74% compared to 45% of players in TS. Thus, players in SS with rational expectations about their partners’ behavior should prefer option X ($EU(Y) = 7 < EU(X) = 8.4 = 0.74*11+0.26*1$), whereas players in TS should prefer option Y ($EU(Y) = 7 > EU(X) = 5.5 = 0.45*11+0.55*1$). However, we do not observe that players in SS choose the ‘better’ option more often than players in TS (74.1% in SS vs. 55.2% in TS, FET: $p < .052$). Consequently, we conclude that there is no evidence that subjects in treatments SS/TT have a learning advantage over subjects in ST/TS.

4.5 Conclusion

It is well-known that the language which describes a given game provides social clues that affect and guide people’s behavior. We show that the game’s strategies and payoffs *per se* also provide such clues. There is no such thing as a ‘neutral’ game. Games with aligned versus misaligned incentives inevitably frame the decision context in different ways, with important consequences for subsequent beliefs and behavior. The different games seem to evoke different

[14] See Supplementary Material for more information on belief accuracy.

social clues, which in turn differently affect information processing modes in terms of a trust or distrust mindset, as suggested by cognition research (Schul et al., 2004). Second-movers in our trust game have more pessimistic beliefs about others after deliberating on misaligned interests of trust games (Study 1). Moreover, they behave less trustworthily (Study 2) even though they do not face any strategic uncertainty. This suggests that trust and distrust mindsets do not only affect beliefs about others' behavior but also preferences. The conclusion is that, at least sometimes, games are inherently intertwined with belief and preference formation in ways not studied before.

References

Abbink, K., and Hennig-Schmidt, H. (2006). Neutral versus loaded instructions in a bribery experiment. *Experimental Economics, 9*(2), 103-121.

Andreoni, J. (1995). Warm glow versus cold prickle – the effects of positive and negative framing on cooperation in experiments. *Quarterly Journal of Economics, 110*, 1-21.

Balliet, D., Wu, J., and De Dreu, C. K. W. (2014). Ingroup favoritism in cooperation: a meta-analysis. *Psychological Bulletin, 140*(6), 1556-1581.

Bohnet, I., and Huck, S. (2004). Repetition and reputation: Implications for trust and trustworthiness when institutions change. *American Economic Review, 94*(2), 362-366.

Bohnet, I., and Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior and Organization, 55*, 467-484.

Bolton, G. E., Feldhaus, C., and Ockenfels, A. (2016). Social interaction promotes risk taking in a stag hunt game. *German Economic Review, 17*(3), 409-423.

Bolton, G. E., and Ockenfels, A. (2005). A stress test of fairness measures in theories of social utility. *Economic Theory, 25*(4), 957-982.

Branas-Garza, P. (2007). Promoting helping behavior with framing in dictator games. *Journal of Economic Psychology, 28*(4), 477-486.

Burnham, T., McCabe, K., and Smith, V. L. (2000). Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavior and Organization, 43*(1), 57-73.

Cartwright, E. (2016). A comment on framing effects in linear public good games. *Journal of the Economic Science Association, 2*, 73-84.

Cooper, R. W., DeJong, D. V., Forsythe, R., and Ross, T. W. (1992). Communication in coordination games. *Quarterly Journal of Economics, 107*(2), 739-771.

Deutsch, M. (1958). Trust and suspicion. *Journal of Conflict Resolution, 2*, 265-279.

Dufwenberg, M., Gächter, S., and Hennig-Schmidt, H. (2011). The framing of games and the psychology of play. *Games and Economic Behavior, 73*, 459-478.

Eiser, J. R., and Bhavnani, K. (1974). The effect of situational meaning on the behavior of subjects in the Prisoner's Dilemma Game. *European Journal of Social Psychology, 4*, 93-97.

Ellingsen, T., Johannesson, M., Mollerstrom, J., and Munkhammer, S. (2012). Social framing effects: Preferences or beliefs? *Games and Economic Behavior, 76*, 117-130.

Engel, C., and Rand, D. G. (2014). What does ‘‘clean’’ really mean? The implicit framing of decontextualized experiments. *Economics Letters, 122*, 386-389.

Fetchenhauer D., and Dunning, D. (2009). Do people trust too much or too little. *Journal of Economic Psychology, 30*, 263-276.

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

Gächter, S., Orzen, H., Renner, E., and Starmer, C. (2009). Are experimental economists prone to framing effects? A natural field experiment. *Journal of Economic Behavior and Organization, 70*, 443-446.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114-125.

Hoffman, E., McCabe, K., and Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games. *American Economic Review*, *86*(3), 653-660.

Kay, A. C., and Ross, L. (2003). The perceptual push: The interplay of implicit cues and explicit situational construals on behavioral intentions in the Prisoner's dilemma. *Journal of Experimental Social Psychology*, *39,* 634-643.

Larrick, R. P., and Blount, S. (1997). The claiming effect: Why players are more generous in social dilemmas than in ultimatum games. *Journal of Personality and Social Psychology, 72*(4), 810-825.

Levin, I. P., Schneider, S. L., and Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, *76*(2), 149-188.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, *115*(3), 482-493.

Martinez, L. F., and Zeelenberg, M. (2015). Trust me (or not): Regret and disappointment in experimental economic games. *Decision, 2*(2), 118-126.

Mayer, J., and Mussweiler, T. (2011). Suspicious spirits, flexible minds: when trust enhances creativity. *Journal of Personality and Social Psychology*, *101*, 1262-1277.

Palfrey, T. R., and Wang, S. W. (2009). On eliciting beliefs in strategic games. *Journal of Economic Behavior & Organization, 71*, 98-109.

Peysakhovich, A., and Rand, D. G. (2016). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science, 62*(3), 631-647.

Posten, A.-C., and Mussweiler, T. (2013). When distrust frees your mind: the stereotype-reducing effects of distrust. *Journal of Personality and Social Psychology*, *105*, 567-584.

Rick, S., and Weber, R. A. (2010). Meaningful learning and transfer of learning in games played repeatedly without feedback. *Games and Economic Behavior*, 68, 716-730.

Schier, U. K., Ockenfels, A., and Hofmann, W. (2016). Moral values and increasing stakes in a dictator game. *Journal of Economic Psychology, 56*, 107-115.

Schul, Y., Burnstein, E., and Bardi, A. (1996). Dealing with deceptions that are difficult to detect: encoding and judgment as a function of preparing to receive invalid information. *Journal of Experimental Social Psychology, 32*, 228-253.

Schul, Y., Mayo, R., and Burnstein, E. (2004). Encoding under trust and distrust: The spontaneous activation of incongruent cognitions. *Journal of Personality and Social Psychology, 86*(5), 668-679.

Schul, Y., Mayo, R., and Burnstein, E. (2008). The value of distrust. *Journal of Experimental Social Psychology, 44*, 1293-1302.

Schwerter, F., and Zimmermann, F. (2016). How stable is trust? The case of personal experience of unfair and fair treatment (Doctoral Dissertation). University of Bonn.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics, 1*, 43-62.

Tversky, A., and Kahneman, D. (1981). The framing of decisions and the psychology of play. *Science, 211*, 453-458.

Supplementary Material

Table 8: Determinants of beliefs about others' behavior in previous experiments (Study 1).

Dependent = beliefs about previous cooperation rate	**(1)**	**(2)**	**(3)**	**(4)**
Trust mindset (playing SHGs)	23.2*** (4.00)	22.6*** (3.86)	23.4*** (4.46)	26.0*** (4.45)
Estimating previous SHG		14.4*** (3.80)	14.7*** (3.91)	13.3*** (3.87)
Session size			-0.2 (0.38)	-0.2 (0.38)
Mean profit from games			0.0 (0.90)	0.1 (0.89)
Game theory knowledge				-12.2** (4.27)
Prior participation in similar studies				3.3 (4.38)
Female				-1.7 (3.93)
Age				1.2 (0.66)
Constant	51.5*** (3.07)	44.4*** (3.51)	47.7*** (8.14)	22.4 (18.46)
Observations	183	183	183	183
R-squared	0.16	0.22	0.22	0.27

Notes. Results from OLS regressions with standard errors in parentheses. Trust mindset = dummy variable for playing stag-hunt games in the first part (baseline: trust game); Estimating previous SHG = dummy variable for estimating behavior of previous stag-hunt game players in second part (baseline = of trust game players); Mean profit = average profit from playing games in the first part (revealed to subjects after belief elicitation); Game theory knowledge = dummy variable with no previous knowledge of game theory as baseline; Prior participation = dummy variable with no prior experience in playing stag-hunt or trust games as baseline; Female = dummy variable with male as baseline; Age = continuous self-reported variable. P-values: *** ≤ .001, ** ≤.01, * ≤.05.

Table 9: Determinants of cooperation behavior (Study 2).

Dependent = cooperation in stag-hunt or trust game	**(1)**	**(2)**	**(3)**	**(4)**
Trust mindset (estimating SHGs)	0.75*** (0.18)	0.75*** (0.18)	0.64*** (0.19)	0.61*** (0.19)
Playing a subsequent SHG		0.17 (0.18)	0.30 (0.23)	0.40 (0.23)
Session size			-0.01 (0.38)	-0.01 (0.38)
Belief inaccuracy			-0.55*** (0.16)	-0.57*** (0.17)
Game theory knowledge				-0.47* (0.21)
Prior participation in similar studies				-0.15 (0.20)
Female				0.11 (0.20)
Age				-0.01 (0.02)
Constant	-0.19 (0.13)	-0.28 (0.16)	0.62 (0.19)	1.03 (0.83)
Observations	205	205	205	205
Pseudo R-squared	0.06	0.06	0.11	0.13

Notes. Results from Probit regressions with standard errors in parentheses. Trust mindset = dummy variable for estimating others' behavior in previous stag-hunt games in the first part (baseline: previous trust games); Playing a subsequent SHG = dummy variable for playing a stag-hunt game in the second part (baseline = a trust game); Session size = controlling for number of players in each session; Belief inaccuracy = sum of belief deviations in first part; Game theory knowledge = dummy variable with no previous knowledge of game theory as baseline; Prior participation = dummy variable with no prior experience in playing stag-hunt or trust games as baseline; Female = dummy variable with male as baseline; Age = continuous self-reported variable. P-values: *** ≤ .001, ** ≤.01, * ≤.05.

