

Tilburg University

Essays on behavioral responses to dishonest and anti-social decision making

Brouwer, Thijs

DOI:
[10.26116/center-lis-2107](https://doi.org/10.26116/center-lis-2107)

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):

Brouwer, T. (2021). *Essays on behavioral responses to dishonest and anti-social decision making*. CentER, Center for Economic Research. <https://doi.org/10.26116/center-lis-2107>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.



CentER

Essays on Behavioral Responses
to Dishonest and Anti-Social
Decision-Making

THIJS BROUWER

ESSAYS ON BEHAVIORAL RESPONSES TO DISHONEST AND ANTI-SOCIAL DECISION- MAKING

Proefschrift ter verkrijging van de graad van doctor aan
Tilburg University op gezag van de rector magnificus,
prof. dr. W.B.H.J. van de Donk, in het openbaar te
verdedigen ten overstaan van een door het college voor
promoties aangewezen commissie in de Aula van de
Universiteit op woensdag 19 mei 2021 om 16.30 uur door

THIJS BROUWER,

geboren op 5 februari 1992 te Nijmegen, Nederland.

Promotores: prof. dr. J.J.M. Potters (Tilburg University)
prof. dr. E.E.C. van Damme (Tilburg University)

Leden prof. dr. G. Kirchsteiger (Université Libre de Bruxelles)
promotiecommissie: dr. J.T.R. Stoop (Erasmus Universiteit)
prof. dr. E.C.M. van der Heijden (Tilburg University)
prof. dr. S. Suetens (Tilburg University)

Chapter 3 of this doctoral thesis has been funded by LABEX CORTEX of Université de Lyon, within the program Investissements d’Avenir operated by Agence Nationale de la Recherche (ANR), and by the INDEPTH program from IDEXLYON.

©2021 Thijs Brouwer, The Netherlands. All rights reserved. No parts of this thesis may be reproduced, stored in a retrieval system or transmitted in any form or by any means without permission of the author. Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd, in enige vorm of op enige wijze, zonder voorafgaande schriftelijke toestemming van de auteur.

To Marianne

Acknowledgements

Als de Waal
uit het zicht is
stroomt
de verbeelding

T. NIESKEN

This doctoral thesis marks the end of my 11-year spell at Tilburg University. When you would have told me the length of my stay when I first arrived in Tilburg for the TOP-week in 2010, I would probably not have believed you. Study, graduate, and return to Nijmegen, that was the plan. Gradually, however, I learned that living in Noord-Brabant wasn't so bad after all.

It would be a bit of an exaggeration to call my program choice a Hail Mary, but it was far from a well-informed decision. A high-school teacher once advised me to choose what I liked, not what I thought I ought to choose. This gut feeling first drew me to Economie & Bedrijfseconomie in Tilburg, then to the Research Master, and ultimately it brought me here. Over the course of this journey, I have received insurmountable support from many wonderful people. Please allow me to express my sincere gratitude to them.

To begin with, I dedicate this thesis to my mother, Marianne, who passed away just after I had started the research master in 2014. When it came to my education, she always trusted my judgment and encouraged me to continue studying. I am also grateful for the support received from my father Egbert and my brother Michiel. A couple of years ago, my girlfriend Evelien joined my personal support team. Eef, I cannot describe how much I love you and what you mean to me. We never really talk about research, and I would like to keep it that way. Thank you for your understanding when I want to send an email at 10 pm or grade exams on a Sunday. Furthermore, I thank Jeanette, Vries Sr., Marijke, Judith and Ruud for their support, especially in the aftermath of my mother's death. In fact, I still consider passing all

my courses in the first year of the research master as my biggest achievement.

In the research realm, I am indebted to Jens Prüfer for encouraging me to apply for the research master program in Tilburg. Back in 2013, Jens was the second-reader of my BSc thesis (on the use of performance-enhancing drugs in cycling), and he made me aware of the existence of the research master program. After having finished this program in 2016, I still was not done in academia and started my PhD.

During this PhD, I enjoyed a lot of support from my supervisors. My first supervisor Jan Potters and I bonded over our interest in cycling, which sparked our first paper together about breakaways in cycling. I very much appreciate his ever constructive feedback, which never disqualified my ideas, but instead refined and reshaped them. Moreover, I am grateful for his confidence in me when it came to teaching a course in the MSc Economics program and organizing the TIBER Symposium. I am also indebted to the crucial feedback received from Eric van Damme, my second supervisor, who always challenged my assumptions by offering an alternative perspective.

This doctoral thesis would not have been completed without the members of my committee. I thank Prof. Georg Kirchsteiger, Dr. Jan Stoop, Prof. Eline van der Heijden, and Prof. Sigrid Suetens for their willingness to be a member of this committee and their incredibly valuable comments which have definitely improved my dissertation. I remember being rather nervous before the pre-defense, but I actually experienced it as a pleasant academic discussion.

Over the years, I enjoyed a fruitful collaboration teaching microeconomics to first-year students with Boris van Leeuwen and Eline van der Heijden. Like Jan and Eric, they belong to the “Experimental and Behavioral Economists at Tilburg.” I am grateful for the feedback received from the other current and former faculty members and PhD students in this group: Pascal Achard, Paul van Bruggen, Elena Cettolin, Patricio Dalton, Sebastian Dengler, Riccardo Ghidoni, Victor González, Lenka Fiala, Gijs van de Kuilen, Manwei Liu, Wieland Müller, Phúc Phùng, Julius Rüschenpöhler, David Schindler, Gyula Seres, Yi Sheng, Eli Spiegelman, Daan van Soest, Sigrid Suetens, Stefan Trautmann, Ben Vollaard, Yilong Xu, Jierui Yang, Yadi Yang, and Wanqing Zhang. I have always experienced the meetings of this group as a safe environment to present my research ideas and discuss those of others. During my time at the GATE institute in Lyon, I had the pleasure to collaborate with Fabio Galeotti and Marie Claire Villeval, whom I thank for welcoming me so hospitably and sharing their research ideas with me. Furthermore, I have received valuable comments on my chapters throughout the years from Julien Bénistant, Raphael Epperson, Simon Gächter, Thomas Garcia, Clément Gorin, John Hamman, Václav Korbek, Peter Mof-

fatt, Jens Prüfer, Louis Raes, Hannah Schildberg-Hörisch, Claire Rimbaud, Florian Schütt, Florian Sniekers, Alice Solda, Rémi Suchon, Robert Sugden, Matthias Sutter, Barbora Sýkorová, Vincent Théroude, Bertil Tungodden, Morgan Ubeda, and Adam Zylbersztejn. Outside of my direct field of interest, I enjoyed support from my education coordinator Burak Uras and the heads of department Jan Boone, Reyer Gerlagh and Sjak Smulders, especially in relation to my contract extension and teaching load. Finally, I am indebted to the secretarial support team at CentER and the Department of Economics consisting of Cecile, Aislinn, Ella, Renée, Corina, Ank, Bibi, and Korine.

From a social point of view, I enjoyed a lot of support from my fellow students. First and foremost, I am extremely grateful having had Lenka as my office mate over the last five years. I am going to miss mocking student answers or bashing Donald Trump. Second, I enjoyed sharing an apartment with Santiago for three years. I thank him for being such a pleasant and agreeable housemate. Furthermore, I enjoyed having lunch, drinking beers at Kandinsky, playing D&D, canoeing in the Czech Republic, and simply spending time on- and off-campus with Albert, Ana, Clemens, Dorothee, Frank, Freek, Hanan, Hugo, Jantsje, Laura, Liz, Lucas, Madina, Manuel, Manwei, Marie, Mirthe, Oliver, Pepijn, Peter, Ricardo, Richard, Roweno, Sebastian, Shan, Sophie, and Takumin.

Many others have enabled me to switch off from research and enjoy myself in Tilburg. I thank the group of friends from the Bachelor program which so humbly calls itself *Bazen*. In particular, I would like to thank Yuri, Vincent, and Sander for making my early years in Tilburg so great. Furthermore, I couldn't have wished for a better place to live than the Godefridus Mansion: Danny, Jolijn, Lisa, Lisanne, Michiel Noortje, and Quinn, you guys made my life bearable every time I was fighting the multi-headed monster that is the research master. I happily drank a few beers after a horrendous exam or when I could not solve some stupid assignment. Undoubtedly, you have become close friends over the years, and I am happy to say that we are still in touch years after I have left the house.

As the epigraph at the beginning of these Acknowledgements suggests, I never forgot my roots. Back home in Nijmegen, I had already established a valuable social network before moving to Tilburg. My best friends are still those with whom I connected in high school: Ivo, Jimmy, Joost, Koen, Leon, Mart G., Mart P., Micky, Nicky, Rik, Sebastiaan, and Timo. During our regular meet-ups all across the country as *Oppidum Batavorum*, I am reminded again and again that, even though it has been 10 years since we graduated from high school, fun is brought to another level with you. During this lockdown, I miss hanging out with you and notice that absence does

make the heart grow fonder. I look forward to creating more stories and adventures to tell our kids once the pandemic is over.

Finally, I have always continued to play football and have spent the better part of my Sundays and Saturdays playing games in and around Nijmegen. For the last five years, I am glad to have been part of RKS V Brakkenstein 5, better known as Brakka 5. It appears that there is no better way to get your head off research than playing a nice game of football. And although I absolutely love the time spent on the pitch, I just as much enjoy our time spent off the pitch. Hopefully, by the time you read this, we have returned to playing games and enjoying some *pels* and *sneks* afterwards.

For now, I am curious for what the future has in store for me. As I know that most of you only read this part, thank you for your attention. For my true fans, happy reading!

Thijs Brouwer
Tilburg, February 23, 2021

Contents

List of Figures	x
List of Tables	xii
1 Introduction	1
2 Would You Trust Someone Who Cheats in Your Favor?	5
2.1 Introduction	5
2.2 Related Literature	8
2.3 Experimental Design	10
2.3.1 Baseline Treatment	10
2.3.2 No Pro-Sociality Treatment	13
2.3.3 Experimental Procedure	14
2.3.4 Theoretical Framework and Hypotheses	15
2.4 Data Description	18
2.5 Results	19
2.5.1 Behavior of S2	19
2.5.2 Believed and Actual Behavior of S1	29
2.6 Concluding Discussion	29
2.A Experimental Material	35
2.A.1 Instructions	35
2.A.2 Survey Questions	40
2.A.3 Selection of z-Tree Screens	43
2.B Supplementary Tables & Figures	49
2.B.1 Tables	49
2.B.2 Figures	51
2.B.3 Robustness: Restricted Sample	51

3	Teaching Children Norms in the Streets	55
3.1	Introduction	55
3.2	Related Literature	59
3.3	Experimental Design	61
3.3.1	Conditions and Conjectures	61
3.3.2	Procedures	65
3.4	Data Description	70
3.5	Results	71
3.5.1	Main Results	71
3.5.2	Robustness Tests	75
3.6	A Vignette Study for Norm Elicitation	77
3.7	Concluding Discussion	81
3.A	Experimental Material	84
3.A.1	Instructions of the Experiment	84
3.A.2	Instructions of the Vignette Study	89
3.A.3	Additional Material	93
3.B	Supplementary Analyses	94
3.B.1	Summary Statistics of Survey Respondents	94
3.B.2	Effect of Parent's Gender on Punishment in Violation	95
3.B.3	Alternative Estimation Models	96
3.B.4	Analysis of Heterogeneous Effects	97
3.B.5	Analysis of Timing Effects	100
4	An Eye for a Tooth	105
4.1	Introduction	105
4.2	Related Literature	109
4.3	Experimental Design	111
4.3.1	The Production Game	111
4.3.2	Preference-Elicitation Tasks	116
4.3.3	Survey and Procedures	116
4.4	Model and Hypotheses	120
4.4.1	Set-Up	120
4.4.2	Utility	120
4.5	Data Description	126
4.6	Results	128
4.6.1	Manipulation Check	128
4.6.2	Result 4.1	131

4.6.3	Result 4.2	133
4.6.4	Regression Analysis	133
4.6.5	Stage 1 Outcomes following Option B	135
4.6.6	Other Results	138
4.7	Concluding Discussion	140
4.A	Theoretical Derivations	143
4.A.1	Derivation of Kindness Term	143
4.A.2	Worker Choice of Alternative	144
4.A.3	Employer Behavior	145
4.A.4	Risk Preferences	146
4.B	Experimental Materials	148
4.B.1	Instructions	148
4.B.2	Survey Questions	156
4.B.3	List of Red Cross Projects in the Experiment	159
4.B.4	Selection of z-Tree Screens	166
4.C	Supplementary Analyses	174
4.C.1	Distribution of Performance	174
4.C.2	Mood	175
4.C.3	Mistakes	176
4.C.4	Outliers	177
4.C.5	Worker's Stage 1 Choice	177
4.C.6	Effects of CoViD-19 Measures on Sample and Results	181

Bibliography

185

List of Figures

2.1	Comparison of Trust between Cases in Baseline	22
2.2	Distributions of Trusting Decisions across Cases in Baseline	23
2.3	Distributions of Trustworthiness across Cases in Baseline	24
2.4	Trust Dependent on Own Stage 1 Choice in Baseline	25
2.5	Comparison of Trust between Baseline and No-Pro	26
2.6	Actual and Believed Trust of S1 in Baseline	30
2.7	Die Roll (Discolored)	43
2.8	Die Report (Discolored)	44
2.9	Summary Senders (Discolored)	45
2.10	Summary Receivers (Discolored)	46
2.11	Choosing a (Discolored)	47
2.12	Choosing b Using Strategy Method (Discolored)	48
2.13	Comparison of Trust across Cases in Baseline	51
2.14	Trustworthiness for Each Level of Trust in Baseline	51
2.15	Comparison of Trust between Cases in Baseline with Restricted Samples	53
3.1	Behavior of Parents, by Treatment and Condition	72
3.2	Materials Used in the Experiment and Scenes	94
4.1	Extensive-Form Depiction of Stage 1 in Intentions	114
4.2	Experimental Design Summary	119
4.3	Average Mood under Option A and Option B	131
4.4	Worker Performance across Options and Treatments	134
4.5	Net Productivity per Stage 1 Outcome	137
4.6	Production Game	150
4.7	Page 1 of Catalogue (Discolored)	151
4.8	Page 2 of Catalogue (Discolored)	152
4.9	Page 3 of Catalogue (Discolored)	153

4.10	Field of Boxes in Part 2 (Discolored)	154
4.11	Matching of Participants in Part 3	155
4.12	Stage 1 Decision of Employer (Discolored)	166
4.13	Stage 1 Expectation of Worker (Discolored)	167
4.14	Alternatives of Worker (Discolored)	168
4.15	Random Draw in Alternative 2 (Discolored)	169
4.16	Random Draw in Alternative 2 - Outcome (Discolored)	170
4.17	Coin Identification Task - Denomination (Discolored)	171
4.18	Coin Identification Task - Country-of-Origin (Discolored)	172
4.19	Coin Identification Task - Performance (Discolored)	173
4.20	Distribution of Performance across Options and Treatments	174
4.21	Excitement of Workers in Stage 1	175
4.22	Degree of Upset of Workers in Stage 1	175
4.23	Shame of Workers in Stage 1	175
4.24	Hostility of Workers in Stage 1	176
4.25	Determination of Workers in Stage 1	176
4.26	Worker Performance Before and After CoViD-19 Outbreak	183

List of Tables

2.1	Payoff Structure in Stage 1 of Baseline and No-Pro	12
2.2	Number of S2s per Case in Baseline and No-Pro	19
2.3	Summary Statistics and Balance of S2s across Cases	20
2.4	Summary Statistics and Balance of S2s across Treatments	21
2.5	Linear Regressions on Trust	28
2.6	Mover 1's Actions and Consequences for Both Accounts	37
2.7	Mover 2's Actions and Consequences for Both Accounts	38
2.8	Trust of S2s in Baseline	49
2.9	Trust of S2s in Baseline When Conditioning on Own Behavior	49
2.10	Comparison of S2 Trust between Baseline and No-Pro	50
2.11	Beliefs of S2 about S1's Degree of Trust	50
2.12	Comparison of Responsiveness to a across Rolls	50
2.13	Trust of S2 When Restricting Sample	52
3.1	Treatments	62
3.2	Summary Statistics	71
3.3	Regression Analyses of Punishment Rate and Helping Rate	73
3.4	Robustness Checks for Punishment and Helping	76
3.5	Social Appropriateness of Passerby Behavior	80
3.6	Summary Statistics of Survey Respondents	95
3.7	Punishment in the Violation Treatment	96
3.8	Marginal Effects of Logit Estimations	97
3.9	Secondary Analyses of Punishment Behavior	101
3.10	Secondary Analyses of Helping Behavior	102
3.11	Role of the Timing of Scenes on Punishment and Helping	103
4.1	Number of Workers in Each Cell	126
4.2	Summary Statistics of Workers and Balance across Treatments	129

4.3	Summary Statistics of Workers and Balance across Options	130
4.4	Average Performance in Each Cell	131
4.5	Regression Analysis of Net Productivity	136
4.6	Worker Behavior in Intentions Treatment by Expectations	139
4.7	Employer Outcomes in Each Cell	140
4.8	Red Cross Projects	159
4.9	Regression Analysis of Mistakes	178
4.10	Regression Analysis of Winsorized Net Productivity	179
4.11	Regression Analysis of Winsorized Mistakes	180
4.12	Multinomial Logit Analysis of Alternative Chosen	181
4.13	Summary Statistics Before and After CoViD-19 Outbreak	182

Introduction

As can be inferred from its title, *Essays on Behavioral Responses to Dishonest and Anti-Social Decision-Making*, this doctoral thesis presents three essays that discuss an experimental approach to the punishment of different types of undesirable behavior. Throughout this thesis, these are deemed undesirable because the prospective punisher considers them as unkind, norm-violating, or both. For example, dishonest behavior violates the universally-accepted norm of honesty, but may also come at the expense of others' well-being. The commonality of the types of undesirable behavior is that they do not necessarily hurt the punisher *directly*. Instead, they include (the threat of) harming a passive third party or simply constitute behavior of which the punisher may disapprove, like littering, speeding, or vandalism.

My motivation for studying the punishment of undesirable behavior is twofold. First, this doctoral thesis aims to identify important potential externalities and spillovers of undesirable behavior that may have been overlooked in previous research. For example, while an employee misleading a customer obviously hurts said customer, such practices may also harm trust and cooperation within the organization when other employees frown upon them. To this end, Chapters 2 and 4 examine the effect of dishonest and anti-social behavior on others' social preferences. Second, this doctoral thesis aims to shed light on the transmission of compliance with social norms to future generations. Since punishment may deter norm violations in the future, Chapter 3 assesses whether parents punish more often in front of their children, in order to educate them about norms. Importantly, no one is hurt by the norm violation in this chapter, and punishment is assumed to be motivated by a desire

to promote future norm compliance in society. Whichever of these two motivations you consider, punishment is costly in terms of time, resources, or the threat of being counter-punished. This may inhibit the use of punishment in practice and makes it an interesting domain to study. Its relevance has increased in light of the current CoViD-19 pandemic, during which behavioral responses to undesirable behavior may play an important role in encouraging compliance with government measures and identifying indirect effects of the crisis. I now preview my three core chapters below.