Study 2: Analysis of belief inaccuracy in the first part

Subjects in Study 2 estimated previous stag-hunt game behavior more accurately than previous trust game behavior ($average\_inaccuracy_{SS/ST}$ = 21, $average\_inaccuracy_{TT/TS}$ = 27, MWU: $p < .0001$). When we regress behavior in the second part on the inaccuracies of beliefs in the first part, a Probit regression suggests that less accurate beliefs decrease cooperation in the second part (β_0= 0.99, $p < .001$, β_1 = -0.67, $p < .001$). Thus, it seems that the better subjects are in estimating other players' behavior in previous games, the more cooperative they behave afterwards in a stag-hunt or trust game. Moreover, the accuracy of beliefs significantly interacts with our games-as-frames effect, in the sense that less accurate beliefs in the first part seem to decrease the games-as-frames effect (β_0= -0.31, $p < .389$, β_{belief_inacc} = 0.09, $p < .718$, $\beta_{estimatingSHGs}$ = 2.09, $p < .001$, $\beta_{interaction\_effect}$ = -1.22, $p < .002$).

Appendix A1. Experimental Instructions for Study 1

GENERAL INSTRUCTIONS

Welcome to the experiment!

Please read the following instructions carefully and from now on, please do not communicate with other participants. If you have any questions during the experiment, please raise your hand and we will come to you and answer your questions.

In this experiment, you can earn money. How much you can earn, depends on your decisions and the decisions of other participants. Regardless of this, you will receive 4 EUR for showing-up to the experiment. At the end of the experiments, all earnings from the experiment will be added, and paid out to you in EUR.

Your decisions during the experiment are anonymous. In addition, your earnings will be kept in confidence.

During the experiment, we ask you to turn off your cell phone and to store it out of reach. All documents not related to the experiment (lecture notes, books, etc.) must not be used. Breach of the rules can lead to exclusion from the experiment and all payments.

On the next page, you are given instructions on the experimental procedure. The experiment starts, as soon as all participants have read and understood the instructions.

INSTRUCTIONS ON THE EXPERIMENTAL PROCEDURE

In this experiment, there are two types of participants: player A and player B. At the beginning, it will be determined randomly whether you are player A or player B and you will be informed about this.

Every player A will be randomly assigned a player B. Both players, each have to choose between two options, X or Y. [*added for treatments TT/TS:* First, player A decides. Then, player B decides, without knowing player A's decision.]

For each pair of players, payoffs depend on which options both players choose.

For example:

Payoff table *[for treatments SS/ST]*	
If both player choose option X:	Both players get …€
If player A chooses option X, and player B chooses option Y:	A gets …€ and B gets …€
If player A chooses option Y, and player B chooses option X:	A gets …€ and B gets …€
If both player choose option Y:	Both players get …€

Payoff table *[for treatments TT/TS]*	
If player A chooses option Y (independent of player B's choice):	Both players get …€
If both player choose option X:	Both players get …€
If player A chooses option X, and player B chooses option Y:	A gets …€ and B gets …€

The exact amounts paid […€] vary per round. In total, there are five rounds.

At the beginning of each round, you will see a payoff table, as the one displayed above, that will inform you about the exact amounts paid. After that, you have to decide whether you want to choose option X or option Y.

The decisions from your partner will only be revealed to you at the end of the experiment. This means between rounds, you do not receive any information about your payoff from the previous round.

For your earnings, one round will be randomly selected at the end of the experiment.

At the end of the experiment, you will receive another question. You will see a similar payoff table to the one above and have to estimate how players from a previous experiment behaved in this game. This experiment was conducted last month in this laboratory with different participants.

For your estimation you have to choose an integer between 0 and 100%. Your payoff for your estimation depends on how accurate your estimation was:

$$Your\ Payoff = 1 - 0{,}8 * Deviation^2$$

The deviation is the difference between your estimation and the actual behavior of players from the previous experiment:

$$Deviation = |Your\ Estimation - Actual\ Behavior|$$

This means that the better you estimate the behavior of the previous players, the higher your payoff will be. Your payoff will be between 0.20€ and 1.00€.

Finally, we ask you to fill out a short questionnaire. After that you will learn about the results of the five games, as well as about how accurate your estimation was. At the end, you will receive an overview of your total earnings from the experiment.

If you have any questions about the instructions, please raise your hand and we will come to you.

Appendix A2. Experimental Instructions for Study 2

GENERAL INSTRUCTIONS

Welcome to the experiment!

Please read the following instructions carefully and from now on, please do not communicate with other participants. If you have any questions during the experiment, please raise your hand and we will come to you and answer your questions.

In this experiment, you can earn money. How much you can earn, depends on your decisions and the decisions of other participants. Regardless of this, you will receive 4 EUR for showing-up to the experiment. At the end of the experiments, all earnings from the experiment will be added, and paid out to you in EUR.

Your decisions during the experiment are anonymous. In addition, your earnings will be kept in confidence.

During the experiment, we ask you to turn off your cell phone and to store it out of reach. All documents not related to the experiment (lecture notes, books, etc.) must not be used. Breach of the rules can lead to exclusion from the experiment and all payments.

On the next page, you are given instructions on the experimental procedure. The experiment starts, as soon as all participants have read and understood the instructions.

INSTRUCTIONS ON THE EXPERIMENTAL PROCEDURE

This experiment consists of two parts.

In the first part, you are asked to make estimations about how players behaved in five previous experiments. These experiments were conducted last month in this laboratory with different participants.

In each experiment, there were two types of participants: player A and player B. Types were randomly determined and participants were informed about their type at the beginning of the experiment. Every player A was randomly assigned a player B. Both players, each had to choose between two options, X or Y. [*added for treatments TT/TS:* First, player A decided. Then, player B decided, without knowing player A's decision.]

For each pair of players, payoffs depended on which options both players chose:

For example:

Payoff table *[for treatments SS/ST]*	
If both player choose option X:	Both players get …€
If player A chooses option X, and player B chooses option Y:	A gets …€ and B gets …€
If player A chooses option Y, and player B chooses option X:	A gets …€ and B gets …€
If both player choose option Y:	Both players get …€

Payoff table *[for treatments TT/TS]*	
If player A chooses option Y (independent of player B's choice):	Both players get …€
If both player choose option X:	Both players get …€
If player A chooses option X, and player B chooses option Y:	A gets …€ and B gets …€

Your task is to estimate how player A in five of these previous experiments behaved. The exact amounts paid […€] varied between the five experiments. Before each estimation, you will see a payoff table that will inform you about the exact amounts paid.

For each estimation, you have to choose an integer between 0 and 100%. Your payoff for an estimation depends on how accurate your estimation was:

$$Your\ Payoff = 1 - 0{,}8 * Deviation^2$$

The deviation is the difference between your estimation and the actual behavior of players from the previous experiment:

$$Deviation = |Your\ Estimation - Actual\ Behavior|$$

This means that the better you estimate the behavior of the previous players, the higher your payoff will be. Your payoff can be between 0.20€ and 1.00€.

For your earnings, one of the five estimations will be randomly selected at the end of the experiment.

In the second part of the experiment, you will play a game yourself. Again, there will be player A and player B. You will be informed about your type at the beginning of the second part. Types are randomly determined, and each player A is randomly assigned a player B. Both, player A and B, have to choose between two options, X and Y. For each pair of players, payoffs depend on which options both players choose. For this, you will see a payoff table, similar to the one illustrated above that will inform you about the exact payoffs for the second part.

\- -

Finally, we ask you to fill out a short questionnaire. After that you will learn about the results of the game in the second part, as well as about how accurate your estimations in the first part were. At the end, you will receive an overview of your total earnings from the experiment.

If you have any questions about the instructions, please raise your hand and we will come to you.

Appendix B. Experimental Screens

Figure 10: Exemplary screens for the first part of Study 1 (left: treatments SS and ST, right: treatments TT and TS).

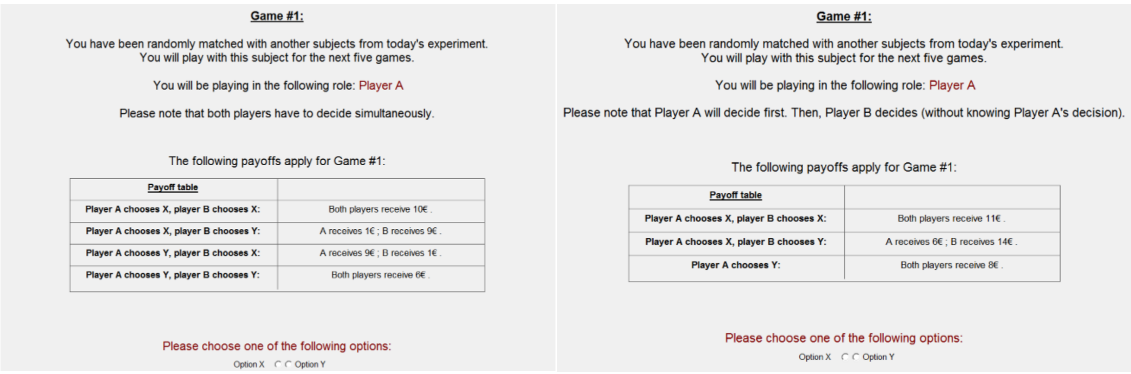
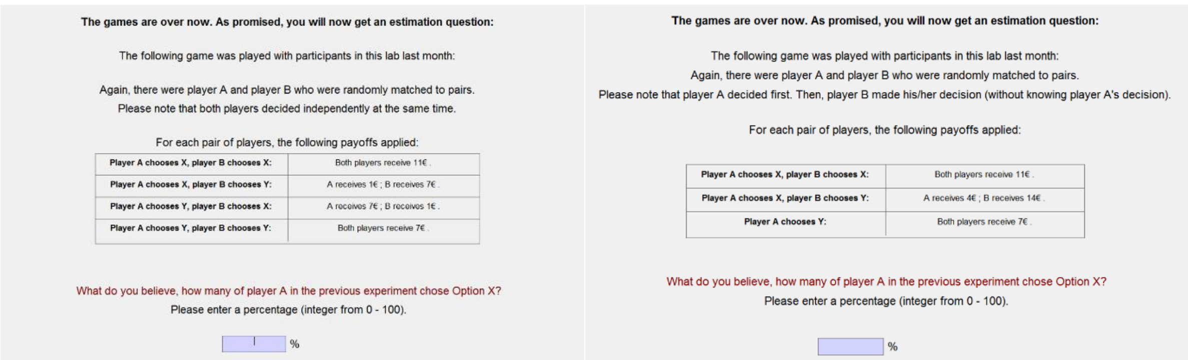
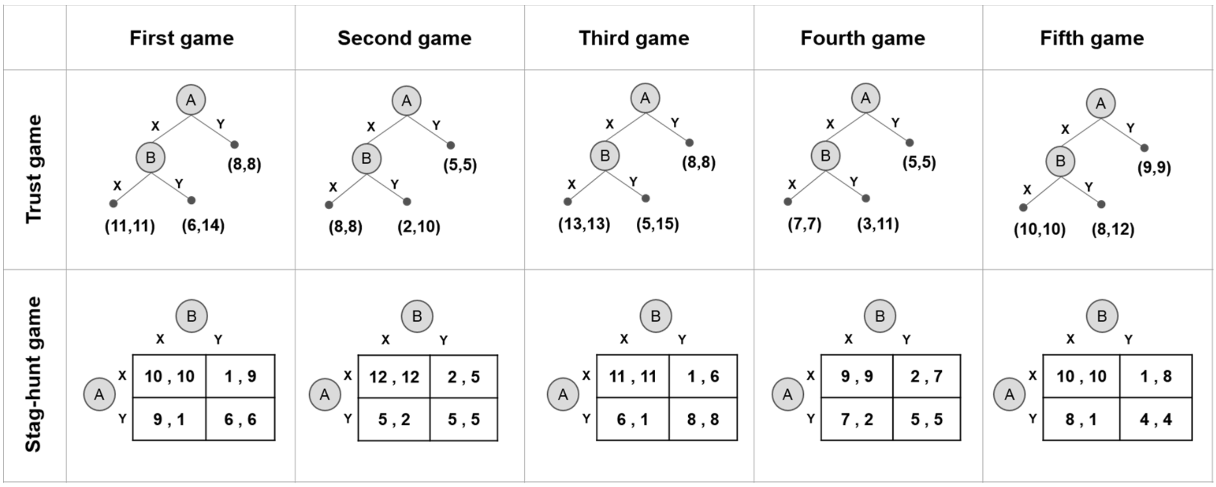


Figure 11: Exemplary screens for the second part of Study 1 (left: treatments SS and TS, right: treatments TT and ST).



For Study 2, the zTree-screens looked similar, but the order of the two parts was reversed. In the first part, subjects estimated behavior from the five games played in part I of Study 1. In the second part, subjects made a choice for the games displayed in Figure 7.

Figure 12: Exact payoffs for the games played in the first part of Study 1 and estimated in the first part of Study 2.



CHAPTER 5:
Female and male role models and competitiveness*

Abstract

We investigate how role models affect the competitiveness of men and women. A strand of experimental literature shows that women shy away from competition. In a laboratory experiment, we test how observing a woman or a man competing and succeeding (a female or a male role model) affects women's and men's competitiveness. We find that a female role model increases women's self-confidence and their competitiveness. In comparison, men's competitiveness is not affected by a male or female role model, nor do women significantly respond to male role models.

* Uta K. Schier, Department of Economics, University of Cologne, Albert-Magnus-Platz, 50937 Cologne, Germany, e-mail: schier@wiso.uni-koeln.de. Financial support of the German Research Foundation (DFG) through the Research Unit "Design & Behavior: Economic Engineering of Firms and Markets" (FOR 1371) is gratefully acknowledged. I thank Sebastian Berger, Gary Bolton, Kevin Breuer, Lea Cassar, Christoph Feldhaus, Uri Gneezy, Christiane Hartog, Kiryl Khalmetski, Felix Kölle, Axel Ockenfels, Peter Werner and seminar participants for helpful comments, and Tobias Dahint and Lea Mohnen for research assistance.

5.1 Introduction

Role models might inspire young people to seek more ambitious goals. This might be particularly interesting nowadays, as a number of powerful female leaders has emerged around the world (e.g. Angela Merkel, Christine Lagarde or Federica Mogherini). Simultaneously, a growing experimental literature suggests that besides differences in human capital (Blau and Kahn, 2000), discrimination or stereotypes (Reuben et al., 2014), women themselves seem to shy away from competitive environments (Niederle and Vesterlund, 2011). As such, the question arises whether successful women can serve as role models for other women which is still an open question in the literature. The goal of this paper is to address this question by providing experimental evidence regarding the importance of role models in encouraging competitiveness. In particular, how observing a woman competing and succeeding (i.e. a female role model) affects women's and men's competitiveness and how this differs from observing a similar male role model. In addition, we examine how people update their beliefs about own ability and others' ability in face of a female or male role model.

We test the importance of role models, and the gender interaction between the role model and the observer, in a laboratory experiment in which subjects are randomly assigned either to a control treatment with no role model, a female-role-model treatment or a male-role-model treatment. We measure subjects' willingness to compete (Saccardo et al., 2017) in a simple math task (Niederle and Vesterlund, 2007) in response to information about the competitiveness of a successful woman or man. We find that women's confidence regarding their ability, as well as their willingness to compete, increases after observing a female role model. We do not observe any other effect of role models on behavior: women do not change their beliefs or behavior after observing a male role model, and men do not respond to either male or female role models. Our findings are important, as the effect of role models on behavior, and not only on aspirations, has hardly been studied in economics.

5.2 Related literature

5.2.1 Role models in economics

The economic literature suggests that sharing a common identity, for example gender, can affect risk, time and other economic preferences (Akerlof and Kranton, 2000; Benjamin et al., 2010; Bertrand et al., 2015). Following this research, and given the underrepresentation of women in STEM subjects and leadership positions, a couple of economic studies investigated the effect of female role models on career choices of women. Bettinger and Long (2005) present experimental evidence that female lecturers at university positively influence course selection of female students in some disciplines. Carrell et al. (2010) show that in the US Air Force Academy female professors increase the likelihood of female students to major in STEM subjects. Similarly, Porter and Serra (2017) find that exposing university students in principles of economics classes to successful female alumni increases the likelihood of female students to major in economics.

Outside of an educational environment, experimental evidence from the field shows that having female leaders in politics can increase career expectations of young girls (Beaman et al., 2012). In all of these studies, the authors argue that the exposure to female role models might encourage other women to seek more challenging and competitive jobs, particularly in math-related domains that are associated with gender stereotypes (see Alan et al., 2004; Guiso et al., 2008; Nosek at al., 2009; Coffman, 2014; etc.).

This is the starting point for our study. We aim to test how role models affect competitiveness in a controlled laboratory environment (Niederle and Vesterlund, 2007; 2011). The main hypothesis we test is whether female role models increase other women's competitiveness. In addition, we study how male role models affect women's competitiveness, how men react to female and male role models with regard to their competitiveness, and how women and men update their beliefs about their own absolute ability and others' absolute ability when they see a female or male role model. In that, our study is closely related to the one by

Meier, Niessen-Ruenzi and Ruenzi (2017) who examine a similar research question but use a less controlled and precise study design.[1] Their study was independently developed.

5.2.2 Gender gap in competitiveness

An array of experimental literature has shown that women shy away from competitive environments, whereas men seek competitive settings (Gneezy et al., 2003; Niederle and Vesterlund, 2007). Laboratory experiments (Gupta et al., 2011) as well as field experiments (Flory et al., 2015) robustly show that women avoid competition even after controlling for risk preferences, ability and confidence. Overall, the literature suggests that high-performing women do not enter competition often enough and low-performing men enter it too often (Niederle and Vesterlund, 2007). [2] This gender gap seems to emerge in young children at the age of three (Sutter and Glätzle-Rützler, 2015), and to be partially socially-learned and driven by culture (Gneezy et al., 2009). These differences in preferences for competition have been shown to predict labor market outcomes, such as annual earnings or specialization on STEM subjects in high school (Buser et al., 2014; Reuben et al., 2017).

Different institutional changes have been tested to decrease the gender gap: quotas or preferential treatment reduce the gap without hampering efficiency (Balafoutas and Sutter, 2012); feedback on relative performance deters low-performing men from entering competition (Wozniak et al., 2014); and advisors recommending subjects to enter the tournament based on their relative performance decreases the gap (Brandts et al., 2015). We seek to complement this

[1] Whereas we stay in the paradigm of Niederle & Vesterlund (2007) by relying on a controlled laboratory environment, Meier et al. (2017) use an online sample (Amazon Mturk), in which subjects are allowed to use calculators which confounds math ability. Further, Meier et al. (2017) focus on the binary decision to enter or not to enter competition. We use a more precise quasi-continuous measure of competitiveness (Saccardo et al., 2017) which allows us to learn more about the distribution of competitiveness by gender and to detect possible differences in the intensity of preferences for competition by women and men. Lastly, Meier et al. (2017) do not study possible channels like belief updating and do not state complete hypotheses for their experiment. Our study contributes by measuring how beliefs about own ability and others' ability are updated before and after seeing a female or male role model. Nevertheless, the results by Meier et al. (2017) are in part consistent with our findings, in that a female role model increases competitiveness of women.

[2] In their study, Niederle and Vesterlund (2007) suggest that 57% of the gender gap in competitiveness can be explained by overconfidence and risk aversion, whereas ability does not seem to explain competitiveness.

literature by investigating how female and male role models affect the gender gap in competitiveness and how they influence belief updating.

5.3 Methods and hypotheses

5.3.1 Sample and experimental design

258 students from the University of Cologne were recruited via ORSEE (Greiner, 2015) and participated in a lab experiment in October and November 2017, programmed with z-Tree (Fischbacher, 2007). Students came from various disciplines, with a mean age of 25 (*sd* = 6.1). Sessions were gender-balanced, with an equal share of female and male participants across treatments. Subjects earned on average 10.15 EUR for 40 minutes of participation, including a 4 EUR show-up fee.

The experiment consisted of three treatments, a *Baseline*, a *Female Role Model* treatment (hereafter: *FRM*), and a *Male Role Model* treatment (hereafter: *MRM*), to which subjects were randomly allocated. The *Baseline* consisted of three rounds in which subjects had to perform simple math tasks under different compensation schemes (Niederle and Vesterlund, 2007; Apicella et al., 2017).[3] The timeline of tasks and the compensation schemes are outlined in Figure 13. In each round, subjects had to add as many sets of five two-digit numbers in three minutes as possible. Subjects were not allowed to use calculators but scratch paper was provided. In round 1, participants were incentivized at a piece-rate of 50 Cent per correctly solved math question. In round 2, subjects were anonymously and randomly matched in pairs and whoever solved more question within a pair received 150 Cent per correctly solved question, whereas the other one received nothing. In case of a tie, the winner was determined randomly. The winner of round 2 was only revealed at the end of the experiment. Before the start of round 3, subjects could choose how they want to be incentivized for their performance (our measure of *competitiveness*). For this, we endowed them with 100 tokens which they could

[3] See Appendix A for written instructions for the experiment.