Chapter 2, entitled *Would You Trust Someone Who Cheats in Your Favor? An Experimental Study*, examines a situation in which dishonest behavior by one individual is accompanied with a benefit to the prospective punisher. In a laboratory experiment, I create a situation in which one participant (Sender 1) can increase the payoff to himself and another participant (Sender 2) by misreporting the outcome of a die-roll, which hurts a third subject in the experiment (Receiver). Sender 2 observes both the die-roll and the report, and thus knows whether Sender 1 was honest or dishonest. While reporting dishonestly obviously violates the norm of honesty, it could be appreciated by Sender 2 because it increases her payoff. Analogously, while reporting honestly could be deemed praiseworthy, it could also be frowned upon by Sender 2 because it fails to increase her payoff. I examine whether reporting dishonestly is appreciated more than reporting honestly by examining the level of trust displayed by Sender 2 in a subsequent economic game in which trust plays an important role. Since trust is an important factor fostering cooperation within organizations and increases efficiency, there could be important and overlooked positive or negative externalities from behaving dishonestly in a team setting.

I find that participants in the role of Sender 2 in my experiment do not display a higher level of trust following having observed either a dishonest Sender 1 or an honest Sender 1. This does not change when I split the sample based on the Sender 2's own, non-materialized choice. At the same time, trust is higher in the situation where the die-roll yields the highest possible outcome by chance and there is no incentive for Sender 1 to report dishonestly to begin with. Moreover, I find trust to remain unaffected when I run a different treatment in which the dishonest choice no longer increases the payoff of the other participant. Together, these results suggest that the mere presence of conflicting motives suffices to erode trust. This interpretation may thus have important implications for organizations and other entities within which individuals work in teams. In particular, organizations would better expend efforts to reduce as much as possible the extent to which conflicting motives are present, rather than instructing employees how to behave in the face of such situations.

Chapter 3, entitled *Teaching Children Norms in the Streets* (co-authored with

Fabio Galeotti and Marie Claire Villeval), examines the role that punishment and reward play in the transmission of social norms from one generation to another. After all, social learning theory (Bandura, 1977) posits that children internalize norms even at a young age through the observation and subsequent imitation of parents. In this chapter, we show that parents may not only teach their children by displaying desired behavior, but also by punishing undesirable behavior of others in front of them. By doing so, the child learns that violations of the norm will not go unpunished, which should in turn promote future norm compliance. To examine this, we conduct a field experiment in the proximity of French elementary schools, which allows us to exogenously vary the presence of a child for an otherwise comparable set of parents. We hire an actor to play one of three different scenes in front of parents with or without their child(ren). In the first scene, the actor litters by throwing away a banana peel, which presents the parent with an opportunity to punish the actor. In the second scene, the actor accidentally drops his/her bag, which presents the parent with an opportunity to provide help. Finally, in the third scene, the actor first litters and then drops his/her bag, which presents the actor with an opportunity to punish directly *or* withhold help as a means of indirect punishment.

We find that parents are more likely to punish the littering violation in the presence of the child, as to compared to parents who are alone. They do so by confronting the actor verbally more often, but they are no more likely than parents alone to withhold help when the opportunity to help is preceded by a littering violation. In addition, we also find that parents are more likely to display desirable behavior in front of their child(ren), as exemplified by the higher likelihood that a parent helps the actor in the absence of any violation. Our results, a complementary vignette study, and our accompanying discussion isolate the parent's teaching motive as the interpretation of our results and we are able to discard alternative explanations relating to the parent's fear of retaliation, the parent's perception of the norm violation, or the parent's social image concerns. Our study contributes to the understanding of how social norms are transmitted from one generation to another and identifies an additional way in which parents can teach their children. Our experimental setting in which parents are not aware of being part of an experiment constitutes a major advantage of our study.

Finally, Chapter 4, entitled *An Eye for a Tooth: The Effects of Employer Pressure on Worker Productivity*, examines the effect of imposing a psychological cost onto a Worker by requiring her to hurt an innocent outside party. I motivate the topic of this chapter by referring to survey evidence and case studies showing the existence of unethical pressures within organizations which substantially reduce worker

motivation and engagement. Lately, this problem has resurfaced with some organizations allegedly forcing employees under the threat of being fired to travel to the office contrary to the government's CoViD-19 regulations. I capture this situation in a two-stage laboratory experiment where, in the first stage, one participant (the Employer) can enrich himself, which forces another participant (the Worker) to make a trade-off between her own payoff and a donation to a charitable organization. When the Employer abstains from enriching himself, both the Worker's payoff and the donation remain intact. Since the Worker dislikes making the trade-off, she may resent the Employer when he enriches himself and puts the Worker on the spot. As a result, the Worker may want to retaliate this action. She receives an opportunity to do so in the second stage of the experiment, where the Worker is asked to perform a task where the quality of her performance benefits (or hurts) the Employer *only*. I hypothesize that the Worker will perform worse in this task when the Employer put her on the spot. Note that this hypothesis is not necessarily trivial, since the Worker still holds her destiny and that of the charitable organization in her own hands. Hence, she could just as much blame herself for the outcome of the first stage.

Still, I find that Workers perform worse in the task when the Employer put them on the spot in the first stage. By comparing performance in this baseline version of the experiment to performance in a version of the experiment where a random draw, instead of the Employer's deliberate choice, determines whether or not the Worker has to make the trade-off, I identify the Employer's intentions as the main force driving the results. More specifically, I show that the Worker punishes when the Employer intentionally chooses to impose the trade-off onto the Worker, but that she does not reward the Employer when he abstains from doing so. This suggests that only negative reciprocity is driving the Workers' behavior. In addition, I show that the reciprocal response does not depend on the actual outcome of the trade-off, which further shows that the Worker reciprocates the psychological cost from the trade-off. With this study, I highlight the importance of elements in workplace relationships that go beyond monetary incentives.

Together, the three essays in my doctoral thesis do not simply study undesirable behavior. Instead, they go one step further and examine how people respond to it. At the same time, my persons of interest in the different studies face trade-offs that do not make it trivial that they will respond in the hypothesized way. Therefore, I believe that my doctoral thesis contains insights that may be of use to others in the profession.

Would You Trust Someone Who Cheats in Your Favor? An Experimental Study

Oh, please! Where was all this conscience when I got us in the first-class lounge at the airport [...]? You know what you are? You're like a mob wife. You look down at me and my ways, but you're happy to wear the mink coat that fell off the back of the truck.

MITCHELL PRITCHETT
Modern Family - 'Earthquake' (S2, E3)

2.1 Introduction

In this chapter, I examine the effect on trust of what the literature calls “Pareto white lies” (Erat and Gneezy, 2012) or “self-serving altruism” (Gino et al., 2013): acts of dishonesty that benefit the perpetrator *and* some passive beneficiary. For example, a student could plagiarize a group assignment to the benefit of the entire

group or a salesman could deceive a customer to ensure a bonus for the entire sales team. Importantly, these dishonest acts seldom occur in isolation, since students and co-workers frequently meet again in future interactions. When deciding whether or not to trust someone who cheated in her favor, the passive beneficiary needs to evaluate this “ethically ambivalent act” (Levine and Schweitzer, 2014, p. 108) on two conflicting dimensions. On the one hand, acting dishonestly can be considered kind in the *payoff* dimension, since it increases the beneficiary’s payoff. On the other hand, acting dishonestly sends a bad signal in the *moral* dimension, as it may signal (future) malevolence. The passive beneficiary faces the opposite moral trade-off when confronted with honest behavior: she may perceive the other’s honest behavior as a signal of moral virtue or as a missed opportunity to increase her payoffs. In this chapter, I study the outcome of this moral trade-off and attempt to answer the following research question: What is the effect of dishonesty that benefits both the decision maker and an additional individual on the trust of the latter towards the former? Throughout, trust encompasses both (the act of) *trusting*, defined as the willingness to accept vulnerability based upon positive expectations about another’s behavior (Rousseau et al., 1998), and *trustworthiness*, defined as reciprocating the trusting act with benevolent behavior that is inherently costly to the trustee.

I capture the moral dilemma between honesty and higher payoffs in a two-stage decision-making laboratory experiment in pairs of “Senders”, whose members are called “S1” and “S2”. The experiment consists of two treatments: a Baseline treatment and a No-Prosociality (No-Pro) treatment. In Stage 1 of the Baseline treatment, each Sender independently engages in a modified version of the die-under-the-cup game (Fischbacher and Föllmi-Heusi, 2013) that allows for an opportunity to increase the payoff of *both* Senders by untruthfully reporting the outcome of a die-roll, at the cost of a third subject in the experiment, called the “Receiver”. Only one of the Senders’ choices is implemented, such that only one of the Senders affects the payoff allocation. As opposed to the Baseline treatment, Senders in the No-Pro treatment can only increase their *own* payoffs by untruthful reporting, thereby removing the alignment of payoffs between Senders in Stage 1.

Subsequently, in Stage 2 of *both* treatments, the same two Senders engage in an extension of the Trust Game, called the Moonlighting Game (Abbink et al., 2000). While the Trust Game only measures trusting and trustworthiness, the Moonlighting Game additionally picks up fear of exploitation and negative reciprocity by allowing subjects to *reduce* the other player’s payoff. Trusting the second-mover boils down to a first-mover transferring a positive amount of money to the second-mover. Similarly, trustworthiness is characterized by positive back-transfers from second- to first-mover.

I examine exhibited trust levels of the Sender whose choice was not implemented, whom I assume to be S2. I compare the behavior of S2s paired with an S1 who reported dishonestly in Stage 1 (*i.e.*, the Dishonest case), those paired with an honest S1 (the Honest case), and those paired with an S1 who did not need to misreport the outcome of the die-roll to obtain the highest payoff (the Neutral case). This Neutral case forms a useful benchmark as it contains no moral dilemma and no trade-off between higher payoffs and honesty.

The conflicting motives that characterize the Honest and Dishonest cases ensure that a clear-cut ex-ante prediction cannot be made, but one would expect the Neutral case to be in-between the two. However, in the experiment, I find the highest levels of trust in the Neutral case, while I find no differences in trust between the Honest and Dishonest cases. This result is most pronounced for the act of trusting, with S2s in the Neutral case sending positive amounts to the second-mover, while S2s in the Honest and Dishonest cases seem to inflict punishment on and *take* money from the second-mover, instead. For trustworthiness, I obtain noisier evidence for higher back-transfers in the Neutral case as compared to the Dishonest and Honest cases. That is, differences are insignificant using non-parametric tests, while a random-effects model provides evidence of significant lower trustworthiness as compared to the Neutral case in the Honest case *only*.

In order to assess whether S2's own preferences for higher payoffs and honesty affect trust, I split the sample according to her own, non-materialized choice. One might expect S2 to respond more favourably to an S1 who made the same choice in Stage 1. However, I find this not to be the case, as I find no differences between the Honest and Dishonest cases in both subsamples. This could be in line with Gross et al. (2018), who demonstrate the presence of *ethical free-riding*, *i.e.*, subjects who behave honestly while having no problem profiting from others' dishonest behavior.

Finally, I use the No-Pro treatment to assess what happens to behavior when the pro-social component is removed, *i.e.*, when only S1 gains from behaving dishonestly. This treatment, which effectively eliminates the trade-off, serves as a useful benchmark to which the results in the Baseline treatment can be compared. After all, dishonest acts in the No-Pro treatment should become unambiguously more negatively judged and responded to than dishonest acts in the Baseline treatment, while the opposite applies to honest acts. Surprisingly, however, I find no differences between the Baseline and the No-Pro treatment, suggesting that S2 may not evaluate Stage 1 behavior at all or that the pro-social component is unimportant to her.

My results indicate that the High roll itself increases the trust of S2s. In the Discussion section, I discuss two competing explanations. First, it could be that

S2 requires both the payoff and the honesty dimension to be fulfilled in order to become more trusting of S1, instead of making a simple trade-off between the two. In other words, S2 trusts more if S1 increases her payoff *and* if he does not lie. Since this is only possible when the roll is High, trust is highest in the Neutral case. Second, the moral dilemma imposed on her (*i.e.*, the Low roll) could bring S2 in a negative psychological state that makes her less willing to cooperate. This explanation relates to the literature exploring the relationship between mood and pro-social behavior (Capra, 2004; Dunn and Schweitzer, 2005; Kirchsteiger et al., 2006; Proto et al., 2019). Either mechanism implies that the mere presence of conflicting motives suffices to erode trust. When valid, this interpretation may have important implications for organizations and other entities in which individuals work in teams and implies that efforts should be better expended to preventing these teams from being faced with moral dilemmas in the first place, rather than instructing employees how to behave in the face of such situations. Future efforts could be devoted to exploring this mechanism. Another avenue for future research concerns the role of communication. Being able to explain and/or justify S1's decision may resolve the ambiguity surrounding his choice from S2's perspective. Although unquestionably relevant in actual workplaces, this feature is absent in my experimental design.

The remainder of this chapter is organized as follows. I discuss relevant literature in Section 2.2. In Section 2.3, I introduce the experimental design (Subsections 2.3.1, 2.3.2, and 2.3.3) and provide the theoretical framework (Subsection 2.3.4). The latter subsection also contains the hypotheses. Then, in Section 2.4, I discuss the data, and in Section 2.5, I present the experimental results. Section 2.6 discusses these results and concludes.

2.2 Related Literature

Although the propensity to tell Pareto white lies has been studied extensively in the literature (see *e.g.*, Wiltermuth, 2011; Erat and Gneezy, 2012; Gino et al., 2013; Shalvi and De Dreu, 2014; Weisel and Shalvi, 2015), few studies have examined how it affects others' attitudes towards the decision-maker. Closest related is Levine and Schweitzer (2015), who also examine trust after having observed deceptive behavior in a similar manner. The authors show that mutually beneficial lies on the outcome of a coin flip can actually breed benevolence-based trust (*i.e.*, the type of trust measured by the trust game), as subjects reward the benevolent intentions of the untruthful decision-maker. At the same time, pro-social dishonesty is shown to harm integrity-based trust, which requires one to rely on the integrity of the untruthful

decision-maker. In addition to using different experimental games, my study differs in a few important respects. First, I more explicitly introduce the harmful side of behaving dishonestly by including a third player who is harmed. Second, I examine trustworthiness in addition to the act of trusting, which may exhibit distinct patterns. Third, Levine and Schweitzer (2015) make use of experimental deception by using confederates as the (always) dishonest player and having the coin flip always yielding the same outcome. Fourth, my experimental design allows for the examination of explicitly *honest* behavior that leaves money on the table.

More generally, this study adds to the literature examining behavioral responses to observing, experiencing, or undertaking dishonest behavior. For example, Levine and Schweitzer (2014) demonstrate that pro-social liars are sometimes viewed more moral than honest individuals. They link these results to the two moral foundations of justice and care, and show that when these two clash, the latter tends to prevail. Ohtsubo et al. (2010) and Konishi and Ohtsubo (2015) extend the concept of costly third-party punishment to the realm of dishonest behavior. They show that a considerable share of observers engages in costly punishment of dishonest messages in the Trust Game. Similarly, whistleblowing can be viewed as potentially self-destructive punishment as it may entail the loss of own payoffs or lead to exclusion from future interactions. Reuben and Stephenson (2013) show that although a sufficient number of subjects are willing to report lies so as to render misreporting unprofitable, the authors also find that whistleblowers are more likely to be vetoed from entering an organization. In the same spirit, Bartuli et al. (2016) show that roughly one third of participants blow the whistle on a money-embezzling manager, even though this terminates the employment contract. These studies illustrate the presence of moral preferences among at least a subset of subjects that induce them to punish dishonestly behaving fellow participants, even at the expense of themselves. On the other hand, Gross et al. (2018) document that subjects who themselves behave honestly in collaborative cheating tasks have no issue being paired with a dishonest subject, as exemplified by the former's wish not to switch partners. These contrasting views from the literature validate the current research question.

Since dishonest behavior may signal something about predicted play in the Moonlighting Game, studies examining the correlation between dishonesty and social preferences may be relevant. However, there seems to be little consensus on the relationship between the two. On the one hand, altruistic subjects are less likely to tell a lie that hurts another participant in the experiment (Kerschbamer et al., 2019) or benefits both participants (Biziou-van Pol et al., 2015; Cappelen et al., 2013). On the other hand, they are more likely to tell a lie that hurts themselves to the bene-

fit of another participant (Biziou-van Pol et al., 2015). Thus, in the context of my experiment, it seems hard to determine what precisely is conveyed about a subject's social preferences by his/her (dis)honest behavior.

2.3 Experimental Design

In order to study the conflict between honesty and benevolence, I design a two-stage laboratory experiment. Subjects take on the role of Sender or Receiver, with a Sender being matched to another Sender and one Receiver. Stage 1 of the experiment mimics a situation in which one Sender's dishonest act benefits himself and the other Sender, while hurting the Receiver. Subsequently, in Stage 2 of the experiment, the two Senders engage in an exchange in which trust plays an important role. All payoffs in the experiment are denoted in experimental currency units, with 3 ECU equaling 1 Euro.

I choose to include a Receiver, albeit passive, to make it salient that misreporting is hurting another player. Each session has two Receivers, so that Receivers are matched to several Sender pairs at the same time (Senders are aware of this). This way, I maximize the number of Senders – my subjects of interest – in each session. Moreover, this means that a session can be run with any sufficiently large even number of subjects.

My experiment consists of two treatments: a Baseline treatment and a No Pro-Sociality (No-Pro) treatment. Below, I discuss the two separately.

2.3.1 Baseline Treatment

Stage 1 In Stage 1, subjects play a modified version of the die-under-the-cup game (Fischbacher and Föllmi-Heusi, 2013). The Senders (called S1 and S2) are shown the *same* randomly-selected videotaped die-roll (this is common knowledge) and have to report the outcome *separately* to the Receiver (R3) who has not observed the roll, by sending a message $m_i \in M, i = 1, 2$. Importantly, and crucial to the die-under-the-cup paradigm, payoffs are determined solely by the message and not by the *actual* die-roll. The die-roll thus serves as a way to establish the *true* state, from which a lying-averse individual may be apprehensive to deviate.

I pre-recorded a video for each of the six potential outcomes of the die-roll. Pre-recorded die-rolls have been used in Kocher et al. (2017) and allow me to observe die-rolls on the individual level, as opposed to the traditional approach pioneered by Fischbacher and Föllmi-Heusi (2013) where the die-roll is private information.