Figure 13: Timeline of tasks and compensation schemes

Baseline	**Additional information in the FRM [MRM] treatment**
Priors: Ex ante beliefs about own absolute ability*	
Round 1: Three-minute addition task – Piece-rate (€0.5)	
Posterior1a: Beliefs about own absolute ability in Round 1* Posterior1b: Beliefs about others' average ability in Round 1*	
Round 2: Three-minute addition task – Tournament (€1.5, winner takes all)	
Posterior2a: Beliefs about own absolute ability in Round 2* Posterior2b: Beliefs about others' average ability in Round 2*	
Round 3: First, choice of compensation with token allocation (each token in PR = €0.005; each token in T = €0.015).	Example: This is how a woman [man] from a previous experiment decided: 30 token in PIECE-RATE and 70 token in TOURNAMENT. This woman [man] was one of the most successful (top 15% of highest-earning women [men]) in her [his] experiment, and solved more tasks than her opponent in Round 3.
Then, three-minute addition task	
Posterior3a: Beliefs about own absolute ability in Round 3* Posterior3b: Beliefs about others' average ability in Round 3*	
Risk measure Questionnaire	
Feedback and payment	

* €1 for a correct belief estimation for one randomly selected belief

split between a *piece-rate* incentive scheme (as in round 1) and a *tournament* incentive scheme (as in round 2). If they wanted to be incentivized as in round 1, they should allocate all 100 tokens to *piece-rate*. If they wanted to be incentivized as in round 2, they should allocate all 100 tokens to *tournament*. If they wanted to be incentivized by both incentives schemes, they could split the tokens as preferred. Each token allocated to the *piece-rate* scheme was converted to 0.5 Cent per correctly solved task in round 3.[4] On top, each token allocated to the *tournament* scheme was converted to 1.5 Cent per correctly solved task, if subjects solved more tasks in round 3 than their partner in round 2, otherwise they received nothing.[5] In case of a tie, the winner of the tournament in round 3 was again determined randomly. Importantly, subjects' performance in round 3 was compared to the round 2 performance of their partner to avoid any externalities from subjects' choice of incentive scheme. At the end of the study, we elicited subjects' risk preferences with a simple investment decision in which subjects could earn an

[4] This means 100 tokens in *piece-rate* corresponds to the incentive scheme of round 1 (50 Cent per correct task).
[5] 100 tokens in *tournament* correspond to the incentive scheme of round 2 (150 Cent per correct task, if more than the partner).

additional amount between 0 and 2 EUR (Gneezy and Potters, 1997), and conducted a short post-hoc questionnaire including demographics. Then, subjects received feedback on their performance, the outcome of the rounds and the investment decision, and their respective payoffs. For their payoffs, one of the three rounds was randomly selected and paid out, in addition to what subjects earned for their investment decision. On top, to measure how beliefs about own ability and others' average ability are affected by role models, we collected incentivized belief measures throughout the experiment. Before the start of round 1, we asked subjects how many of the math tasks they think they can solve within three minutes (*prior*). And after each round, we asked subjects how many tasks they think they solved in the last round (*posteriors1a-3a*), and how many tasks they expect other subjects in the session to have solved in the last round (*posterior1b-3b*). Importantly, we did not give subjects any feedback on their performance during the experiment. At the end of the experiment, one of the seven belief measures was randomly chosen and subjects earned an additional 1 EUR for a correct estimation.

Both role model treatments followed the same procedure as the *Baseline*, only that we presented subjects with additional information before they made their choice of payment for round 3.[6] In the *FRM* treatment, we showed them an example of how a woman from a previous experiment allocated the tokens, which was 70 tokens in the *tournament* scheme and 30 tokens in the *piece-rate* scheme. In addition, we told subjects that this woman was among the most successful in her experiment (top 15% of highest earning women) and that she solved more tasks in round 3 than her partner. In the *MRM* treatment, we described a man who made the same choice, was among the top 15% of highest-earning men in his experiment and solved more tasks than his partner. Both observations were identified from the *Baseline* treatment.

[6] See Appendix B for the experimental screens presented to participants.

5.3.2 Hypotheses

Our experiment was guided by the following hypotheses. For our *Baseline* treatment, we expect to replicate previous findings of a gender gap in competitiveness following Saccardo et al. (2017) and Brandts et al. (2015). For the *FRM* and *MRM* treatments, our hypotheses depend on how subjects update information on a successful and competitive, previous player. As we do not reveal any information on ability of this previous player, but solely on his/her competitiveness and the outcome of the competition, subjects do not directly receive feedback on their own relative ability. However, considering the gender of the role model, subjects might infer something about their own ability, as well as average female and average male ability in the task. Thus, our hypotheses depend on whether subjects update information on the role model gender-dependently or not. In other words, they depend on whether subjects believe that female and male ability in a given task correlate perfectly, in which case the role models bears no relevant information other than that someone previously competed and succeeded.

From the previous literature, it is not clear how much male and female ability in the given math task correlate. Some studies find no significant gender gap in performance in the task under piece-rate or tournament incentives (e.g. Niederle and Vesterlund, 2007; Kamas and Preston, 2012; Wozniak et al., 2014; Brandts et al., 2015; Sutter and Glätzle-Rützler, 2015; Buser et al., 2016; Apicella et al., 2017; Kölle, 2017), whereas other studies find that men significantly outperform women in the task (e.g. Balafoutas and Sutter, 2012; Buser et al., 2014; Almas et al., 2016; Baldiga and Coffman, 2016). Based on these findings, we argue that it is reasonable that people have heterogeneous expectations about how men and women perform in the task. Then, depending on what subjects ex ante believe about male and female ability in this task, two different hypotheses emerge: either subjects update gender-independently when they see a female or male role model because they believe that female and male ability in the task correlate perfectly. This implies that subjects' beliefs about own or other's ability and their preferences for competition should not be different in the role model treatments than in the baseline, which is our general null hypothesis. An alternative hypothesis is that subjects update gender-dependently when they see a female or male role model because they believe female and male ability does not correlate perfectly. This implies that subjects increase their beliefs about own ability after observing a same-sex role model, as well as subjects increase their beliefs about average female or male ability, depending on the gender of the role model. Based on this belief updating, subjects should then adjust their competitiveness accordingly, which results in the following hypotheses:

Hypothesis 1 (*Women in FRM)*: Women's competitiveness in *FRM* is higher than in the *Baseline*.

If women consider the gender of the female role model, they should increase their beliefs about own absolute ability, given that they are also female. In addition, they should increase their beliefs about average female ability in the task. We argue that the net effect of this belief updating should be that women increase their competitiveness after seeing a female role model.

Hypothesis 2 (*Men in FRM)*: Men's competitiveness in *FRM* is lower than in the *Baseline*.

Men that consider the gender of a female role model, do not learn anything about own ability but should increase their beliefs about average female ability, which should result in lower competitiveness.

Analogue to the argumentation above, the following hypotheses result from women and men updating gender-dependently in the male role model treatment.

Hypothesis 3 (*Women in MRM)*: Women's competitiveness in *MRM* is lower than in the *Baseline*.

Hypothesis 4 (*Men in MRM)*: Men's competitiveness in *MRM* is higher than in the *Baseline*.

Notice that based on the prediction that women in the *Baseline* compete significantly less than men, the effects predicted in Hypothesis 1-4 should vary in size. For women, the decision to compete can only be reversed for marginal types, i.e. in *FRM* those who ex ante believe to be worse than men and in *MRM* those who ex ante believe to be better than men. At the same time, it is reasonable to assume that there are more women who ex ante believe to be worse than men in the task. As such, the predicted increase in women's competitiveness in *FRM* (Hypothesis 1) should be larger in magnitude than the predicted decrease in women's competitiveness in *MRM* (Hypothesis 3). Similarly, the predicted increase of competitiveness of men in *MRM* (Hypothesis 4) should be smaller in magnitude than the effect of Hypothesis 1, as relatively few men ex ante believe to be worse in the task than women.

In addition, the effects hypothesized might be affected by a *reference group neglect* (Camerer and Lovallo, 1999), in that subjects might primarily consider how seeing a female or male role model affects their own ability, but might fail to consider what this implies for other's average ability. Then, the effect of same-sex role models should be larger in magnitude than the effect of opposite-sex role models, e.g. Hypothesis 1 (effect of *FRM* on women) should be more pronounced than Hypothesis 2 (effect of *FRM* on men). In sum, we predict that the effect of Hypothesis 1 should be largest in magnitude. Nevertheless, we will examine all four hypotheses in the next section.

5.4 Results

In this section, we outline the results from our experimental study. Although our hypotheses about competitive behavior derive from expected belief updating, we will first analyze the effect of role models on behavior in section 4.1. In section 4.2, we will then describe how any behavioral effects from role models can be explained by belief updating about own and other' ability. Throughout the paper, we test differences in behavior and beliefs between subjects using two-sided Mann-Whitney U (MWU) and Wilcoxon signed-rank (WSR) tests at a significance level of 5%.

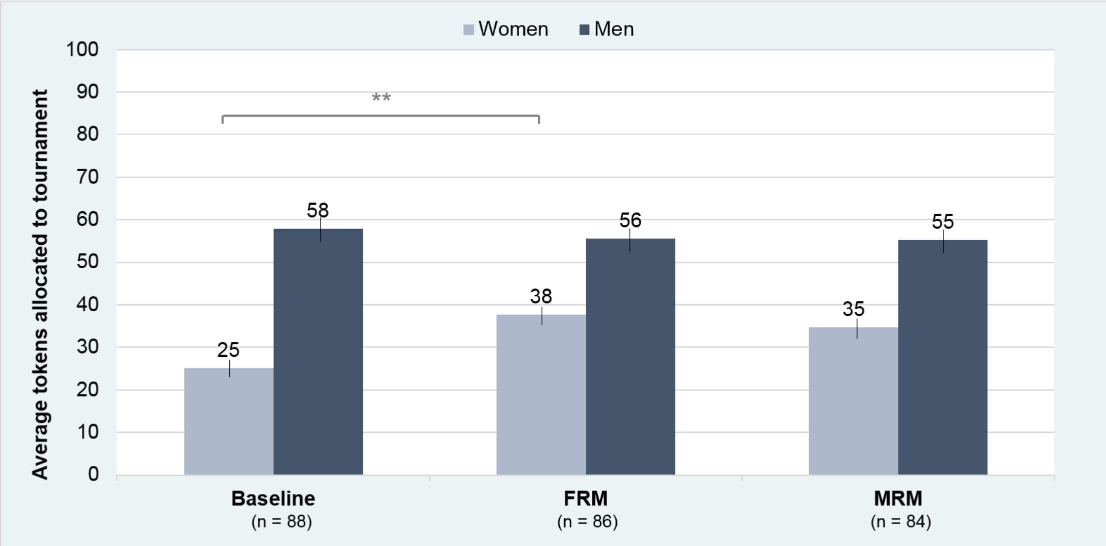
5.4.1 Effect of role models on competitiveness behavior

First, we analyze the performance of subjects in our math tasks. Looking at ability and gender, we observe that men significantly outperform women in our sample. In round 1, men solve on average 7.0 tasks correctly, whereas women solve 5.6 tasks (MWU, $p < 0.001$). This performance difference remains throughout round 2 and round 3 (round 2: men = 7.6, women = 6.0, $p < 0.001$; round 3: men = 8.5, women = 6.7, $p < 0.001$). We control for performance in all our regressions, but ability in the math task does not fully explain the gender gap in competitiveness we observe.[7]

Next, we examine the competitiveness behavior of subjects, measured as the amount of tokens subjects invest in the tournament payment scheme in round 3. As illustrated in Figure 14, we find that men in our sample behave more competitive than women.[8] In the *Baseline*, women invest on average 25 out of 100 tokens in the tournament incentive scheme, whereas men invest 58 tokens. Thus, we are able to replicate the gender gap in competitiveness (MWU, $p < 0.001$) from the previous literature. This gender gap remains significant in the *FRM* treatment (MWU, $p < 0.020$) and in the *MRM* treatment (MWU, $p < 0.008$) but it decreases in size. In specific, we find that women significantly increase their competitiveness in the *FRM* treatment by investing on average 38 tokens in the tournament compared to 25 in the *Baseline*. This increase in competitiveness is statistically significant (MWU: $p < 0.044$) and supports our Hypothesis 1. In line with Hypothesis 2, we observe that men in *FRM* slightly decrease their competitiveness to 55 compared to 58 in the *Baseline*. However, this decrease is not statistically

[7] OLS regression with robust standard errors and competitiveness as outcome variable: β_0=31.0, $p < 0.001$, β_{female}=-18.5, $p < 0.001$, $\beta_{ability}$=3.6, $p < 0.001$.
[8] See Figure 16 in the Supplementary Material for the distribution of competitiveness by gender and treatments.

Figure 14: Effect of role models on competitiveness (N=258).



Notes. The p-value is obtained from a Mann-Whitney U test, with ** < 0.05.

significant (MWU: $p < 0.711$). With regard to Hypothesis 3, we expected women to decrease their competitiveness in face of a male role model. In fact, we find the opposite, in that women increase their competitiveness to 35 in *MRM*. However, compared to the *Baseline* this increase is not statistically significant at a significance level of 5% (MWU, $p < 0.099$). Finally, men slightly decrease their competitiveness to 55 in *MRM* compared to the *Baseline*. This goes against our Hypothesis 4, but is not statistically significant (MWU, $p < 0.636$). As such, we can conclude that men are not affected by observing a female or male role model. In comparison, women significantly increase their competitiveness after observing a female role model, and marginally significantly increase their competitiveness after observing a male role model.

To understand the effect of role models on women's competitiveness better, we conducted multiple OLS regressions, as illustrated in Table 10.[9] In the first regression model (column (1)), competitiveness of women is regressed on an *FRM* treatment dummy (value of 1

[9] See Table 11 in the Supplementary Material for OLS regressions for men's competitiveness. OLS regressions confirm previous findings that men do not respond to female nor male role models.

Table 10: Determinants of competitiveness of women with a female or male role model.

Dependent = competitiveness (tokens allocated to tournament)	**(1)**	**(2)**	**(3)**
FRM	12.56**	10.69*	-24.60*
	(6.14)	(6.12)	(14.24)
MRM	9.65	7.06	7.93
	(5.88)	(6.12)	(6.13)
Confidence		3.63**	2.08
		(1.42)	(1.59)
Ability		-1.06	-0.95
		(1.18)	(1.20)
Risk aversion		-0.12	-0.12*
		(0.07)	(0.07)
FRM * Confidence			5.69**
			(2.27)
Constant	25.11***	15.19	24.67**
	(3.98)	(10.28)	(11.09)
Observations	129	129	129
R-squared	0.04	0.12	0.16

Notes. Results from OLS regressions with robust standard errors in parentheses. FRM = dummy variable for treatment in which female role model was displayed; MRM = dummy variable for treatment in which male role model was displayed; Confidence = incentivized belief measure for expected performance before role model was shown (posterior2a); Ability = number of tasks solved in round 1; Risk aversion = incentivized risk measure (Gneezy and Potters, 1997) from 0 (risk-seeking) to 100 (risk-averse). P-values: *** < 0.01, ** < 0.05, * < 0.1.

for *FRM* and 0 for *Baseline*), and a *MRM* treatment dummy (1 for *MRM* and 0 for *Baseline*). In this model, the *FRM* treatment is a highly significant predictor for competitiveness, with women competing significantly more in face of a female role model (β = 12.56, p < 0.05). In comparison, the *MRM* treatment does not significantly predict women's competitiveness (β = 9.65, p = 0.103).[10]

In model (2), we also control for confidence (subjects' expected performance after round 2 and before the role model is shown), for ability (the number of tasks solved in round 1)[11] and for risk aversion (the amount subjects did not invest from 0 to 100 in an investment game at the end of the experiment). We find that the effect of the *FRM* treatment on competitiveness of

[10] Although we do not find that women respond differently to a female role model than to a male role model in their competitiveness (Wald test: p = 0.6483), we focus on results testing our Hypotheses. In particular, we concentrate on significant evidence for Hypothesis 1, that women increase their competitiveness in *FRM* compared to the *Baseline*.

[11] Results are robust to using performance in round 2 instead.

women is not significant anymore at a 5% significance level (β = 10.69, p = 0.083), when we control for confidence, ability and risk aversion. Rather, confidence in this model is a highly significant predictor of competitiveness of women (β = 3.63, p < 0.013). The more confident women are in their absolute abilities in the math task (before observing the role model), the higher their competitiveness.

In column (3), we include an interaction effect between the *FRM* treatment dummy and confidence (ranging from 1 to 15 for women). Results show that the effect of *FRM* on women's competitiveness significantly interacts with confidence (β = 5.69, p < 0.015). In specific, the more confident women are in their ability in performing the task (before observing the role model), the more they respond to a female role model by increasing their competitiveness. In fact, we observe that above-median confident women (confidence > 6) choose an average competitiveness of 53 in *FRM*, close to men's competitiveness, whereas below-median confident women in *FRM* choose an average competitiveness of 25, similar to women's competitiveness in the *Baseline*. This difference in competitiveness of above- and below-median confident women in *FRM* is highly significant (MWU: p < 0.0047).[12] In comparison, in the *Baseline* we also find that above-median confident women compete more than below-median confident women but this difference is not statistically significant ($mean_{above}$ = 32, $mean_{below}$ = 20, MWU: p = 0.2614). For women in *MRM*, we observe that confidence does not explain mean competitiveness ($mean_{above}$ = 36, $mean_{below}$ = 34, MWU: p = 0.7985). Rather, all women seem to equally respond to a male role model by marginally significantly increasing their competitiveness. In conclusion, we find that particularly highly confident women are affected by a female role model, whereas confidence does not seem to play a role for women's competitiveness in face of a male role model.

[12] Women with a very low confidence might even be deterred by a Female Role Model, as social comparison theory (Mussweiler, 2003) suggests that people can assimilate but also contrast away from others. However, we do not find any evidence that women with a confidence in the lowest quartile (<5) are less competitive in *FRM* than in *Baseline* (MWU: p = 0.7503), i.e. we do not find any backlash effects.

On this note, in the next section we will examine how beliefs (e.g. confidence) are updated in face of a female or male role model. In particular, we want to understand how beliefs are driving the shift in competitive behavior of women in *FRM*.

5.4.2 Effect of role models on beliefs

First, subjects might hold different beliefs about how they will perform in the math task before the start of round 1. We find that men report significantly higher priors about their expected performance in the math task compared to women ($mean_{prior\ of\ men}$ = 8.2, $mean_{prior\ of\ women}$ = 6.6, MWU: p < 0.001). Given that men outperform women in our math task (round 1 performance: $mean_{men}$ = 7, $mean_{women}$ = 5.6, MWU: p < 0.001), this is not surprising. In fact, if we compare the accuracy of priors, in terms of the difference between the prior and actual performance in round 1, men and women both seem slightly overconfident but do not differ much in their overconfidence ($\text{overconfidence}_{men}$ = 1.2, $\text{overconfidence}_{women}$ = 1.1, MWU: p = 0.1892).

More interesting is how subjects, given that they hold certain priors, update their beliefs about their own absolute performance after observing a female or male role model. We refer to subjects' beliefs about their own absolute performance in the math task as self-confidence, and examine whether self-confidence is affected by the role model. For this, we compare our measures *posterior2a* and *posterior3a* which describe subjects' self-confidence before and after observing a role model. If we expect that subjects do not take the gender of the role model into account (our general null hypothesis), we should not observe any updating of self-confidence (see 3.2 Hypotheses). If subjects update gender-dependently (our alternative hypothesis), then we should observe that subjects increase their self-confidence after observing a same-sex role model. In fact, we find that women significantly increase their self-confidence in the *FRM* treatment (posterior2a = 6.2, posterior3a = 6.8, WSR: p < 0.015) but do not change their self-confidence in the *MRM* treatment (posterior2a = posterior3a = 6.9).[13] This is in line with our alternative hypothesis, suggesting that women update gender-dependently when they see a role model by taking its gender into account.

However, we do no find that men's self-confidence is affected by a same-sex role model (*MRM*) nor by the *FRM* treatment (FRM: posterior2a = 8.0, posterior3a = 8.2, WSR: $p = 0.935$; MRM: posterior2a = 7.7, posterior3a = 8.0, WSR: $p = 0.219$). Thus, for men's self-confidence we do not find any evidence that they update gender-dependently, as a same-sex role model does not seem to impact their beliefs about own ability in the task.