The die-roll can be either Low or High. In order to create plenty opportunities for dishonest reporting, a roll is considered High (H) only if it is a 5 or a 6, and Low (L) otherwise. Consequently, the Senders' message space is $M = \{L, H\}$.¹ However, only one of the two messages m_1 and m_2 actually reaches R3. This message m^* is randomly selected from the two messages and determines payoffs to both Senders. I choose to have both Senders select a message in order to be able to relate a Sender's (non-materialized) choice to behavior in the Moonlighting Game. This feature serves as the foundation for my second hypothesis discussed in Subsection 2.3.4. Without loss of generality, I assume S1 to be the player whose report is selected as m^* . As a result, S2 is our player of interest in Stage 2.

Payoffs are depicted in Panel A of Table 2.1. If $m^* = L$, S1 and S2 obtain a payoff of 6 experimental currency units (ECU) and R3 obtains 15 ECU. For $m^* = H$, payoffs are 12 ECU for S1 and S2 and 3 ECU for R3. Hence, whenever the roll is Low, Senders can try to improve their payoff by untruthfully sending the High message instead. Given the definition of Low and High rolls, around two-thirds of Sender pairs are expected to be presented with an incentive to send an untruthful message. Analogously, one-third of Sender pairs observe the High roll and obtain the highest outcome without having to lie. This provides me with a benchmark case to which I can compare behavior.

R3, who does not know the payoff structure, receives one m^* from each pair he is matched with (between 2 and 5 pairs) and one of these messages is selected at random for payment. This is done to avoid any concerns for severe disadvantageous inequality among Senders, which may occur if Receivers receive their payoffs from all pairs to which they are matched. An important caveat of this design choice is that sending the High message is efficient since the payoffs for the Senders are certain while the payoff for the Receiver only materializes if this report is actually chosen to be payoff-relevant. Moreover, this set-up leads to a dilution of responsibility from the point of view of the Senders, as it is uncertain whether the Receiver is affected by S1's choice. As a result, this may attenuate the extent to which lying is frowned upon by S2. I return to this issue in Section 2.6.

Transition Senders are informed that they will be matched to the same Sender in Stage 2 of the experiment *after* Stage 1.² Moreover, they are informed of each other's choices and of which choice was implemented. Even though this ensures that S1's

¹In particular, the message reads: "The die-roll was low/high."

²Matching for Stage 2 is never mentioned in the instructions before this point. As such, we do not deceive subjects, although we do withhold information.

Table 2.1: Payoff Structure in Stage 1 of Baseline (Panel A) and No-Pro (Panel B)

A. Baseline				B. No-Prosociality			
		MESSAGE (m^*)				MESSAGE (m^*)	
		Low (1-4)	High (5-6)			Low (1-4)	High (5-6)
ROLL	Low (1-4)	6, 6, 15	12, 12, 3	ROLL	Low (1-4)	6, 12, 15	12, 12, 3
	High (5-6)	6, 6, 15	12, 12, 3		High (5-6)	6, 12, 15	12, 12, 3

Note: The table contains the payoff allocations in the Baseline (Panel A) and the No-Prosociality (Panel B) treatment. In each cell, the first entry denotes the payoff of the deciding Sender (S1), the second that of the other Sender (S2), and the third that of the Receiver (R3). Payoffs in ECU.

beliefs about S2's choice in Stage 1 are controlled for, this informational symmetry may create additional reciprocity between Senders in the Moonlighting Game. I discuss this potential confound at more length in Subsection 2.3.4.

Stage 1 could have four outcomes: (1) the die-roll was low and S1 reported honestly (referred to as the Honest case); (2) the die-roll was low and S1 reported dishonestly (Dishonest); (3) the die-roll was high and S1 reported honestly (Neutral); or (4) the die-roll was high and S1 misreported. The latter case is unlikely to occur and indeed I do not find a single instance of case (4).³ As a result, I feel comfortable dubbing case (3) the Neutral case. Since matching to S1 is random, one, and only one, of the three cases is imposed on S2 independently from her own choices and characteristics. This ensures a between-subjects design in which S2s can be expected to be similar across the three cases.

Stage 2 Senders play a Moonlighting Game (Abbink et al., 2000; Falk et al., 2008) in which S1 and S2 make decisions as both first- and second-mover using the strategy method.⁴ The Moonlighting Game has the following structure (I use the parameters of Falk et al., 2008). Both players are endowed with 12 ECU. The first-mover can choose to either take money from or send money to the second-mover, that is, his actions entail $a \in \{-6, -5, \dots, 5, 6\}$. Any money taken from the second-mover ($a < 0$) is added to the first-mover's endowment. Any money sent ($a > 0$) is tripled by the experimenter and added to the second-mover's endowment. This embodies the trust aspect of the interaction: trusting another to reciprocate entails the opportunity for mutual benefit. Subsequently, the second-mover can choose to either punish or reward

³However, Utikal and Fischbacher (2013) show that nuns seemingly have a tendency to *underreport* die rolls. To the best of my knowledge, no nuns participated in the experiment.

⁴While Senders engage in the Moonlighting Game with each other in Stage 2, the two R3s in the session are matched to each other and also play the Moonlighting Game. This is mainly to prevent them from being inactive for the remainder of the experiment and I will not examine their behavior. In the following, I describe Stage 2 from the perspective of the Senders, but the same procedures apply to R3s.

the first-mover by choosing $b \in \{-6, -5, \dots, 18\}$. Punishment ($b < 0$) is costly, in that it costs the second-mover 1 ECU to punish the first-mover by 3 ECU. Rewarding the first-mover ($b > 0$) entails transferring money directly from the second- to first-mover. The second-mover can punish by up to six units and reward by as many as 18 ECU, as long as these actions do not yield negative payoffs to either of the subjects. By allowing both positive and negative transfers by the first-mover, I allow subjects to exhibit both trust in reciprocation and fear of exploitation. Similarly, by allowing for both punishments and rewards, I allow for both positive and negative reciprocity on the part of the second-mover. In sum, payoffs are $\pi_1 = 12 - a + \min\{b, 3b\}$ and $\pi_2 = 12 + \max\{a, 3a\} - |b|$ for the first- and second-mover, respectively.

Assuming selfish preferences, the subgame-perfect Nash equilibrium of the Moonlighting Game is for the first-mover to recognize that a second-mover will neither punish nor reward, since both actions are costly to her. As a result, the first-mover takes as much as he can from the second-mover, without facing negative consequences. However, this is generally not what is observed in laboratory experiments. In their baseline treatment, Falk et al. (2008) observe a median a of 1, and a corresponding median b of 2 (strategy method was used). Remarkably, 24 percent of first-movers decide to send everything, and this is reciprocated by second-movers with a median response of 9, *i.e.*, the equal split of the created surplus.

As a primary comparison, I examine differences in average a and b between the Neutral, Honest, and Dishonest cases. Since I use the strategy method, S2 is asked to report a b for every possible a before the actual decision is known. This allows for an appropriate comparison of S2's choice of b across the three cases independent of the actual a chosen.

2.3.2 No Pro-Sociality Treatment

The Baseline treatment suffices to study trust in situations where preferences for honesty and payoffs need to be traded off. A potential concern is that the two dimensions cancel out in the aggregate or even on an individual level when comparing the Honest and Dishonest case. As a result, I would observe no behavioral differences between these cases.

In order to disentangle the the effects of the payoff and honesty dimensions from each other, I modify Stage 1 of the Baseline treatment. In this No Pro-sociality (No-Pro) treatment, misreporting only benefits S1, while it leaves S2's payoff unaffected. As a result, S2 is not affected monetarily, but still has to evaluate S1's (dis)honesty. The payoff structure is depicted in Panel B of Table 2.1. As can be seen, I have

chosen to keep the gains of cheating for S1 and the losses therefrom to R3 identical to the Baseline treatment. Even though this constitutes a minimal modification, it changes the efficient option and may change the reference point from the perspective of Senders. I return to this potential issue in Section 2.6. Stage 2 of the No-Pro treatment is the same as in the Baseline treatment.

2.3.3 Experimental Procedure

The experiment was conducted in the CentERlab at Tilburg University. An even number of subjects was required for each session; in case of an odd number, one subject was randomly selected and asked to leave the laboratory (after having received the show-up fee). Subjects were randomly assigned to their workstations and matched to a fellow subject by the software. Subjects were told that “the experiment consists of two Parts and a survey”. They received instructions (in English) for Stage 1, which were read aloud by the experimenter. On the final screen of Stage 1, Senders were informed that they would play Stage 2 with the other Sender from Stage 1. At the same time, instructions for Stage 2 were distributed, which were again read aloud by the experimenter. Identical copies of the instructions used during the experiment can be found in Appendix 2.A.1.

In Stage 2, subjects first answered control questions about a hypothetical scenario. The experiment did not proceed until all subjects in the session answered all questions correctly. The experimenter was available to explain the dynamics of the game in private in case subjects repeatedly failed to answer the questions correctly. Subsequently, subjects first chose a , then b , and then reported their beliefs by adjusting sliders to their preferred position. While doing so, account totals were automatically updated, such that the consequences of the subject’s actions were made as clear as possible. The belief elicitation was incentivized in the following way: subjects earned 1 ECU per correct guess, 0.5 ECU if they were one off, and 0 ECU otherwise. The experiment was run using the experimental software z-Tree (Fischbacher, 2007). A selection of experimental screens is displayed in Appendix 2.A.3.

Subjects played the game only once. I added a post-experimental survey to obtain the subjects’ background characteristics and (social) preferences. The details of this survey can be found in Appendix 2.A.2. First, I measured subjects’ risk aversion in two distinct ways: I used a hypothetical Multiple Price List (Binswanger, 1981; Holt and Laury, 2002) and I asked subjects to rate their willingness to take risks on a scale from 1 to 10. Second, using the survey questions from the Global Preferences Survey (Falk et al., 2018), I measured subjects’ self-reported degree of trust, reciprocity, and

altruism on a 10-point scale. For reciprocity, three dimensions were assessed: indirect reciprocity, positive reciprocity, and negative reciprocity. In order to collect these dimensions into one variable for the econometric analysis, I average the responses in these three dimensions into one reciprocity variable. Finally, I also measure altruism through a hypothetical dictator game.

Subjects were paid their total earnings over the two stages in cash directly after the experiment, including the belief elicitation and a 3 Euro show-up fee. On average, each session lasted 50 minutes and subjects earned 12.13 Euro.

2.3.4 Theoretical Framework and Hypotheses

I assume that S2 derives utility from her own and S1's payoff, with the weight attached to S1's payoff determined by altruistic *and* reciprocal tendencies (akin to Charness and Rabin, 2002). That is, S2 evaluates S1's behavior in Stage 1 and consequently increases or reduces the weight attached to S1's payoff, depending on whether she evaluates Stage 1 behavior positively or negatively, respectively. S2 evaluates behavior along two preference dimensions: payoffs and honesty. In turn, these dimensions affect the weight attached to S1's payoff. On the one hand, S2 appreciates receiving higher payoffs due to S1's actions and this increases her weight attached to S1's payoff. On the other hand, S2 dislikes outcomes that are obtained through dishonest choices and this decreases her weight attached to S1's payoff.⁵ Then, when evaluating Honest or Dishonest behavior in the Baseline treatment, the outcome depends on the net result of the two dimensions: if she cares more about payoffs than honesty, then the evaluation of Dishonest behavior is positive and the weight put on S1's payoff is increased. Consequently, a larger weight translates into higher choices for a and b in the Moonlighting Game on average. The opposite reasoning applies when S2 cares more about honesty than payoffs. Hence, positively evaluated behavior in Stage 1 leads to more trusting and trustworthy behavior in Stage 2 from the side of S2, whereas negatively evaluated behavior achieves the opposite. I assume that S2 does not evaluate S1 in any way in the Neutral case, since the latter's choice to report the High roll reveals nothing about his predisposition and potential behavior in the Moonlighting Game.

⁵The relevance of such a predisposition has been illustrated by the literature on costly third-party punishment of norm violations (Fehr and Fischbacher, 2004), the insights of which have been shown to extend to violations of the truth-telling norm (Ohtsubo et al., 2010; Konishi and Ohtsubo, 2015). Whistleblowing (Reuben and Stephenson, 2013; Bartuli et al., 2016) can also be interpreted as behavior displaying an aversion to dishonest behavior.

Main Hypothesis Following the above discussion, it proves difficult to sign the predicted effect of Stage 1 outcomes. What is more, this sign is likely to be individual-specific, and there is no way of predicting which of the two cases will display higher levels of trust. This ex-ante ambiguity is also reflected in the literature studying both dimensions either jointly or independently. On the one hand, Levine and Schweitzer (2014) find that pro-social liars are perceived as more moral and Levine and Schweitzer (2015) find that pro-social lies tend to breed trust. This implies positive evaluations in the Dishonest case and negative evaluations in the Honest case. On the other hand, Cappelen et al. (2013), Bizziou-van Pol et al. (2015), and Kerschbamer et al. (2019) found that altruists are less inclined to behave dishonestly, even if this would benefit someone else in addition to themselves. Hence, honest behavior in Stage 1 could tell S2 that she is dealing with an altruist, which may provide her with a reason to evaluate honesty *positively*. I therefore formulate an undirected hypothesis regarding the difference between the Honest and Dishonest case. However, the discussion above predicts that the Neutral case, in which S1 does not go out of her way to increase S2's payoff *and* does not lie, should be in-between the other two. Hence:

Hypothesis 2.1 *The average level of trusting (a) and trustworthiness (b) of S2s differs between the Honest and Dishonest cases, with the Neutral case in-between.*

Own Message One might argue that S2's own choice in Stage 1 is informative of which of the two preference dimensions dominates. For example, if S2 chooses the Dishonest message, this may imply that when evaluating S1's choice she values the payoff dimension more than the honesty dimension. This is loosely in line with models like the one developed by Ellingsen and Johannesson (2008) and Levine (1998), which posit that subjects may be more inclined to behave pro-socially towards subjects that are similar to them. Reflecting this, Hypothesis 2.2 predicts that S2s who choose the same message in Stage 1 exhibit higher levels of trust than their counterparts whose message differed from the one sent by S1.

Hypothesis 2.2 *When the roll is low,*

- a. S2s who send the Honest message themselves exhibit higher trust in the Honest case as compared to the Dishonest case.*
- b. S2s who send the Dishonest message themselves exhibit higher trust in the Dishonest case as compared to the Honest case.*

As explained in Subsection 2.3.1, both Senders are informed of each other's choice. Admittedly, this informational symmetry may invite additional reciprocity that would be absent if S1 would not know S2's non-materialized choice. Potentially, S1 could realize that S2 would have acted identically in Stage 1 and become more trusting of him, which S2 could realize in turn. This may reinforce reciprocity between Senders and yield higher levels of trust when the Senders' messages are identical, especially when they are both Dishonest and display the desire to increase each other's payoff. This mechanism may cloud any results related to Hypothesis 2.2, but also to Hypothesis 2.1. After all, if this additional reciprocity is indeed created in the Dishonest case, and not (as much) in the Honest case, then it affects the difference between the cases to the advantage of the Dishonest case. Hence, results should be interpreted with a bit of caution, especially when trust appears larger in the Dishonest case than in the Honest case.

No-Pro Finally, the simple framework can be adapted easily to incorporate the No-Pro treatment in which S1's decision to misreport does not have beneficial consequences for S2. Since S2 trades off the payoff and honesty dimensions, the evaluation should become less complicated in the No-Pro treatment. After all, payoff differences are absent for S2 in this treatment, which causes the preference for honesty to be the only relevant dimension of the evaluation. Compared to the Baseline treatment, S2 is expected to dislike Dishonest behavior more in the No-Pro treatment, since S2 no longer feels appreciation for S1 for having increased S2's payoff. Similarly, compared to the Baseline treatment, S2 is expected to approve of Honest behavior more in the No-Pro treatment, since S2 no longer feels resentment for S1 having failed to improve S2's payoff. As a result, the Honest case in the No-Pro treatment is predicted to feature higher levels of trust in the Moonlighting Game than the Honest case in the Baseline treatment. The opposite prediction can be made for the Dishonest case across treatments.

Hypothesis 2.3 *When comparing the No-Pro and Baseline treatments:*

- a. *S2s in the No-Pro Honest case display higher levels of trust than S2s in the Baseline Honest case.*
- b. *S2s in the No-Pro Dishonest case display lower levels of trust than S2s in the Baseline Dishonest case.*

Preferences vs. Beliefs The above discussion focuses on S2's other-regarding preferences vis-à-vis S1. That is, S1's behavior in Stage 1 affects the parameters of

S2's utility function and in particular the weight that S2 attaches to S1's payoffs. Alternatively, S1's behavior could affect S2's *beliefs* about S1's trustworthiness. For example, Dishonest behavior may lead S2 to believe that S1 is untrustworthy because he has violated a moral norm. Such considerations are exclusively important for the act of trusting. Hence, differences in a can be driven by preferences ("I (dis)like this person") and beliefs ("I (do not) expect this person to be trustworthy"), while differences in b can only be driven by preferences. This implies that a comparison between a , where beliefs play a role, and b , where beliefs are absent, can help to identify the role of beliefs. Moreover, since I also elicit beliefs explicitly, I can directly examine whether expected back-transfers differ between cases in a way that is in line with observed behavior. That is, if beliefs play a role, I should observe that differences in a across the cases trace differences in beliefs. I assess this prediction in Subsection 2.5.2.

2.4 Data Description

In total, 264 subjects participated in the experiment, divided over 17 sessions featuring between 12 and 20 subjects each. The data from one session needed to be discarded due to software problems (12 subjects).⁶ Furthermore, I dropped all subjects playing the role of R3 (32 subjects), since I am not interested in the behavior of these subjects. Finally, I identified three subject pairs that contained a subject who erroneously participated twice in the experiment. I maintain these in the sample, since excluding these would strengthen the results presented in Section 2.5. In total, my sample contains 220 subjects and 110 independent observations to be analyzed.

Of the 220 subjects, 120 are in the Baseline treatment and 100 in the No-Pro treatment. By construction, exactly half of these subjects take up the role of S2 (60 and 50, respectively). As could be expected, around two thirds of the rolls (74 out of 110) are Low. Of the 148 messages following a Low roll, 100 (68%) state that the roll was High instead. The fact that not all subjects are honest or dishonest is reassuring since it means that there exists variation that can be exploited. Only 1 subject observing the High roll, in the role of S2, sends a Low message. Interestingly, when looking at the choices of all Senders observing a Low roll, 56 out of 84 Senders (67%) in the Baseline treatment misreport, while 45 out of 64 (70%) misreport in the No-Pro treatment. It thus seems that the pro-social component does not induce Senders to lie much more in the Baseline treatment, which suggests that this motive does not play an important role for Senders. I return to this later.

⁶Erroneously, the videotaped die-rolls were not showing on the screen.