Apart from the effect that a role model can have on self-confidence, it can also affect how subjects update their beliefs about others' performance in the math task. We do not have a direct measure of beliefs about average female and average male ability in the task, as we did not want to prime subjects with gender during the experiment. However, we elicited subjects' beliefs about others' average absolute ability before and after seeing a role model, namely *posterior2b* and *posterior3b*. If subjects update their beliefs about average female and male ability in the task by considering the gender of the role model, we might observe that *posterior3b* changes compared to *posterior2b*. However, we do not find any significant changes in *posterior2b* to *posterior3b*: women's beliefs about other's ability remain constant in *FRM* (posterior2b = posterior3b = 6.9, WSR: $p = 0.823$), as well as in *MRM* (posterior2b = 7.8, posterior3b = 7.7, WSR: $p = 0.664$). Similarly, men's beliefs do not change significantly in *FRM* (posterior2b = 7.9, posterior3b = 7.7, WSR: $p = 0.13$) or *MRM* (posterior2b = 6.8, posterior3b = 7.1, WSR: $p = 0.105$). As such, we do not find any evidence that female or male role models affects subjects' beliefs about other's ability in the math task.

In sum, we find evidence for women increasing their self-confidence when they see a female role model, which increases their competitiveness as predicted by Hypothesis 1. Thus,

[13] Comparing both role model treatments, we do not find that women's increase in self-confidence is larger in *FRM* than in *MRM* ($increase_{FRM} = 0.60$ vs. $increase_{MRM} = 0.05$, MWU: $p = 0.1439$).

we argue that the main channel for out treatment effect of *FRM* increasing women's competitiveness is through belief updating, i.e. through an increase in self-confidence. We do not observe any effects for men updating their beliefs or changing their behavior after seeing a female or male role model, nor do women significantly respond to a male role model in their beliefs or their behavior. In the next section, we discuss reasons for this asymmetry and potential confounds, as well as examine whether behaving more competitive is desirable.

5.5 Discussion

In our experiment, we find that women increase their self-confidence and their competitiveness after observing a female role model. An alternative explanation for any effect we observe might be an anchoring effect (e.g. Ariely et al., 2003). Presenting subjects with an example of a previous participant that invested 70 tokens in the competitive payment scheme might simply anchor subjects to choose 70, too. First, observe that the modal value of competitiveness in all three treatments is 0 (and not 70). Second, we compare the share of subjects choosing a competitiveness of 70 in the *Baseline* with the role model treatments. We find that 4.6% of subjects choose a competitiveness of 70 in the *Baseline* compared to 6.5% in both role model treatments which is not statistically significant (Fisher's exact test: p = 0.780).[14] Both is indicative evidence that anchoring is not a convincing explanation for our findings.

Although we find that women respond to a same-sex role model, we do not observe the same effect for men. First, we expected to find the strongest effect for women in the female role model treatment. Also, the psychological literature suggests that same-sex role models are more relevant for women (Lockwood, 2006). However, an additional explanation for why men do not increase their competitiveness after seeing a male role model in our experiment might be that the male role model we present is rather an "anti-role model" to some men. Men in the *Baseline* on average already invest an amount close to what the male role model invests (58 vs.

[14] This finding is robust to extending the anchor to a competitiveness of 70 ± 10 (Fisher's exact test: p = 0.395).

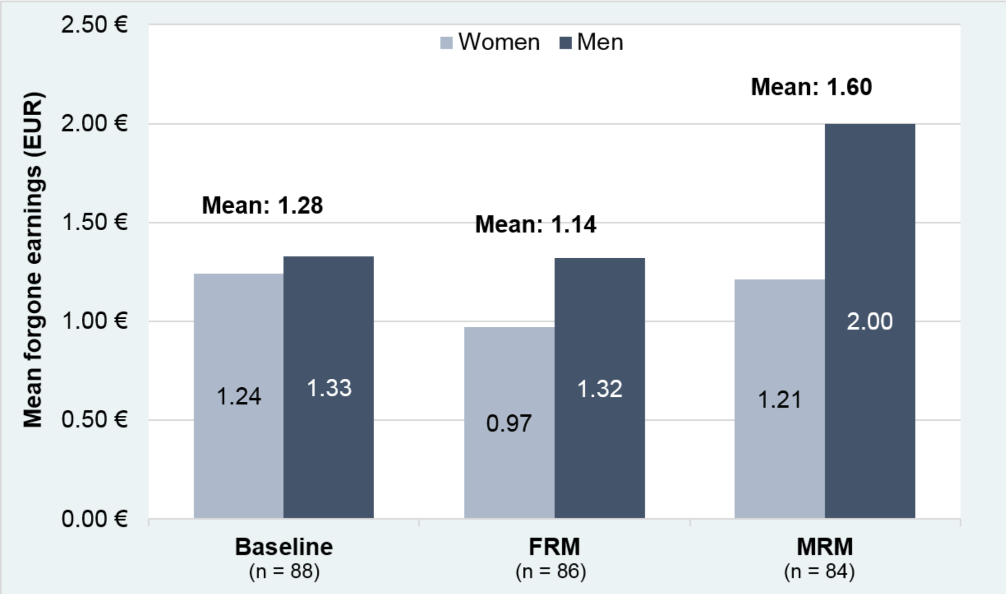
70). In contrast, the female role model invest more than twice as much as the average woman in the *Baseline* (25 vs. 70). With our experimental design, we intentionally took this into account to have cleaner treatment manipulations, in which we only vary one factor (i.e. gender of the role model) while keeping all other information the same between treatments. Alternatively, we could have presented subjects with role models that competed as much as for example the top 10 percent of their gender. However, we opted against this, as competing in the top 10 percent does not mean one is also successful. We rather aimed to present subjects with two role models that competed the same but were both equally successful in earnings compared to their own sex (i.e. among the top 15% of highest-earning men or women). Even though we argue that our design is a cleaner treatment manipulation, this might lead to the male role model treatment in our study having a weaker effect on men's behavior than the female role model treatment might have on women. Nevertheless, this does not explain, why men do not respond as much to the female role model treatment by updating their beliefs about average female ability. This might rather be explained by the literature on *reference group neglect* (Camerer and Lovallo, 1999) which suggests that people sometimes ignore their reference group. Interestingly, subjects in our study do not seem to derive any valuable information about other's ability from seeing a female or male role model.

After subjects made all payoff-relevant decisions but before we gave them any feedback on their performance, participants self-reported their beliefs about average female and male competitiveness (unincentivized). In specific, we asked them to estimate the number of tokens men and women in the experiment on average invested in the tournament scheme, respectively. Average beliefs about male competitiveness do not differ much between treatments ($m_{Baseline}$ = 59, m_{MRM} = 54, m_{FRM} = 61, Kruskal-Wallis test: p = 0.06). In comparison, beliefs by men and women about female competitiveness are significantly higher in the *FRM* treatment than in the *Baseline* ($m_{Baseline}$ = 36 vs. m_{FRM} = 43, MWU*:* $p < 0.02$). Finally, beliefs about average female competitiveness in *MRM* do not differ significantly from the *Baseline* (m_{MRM} = 42, MWU: $p = 0.07$). In summary, it seems that subjects have relatively accurate expectations about how men and women respond to our female and male role models, and that both are aware that female role models might have a positive effect on women's competitiveness.

Moreover, one might argue that role models do not affect competitiveness behavior or beliefs but actually shift risk preferences. First, when regressing risk-aversion and a gender-dummy variable on competitiveness, we do not find that risk-aversion is a significant driver of competitiveness (β_0 = 58.0, p < 0.001, $\beta_{riskaverse}$ = -0.06, p = 0.316, β_{female} = -22.6, p < 0.001). Secondly, we do not observe that our role model treatments significantly affect risk preferences. In a simple base OLS regression with risk aversion as dependent variable, a dummy for both role model treatments does not affect risk-aversion (β_0 = 45.3, p < 0.001, β_{RM} = -6.1, p = 0.248). The treatment coefficient remains a non-significant predictor for risk-aversion, even when controlling for gender and the interaction of gender and the role model treatments (β_0 = 30.0, p < 0.001, β_{RM} = 1.0, p = 0.894, β_{female} = 30.5, p < 0.001, $\beta_{female*RM}$ = -14.2, p = 0.159). Hence, we conclude that our role model treatments do not shift risk preferences. Instead, we argue that female role models shift women's beliefs about their own ability, and thereby impact women's competitiveness.

Finally, we analyze whether competitiveness decisions maximize expected earnings. Behaving more competitive does not necessarily increase subjects' expected earnings which also depend on subjects' relative performance. Therefore, we analyze subjects' *forgone earnings* between treatments, to understand in which treatment subjects' competitiveness decisions yield the most or least in terms of expected earnings. We define *forgone earnings* as the difference in expected earnings between a subject's optimal token allocation and his/her actual token allocation. For subjects' optimal token allocation, we estimate subjects' expected probability of winning the tournament in round 3, given his/her performance in round 2 and

Figure 15: Average forgone earnings by subjects in each treatment.



Notes. Forgone earnings is the difference in expected earnings from optimal competitiveness and actual competitiveness (tokens allocated to tournament incentive scheme), based on subjects' round 2 performance and their expected probability of winning the tournament in round 3.

being randomly paired to any other subject in the experiment.[15] As such, *forgone earnings* can be interpreted as the difference between opportunity costs and actual earnings, and are suggested to be a measure of efficiency gains in economic terms (Brandts et al., 2015). Given our design, subjects with an expected probability of 33% of winning the tournament in round 3 should be indifferent between piece-rate and tournament payment, as the latter pays subjects thrice as much per correctly solved task. For our sample, we find that subjects who solved 6 tasks in round 3 have an expected probability of at least 33% of winning the tournament in round 3. This means, subjects who solved *less* than 6 tasks in round 2 should expect to lose the tournament in round 3 and should therefore invest all 100 tokens in the piece-rate incentive scheme to maximize expected earnings. Similarly, subjects that solved *more* than 6 tasks in round 2 should invest all 100 tokens in the tournament scheme to maximize expected earnings,

[15] We suggest that subjects' best measure for their expected performance in round 3 is their performance in round 2. Although some subjects might perform better under piece-rate than in a tournament, the majority of subjects seems to improve their performance from round 1 to round 2 (WSR: $p < 0.0002$). Also, performance in round 2 and round 3 correlate strongly (Spearman's rho: 0.80, $p < .0001$).

as in expectation they win the tournament in round 3. Based on these estimates, we calculate *forgone earnings* for each subject and compare average forgone earnings by men and women between treatments (see Figure 15).