Table 2.2: Number of S2s per Case in Baseline (Panel A) and No-Pro (Panel B)

A. Baseline (60 obs.)					B. No-Pro (50 obs.)				
		MESSAGE (m^*)					MESSAGE (m^*)		
		Low	High	Total			Low	High	Total
ROLL	Low	10 (24%)	32 (76%)	42 (100%)	ROLL	Low	9 (28%)	23 (72%)	32 (100%)
	High		18 (100%)	18 (100%)		High		18 (100%)	18 (100%)

Note: Entries denote the number of S2s in each of the Neutral (lower-right cell), Honest (upper-left) or Dishonest (upper-right) cases as induced by the outcome of Stage 1. Percentages denote the share of S1s sending the particular message *conditional on the roll*.

2.5 Results

2.5.1 Behavior of S2

In this subsection, I focus on the 110 subjects in the role of S2. Remember that S2 could have observed a High roll, which I dubbed the Neutral case, or a Low roll. In case of a Low roll, S2 could have observed S1 report dishonestly, *i.e.*, the Dishonest case, or honestly, *i.e.*, the Honest case. Table 2.2 shows the distribution of S2s over the three cases in the Baseline and No-Pro treatment separately. There are 60 S2s in the Baseline treatment, of whom 18 are in the Neutral case, 32 are in the Dishonest case, and 10 are in the Honest case. Similarly, there are 50 S2s in the No-Pro treatment, of whom 18 are in the Neutral case, 23 are in the Dishonest case, and 9 are in the Honest case. Tables 2.3 and 2.4 contain summary statistics across cases and treatments, respectively. For the background characteristics, I include Age, a Male dummy, a European dummy, and a dummy for being in an Economics or Business program. In total, 34% of S2s is male, 55% is European, and 71% studies Economics or Business. Moreover, I control for standardized and self-reported Risk Tolerance, Altruism, and Reciprocity. The latter is a simple average of self-reported measures of positive, negative, and indirect reciprocity. The last column of both tables reports the result of a balance test across cases and treatments, respectively. As can be seen, it seems that the sample is balanced on most observables, with the exception of Risk Tolerance across the three cases.

In the following, I examine the decisions of S2s in the Moonlighting Game by employing non-parametric tests in the same order as the hypotheses formulated above. This means that I start by comparing average trust levels across cases in the Baseline treatment, then move to splitting the sample according to S2's own choice, and finish by comparing the Baseline treatment to the No-Pro treatment. I have chosen to illustrate the results using figures; the corresponding tables can be found in Appendix

Table 2.3: Summary Statistics and Balance of S2s across Cases

	CASES				
	(1) All	(2) Neutral	(3) Dishonest	(4) Honest	(5) χ^2
A. Personal Characteristics					
Age	22.04 (2.70)	22.25 (1.90)	22.04 (3.24)	21.63 (2.27)	1.37 (0.50)
Male	0.34 (0.47)	0.28 (0.45)	0.35 (0.48)	0.42 (0.51)	1.18 (0.55)
European	0.55 (0.50)	0.56 (0.50)	0.56 (0.50)	0.53 (0.51)	0.08 (0.96)
Economics & Business	0.71 (0.46)	0.75 (0.44)	0.71 (0.46)	0.63 (0.50)	0.85 (0.66)
B. Preferences					
Risk Tolerance	-0.03 (1.07)	0.23 (0.98)	-0.25 (1.10)	0.14 (1.04)	5.85* (0.05)
Altruism	0.03 (1.04)	-0.15 (1.22)	0.17 (0.94)	-0.03 (0.96)	1.18 (0.55)
Reciprocity	-0.07 (1.14)	-0.08 (1.14)	-0.09 (1.21)	0.02 (0.97)	0.12 (0.94)
Observations	110	36	55	19	110

Note: Columns (1)-(4) contain (sub)sample averages across S2s, with standard deviations in parentheses. Preferences are standardized across the entire sample (including S1s and R3s) to have mean 0 and standard deviation 1. Column (5) shows the χ^2 -test statistic ($df = 2$) of a chi-squared (for the binary variables) or Kruskal-Wallis (for the continuous variables, adjusted for ties) test for differences across the three cases, with p -values in parentheses. For the exact survey questions, see Appendix 2.A.2.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

2.B.⁷

2.5.1.1 Result 2.1: Trust Is Highest in Neutral, and Similar in Honest and Dishonest

I begin by examining S2's level of trust in the three cases in the Baseline treatment. Figures 2.1a and 2.1b show average levels of trusting and trustworthiness, respectively, in the Neutral (green), Dishonest (yellow), and Honest (blue) cases, while Figures 2.2

⁷In addition, I compared behavior of the subjects in my Moonlighting Game with that of Falk et al. (2008), from which the game's parameters are taken. Taking all subjects in the Neutral case – S1 and S2, Baseline and No-Pro (66 subjects) – I find a median a of 1, to which subjects respond with a median b_1 of 0. Falk et al. find an identical median a to which subjects respond with a median b_1 of 2, instead. Moreover, 20% of subjects send everything to the second-mover (24% in Falk et al.).

Table 2.4: Summary Statistics and Balance of S2s across Treatments

	TREATMENTS			
	(1) All	(2) Base	(3) No-Pro	(4) χ^2
A. Personal Characteristics				
Age	22.04 (2.70)	21.78 (2.55)	22.34 (2.86)	1.04 (0.31)
Male	0.34 (0.47)	0.32 (0.47)	0.36 (0.48)	0.23 (0.63)
European	0.55 (0.50)	0.48 (0.50)	0.64 (0.48)	2.71 (0.10)
Economics & Business	0.71 (0.46)	0.72 (0.45)	0.70 (0.46)	0.04 (0.85)
B. Preferences				
Risk Tolerance	-0.03 (1.07)	0.00 (1.02)	-0.07 (1.13)	0.03 (0.86)
Altruism	0.03 (1.04)	0.12 (1.00)	-0.08 (1.08)	0.76 (0.38)
Reciprocity	-0.07 (1.14)	0.08 (1.13)	-0.24 (1.13)	1.87 (0.17)
Observations	110	60	50	110

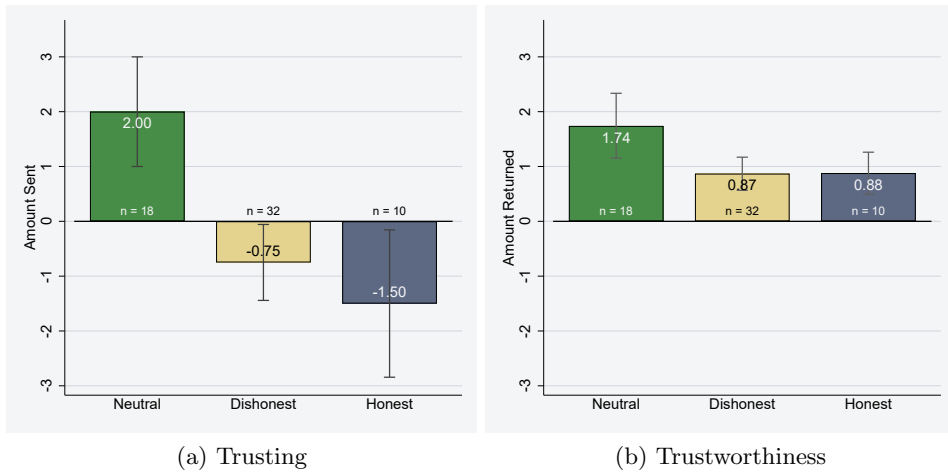
Note: Columns (1)-(3) contain (sub)sample averages across subjects, with standard deviations in parentheses. Preferences are standardized across the entire sample (including S1s and R3s) to have mean 0 and standard deviation 1. Column (4) shows the χ^2 -test statistic ($df = 1$) of a chi-squared (for the binary variables) or Kruskal-Wallis (for the continuous variables, adjusted for ties) test for differences across the two treatments, with p -values in parentheses. For the exact survey questions, see Appendix 2.A.2.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

and 2.3 show their distributions. Trust levels are also summarized in Table 2.8 in Appendix 2.B.

Act of Trusting To begin with the act of trusting, Figure 2.1a shows that S2s in the Neutral case send on average 2.00 ECU as a first-mover (the median is 3.00 ECU and the mode is to send 6.00 ECU). Quite surprisingly, this number is significantly lower in both the Dishonest (at a 5%-level, Mann-Whitney U: $U = 2.151, p = 0.031$) and Honest case (at a 10%-level, MWU: $U = 1.874, p = 0.062$) with an amount sent of -0.75 and -1.50 ECU, respectively (the medians are -1.50 and -3.00, respectively). As can be seen in the distributions across cases in Figure 2.2, a substantial mass is shifted towards higher values of a in the Neutral case as compared to the Honest and

Figure 2.1: Comparison of Trust between Cases in Baseline



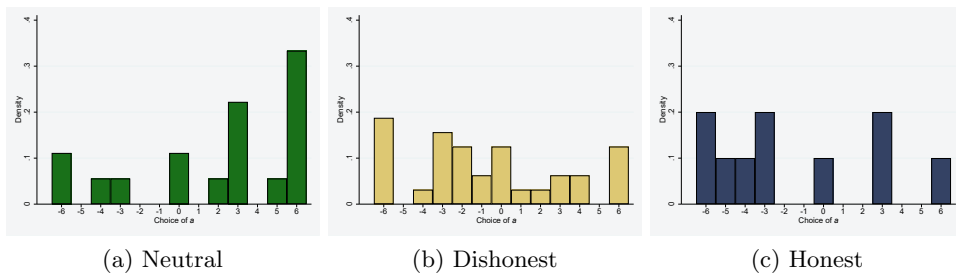
Note: Panel (a) depicts trusting levels in the three cases in Baseline, while Panel (b) contains levels of trustworthiness in Baseline. For Panel (b), trustworthiness levels are averaged over all 13 possible values of trusting. The spikes represent the standard error around the mean.

Dishonest cases. To illustrate, a significantly higher share of S2s transfer a strictly positive amount to the second mover in the Neutral case (12 out of 18), as compared to the Dishonest (10 out of 32, two-sided Fisher exact test $p = 0.036$) and Honest case (3 out of 10, $p = 0.097$). In other words, more people decide to trust S1 in the Neutral case, as compared to both the Honest and Dishonest case. In contrast, the majority of S2s in the Honest (6 out of 10) and Dishonest (18 out of 32) cases display distrust and choose an $a < 0$.⁸ As opposed to the substantial differences with the Neutral case, I observe no differences between the Honest and Dishonest cases (MWU: $U = -0.67, p = 0.51$). Taken together, these results suggest that S2s decide to trust less as a result of the moral dilemma imposed on them by the low die-roll, while the particular behavior of S1 does not differentially affect it. Thus, behaving in a pro-social, yet dishonest, manner seems to neither breed nor harm the extent to which S2 decides to trust S1.

Trustworthiness Subsequently, I examine S2's trustworthiness by studying her choices of b . Remember that subjects were asked to submit a choice for b following

⁸The distributions are significantly different from the Neutral case for the Dishonest case only (Dishonest Kolmogorov-Smirnov: $p = 0.047$, Honest KS: $p = 0.25$). Moreover, I perform bootstrapped t -tests and a Kruskal-Wallis test as an alternative to the Mann-Whitney U test and find similar results.

Figure 2.2: Distributions of Trusting Decisions across Cases in Baseline



Note: The above figures plot case-by-case histograms for trusting decisions in Baseline. Each bar represents the share of S2s in the corresponding case who select that particular a .

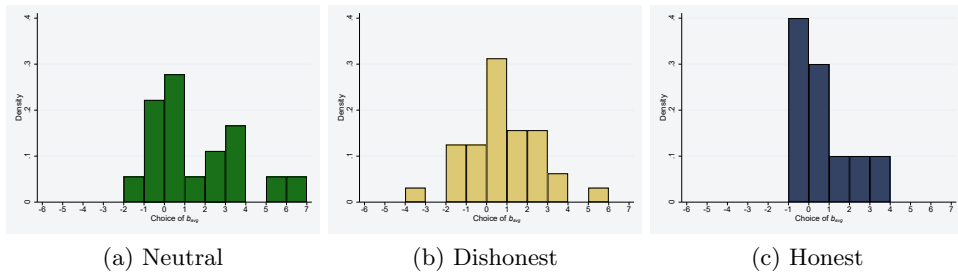
each possible choice of a by the first-mover (strategy method), meaning that b is not affected by the actual choice of a . As one would expect, there exists a positive relationship between S2's choice of b and S1's choice of a in all cases. Figure 2.14 in Appendix 2.B displays this in more detail. Here, for ease of exposition, I focus on the average choice of b over all a , which I dub b_{avg} .

As with the act of trusting, trustworthiness is highest in the Neutral case and there seem to be only small differences between the Dishonest and Honest cases (see Figure 2.1b). S2s in the Neutral case send back 1.74 ECU on average, as compared to 0.87 in the Dishonest and 0.88 in the Honest case. However, the median b_{avg} appears to be similar in the three cases (Neutral: 0.57; Dishonest: 0.62; Honest: 0.50) and pairwise comparisons show that the differences between cases are insignificant (MWU: all $p > 0.38$). The distribution of b_{avg} is displayed in Figure 2.3, where each bar represents the share of S2s who have a b_{avg} in the unit interval on the horizontal axis. In contrast to the act of trusting, the distributions look similar across cases, with most S2s bunching around zero.⁹ In other words, there seem to be no significant differences in trustworthiness across the three cases.

Robustness In Appendix 2.B, I discuss a robustness check in which I restrict the sample based on the apparent level of understanding of the Moonlighting Game. I base this restriction on the consistency of choices in the game and the time that subjects need to answer the trial questions. This exercise shows the difference in the act of trusting between the Neutral case on the one side, and the Dishonest and Honest case on the other, to become more pronounced, while the differences in

⁹Again, I perform a Kruskal-Wallis test comparing the three cases ($\chi^2 = 3.34, p = 0.31$) and Kolmogorov-Smirnov tests comparing distributions (all $p > 0.58$), and find identical results.

Figure 2.3: Distributions of Trustworthiness across Cases in Baseline



Note: The above figures plot case-by-case histograms for trustworthiness in Baseline. Each bar represents the share of S2s in the corresponding case who have an average level of trustworthiness in-between the unit interval on the x -axis. Trustworthiness is truncated at 6 (1 observation).

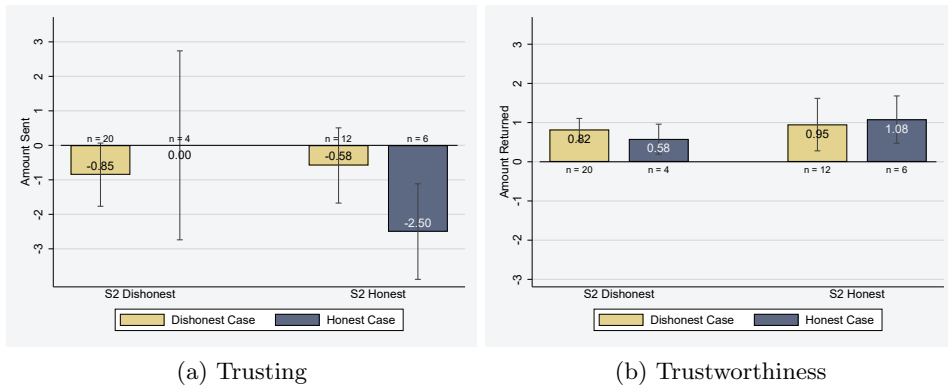
trustworthiness remain insignificant.

2.5.1.2 Result 2.2: Trust Is Independent from S2's Own Choice

Result 2.1 shows no differences in trust between the Honest and Dishonest cases. However, one may naturally expect S2's act of trusting to depend on her own Stage 1 choice. As stated above, Hypothesis 2.2 predicts that S2s who send the same message as S1 display higher levels of trust than those who send a different message. In order to assess this, I split the sample into Honest and Dishonest S2s and see if they respond differently to S1 behavior. This yields Figure 2.4 (see also Table 2.9 in Appendix 2.B). Looking at the act of trusting in Figure 2.4a, both Honest and Dishonest S2s take on average (slightly) more from S1s sending the same message than from S1s sending the opposite message: Dishonest S2s take 0.85 from Dishonest S1s and 0 from Honest ones, while Honest S2s take 2.50 from Honest S1s and 0.58 from Dishonest ones. The differences go against the hypothesized result, although they do not achieve statistical significance (MWU: both $p > 0.22$). This may be caused by the small number of honest subjects in particular, with the smallest subsample (a Dishonest S2 paired with an Honest S1) containing only four observations.

Then, I assess trustworthiness in Figure 2.4b. The average amount sent back is, on average, positive for all subsamples and neither Dishonest nor Honest S2s display significantly higher levels of trustworthiness when matched with S1s who send the same message (MWU: both $p > 0.2$). Taking this together with the results for the act of trusting leads me to conclude that S2s do not respond more favorably to S1s who made the same choice in Stage 1. Hence, I find no evidence in line with Hypothesis 2.2. This also suggests that the informational symmetry between S1 and S2 in Stage

Figure 2.4: Trust Dependent on Own Stage 1 Choice in Baseline



Note: Panel (a) depicts trusting levels in the three cases in Baseline, while Panel (b) contains levels of trustworthiness in Baseline. In both Figures, the sample is split according to S2's own choice in Stage 1. For Panel (b), trustworthiness levels are averaged over all 13 possible values of trusting. The spikes represent the standard error around the mean.

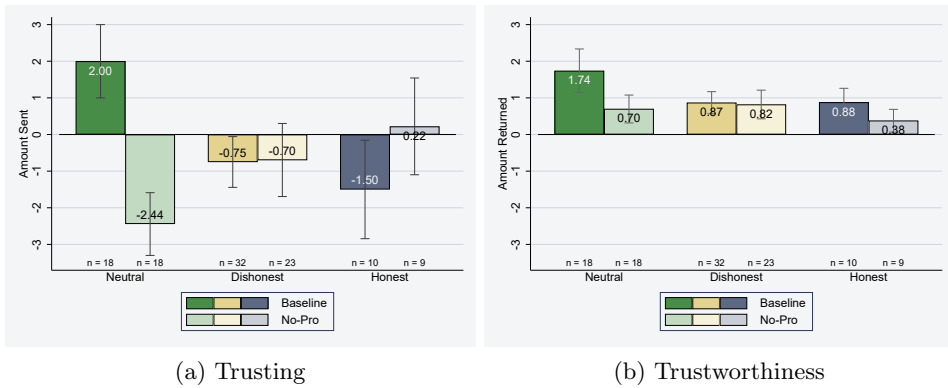
1 about each other's choices in Stage 1, as discussed in Subsection 2.3.4, does not induce any additional reciprocity between them.

2.5.1.3 Result 2.3: Trust Is the Same across Treatments for the Dishonest and Honest Case

My results so far beg the question whether S2s respond at all to Stage 1 behavior or whether the conflicting motives cancel each other out. As explained, the No-Pro treatment serves to disentangle these potential explanations from each other, with Hypothesis 2.3 predicting higher trust in the No-Pro Honest case and lower trust in the No-Pro Dishonest case, as compared to their Baseline counterparts.

Results are depicted in Figure 2.5a for trusting and Figure 2.5b for trustworthiness (see also Table 2.10 in Appendix 2.B). Comparing the different cases across treatments yields no results in line with Hypothesis 2.3. For both the Dishonest and Honest case, differences in trust between treatments are insignificant (MWU: all $p > 0.23$). Moreover, there is no evidence that trust is higher in the No-Pro Honest case as compared to the No-Pro Dishonest case. Thus, it appears S1's choice in Stage 1 does not affect S2's trust towards the former in Stage 2 in the No-Pro treatment. Intriguingly, behavior differs between the Neutral cases in both treatments, with higher levels of trusting recorded in the Baseline treatment: on average, S2s in the No-Pro Neutral case take 2.44 ECU, which is significantly lower than trusting in the

Figure 2.5: Comparison of Trust between Baseline and No-Pro



Note: Panel (a) depicts trusting levels in the three cases, while Panel (b) contains levels of trustworthiness. For Panel (b), trustworthiness levels are averaged over all 13 possible values of trusting. The lighter shaded bars represent the No-Pro treatment. The spikes represent the standard error around the mean.

Baseline Neutral case (MWU: $U = 3.032, p = 0.002$). No such differences are found for trustworthiness. I discuss this unexpected result further in Section 2.6.

In sum, if S2s indeed trade off preferences for honesty and higher payoffs when evaluating S1's decision, I should have seen lower levels of trust in the Dishonest case when I remove the pro-social component in the No-Pro treatment. However, my results fail to find support for this conjecture. This suggests that S2s are insensitive to the pro-social component that is present in Baseline and absent in No-Pro. This is also exemplified by the fact that the share of High reports is similar in both treatments (see Table 2.2). Hence, the pro-social component does not seem to be an important motive for the lying decision in the first place, and therefore it is no surprise that removing it does not affect behavior.

2.5.1.4 Parametric Regressions

In addition to the non-parametric tests above, I also assess trust levels using parametric regression models. Doing so allows me to control for individual characteristics of the subjects. In particular, I control for the variables displayed in Tables 2.3 and 2.4: Age, Male, European, Economics & Business, Risk Tolerance, Altruism, and Reciprocity. Moreover, for trustworthiness only, I can assess how it depends on the act of trusting a . Table 2.5 presents the results of this exercise. In all regressions, the Neutral case in the Baseline treatment forms the reference category. Since one sub-

ject left the lab without completing the survey, I drop this subject from the analysis and I am left with 109 subjects in the role of S2.

Columns (1) and (2) focus on the the act of trusting and employ ordinary least-squares models with robust standard errors. Column (1) simply compares the different cases and reproduces the main results from above. That is, (i) trusting is significantly lower (at a 10%-level) in the Dishonest and Honest case in the Baseline treatment only, (ii) subjects in the No-Pro Neutral case exhibit significantly lower levels (at a 1%-level) of trusting than subjects in the Baseline Neutral case, and (iii), within treatments, there are no differences between the Dishonest and Honest cases. In Column (2), I split the S2s facing a low roll in those who send the same message as S1 and those who send a different message. For the Baseline treatment, coefficient estimates are significantly negative (at a 10%-level) for both and nearly identical to each other. This implies that a low roll in itself generates a drop in trusting, while it does not matter whether S2 sends the same message as S1 or a different one. This latter observation also applies to the No-Pro treatment.

Columns (3) and (4) present a random-effects model with trustworthiness b as the dependent variable. Since I employ the strategy method, each of the 109 subjects in the role of S2 makes a choice for each of the 13 potential values of a . Thus, I have a balanced panel of 1417 observations over 109 clusters. I observe a consistently positive and significant effect of the first-mover's choice of a , which indicates that second-movers do reciprocate trusting decisions. The low magnitude (around 0.36 ECU returned for each ECU sent) also implies, however, that trusting does not pay off on average. Moreover, the insignificant interaction terms $a \times \textit{Dishonest Case}$ and $a \times \textit{Honest Case}$ indicate that the responsiveness of trustworthiness to different levels of trusting is the same across cases. Column (3) shows that trustworthiness is significantly lower (at a 10%-level) in the Honest case, only. However, the difference in magnitude *and* significance between the coefficients on the Dishonest and Honest cases is small. Hence, I cannot reject the null that trustworthiness is affected in the same way in the two cases. In Column (4), I again split the sample in S2s who send the same message and S2s who send a different one. Even though the coefficient is significantly negative for Different Message, only, the difference between Same Message and Different Message is small and insignificant, which again indicates that whether or not S2 sends the same message as S1 does not affect trustworthiness. In the No-Pro treatment, the difference between pairs who send the same message and pairs who send a different message appears to be larger, but again, this difference is insignificant.

Table 2.5: Linear Regressions on Trust

	TRUST (OLS)		TRUSTWORTH. (RE)	
	(1)	(2)	(3)	(4)
a			0.362*** (0.090)	0.362*** (0.090)
$a \times$ Dishonest Case			-0.011 (0.112)	
$a \times$ Honest Case			-0.001 (0.148)	
$a \times$ Different Message				0.043 (0.134)
$a \times$ Same Message				-0.045 (0.114)
Dishonest Case	-2.349* (1.286)		-0.959 (0.615)	
Honest Case	-3.151* (1.750)		-1.211* (0.697)	
No-Pro	-4.195*** (1.414)	-4.200*** (1.409)	-0.803 (0.637)	-0.787 (0.635)
No-Pro \times Dishonest Case	2.379* (1.396)		-0.349 (0.525)	
No-Pro \times Honest Case	2.187 (1.444)		-0.479 (0.482)	
Same Message		-2.579* (1.371)		-0.917 (0.600)
Different Message		-2.508* (1.408)		-1.209* (0.704)
No-Pro \times Same Message		4.443** (1.997)		-0.008 (0.697)
No-Pro \times Different Message		5.337*** (1.920)		1.361 (0.875)
Constant	1.032 (4.330)	1.183 (4.357)	1.301 (1.598)	1.366 (1.542)
Controls	✓	✓	✓	✓
Observations	109	109	1417	1417
Clusters			109	109
ρ			0.30	0.29
χ^2			66.22	65.88
F	2.72	2.71		
df	96, 12	96, 12	15	15
R^2	0.21	0.21	0.22	0.23

Note: Columns (1) and (2) contain the results of an OLS model with a as the dependent variable and robust standard errors. Columns (3) and (4) contain a random-effects panel estimation using b as the dependent variable, treating each choice as an observation and controlling for correlation within the same cluster (*i.e.*, subject) by clustering standard errors. In all columns, the Neutral case (in the Baseline treatment) is the reference category. Controls include all variables from Tables 2.3 and 2.4, but are omitted from the table. One subject was dropped from the analysis for not having completed the post-experimental survey. The parameter ρ denotes the share of the variance due to the subject-specific error term u_i .

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

2.5.2 Believed and Actual Behavior of S1

Let me finalize the Results section by discussing the behavior of S1 in the Baseline treatment of the Moonlighting Game, and S2's beliefs about this.

Behavior of S1 To begin with, Figures 2.6a and 2.6b show that I am unable to replicate Result 2.1 for S1, even though the ordering across cases is the same. In particular, I find a negative average amount sent a in all three of the cases and no significant differences across cases for the act of trusting or trustworthiness. This shows that S1's choice in Stage 1 is not predictive of the amount of trust that S1 exhibits in the Moonlighting Game. Moreover, it shows that obtaining a high roll only affects S2's act of trusting in a positive way. I refer back to this finding in the Discussion section when discussing potential explanations.

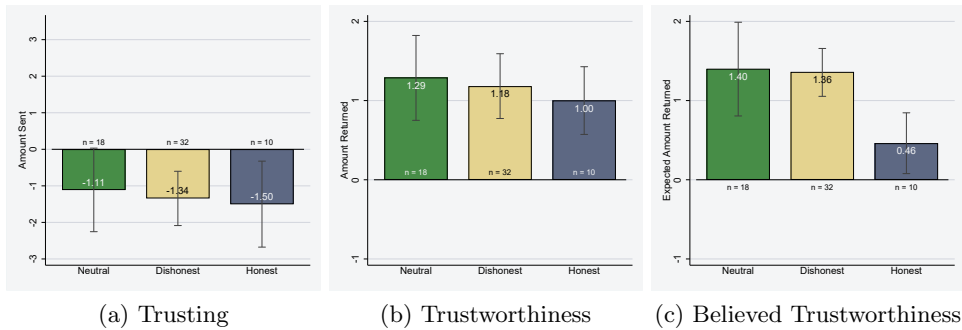
Beliefs of S2 In order to examine the mechanism behind the result for a higher amount sent in the Neutral case, I examine S2's beliefs about S1's trustworthiness. If an S2 in the Neutral case somehow expects her S1 to return more, this may be a reason for her to choose higher a . In Figure 2.6c (see also Table 2.11 in Appendix 2.B), average beliefs about S1's choice of b are split into the three familiar cases Neutral, Dishonest and Honest. I observe a similar pattern as shown in Figure 2.1a for the act of trusting, with the expected back-transfer highest in the Neutral case (1.40 ECU), followed by the Dishonest (1.36 ECU) and Honest case (0.46 ECU). However, these differences appear to be too small in magnitude compared to the differences in trust in order to explain this result. Indeed, all pairwise differences are insignificant (MWU: all $p > 0.25$). It thus seems that beliefs cannot explain the observed differences in the act of trusting.

As before, I also tested for differences with the No-Pro treatment. Since behaving dishonestly in the No-Pro treatment entails no pro-social component, it can be seen by S2 as worse of a behavior and more predictive of untrustworthy behavior. I do not find such differences, and therefore these results are not reported (available upon request).

2.6 Concluding Discussion

In this section, I provide an extensive discussion of my results. Summarizing the results from Section 2.5, I find that (1) S2s do not differentiate between Honest or Dishonest S1s in terms of trust, but does exhibit decreased levels of trust in both

Figure 2.6: Actual and Believed Trust of S1 in Baseline



Note: Panel (a) depicts trusting levels of S1 in the three cases in Baseline, Panel (b) contains levels of trustworthiness of S1 in Baseline, and Panel (c) contains S2's beliefs about S1's trustworthiness in Baseline. For Panel (b) and (c), (believed) trustworthiness levels are averaged over all 13 possible values of trusting. The spikes represent the standard error around the mean.

cases as compared to a Neutral case; (2) the congruence of both Senders' messages does not matter for the trust exhibited in the Moonlighting Game; (3) removing the pro-social component does not yield more trust in the Honest case and does not yield less trust in the Dishonest case.

My results are at odds with some earlier findings and adds new insights to the existing literature. I show that the very instance of being faced with a moral dilemma erodes trust. This has not been found by Levine and Schweitzer (2015), for instance, who showed that dishonesty, rather than honesty, breeds trust. In addition, they did not include the possibility for a Neutral case where cheating opportunities are absent. My results are not in line either with the notion that benevolent cheaters are perceived as more moral than selfish truth-tellers, as reported by Levine and Schweitzer (2014), although it should be noted that truth-tellers in my experiment are not "selfish" as they do not gain materially from behaving honestly. In turn, although Result 2.2 would be hard to justify in, for example, an Ellingsen and Johannesson (2008) framework, it is in line with the paper by Gross et al. (2018) on ethical free-riding. As in their study, I show that subjects may have a preference to behave honestly, while they have no problem profiting from a dishonest fellow subject.

Result 2.1 may seem puzzling at first sight and merits further discussion. I am able to discard explanations related to beliefs, as I find no significant differences in believed trustworthiness across cases. Instead, the results seem to be driven by differences in preferences originating from what happened in Stage 1. As the most plausible explanation, I posit that the High roll in itself induces higher levels of trust

among S2s in the Baseline treatment. I identify two competing mechanisms for this pattern. First, the most straightforward explanation posits that trust is generated if and only if an S1 increases S2's payoffs *and* does not need to lie. Whenever one of these conditions is not fulfilled, trust is not increased. In a sense, either message following a Low roll provides S2 with a reason not to increase her trust: a Dishonest S1 misreported, while an Honest S1 failed to increase S2's payoff. In relation to the theoretical framework discussed in Subsection 2.3.4, this suggests that preferences for payoffs and honesty do not compensate each other, but actually complement and reinforce each other. This would also explain why I fail to observe the same pattern in the No-Pro treatment (Result 2.3). After all, since S1's choice does not affect S2's payoffs, S2 never sees a reason to become more trusting of S1 and trust in the No-Pro Neutral case is similar to the No-Pro Dishonest and No-Pro Honest cases. What is more, S2 may even punish S1 in the No-Pro Neutral case, because sending the High message yields an inefficient outcome in terms of total payoffs.

Alternatively, S2 may incur a psychological cost from the moral dilemma imposed on her by the Low roll, which brought her in such a negative psychological state, that she feels less cooperative and act more selfishly as a result. This relates to the literature studying the effect of mood on behavior, assuming that a high roll induces a good mood and a low roll induces a bad mood. This literature documents both positive and negative effects of good mood (Capra, 2004; Dunn and Schweitzer, 2005; Kirchsteiger et al., 2006; Proto et al., 2019). My results resonate the findings of Capra (2004) and Dunn and Schweitzer (2005) who find a good mood to induce higher levels of trust, and those of Kirchsteiger et al. (2006) who find that a good mood increases trustworthiness *levels* overall.¹⁰ On the other hand, I do not find a similar pattern of higher trust in the Neutral case for S1s, which one would expect as the roll could be assumed to affect both Senders' mood in the same way (see Subsection 2.5.2). Future efforts could be invested in discriminating between the two mechanisms described above, for example by conducting an experiment in which S2 would be paired with a *different* S1 for the Moonlighting Game, about whom she receives no Stage 1 information. If the roll is sufficient to induce the negative state of mind, we should observe the same pattern of choices in this alternative design. In addition, one could actively measure (self-reported) mood of Senders in the experiment.

Admittedly, my experimental design contains a few limitations and design choices that may have affected (the validity of) my results. First and foremost, while running

¹⁰Kirchsteiger et al. (2006) also find that a bad mood increases the *responsiveness* of trustworthiness to changes in a . I tested this claim by regressing b on a for each S2 and comparing slope coefficients between rolls using Mann-Whitney U tests. In fact, I find the coefficient to be insignificantly higher following a high roll. See Table 2.12 in Appendix 2.B.

the experiment, it became apparent that a non-negligible share of subjects deem the Moonlighting Game complicated. The game's structure in combination with the use of the strategy method requires subjects to think carefully about all contingencies that could occur. This causes less than half of the subjects to submit choice of b that are weakly monotonically increasing in a . Admittedly, my experiment could benefit from a simpler design with fewer possible actions for each role. Relatedly, the use of the strategy method in the Moonlighting Game may have attenuated the intensity of the response b to different levels of a . Emotions like anger and frustration, which may intensify feelings of reciprocity, arguably are much less present in the strategy method as compared to the direct response method (Aina et al., 2020, show this for the ultimatum game). After all, the strategy method requires S2 to imagine, rather than actually experience, how she would feel about each potential a . As a result, one could expect differences in trustworthiness to be larger across levels of trusting when using the direct response method. Even though this consideration applies equally to all cases (the response to the cases is already elicited in a "direct response" manner), the direct response method could then result in more pronounced differences in trustworthiness across cases if the reciprocal response towards Stage 1 behavior and the reciprocal response towards the first-mover's choice of a tend to reinforce each other. At the same time, the direct response method would pose a challenge as to how to cleanly compare trustworthiness levels across cases with potentially different levels of trusting.

Second, there are generally few Honest Senders to study, which makes the examination of this type of Senders tricky. This mainly hampers a reliable statistical analysis when the sample is divided according to S2's own, non-materialized choice and different results might have been found regarding Hypothesis 2.2 with a sufficiently large sample size. In order to achieve a more equal distribution of Honest and Dishonest Senders, one could make the difference in payoffs between the High and Low messages smaller, so as to discourage sending the High message when the roll was actually Low. At the same time, this may affect how dishonesty is perceived by S2.

Third, the different treatments and cases may not be perfectly comparable due to different payoff allocations resulting from them. For example, the Honest case features a lower payoff by construction. The welfare difference (6 ECU) is not negligible, as it equals half of the endowment in the Moonlighting Game. At the same time, this 'loss' of income forms the feature that is supposed to negatively affect S2's behavior. It is thus crucial and inherent to the design. One could potentially solve this issue by endowing Senders in the Honest case with 18 rather than 12 ECU in the

Moonlighting Game. Similarly, the only concrete difference between the Baseline and No-Pro treatments is the payoff to S2 associated with the Low message. However, this modification changes the reference point for each of the cases and the motives for (mis)reporting: the Low message in the No-Pro treatment yields an efficient outcome in terms of total surplus, but it also yields a lower payoff to S1 as compared to S2 and R3. This may have contributed to the unexpected results found in relation to the No-Pro treatment (Result 2.3). Future efforts could be invested in designing an alternative treatment where the pro-social component is removed without changing the reference point (as much). For example, S1 and S2 could engage in a Stage 1 identical to the Baseline treatment, after which S2 is paired to a different S1 for Stage 2 while learning this new S1's Stage 1 choice.

Fourth, sessions contained two Receivers who were matched to multiple pairs, with one of these pairs to be randomly selected to determine the Receiver's payoff. Since Senders are aware of this structure, it may dilute the motive to behave honestly. After all, the consequences to the Receiver are uncertain and only small in expected value. This consideration is amplified by the fact that the Receiver does not need to do anything and does not know how payoffs are related to the Senders' messages. Arguably, being dishonest is considered worse in a situation where each Receiver is matched to only one Sender pair. This could then lead to lower trust levels in the Dishonest case.

Fifth, both Senders make a choice in Stage 1 in order to be able to relate S2's own choice to her behavior in the Moonlighting Game. In addition, S1 is informed of this non-materialized choice. While the former design choice is necessary to test Hypothesis 2.2, the latter feature could have been omitted to avoid any concerns that this informational symmetry reinforces feelings of reciprocity, especially between Senders who both choose to be Dishonest as they realize that they both had a desire to increase each other's payoff. Even though it does not seem that this mechanism is at work, as exemplified by the absence of higher trust between Senders who are both Dishonest in Stage 1 as compared to Senders who send different messages, it could potentially have clouded our results in relation to Hypothesis 2.1, too, and a setting in which S1 is not informed of S2's choice would have made for a cleaner design.

In sum, the current study provides evidence that the mere presence of an adverse outcome suffices to erode trust. My results may have important implications for organizations in which individuals work in teams. It suggests that the presence of moral dilemmas may induce negative spillovers to the internal work environment. As a result, efforts should better be expended to preventing teams from being faced with

a moral dilemma in the first place, rather than instructing them how to behave when faced with one. Transparency and appropriate design of incentive schemes, such as rewards that do not depend on performance dimensions that cannot be monitored easily, can help in limiting the opportunities for employees to gain from dishonest acts and could, in addition to their obvious advantages, limit the extent to which the quality of the work environment is eroded. The potential importance of adverse events and moral dilemmas in affecting trust may also be relevant in light of the current CoViD-19 pandemic. When employees are repeatedly being faced with trade-offs between obeying by government regulations and doing what is best for their organization, and this trade-off *per se* indeed decreases trust, this decline in trust may constitute an overlooked indirect consequence of this health crisis.