We find that subjects on average forgo 1.28 EUR in terms of expected earnings in the *Baseline*, 1.14 EUR in *FRM* and 1.60 EUR in *MRM* but the differences between treatments are not statistically significant (Kruskal-Wallis-test, $p = 0.497$).[16] Considering that women in *FRM* compete more than women in the *Baseline*, we also observe that they maximize expected earnings more in *FRM*, by forgoing on average 0.97 EUR in expected earnings in *FRM* compared to 1.24 EUR in the *Baseline*. However, this difference is not statistically significant (MWU: $p = 0.397$). Overall, we do not find that competitiveness decisions of men and women across treatments maximize expected earnings differently. As such, we can conclude that we do not find evidence that a female or male role model has a negative impact on men's or women's competitiveness behavior.

5.6 Conclusion

This study investigates how female and male role models affect beliefs and competitiveness of men and women. We find evidence that women increase their competitiveness after seeing a female role model. Moreover, we identify an increase in women's self-confidence as the main channel. In comparison, we do not find that women significantly react to male role models, nor that men respond to male or female role models in their competitiveness or in their beliefs about own and other's behavior. With regard to our hypotheses, we thus find partial evidence for subjects updating their beliefs gender-dependently when they observe a same-sex role model. However, this seems to apply to women only. On a side note, the increase in self-confidence by women observing a female role model might be either because they learn that women are on

[16] In comparison, the total sum of forgone earnings by all subjects in each treatment are 113 EUR in the *Baseline*, 98 EUR in *FRM* and 135 EUR in *MRM*.

average better than they previously thought, or because they learn that average women are more confident than they previously thought. However, this is not the focus of the current study but further research on this would be needed.

To conclude, this research provides some evidence that female role models increase women's competitiveness. Ironically, in our setting it seems that the women affected most by a female role model are the ones which are already relatively confident. Nevertheless, for our setting we learn that female role models do not have a significant impact on men, neither does the presence of a male role model significantly affect men or women which might also be a valuable insight for future research and policy-makers. From a general economic perspective, our findings raise the question whether role models can also be effective in increasing the competitiveness of other minority groups that are currently underrepresented in politics, management and academia, which could be promising future research topic.

References

Akerlof, G. A., and Kranton, R. E. (2000). Economics and Identity. *Quarterly Journal of Economics, 115*(3), 715-753.

Alan, S., Ertac, S., and Mumcu, I. (2017). Gender stereotypes in the classroom and effects on educational outcomes. *Review of Economics and Statistics*, forthcoming.

Almas, I., Cappelen, A. W., Salvanes, K. G., Sorensen, E. O., and Tungodden, B. (2016). Willingness to compete: family matters. *Management Science, 62*(8), 2149-2162.

Apicella, C. L., Demiral, E., and Mollerstrom, J. (2017). *No gender difference in willingness to compete when competing against self*. Discussion Papers DIW 1638.

Ariely, D., Loewenstein, G., and Prelec, D. (2003). "Coherent arbitrariness": stable demand curves without stable preferences. *Quarterly Journal of Economics*, *118*(1), 73–105.

Balafoutas, L., and Sutter, M. (2012). Affirmative action policies promote women and do not harm efficiency in the lab. *Science*, *335*(6068), 579-582.

Baldiga, N. R., and Coffman, K. B. (2016). Laboratory evidence on the effects of sponsorship on the competitive preferences of men and women. *Management Science*, 64(2), 888-901.

Beaman, L., Duflo, E., Pande, R., and Topalova, P. (2012). Female leadership raises aspirations and educational attainment for girls: a policy experiment in India. *Science*, *335*(6068), 582-586.

Benjamin, D. J., Choi, J. J., and Strickland, A. J. (2010). Social identity and preferences. *American Economic Review, 100*(4), 1913-1928.

Bertrand, M., Kamenica, E., and Pan, J. (2015). Gender identity and relative income within households. *Quarterly Journal of Economics, 130*(2), 571-614.

Bettinger, E. P., and Long, B. T. (2005). Do faculty serve as role models? The impact of instructor gender on female students. *American Economic Review, 95*(2), 152-157.

Blau, F., and Kahn, L. M. (2000). Gender differences in pay. *Journal of Economic Perspectives, 14*(4), 75-99.

Brandts, J., Groenert, V., and Rott, C., (2015). The impact of advice on women's and men's selection into competition. *Management Science, 61*(5), 1018-1035.

Buser, T., Dreber, A., and Mollerstrom, J. (2016). The impact of stress on tournament entry. *Experimental Economics, 20*(2), 506-530.

Buser, T., Niederle, M., and Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *Quarterly Journal of Economics, 129*(3), 1409-1447.

Camerer, C., and Lovallo, D. (1999). Overconfidence and excess entry: an experimental approach. *American Economic Review, 89*(1), 306-318.

Carrell, S. E., Page, M. E., and West, J. E. (2010). Sex and science: how professor gender perpetuates the gender gap. *Quarterly Journal of Economics, 125*(3), 1101-1144.

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *Quarterly Journal of Economics, 129*(4), 1625-1660.

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

Flory, J. A., Leibbrandt, A., and List, J. A. (2015). Do competitive workplaces deter female workers? A large-scale natural field experiment on job-entry decisions. *Review of Economic Studies, 82*(1), 122-255.

Gneezy, U., and Potters, J. (1997). An experiment on risk taking and evaluation periods. *Quarterly Journal of Economics, 112*(2), 631-645.

Gneezy, U., Leonard, K. L., and List, J. A. (2009). Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica, 77*(5), 1637-1664.

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

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association, 1*(1), 114-125.

Guiso, L., Monte, F., Sapienza, P., and Zingales, L. (2008). Culture, gender, and math. *Science, 320*(5880), 1164-1165.

Gupta, N. D., Poulsen, A., and Villeval, M. C. (2011). Gender matching and competitiveness: Experimental evidence. *Economic Inquiry, 51*(1), 816-835.

Kamas, L., and Preston, A. (2012). The importance of being confident; gender, career choice, and willingness to compete. *Journal of Economic Behavior and Organization, 83*(1), 82-97.

Kölle, F. (2017). Affirmative action, cooperation, and the willingness to work in teams. *Journal of Economic Psychology, 62*, 50-62.

Lockwood, P. (2006). Someone like me can be successful: Do college students need same-gender role models? *Psychology of Women Quarterly, 30*, 36-46.

Meier, K., Niessen-Ruenzi, A., and Ruenzi, S. (2017). *The impact of role models on women's self-selection in competitive environments.* Working paper, University of Mannheim.

Mussweiler, T. (2003). Comparison process in social judgment: Mechanisms and consequences. *Psychological Review, 110*(3), 472-489.

Niederle, M., and Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics, 3*, 601-630.

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.

Nosek, B. A, Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., Bar-Anan, Y., Bergh, R., Cai, H., …Greenwald, A. G. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceeding of the National Academy of Sciences, 106*(26), 10593-10597.

Porter, C., and Serra, D. (2017). *Gender differences in the choice of major: the importance of female role models.* Working paper, Southern Methodist University.

Reuben, E., Sapienza, P., and Zingales, L. (2014). How stereotypes impair women's career in science. *Proceeding of the National Academy of Sciences, 111*(12), 4403-4408.

Reuben, E., Wiswall, M., and Zafar, B. (2017). Preferences and biases in educational choices and labour market expectations: shrinking the black box of gender. *Economic Journal, 127*(604), 2153-2186.

Saccardo, S., Pietrasz, A., and Gneezy, U. (2017). On the size of the gender difference in competitiveness. *Management Science*, *64*(4), 1541-1554.

Sutter, M., and Glätzle-Rützler, D. (2015). Gender differences in competition emerge early in life and persist. *Management Science, 61*(10), 2339-2354.

Wozniak, D., Harbaugh, W. T., and Mayr, U. (2014). The menstrual cycle and performance feedback alter gender differences in competitive choices. *Journal of Labor Economics, 32*(1), 161-198.

Supplementary Material

Figure 16: Cumulative distribution functions for competitiveness by gender and treatment.

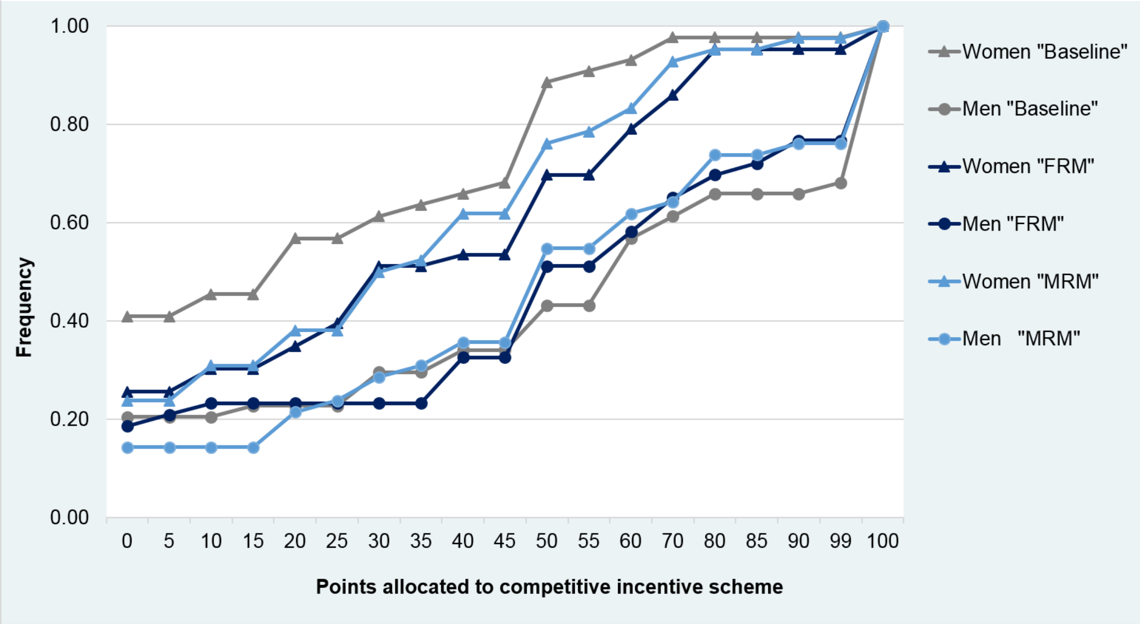


Table 11: Determinants of men's competitiveness with a female or male role model.