Appendices

2.A Experimental Material

2.A.1 Instructions

You are participating in an economic decision-making experiment. Your final payoffs will depend on the choices you make, as well as on those of others. Therefore, it is important that you read these instructions carefully. Please be reminded that any form of communication with other participants is prohibited and will lead you to being removed from the experiment. If you have a question, raise your hand and the experimenter will come to you to answer your question in private. *All decisions you make within the walls of the experimental lab will be anonymized to both the experimenter and the other participants.*

Throughout the experiment, your earnings will be denoted in experimental currency units (ECU). At the end of the experiment, you will be paid according to the amount of ECU that you collected. *The conversion rate is 1 Euro per 3 ECU.* So, for example, if you end up with 18 ECU, you will be paid out 6 Euro. On top of your earnings, *you will receive a show-up fee of 3 Euro.* You will be paid out in cash directly after the experiment. [The cash will be put in envelopes. Your envelope will be indicated with the same number as your PC. After checking (privately) whether the envelope contains the correct amount, you sign a proof of payment.]

The experiment will consist of two Parts and a survey. *You will be paid your total earnings over Parts 1 and 2.* You will receive instructions for Part 2 once you have completed Part 1.

INSTRUCTIONS PART 1

In Part 1, you will play one of two roles, Role A (referred to as Player A) or Role C (Player C). As a Player A, you will be matched to one participant with the same Role to form a pair. We refer to this matched player as *your Partner*. The two Players A will be shown a video of *the same die-roll* and have to report the outcome of the roll to Player C, *who has not observed the die-roll and will never know its outcome.*

The die-roll can have two outcomes, which are explained to Players A in more detail

later on, and thus two different messages of the form "The die-roll was X". Both Players A independently choose which Message to send. Once both Players A have done so, one of the two Messages is randomly picked and actually sent to Player C. Hence, per pair of Players A, only one Message is sent to Player C and ***only this message determines payoffs*** to both Players A. ***It is important to stress that the Message, instead of the actual roll, determines the payoffs in Part 1 to all players.*** Moreover, the payoffs corresponding to the different Messages will be shown on the screen only to Players A.

The session is structured such that there are exactly two players with Role C, while the rest of the participants are divided into pairs of Players A. A Player C is passive in Part 1 and is matched to half of the pairs, from each of whom they receive one Message. Hence, Players A send a Message to ***only one of the two Players C.*** Of these Messages, one is selected independently for each Player C to determine their payoffs for this part. Importantly, even if a Message that is sent to Player C is not picked to determine his payoffs, ***it still determines payoffs to Players A.***

YOU WILL NOW HAVE THE OPPORTUNITY TO ASK QUESTIONS. PLEASE RAISE YOUR HAND AND THE EXPERIMENTER WILL COME TO YOU TO ANSWER YOUR QUESTION IN PRIVATE.

[Subjects received instructions for Part 2 on a separate sheet and only after the completion of Part 1]

INSTRUCTIONS PART 2

In Part 2, *Players A are matched with the same partner as in Part 1*. The two Players C will be matched to each other. You play the following game:

Both Players are given an account containing *12 ECU* and you will be randomly assigned the roles of Mover 1 and Mover 2.

First, Mover 1 chooses an action that affects both accounts. That is, he can either choose an action that decreases his own account and increases Mover 2's account by the tripled amount, or an action that increases his own account and decreases Mover 2's account by the same number. The actions that Mover 1 can choose include all integer numbers between -6 and 6, *where negative numbers decrease Mover 2's account to the benefit of Mover 1's account and positive numbers increase it at the expense of Mover 1's account*. Hence, Mover 1 has to choose one of the following options:

Table 2.6: Mover 1's Actions and Consequences for Both Accounts

(1) Action Mover 1	CHANGE IN ACCOUNT		NEW TOTALS	
	(2) Mover 1	(3) Mover 2	(4) Mover 1	(5) Mover 2
-6	+6	-6	18	6
-5	+5	-5	17	7
-4	+4	-4	16	8
-3	+3	-3	15	9
-2	+2	-2	14	10
-1	+1	-1	13	11
0	0	0	12	12
1	-1	+3	11	15
2	-2	+6	10	18
3	-3	+9	9	21
4	-4	+12	8	24
5	-5	+15	7	27
6	-6	+18	6	30

Note that Columns (2) and (3) contain *changes in Mover 1's and Mover 2's account*, while Columns (4) and (5) contain *their new account totals*.

Subsequently, Mover 2 has to decide on an action that affects both accounts. That is, Mover 2 can decrease her new account total by up to 18 points and increase Mover 1's account by the same amount. Alternatively, Mover 2 can decrease her account by up to 6 points and decrease Mover 1's account by the tripled amount.

So, decreasing Mover 1's account is also costly to Mover 2. *Note that neither of the accounts can become negative, and possible actions are adjusted accordingly.* Summarizing, Mover 2 has to choose one of the following options, provided that both accounts remain non-negative:

Table 2.7: Mover 2's Actions and Consequences for Both Accounts

(1) Action Mover 1	CHANGE IN ACCOUNT	
	(2) Mover 1	(3) Mover 2
-6	-6	-18
-5	-5	-15
-4	-4	-12
-3	-3	-9
-2	-2	-6
-1	-1	-3
0	0	0
1	-1	+1
2	-2	+2
3	-3	+3
4	-4	+4
5	-5	+5
6	-6	+6
7	-7	+7
8	-8	+8
9	-9	+9
10	-10	+10
11	-11	+11
12	-12	+12
13	-13	+13
14	-14	+14
15	-15	+15
16	-16	+16
17	-17	+17
18	-18	+18

Note that Columns (2) and (3) contain *changes in accounts of Mover 2 and Mover 1*, respectively. New account totals for both players also depend on what Mover 1 did.

Importantly, you are asked to make decisions for both roles, before you know what role you will actually play. Hence, you make one decision as Mover 1, and thirteen decisions (one following each possible Mover 1 decision) as Mover 2. After that, it will be randomly determined which role you will play. So, *all decisions* you make may have monetary consequences for you and your partner. On the screen, you will receive more instructions about how to report your decisions.

Your payoffs for Part 2 will be determined by the amount in your account after the end of the game. On your screen, Tables 1 and 2 will be depicted again. Before you proceed to the game, you will be asked some trial questions to test your understanding. You need to answer these correctly before being able to play the game.

After the end of Part 2, you will be asked to fill in a short survey. After the survey, you will be paid. Please remain seated once you have finished.

2.A.2 Survey Questions

After Stage 2, subjects are asked to fill in a short survey, containing the following questions:

- What is your age?
- What is your major/study program?
- Are you in your Master or Bachelor?
- For how many years have you been doing your current program (including this academic year)?
- What is your gender?
- What is your nationality?
- (Hypothetical Dictator Game) Suppose that you are matched with a random other participant in this experiment. You (singular, not plural) are given 10 Euro and have to decide how much to keep for yourself and how much to give to the other person. Any proposal that you make is automatically implemented, without the other participant having the opportunity to reject or accept the proposal. How many person would you allocate to **the other person**?
- (Hypothetical Multiple Price List) 10 Euro for sure or (x chance of 20, $1 - x$ chance of 0) with $x \in \{1, 0.95, 0.90, \dots, 0.10, 0.05, 0\}$
- (Self-reported risk tolerance on a scale from 1 to 10) How willing are you to take risks?
- (Self-reported indirect reciprocity on a scale from 1 to 10) How willing are you to punish someone who treats others unfairly, even if there may be costs for you?
- (Self-reported altruism on a scale from 1 to 10) How willing are you to give to good causes without expecting anything in return?
- (Self-reported positive reciprocity on a scale from 1 to 10) When someone does me a favor I am willing to return it.
- (Self-reported negative reciprocity on a scale from 1 to 10) If I am treated unfairly, I will take revenge at the first occasion, even if there is a cost to do so.

- (Self-reported trust on a scale from 1 to 10) I assume that people only have the best intentions.

All questions regarding self-reported preferences are based on the questionnaire used for the Global Preferences Survey (Falk et al., 2018).

2.A.3 Selection of z-Tree Screens

Figure 2.7: Die Roll (Discolored)

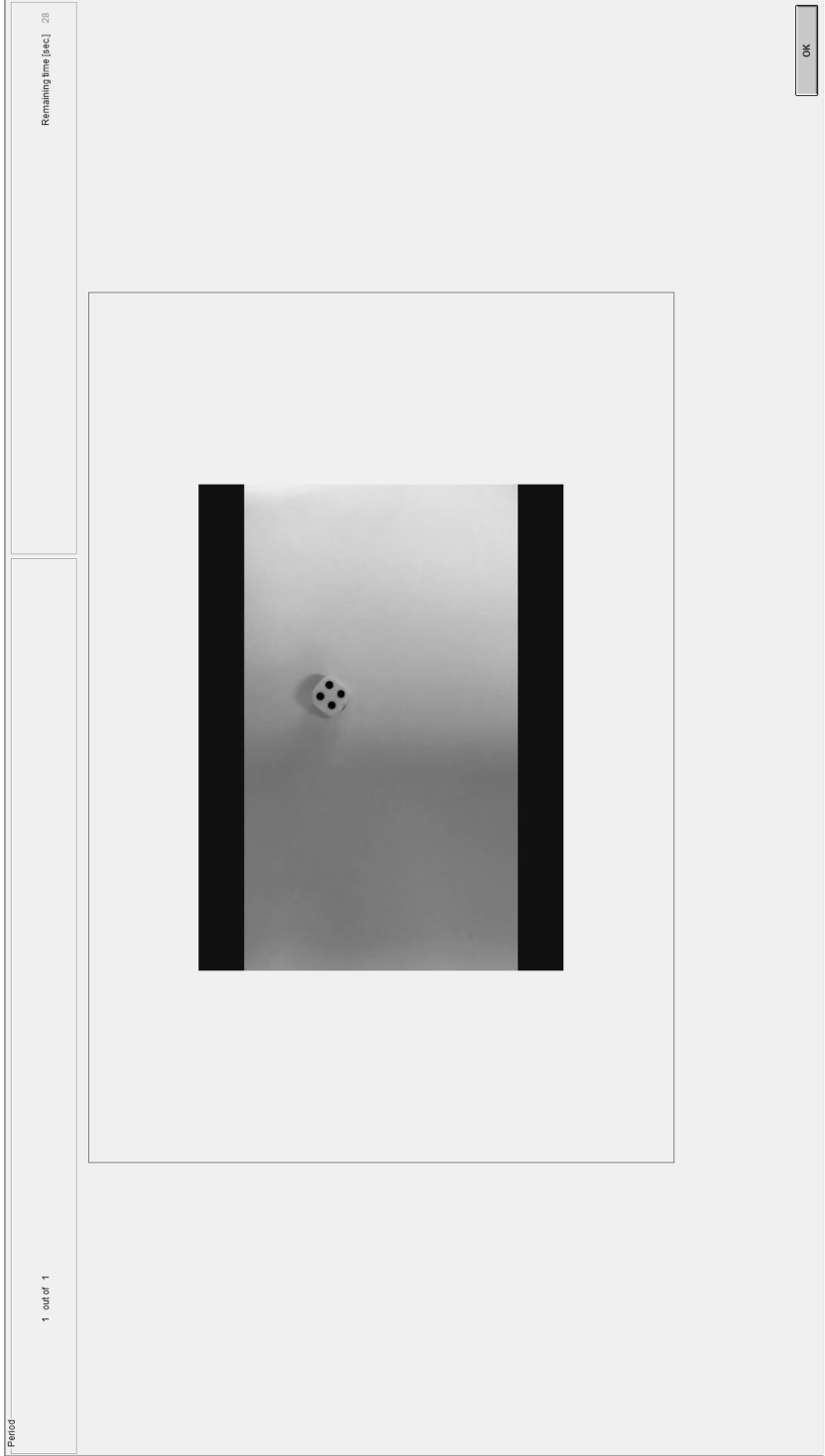


Figure 2.8: Die Report (Discolored)

Period
1 out of 1
Remaining time (sec) 0

PLEASE REPLY A DECISION

Please select the message that you want to send to Player C:

The die-roll was Low
 The die-roll was High

SEND

Below, we depict the payoff matrix associated with this task.

Each cell contains the payoffs in ECUs for all three players for each combination of Die-Roll and Message. Every cell contains the payoff for you (Player A), the payoff for your Partner (Player B), and the third that of Player C. Note that the payoff of the three players in your group **depends solely on your message sent** (conditional on it being chosen to be sent).

Hence, sending the message High will give you and your Partner both a payoff of 12, while giving Player C a payoff of 3. Alternatively, sending the message Low will give you and your Partner, while it yields Player C 15.

Study the payoff matrix carefully before moving to the video.

	Message Low (1, 2, 3 or 4)	Message High (5 or 6)
Roll Low (1, 2, 3 or 4)	6,6;15	12,12;3
Roll High (5 or 6)	6,6;15	12,12;3

Note: Player C is not aware of this payoff matrix.

Figure 2.9: Summary Senders (Discolored)

Period

1 out of 1

Remaining time [Sec.] 41

YOUR PARTNER'S MESSAGE has been sent to Player C

The die-roll was: High

Your Partner's message: The die-roll was Low
 Your message: The die-roll was Low
 Your payoff: 6.0

Total payoff for Part 1 (ECU): 6.0

We will now proceed to Part 2 of the experiment. For this second part, **you will again be matched with your Partner.**

TO PART 2

	Message Low (1, 2, 3 or 4)	Message High (5 or 6)
Roll Low (1, 2, 3 or 4)	6,6;15	12,12;3
Roll High (5 or 6)	6,6;15	12,12;3

Figure 2.10: Summary Receivers (Discolored)

Period
1 out of 1
Remaining time [Sec.] 52

You have received a Message from **Groups 2 through 5**

Group no.	Message
2	The die-roll was High
3	The die-roll was High
4	The die-roll was Low
5	The die-roll was High

You have been matched to Group: 2

Message sent: The die-roll was High

Payoff: 3.0

Total payoff for Part 1 (ECU): 3.0

We will now proceed to Part 2 of the experiment. For this second part, **you will be matched with the other player in Role C.**

TO PART 2

Figure 2.11: Choosing a (Discolored)

Period

1 out of 1

Remaining time (sec): 114

PART 2

You are now asked to make your decisions for this game. **Importantly, you are asked to make decisions for both roles, before you know what role you will actually play.** Hence, you make **one decision as Mover 1** and **thirteen decisions (one following each possible Mover 1 decision) as Mover 2.** After having made your decisions, it will be randomly determined which role you will play.

DECISION AS MOVER 1

Below, you are first asked to make a decision as **Mover 1**. You make a decision by **adjusting (click and drag) the slider in the first column to your desired action.** Your choice of action is depicted in the second column, next to the slider. The new account totals for you and Mover 2 are shown in the third and fourth column, respectively, and are updated as you change your action.

Remember that if you are chosen to be Mover 1, your choice will be implemented and will have payoff consequences for you and your partner, also depending on the actions chosen by Mover 2. Press the SUBMIT-button when you have decided on your action.

What would you choose if you were to be Mover 1?

	YOUR ACTION	YOUR ACCOUNT	MOVER 2'S ACCOUNT
CHOOSE YOUR ACTION (MOVER 1) -6 6	0	12	12

SUBMIT

Figure 2.12: Choosing *b* Using Strategy Method (Discolored)

Period

1 out of 1

Remaining time (sec): 2:38

DECISIONS MOVER 2

Now, you are asked to make decisions as Mover 2. That is, we ask you which action you would pick as Mover 2 following each possible choice of Mover 1. Hence, below you are asked to make thirteen decisions. In the grid below, the first column describes the action taken by Mover 1. The second column then shows your new account total following Mover 1's action. In the third column, you are asked to choose your action by adjusting the sliders to your preferred action. The final three columns contain your chosen action, Mover 1's final payoffs following your chosen action and your final payoff, respectively. Again, these numbers are updated when you change your action. Note that the set of actions that you can choose differs between the rows, since neither of the two accounts can become negative. All defaults are set to zero.

Remember that when you are chosen to be Mover 2, your choices will be implemented and will have payoff consequences for both you and your partner, also depending on what your partner did as Mover 1. Press the SUBMIT-button when you have made your decisions.

What would you choose in each of the following cases if you were to play the role of Mover 2?

MOVER 1'S ACTION	YOUR NEW ACCOUNT TOTAL	CHOOSE YOUR ACTION (MOVER 2)	YOUR ACTION	MOVER 1'S PAYOFF	YOUR PAYOFF
Mover 1 chooses -6	12 - 6 = 6	-6 / 6	0	18	6
Mover 1 chooses -5	12 - 5 = 7	-5 / 7	0	17	7
Mover 1 chooses -4	12 - 4 = 8	-5 / 8	0	16	8
Mover 1 chooses -3	12 - 3 = 9	-5 / 9	0	15	9
Mover 1 chooses -2	12 - 2 = 10	-4 / 10	0	14	10
Mover 1 chooses -1	12 - 1 = 11	-4 / 11	0	13	11
Mover 1 chooses 0	12	-4 / 12	0	12	12
Mover 1 chooses 1	12 + 3*1 = 15	-3 / 15	0	11	15
Mover 1 chooses 2	12 + 3*2 = 18	-3 / 18	0	10	18
Mover 1 chooses 3	12 + 3*3 = 21	-3 / 18	0	9	21
Mover 1 chooses 4	12 + 3*4 = 24	-2 / 18	0	8	24
Mover 1 chooses 5	12 + 3*5 = 27	-2 / 18	0	7	27
Mover 1 chooses 6	12 + 3*6 = 30	-2 / 18	0	6	30

SUBMIT

2.B Supplementary Tables & Figures

2.B.1 Tables

Table 2.8: Trust of S2s in Baseline

	CASE			MANN-WHITNEY U Tests		
	(1) Neutral	(2) Dishonest	(3) Honest	(4) N vs. D	(5) N vs. H	(6) D vs. H
a	2.00 (4.24)	-0.75 (3.92)	-1.50 (4.25)	2.151** (0.031)	1.874* (0.062)	0.670 (0.514)
b_{avg}	1.74 (2.51)	0.87 (1.71)	0.88 (1.21)	0.851 (0.402)	0.894 (0.384)	0.089 (0.936)
Obs.	18	32	10	50	28	42

Note: Columns (1) to (3) contain sample averages, with standard deviations in parentheses. Columns (4) to (7) contain Mann-Whitney U test-statistics, with p -values in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.9: Trust of S2s in Baseline When Conditioning on Own Behavior

	S2 DISHONEST			S2 HONEST		
	(1) S1 D	(2) S1 H	(3) MWU	(4) S1 D	(5) S1 H	(6) MWU
a	-0.85 (4.09)	0.00 (5.48)	0.274 (0.807)	-0.58 (3.78)	-2.50 (3.39)	1.236 (0.230)
b_{avg}	0.82 (1.28)	0.58 (0.77)	0.467 (0.663)	0.95 (2.32)	1.08 (1.48)	0.375 (0.750)
Obs.	20	4	24	12	6	18

Note: Columns (1)-(2) and (4)-(5) contain sample averages, with standard deviations in parentheses. Columns (3) and (6) contain Mann-Whitney U test-statistics, with p -values in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.10: Comparison of S2 Trust between Baseline and No-Pro

	NEUTRAL			DISHONEST			HONEST			(10) D vs. H
	(1) Base	(2) NP	(3) MWU	(4) Base	(5) NP	(6) MWU	(7) Base	(8) NP	(9) MWU	
a	2.00 (4.24)	-2.44 (3.63)	3.032*** (0.002)	-0.75 (3.92)	-0.70 (4.78)	0.017 (0.989)	-1.50 (4.25)	0.22 (3.96)	0.784 (0.453)	0.430 (0.682)
b_{avg}	1.74 (2.51)	0.70 (1.61)	0.944 (0.354)	0.87 (1.71)	0.82 (1.88)	0.883 (0.383)	0.88 (1.21)	0.38 (0.93)	0.713 (0.501)	0.323 (0.761)
Obs.	18	18	36	32	23	55	10	9	19	32

Note: Columns (3), (6), (9) and (10) contain Mann-Whitney U test-statistics, with p -values in parentheses. All other columns contain sample averages, with standard deviations in parentheses. Column (10) tests for significant differences between the Honest and Dishonest cases *within* the No-Pro treatment.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.11: Beliefs of S2 about S1's Degree of Trust

	CASE			MANN-WHITNEY U Tests		
	(1) Neutral	(2) Dishonest	(3) Honest	(4) N vs. D	(5) N vs. H	(6) D vs. H
\hat{a}	1.28 (3.98)	-1.34 (3.68)	-0.20 (4.29)	2.098** (0.035)	0.815 (0.424)	0.780 (0.445)
\hat{b}_{avg}	1.40 (1.89)	1.36 (2.66)	0.46 (2.26)	0.031 (0.980)	1.153 (0.259)	0.807 (0.431)
Obs.	18	32	10	50	28	42

Note: Columns (1) to (3) contain sample averages, with standard deviations in parentheses. Columns (4) to (6) contain Mann-Whitney U test-statistics, with p -values in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.12: Comparison of Responsiveness to a across Rolls

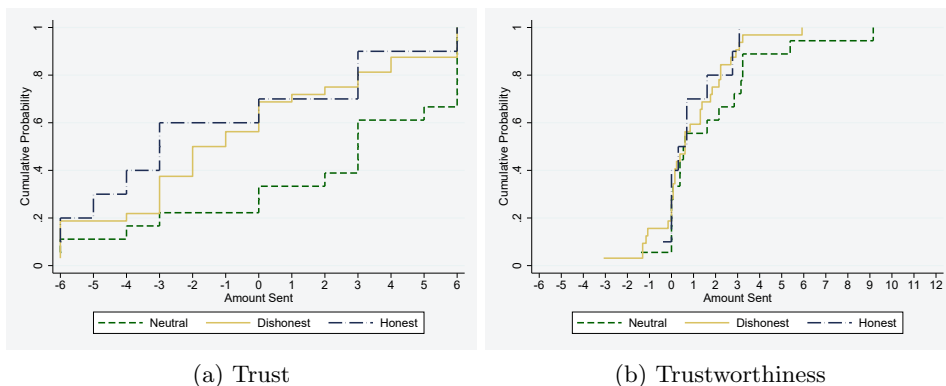
	ROLL		
	(1) High	(2) Low	(3) MWU
β	0.45 (0.63)	0.43 (0.53)	0.420 (0.680)
Obs.	18	42	60

Note: Columns (1) and (2) contain sample averages, with standard deviations in parentheses. Column (3) contains Mann-Whitney U test statistics, with p -values in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

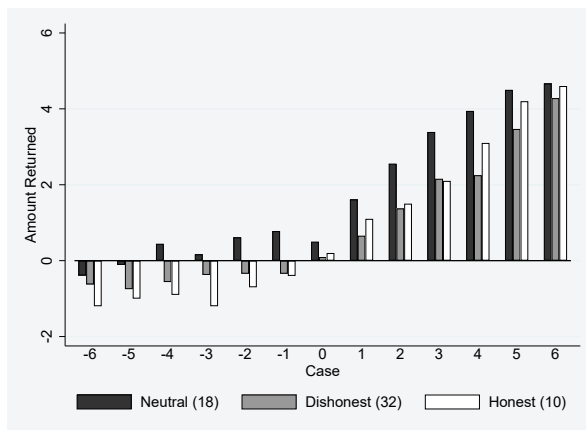
2.B.2 Figures

Figure 2.13: Comparison of Trust across Cases in Baseline



Note: Empirical cumulative distribution functions of trust and trustworthiness across the three cases in Baseline. Trustworthiness levels are averaged over all 13 possible values of trust.

Figure 2.14: Trustworthiness for Each Level of Trust in Baseline



2.B.3 Robustness: Restricted Sample

I restrict the sample in two distinct ways. First, a relatively large number of subjects submit a non-monotone sequence of b -choices.¹¹ I count the number of times each subject reports a b_a that is *lower* than b_{a-1} and restrict the sample to those subjects who

¹¹To some extent, these can be explained by errors on the subjects' part while adjusting the sliders or by psychological narratives that justify giving more after lower choices of a . For example, second-movers could view transferring 1 or 2 as nit-picky or petty, to which they respond less benevolently

report at most one such violation. This yields a sample of 34 S2s.¹² As an alternative restriction, I drop the 25% slowest subjects in the trial questions, as they may have had the hardest time getting to understand the dynamics of the Moonlighting Game. As a result, 43 S2s remain. Both measures correlate positively ($\rho = 0.38, p < 0.001$). As can be seen from Figure 2.15 (and from Table 2.13 in Appendix 2.B), the two sample restrictions render the differences in trust more pronounced, as they increase in magnitude and significance. To illustrate, among consistent subjects (Figure 2.15a), those in the Neutral case send on average 4.0 ECU, while those in the Dishonest and Honest case *take* on average 0.8 and 1.3, respectively (N vs. D: $U = 2.588, p = 0.008$; N vs. H: $U = 2.390, p = 0.019$). Similar results are found for subjects who solve the trial questions quickly. However, this exercise yields no additional insights regarding trustworthiness, as differences remain insignificant in this dimension.

Table 2.13: Trust of S2 When Restricting Sample

	CASE			MANN-WHITNEY U Tests		
	(1) Neutral	(2) Dishonest	(3) Honest	(4) N vs. D	(5) N vs. H	(6) D vs. H
A. Subjects with at most one error						
a	4.00 (2.49)	-0.81 (4.68)	-1.29 (4.46)	2.588*** (0.008)	2.390** (0.019)	0.101 (0.936)
b_{avg}	2.12 (2.76)	1.15 (1.81)	0.77 (1.06)	0.449 (0.676)	1.027 (0.335)	0.506 (0.634)
Obs.	11	16	7	27	18	23
B. Subjects who solve trial questions quickly						
a	2.47 (4.32)	-1.09 (4.12)	-2.50 (3.39)	2.288** (0.021)	2.177** (0.030)	0.763 (0.463)
b_{avg}	1.73 (2.50)	0.72 (1.24)	0.90 (1.10)	1.255 (0.215)	0.821 (0.434)	0.168 (0.880)
Obs.	15	22	6	37	21	28

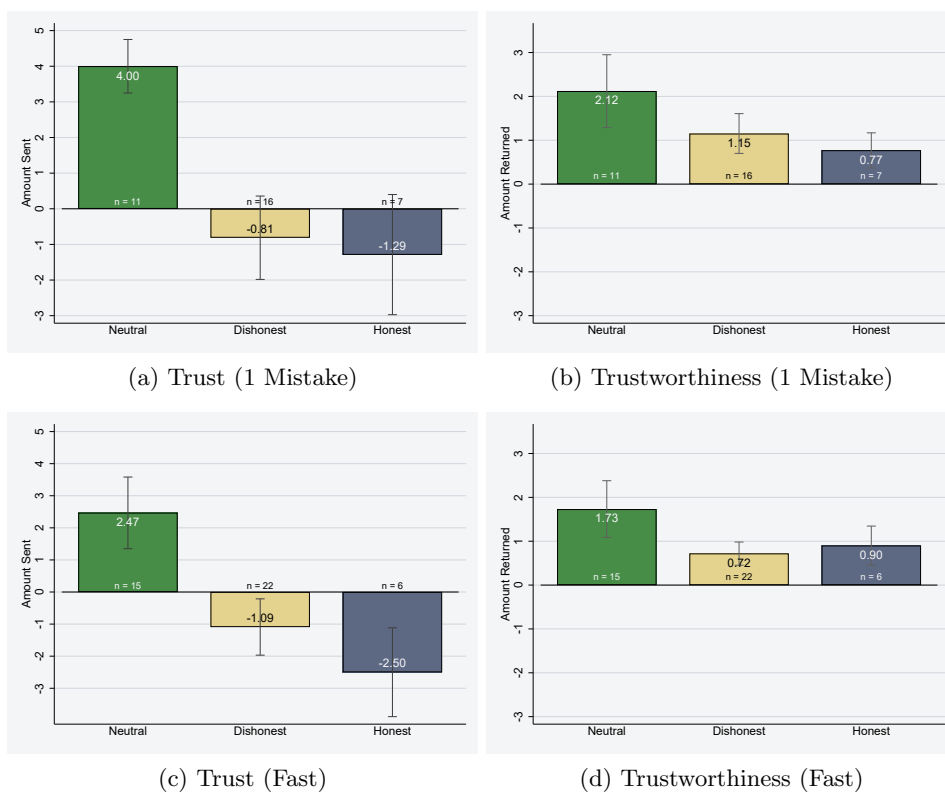
Note: Columns (1) to (3) contain sample averages, with standard deviations in parentheses. Columns (4) to (6) contain Mann-Whitney U test-statistics, with p -values in parentheses. Panel A uses a sample of subjects who record at most one violation, where a violation is defined as choosing a strictly lower b_a as compared to b_{a-1} . Panel B uses a sample that only includes the 75% fastest subjects in the trial questions.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

than not sending anything at all. Alternatively, small amounts could be seen as acts of altruism, whereas large amounts may imply a hidden agenda of expecting a large back-transfer in return, which may crowd-out a second-mover's intrinsic motivation to reciprocate. Yet, it remains a concern.

¹²Results are similar when either restricting the sample based on zero or at most two violations.

Figure 2.15: Comparison of Trust between Cases in Baseline with Restricted Samples



Note: Panels (a) and (b) depict average levels of trust in Baseline for a sample that consists only of subjects who violate the monotonicity requirement at most once. Panels (c) and (d) show the results after dropping the 25% slowest subjects in the trial questions. Trustworthiness levels are averaged over all 13 possible values of trust. The spikes represent the standard error around the mean.

Teaching Children Norms in the Streets*

Children have never been very good
at listening to their elders, but they
have never failed to imitate them.

JAMES BALDWIN

3.1 Introduction

Social norms, as shared understandings of what is acceptable or forbidden (Ostrom, 2000), play a major role in governing daily social interactions in a variety of economically relevant settings. While there is abundant literature on how they are enforced, we still know very little about how they are transmitted from one generation to another. Research in developmental psychology shows that children tend to acquire social behavior and internalize norms even at a young age through the observation and subsequent imitation of parents.¹ This process has been dubbed observational or vicarious learning (Bandura, 1965, 1977) and parents play a major role in it (MacCoby, 1992; Bauer et al., 2014). In particular, they may teach social norm compliance

*Co-authored with Fabio Galeotti and Marie Claire Villeval

¹See *e.g.* Berkowitz and Grych (1998); Smetana (1999); White and Matawie (2004); Hardy et al. (2008); Roest et al. (2009); Degner and Dalege (2013) for papers studying the transmission of (moral) values from parents to children.

to their child through leading-by-example: they transmit social norms through modeling desired behavior to the child. At the same time, norms are also internalized by experiencing punishment of norm violations (Sugden, 1986; Coleman, 1994; Fehr and Gächter, 2000; Fehr and Gächter, 2002; Fehr and Fischbacher, 2004; Young, 2008) and children at a relatively young age understand the fairness of such enforcement of norms (Piaget, 1932). This implies an alternative way through which norms can be transmitted to a child: parents may teach their child about norms by *punishing* norm violations of others in the presence of their child. By doing so, the child learns vicariously that violations of the norm will not go unpunished, which should in turn promote future norm compliance.

As the primary aim, we examine whether parents are more likely to punish norm violations, also referred to as *norm enforcement* in this chapter, in the presence of the child in order to teach the child about the importance of complying with social norms. Previous literature distinguishes between two forms of norm enforcement: direct punishment by verbally confronting the violator or indirect punishment by withholding *nice* behavior towards the violator when given the chance (Balafoutas et al., 2014, 2016). Since the threat of counter-punishment may loom larger with the former, the two types of punishment have been shown to be substitutes. Therefore, as the secondary aim, we study whether the type of punishment inflicted upon a norm violator changes in the presence of the child. As the final aim, although less novel and not the initial aim of this chapter, we examine the extent to which parents model behavior in accordance with a norm, also called *norm compliance*, more often in the presence of the child. Taken together, we aim to answer two research questions: Do parents engage more often in norm enforcement and norm compliance in the presence of their child, with the aim of educating the child? And regarding norm enforcement, does the nature of punishment change in the presence of the child?

To address our research questions, we conduct a field experiment in the vicinity of 30 public elementary schools, chosen randomly, in the municipality of Lyon, France, involving 601 parents of children aged 3 to 12. We focus on the norm of non-littering and the violation thereof. The importance of compliance with this norm is universally acknowledged in the setting of our experiment.² We employ a 3×2 design in which

²As an illustration, almost all respondents (85%) of a national survey conducted by TNS Sofres, using a representative sample of the French population over 15 years old ($N = 1060$), considered littering unacceptable, and more deplorable than other acts such as vandalism (74%) and speeding at 160 km/h on the highway (44%). The respondents also rated educating children as the best policy intervention to deter littering (65%), much higher than more coercive measures (higher detection probability or fines, 49%), better information provision to citizens (32%), or higher budgets devoted to cleaning (13%), which highlights the importance of teaching for the maintenance of the norm. The survey is available at <https://bit.ly/2X0grVK>.

we vary the opportunity to enforce the norm and the presence of the child. The design has been inspired by Balafoutas et al. (2014) to elicit the direct and indirect punishment of the norm violation. We recruit two trained actors, one male and one female, at an actor studio. Each actor plays one of three different scenes in front of the parent, accompanied or not by his or her child. We target parents before *or* after they had visited the respective schools at the beginning *or* end of the school day. This set-up provides natural variation in whether or not the child was present for an otherwise comparable sample of parents. We call these conditions “Child” and “Alone”, respectively. Importantly, parents and their children are not aware of being part of an experiment, which constitutes a major advantage of our study.³ Depending on the scene, (i) the actor intentionally violates the non-littering norm (“Violation”), (ii) the actor drops, seemingly accidentally, the content of his or her bag on the ground, suggesting a need for spontaneous help (“Help”), or (iii) the actor first violates the non-littering norm and subsequently drops the content of his or her bag on the ground (“Violation + Help”). The Violation scene provides the targeted parent with an opportunity to sanction the norm violation directly and informs us on the prevalence of direct punishment. Analogously, the Help scene provides the targeted parent with an opportunity to help a stranger in need and the Violation + Help scene provides him or her with both an opportunity to punish and an opportunity to help. Then, we identify indirect punishment as the drop in Helping between the Help and Violation + Help scenes.

Regarding parents’ willingness to engage in direct punishment of a norm violator in the presence of their child, social learning theory (Bandura, 1977) suggests that parents should be more inclined to enforce a norm from an educative perspective. Our results show that, indeed, twice as many parents (22 vs. 11 percent) engage in direct punishment of the norm violation when accompanying one or more children, as compared to being alone. Similarly, parents with children are about twice more likely to help the actor pick up his or her fallen items compared to parents alone (47 vs. 26 percent), confirming the higher willingness of parents to enforce the helping norm when their child is present. Regarding indirect punishment of a norm violator through withholding help, parents may face conflicting motives when the child is present. On the one hand, the fear of retaliation following direct punishment, in combination with the parent’s desire to educate the child, may induce the parent to resort to indirect

³We acknowledge that not all adults accompanying children to school are parents. This task may also be performed by grandparents or other caretakers. Nonetheless, we will refer to our subjects as “parents” throughout the chapter. In the experiment, we adopted some measures to distinguish real parents from other caregivers, estimating that fathers and mothers represent about 90% of our sample (see Section 3.3).

punishment more often when the child is around. On the other hand, the more subtle educative signal of indirect punishment might be missed by the child if the child does not notice that the parent intentionally abstains from helping. What is more, the parent may realize this and educate the helping norm instead even after a violation. In line with this latter argument, our experiment does not provide evidence for a higher tendency to punish indirectly in the presence of a child, as the drop in helping following a violation is not significantly larger for parents accompanying children than for parents alone (16 vs. 10 percentage points). At the same time, helping rates following the violation are still higher when the child is around, which suggests that parents deem teaching the helping norm as relatively more important than teaching that norm violations ought to be punished.

In order to explore the mechanism behind our results, we conduct a follow-up vignette study in which we elicit the social appropriateness of littering, punishment, and helping among a comparable sample of parents. Specifically, the objective is to test whether the perception of the norms is sensitive to the presence or the absence of the child in the vignettes. Importantly, we find that parents did not perceive littering, punishing, helping, and withholding help as more or less appropriate when the child is around. In other words, the social norms are perceived to be the same in both conditions. Thus, if parents punish or help more in the presence of their child, it is not because norms are different. Our analysis and subsequent discussion are also able to discard alternative explanations related to social image concerns, in the eyes of bystanders or of their own children, or a lower fear of retaliation. This lends support to our interpretation that parents become more motivated to enforce or comply with a given social norm from an educative purpose.