Dependent = competitiveness (tokens allocated to tournament)	**(1)**	**(2)**
FRM	-2.24	-2.75
	(8.00)	(6.90)
MRM	-2.58	-1.72
	(7.87)	(7.21)
Confidence		3.66***
		(1.32)
Ability		1.85
		(1.44)
Risk aversion		0.00
		(0.09)
Constant	57.82***	16.06
	(5.77)	(8.53)
Observations	129	129
R-squared	0.00	0.23

Notes. Result from OLS regressions with robust standard errors in parentheses. FRM = dummy variable for treatment in which female role model was displayed (baseline = Baseline treatment); MRM = dummy variable for treatment in which male role model was displayed; Confidence = incentivized belief measure for expected performance after round 2 and before role model was shown (posterior2a); Ability = number of tasks solved in round 1; Risk aversion = incentivized risk measure (Gneezy and Potters, 1997) from 0 (risk-seeking) to 100 (risk-averse). P-values: *** < 0.01, ** < 0.05, * < 0.1.

Appendix A. Experimental Instructions

Welcome to the experiment!

Please read the following instructions carefully. From now on, please do not communicate with other participants. In case you have any questions during the experiment, please raise your hand and we will come to you.

In this experiment, you can earn money. The amount of money you earn depends on your decisions as well as on other participants´ decisions. Independently, you receive 4 Euro as show-up fee. At the end of the experiment, your earnings will be added up and paid to you in cash.

All decisions you make during the experiment are anonymous. Your earnings are confidential as well.

From now on, mobile phones have to be turned off and may not be used during the experiment. You are not allowed to use any reading material not related to the experiment (books, lecture notes, etc.). In case you violate these rules, we can exclude you from the experiment and all respective payments.

On the next page, you find instructions on the procedure of the experiment. The experiment starts, as soon as all participants have read and understood these instructions.

Procedure of the experiment

This experiment consists of three rounds. At the end, one round is randomly selected and paid out. In each round, you have to solve as many math tasks as possible in a given period of time. In each task you have to sum up five two-digit numbers. You are not allowed to use a calculator but you can use the scratch paper in front of you.

For example: 23 + 57 + 67 + 31 + 89 = ?

The correct answer to this example is "267". As soon as you enter your answer in the respective field and click "OK", you will get a new math task. In total, you have <u>3 minutes</u> to solve as many tasks as possible. The remaining time will be displayed to you in the right upper corner of your screen. You will be informed about the total number of correctly solved tasks only at the end of the experiment.

The rounds differ in the way you are paid:

In **round 1,** you receive **50 Cent** for each correctly solved task**.** We call this "INDIVIDUAL" payment.

– –

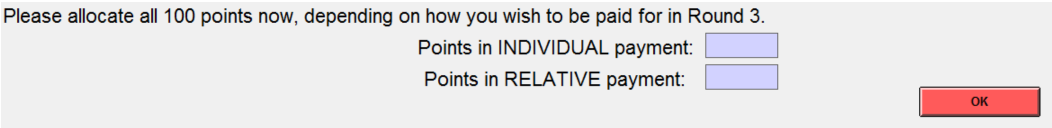
In **round 2,** you are randomly matched with another participant. Whoever solves more tasks correctly receives **150 Cent** for each correct answer. The other person receives 0€. In case of a tie, the winner for round 2 will be randomly determined. We call this "RELATIVE" payment.

– –

In **round 3,** you choose how you want to be paid.

You can choose whether you want to be paid INDIVIDUALLY (as in round 1), whether you want to be paid RELATIVELY (as in round 2), or choose a mix of both payments. For this, you are endowed with **100 points** at the beginning of round 3. The points are converted into Euro at the end of the experiment. Before you start with math tasks in round 3, you have to allocate the 100 points completely to an INDIVIDUAL and/or RELATIVE payment.

In specific, you will see the following decision at the beginning of the third round:



Depending on your point allocation, points will be converted into Euro in the following way:

a) **Each point allocated to INDIVIDUAL payment results in 0.5 Cent** per correct answer in round 3. (E.g., in case you want to be paid as in round 1, you allocate 100 points to INDIVIDUAL payment. This means 100 points = 50 Cent per correct task in round 3.)

b) **Each point allocated to RELATIVE payment results in 1.5 Cent** per correct answer in round 3, if you solve more tasks correctly than your partner. If you answer fewer tasks correctly, you will receive 0€. In case of a tie, the winner for round 3 is determined randomly. (E.g., in case you want to be paid as in round 2, you allocate all 100 points to RELATIVE payment. This means 100 points = 150 Cent for each correctly solved task in round 3, if you solve more tasks than your partner).

Please note: For the RELATIVE payment in round 3, your performance in round 3 is compared to the same partner as in round 2 and her/his respective performance in round 2.

In case you allocate points to both payments, your profit in round 3 is the sum of your earnings from a) and b). (For example: 50 points allocated to INDIVIDUAL = 50 * 0.5 Cent = 25 Cent for each correctly solved task. 50 points allocated to REALATIVE = 50 * 1.5 Cent = 75 Cent for each correctly solved task, if you solve more tasks than your partner.

There will be a trial round before the experiment starts, so that you can familiarize yourself with the task. The trial round takes 1 minute and you will not be paid for the trial round.

In addition, you have to make seven estimations on the number of tasks you and other participants solved correctly. One of these estimations is randomly selected at the end of the experiment. If your estimation was correct, you will receive 1€ additionally.

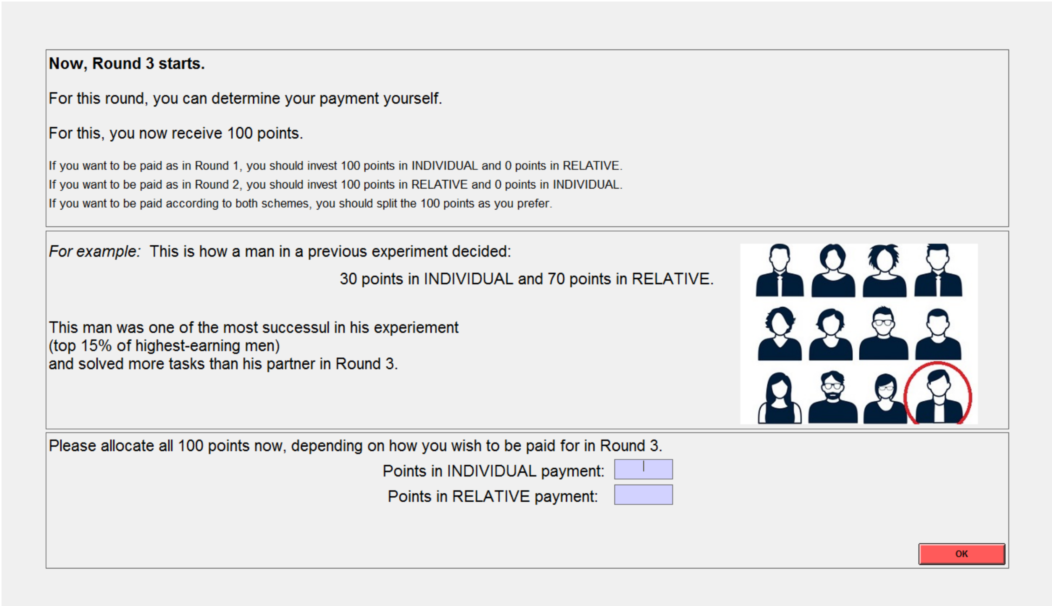
Furthermore, after round 3 you can make an investment, where you can earn between 0€ and 2€.

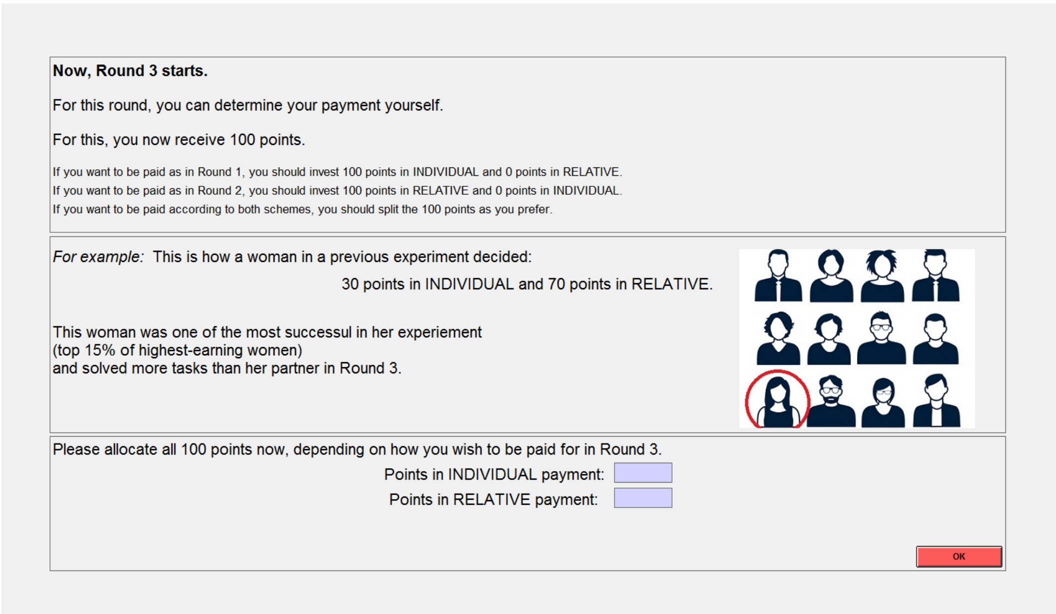
Finally, you have to complete a short questionnaire. Afterwards, you are informed about the results of the three rounds, your estimations and your investment. Then, you will receive an overview of your total payment from the experiment.

In case you have any questions about the instructions, please raise your hand and we will come to you.

Appendix B. Experimental Screens

Figure 17: Experimental screen for the role model manipulations (above: female role model treatment; below: male role model treatment). In the Baseline, subjects saw the same screen without the middle text box.





Curriculum Vitae

Personal Details

Name	Uta Katharina Schier
Place of birth	Berlin-Steglitz

Education

Since 2013	Ph.D. Student in Economics University of Cologne
2012	M.Sc. Management Science (Decision Sciences) London School of Economics and Political Sciences
2010	B.Sc. Management Ludwig-Maximilians-University, Munich
2006	Abitur Christliches Gymnasium Jena