Our field experiment contributes to the literature on the role of parents in the inter-generational transmission of norms and preferences by assessing their norm enforcement and compliance with the aim of educating their child. The development of preferences in children has received a lot of attention in the developmental psychology and economics literature (Sutter et al., 2019), but only few scholars have focused systematically on parental socialization efforts (most notably Houser et al., 2016; Ben-Ner et al., 2017). In contrast to these studies, we are able to assess parents' natural behavior while maintaining experimental control. Since parents are unaware of being part of the experiment, our approach eliminates social desirability bias, which could induce parents to display desirable behavior in the presence of the child, as a potential confound for our results. Moreover, we take a novel perspective by not simply studying the extent to which parents model desired behavior, but also how they teach through punishing others' undesired behavior. This latter feature has, to

the best of our knowledge, not been studied before. Our results therefore contribute to the understanding of the transmission of social norms. Finally, as an additional contribution of our study, our experiment successfully replicates, in both conditions, the results found by Balafoutas et al. (2014), who show that subjects withhold help as a means of indirect punishment and do so as a substitute for direct punishment.

The remainder of this chapter is set up as follows. In the next section, we briefly discuss the relevant literature. Then, we lay out our research design and conjectures in Section 3.3. Section 3.4 describes the data and 3.5 reports our results. Then, Section 3.6 dives deeper into our findings through a vignette study. Finally, Section 3.7 provides a concluding discussion.

3.2 Related Literature

The development of preferences in children and adolescents has constituted an emerging and rapidly expanding topic in economics over the last 15 years (see Sutter et al., 2019, for a review). This growing interest results from the need to better understand the behavior of adults through the development of non-cognitive skills in childhood (*e.g.*, Heckman and Rubinstein, 2001; Heckman et al., 2006), and how early interventions could improve economic and social conditions in adulthood. Controlling for socio-economic status, this literature has evidenced a correlation between parents' and children's preferences in domains such as risk and time preferences (Dohmen et al., 2012; Kosse and Pfeiffer, 2012; Alan et al., 2017; Chowdhury et al., 2018; Brenøe and Epper, 2019). By contrast, the evidence is somewhat mixed for social preferences. In particular, Chowdhury et al. (2020) and Sutter and Untertrifaller (2020) report a positive correlation for distributional preferences and prisoner's dilemma cooperation, respectively, while Cipriani et al. (2013) find no correlation for contributions to a public good.

Previous studies looking at the non-genetic mechanisms behind these correlations attributed an important role to parenting styles and parental investments in terms of time and goods (Cunha and Heckman, 2010; Heckman and Mosso, 2014). Zumbuehl et al. (2013) and Alan et al. (2017) show that the correlation of risk preferences is stronger for parents who invest more time and effort in the upbringing of their child, while Brenøe and Epper (2019) find that the parenting style matters more than the time spent with children for the formation of time preferences. Relatedly, Falk and Kosse (2016) find that a longer duration of breastfeeding, as an instrument for childhood quality, leads to persistently higher rates of patience, altruism, and risk aversion. Exposition to violence during childhood has also been found to increase risk

aversion (Castillo, 2020). In addition, parental behavior has been identified as an important explanation for the gap in preferences and personality traits between children from low and high socio-economic status (SES) families (Benenson et al., 2007; Bauer et al., 2014; Deckers et al., 2017; Falk et al., 2018; Kosse et al., 2020; Sutter and Untertrifaller, 2020). Kosse et al. (2020) show that the gap can be closed by exposure to a mentor, highlighting an additional role for the social environment. Evaluating the effects of early interventions aimed at low SES families, Cappelen et al. (2020) show that preschool and parenting programs have a strong and persistent impact on children's social preferences. Bettinger and Slonim (2006) find a positive effect of education programs on charitable giving that seems to be driven by improvements in parental mental health. Attanasio et al. (2020) conclude that increases in parental investments drive the positive effects of an early childhood intervention on cognitive and socio-emotional skills. Parents' ambitions concerning the professional success of their offspring have also been found to predict the children's degree of competitiveness (Khadjavi and Nicklisch, 2018). Language (e.g., Sutter et al., 2018) and culture (e.g., Gneezy et al., 2009; Andersen et al., 2013; Falk et al., 2018) play a major role in the transmission of preferences, too. From a theoretical perspective, a few models have argued that parents are motivated by so-called "imperfect empathy" or "paternalistic altruism" that induces them to transmit their preferences to their children in an optimal way (Bisin and Verdier, 2001; Tabellini, 2008; Adriani and Sonderegger, 2009; Doepke and Zilibotti, 2017).

Together, the existing evidence shows that children's preferences can be shaped by parental and environmental characteristics, leaving room for parents to actively try to mould these. The current study contributes to this field through a different perspective. Instead of exploring the correlation between parents' and children's preferences, we focus on the identification of parents' socialization efforts to transmit normative preferences to their child. Our study shares this feature with Ben-Ner et al. (2017), who show that fathers and parents of generous children (as measured by a baseline dictator game) behave more generously in the dictator game when they know that their child will observe their choice, and Houser et al. (2016), who show that parents are significantly more likely to report dishonestly (when this benefits the child) when the child is absent than when the child is present. As opposed to these studies that focus on modeling desired behavior, we additionally study whether and how parents teach their children about norms through punishing undesired behavior by another party. In addition, we observe parents' natural behavior instead of behavior in artefactual experiments, while still being able to exploit exogenous variation in socialization motives.

Finally, in addition to the literatures on the transmission of norms and on parental involvement in the development of the non-cognitive skills of their offspring, our study contributes to the small literature exploring the punishment of norm violations in the field (e.g., Balafoutas and Nikiforakis, 2012; Balafoutas et al., 2014, 2016; Berger and Hevenstone, 2016; Przepiorka and Berger, 2016; Artavia-Mora et al., 2017). As explained above, punishment could take a direct form by verbally confronting the violator. However, the deterring effect of direct punishment is seriously hampered by the risk of retaliation and counter-punishment (Denant-Boemont et al., 2007; Janssen and Bushman, 2008; Nikiforakis, 2008), which typically causes direct punishment rates to be low in field studies (Balafoutas and Nikiforakis, 2012; Balafoutas et al., 2014, 2016; Berger and Hevenstone, 2016; Artavia-Mora et al., 2017). Instead, people resort to means of indirect punishment that do not involve the risk of retaliation (Casari, 2012). We add to this strand of literature by considering a setting where the punishment of norm violators could stem from a desire to teach normative preferences to a child observing the violation.

3.3 Experimental Design

We ran our experiment in the vicinity of 30 public elementary schools in Lyon, France. To make sure that the norm violation was introduced exactly the same way throughout the experiment, we recruited two trained actors, one male and one female, from a professional acting school. Teams of four collected the data: a trained actor, two research assistants (RA1 and RA2), and a supervisor (one of the researchers). The actor and the two RAs were blind to the purpose of the study. Ethical approval was obtained from CEEI, the institutional review board of the French National Institute of Medical Research and Health (Inserm, IRB00003888, No. 19-592). This section first introduces the different conditions and our conjectures. Then, it presents the detailed procedures and discusses some aspects of the protocol.

3.3.1 Conditions and Conjectures

The protocol of the scenes is inspired by Balafoutas et al. (2014). In all conditions, the actor wears plain clothes, holds a small plastic bag containing food waste (a banana peel), and carries a cotton shoulder bag containing five file folders and a few pens (see a picture of the materials in Appendix 3.A.3). The actor plays one of three different scenes in front of a targeted parent in the streets surrounding the targeted school. The scenes constitute our three treatments and they are summarized in Table

3.1. They differ in the opportunities that are provided to the parent to enforce the norm, as explained below. We target parents approaching *and* leaving the school in the morning *and* in the afternoon. This varies whether or not the child is present for an otherwise similar sample of parents and creates two conditions, to which we refer as “Child” (C) and “Alone” (A). This results in the 3×2 between-subjects design summarized in Table 3.1.⁴

Table 3.1: Treatments

Step	Violation	Violation + Help	Help
1	Actor approaches targeted parent from the front	Actor approaches targeted parent from the front	Actor approaches targeted parent from the front
2	Actor pauses while going through the bag	Actor pauses while going through the bag	Actor pauses while going through the bag
3	Actor litters	Actor litters	-
4	-	Actor takes out bag contents	Actor takes out bag contents
5	Actor continues moving	Actor continues moving	Actor continues moving
6	Actor pauses a second time	Actor drops bag contents on the floor	Actor drops bag contents on the floor
7	Actor leaves the scene	Actor picks up stuff and leaves the scene	Actor picks up stuff and leaves the scene
N_{Child}	100	100	100
N_{Alone}	100	100	101

Note: Scenario for each of the scenes. The sample size and corresponding power analysis is discussed in Section 3.3.2. We stopped the data collection after having met our objective in each of the cells. A misconception of the state of affairs during the last session accidentally led to a 101st observation in H_A before the 100th observation in VH_A .

The first scene, called the “Violation” treatment (abbreviated V_C and V_A for the Child and Alone condition, respectively), aims at measuring the prevalence of direct punishment of a social norm violation. In this scene, the actor approaches the targeted parent from the front. When the parent is roughly 10 meters away, the actor pauses and goes through the cotton bag. Then, when the parent is roughly 5 meters away, the actor litters in clear sight of the parent by throwing away the plastic bag with the food waste. The actor makes sure to throw the plastic bag on the side of the street and away from the parent. This is to prevent the parent from perceiving the litter as a potential “danger” for the child. Following Balafoutas and Nikiforakis (2012), the

⁴Note that the distinction between morning and afternoon implies that, strictly speaking, we have a $3 \times 2 \times 2$ design. However, during the design of our experiment, we have assumed this distinction to be orthogonal to our research questions and decided to aggregate observations across morning and afternoon sessions. Indeed, patterns and results remain unchanged when assessing morning and afternoon observations separately.

actor makes it as clear as possible that the act is intentional. Subsequently, the actor slowly starts moving again, while still going through the bag, thus clearly showing no intention to pick up the litter. We classify all forms of verbal confrontation aimed directly at the actor and which explicitly address the violation as direct punishment.⁵ In order to avoid different interpretations across research assistants and facilitate a uniform definition of punishment, we do not identify punishment according to its intensity (*e.g.*, a raised voice or use of profanities). Moreover, results are similar when including RA-fixed effects, which further alleviates this concern (available upon request). Additionally, the parent might, as a substitute or complement to direct punishment, address the child by disapproving of the violation. If this happens within earshot of the violator, this is recorded as a separate type of reaction and not as a form of direct punishment.⁶

The difference in direct punishment rates, $p(\cdot)$, after observing the Violation of the social norm between the Child and Alone conditions informs us on the parents' tendency to engage in direct punishment with the goal to teach the child that the norm violation constitutes misbehavior that ought to be punished. Following social learning theory (Bandura, 1977), we conjecture that educative motives spur parents to inflict direct punishment when the child is present:

Conjecture 3.1 (Socializing Direct Punishment) *In reaction to a social norm violation, direct punishment rates are higher for parents accompanying a child than for parents alone: $p(V_C) > p(V_A)$.*

Importantly, an implicit assumption underlying Conjecture 1 is that the fear of retaliation is the same across conditions. This need not be the case. On the one hand, the parent may fear the consequences of retaliation more in the presence of the child. In particular, the parent may be anxious that the child will be involved in some way in the retaliation, even by simply witnessing it. If this is the case, our experiment may underestimate the effect of the child's presence and this would work against Conjecture 1, making any results that support it even more convincing. On the other hand, the parent may deem the violator less likely to retaliate in front of a child, which may

⁵An example of direct punishment in our experiment is: "You should not throw that on the ground; you should throw it in the garbage bin."

⁶When the parent addresses the child directly about the violator within the violator's earshot, this is considered as "indirect reprimand" (Berger and Hevenstone, 2016). Another alternative way of teaching norms is picking-up litter. However, its meaning is ambiguous. Indeed, it is a reprimand if the violator observes the action, but it does not teach the child that violators have to be punished. Finally, the parent can explain the child in private about the violation. This may teach the importance of compliance but again, not that violators have to be punished. Our protocol focuses on direct and indirect punishment, measures indirect reprimand, but does not allow us to measure further explanation to the child in private.

alleviate the fear of retaliation. If so, our experiment may overestimate the effect of the child’s presence. We discuss this issue further in Section 3.5 when presenting our results.

Because of the fear of retaliation discussed above, direct punishment rates are typically low in the field (Balafoutas et al., 2014, 2016). Instead of direct punishment, parents may resort to forms of indirect punishment, for which retaliation is arguably less likely. We provide the opportunity for indirect punishment in the second treatment, called “Violation + Help” and abbreviated VH_C and VH_A for the Child and Alone condition, respectively. This scene starts in a similar way as the Violation scene. However, a few seconds after having littered and *before* the parent reaches the actor, the actor accidentally drops the contents of his bag on the sidewalk. This presents the parent with an opportunity to withhold help as a form of indirect punishment (in addition, the parent can still punish directly). Note that the Violation + Help scene may also provide a longer window to punish *directly*, as the actor is no longer moving. To equalize this window of opportunity across scenes, the actor pauses a second time in the Violation scene around the same time where (s)he would drop the files in the other two scenes.

We define a parent to help if he or she picks up at least one item from the ground, as in Balafoutas et al. (2014). Conveniently, this constitutes a clear and tangible criterion which limits the scope for different interpretations of helping across the different research assistants. In case parents stimulate their children to help, we also count this as helping.⁷ In the rare cases the child helps without any intervention of the parent, this is not counted as helping. In order to measure whether parents indeed withhold help, we introduce a third treatment, called “Help” and abbreviated H_C and H_A for the Child and Alone condition, respectively. In this scene, the helping opportunity is not preceded by a littering violation by the actor. Indirect punishment then shows in the aggregate through significantly lower helping rates $h(\cdot)$ in the presence of a violation: $h(H_i) - h(VH_i), i = A, C$.

Naturally, helping a stranger in need is an example of a social norm in itself that parents may be willing to transmit to their children. Although our primary interest is in the use of punishment to educate children, we believe that we can also contribute in the dimension of parents showing desired behavior in front of their children. Therefore, before moving to indirect punishment, we formulate our hypothesis regarding helping behavior. Based on social learning theory (Bandura, 1977), we expect helping rates in the Help scenes to be higher for parents accompanying their child, as

⁷This is a rare event and treating it as not helping does not change the results, as shown in Section 3.5.

compared to parents alone:

Conjecture 3.2 (Socializing Helping) *In the absence of a littering violation, parents in the presence of their child, in contrast to parents alone, are more likely to provide help: $h(H_C) > h(H_A)$*

Subsequently, we hypothesize that educative motives induce parents to punish indirectly more often when the child is around, meaning that we should observe a larger decrease in helping rates in the presence, rather than the absence, of the child.

Conjecture 3.3 (Socializing Indirect Punishment) *The extent to which parents withhold help after observing the violation of a social norm is larger in the presence of the child: $h(H_C) - h(VH_C) > h(H_A) - h(VH_A)$.*

Although indirect punishment is less likely to evoke retaliation, the educative motive of withholding help is probably weaker than that of direct punishment, as its implicit nature may be harder for the child to grasp. The child may thus not understand completely that the parent is punishing the violator. Hence, the marginal teaching benefit from indirect punishment is likely to be smaller than that of direct punishment. Realizing this, the parent might still want to teach the child that one should help a stranger in need, despite the violation. These considerations would however work against our conjecture. The Violation + Help treatment may also be informative about the fear of retaliation between conditions, as we can observe the extent to which direct punishment is substituted for indirect punishment. In case the drop in direct punishment is larger (smaller) in the Child condition as compared to the Alone condition, this would point to a higher (lower, respectively) fear of retaliation in the presence of the child.

The three conjectures have been pre-registered with AsPredicted (#24270).

3.3.2 Procedures

The three different scenes were played in random order at two different times of the day: in the morning and the afternoon. Schools provide the opportunity to parents to drop off their children from 7:50 AM onward and school starts at 8:30 AM. In the afternoon, school finishes at 4:45 PM and after-school activities finish at 5:30 PM. As a result, we played the scenes roughly between 7:45 and 8:45 AM in the morning and between 4:15 and 6:00 PM in the afternoon.

The actors were randomly alternated across sessions. Targeted parents were identified by the actor and the supervisor. Depending on the condition, they either

scouted the streets for parents approaching the school or parents leaving the school. We restricted ourselves to single parents walking alone or with one or more children. We did not target parents pushing a stroller, riding a bike, walking a dog, holding something with both hands, or accompanying disabled children. This, combined with our definition of helping, should make the cost of helping negligibly small and not related to the presence of the child. Parents visibly in a rush or talking on the phone were also avoided. The actor staged the scene such that ideally only the targeted parent could respond to it and any collective action problem regarding the execution of punishment or helping was extremely limited. In case the parent engaged in direct punishment, the actor always complied without speaking and disposed of the litter. Subsequently, the actor quietly left the scene.

In the meantime, the first research assistant (RA1) observed the scene from a distance and recorded the type of scene being played, whether the parent was accompanied by one or more children, and the parent's response to the scene. RA1 measured up to three outcome variables in the scenes: whether the targeted parent helped the actor, whether the targeted parent confronted the actor verbally regarding the violation, and whether the parent expressed his or her disapproval to the child in a way that could be heard by the actor. RA1 furthermore recorded whether there were witnesses who could possibly have intervened in the scene, the gender of the parent and the child, the weather conditions, the time of the day, the cleanliness of the street, and whether the target actually observed the violation. After the scene had ended, RA1 cleaned up the scene in case of a non-sanctioned littering violation and verified the recorded information with the actor.

After the targeted person had left the scene, the second research assistant (RA2) approached the parent and asked whether he or she was willing to participate in a short, seemingly unrelated, survey. RA2 informed the parent that the survey was for a university project. We made sure to mention nothing about the preceding scene. Our main interest was to assess whether the targeted parent was accompanying or had accompanied (i) his or her own child, (ii) a child that he or she guards, (iii) the child of a relative, or (iv) no child. In cases (i) to (iii), we also asked for the child(ren)'s age and gender. If a parent declined to take the survey, RA2 had to guess the gender and age of the child(ren), if present. In case the target indicated having no child going to the school, we dropped this person from the sample. Moreover, RA2 noted down information on the appearance of the target. The main purpose of this was to verify that no parent had been targeted twice during the day. For an exhaustive list of information recorded and the survey questions, see Appendix 3.A.1.

Challenges We identified three main challenges in relation to our design. First, parents might be more in a hurry when arriving at school, rather than when leaving school. We addressed this issue by running the scenes both in the morning when parents dropped off their kids at school *and* in the afternoon when parents arrived at school to pick up their child. As a result, we equalized parents' "hurry" between child conditions as much as possible. As additional measures to avoid parents in a hurry altogether, we did not stage the scenes in the final five minutes before the beginning or end of the class and avoided parents who were visibly in a rush. Also, we ensured that each of the scenes did not last for more than a couple of seconds and that helping constituted a relatively quick act. Finally, when analyzing the data we examined whether there are timing effects on behavior, specifically whether parents who arrived early or late to pick up or drop their children behave differently. To do so, we included the number of minutes since or until the beginning or end of school, or the observation number within a session (morning or afternoon) in the regression analysis. We found no timing effects, as explained in Appendix 3.B.5.

A second concern was the audience of the interaction. Ideally, we would have liked to have one parent-child pair in an otherwise empty street to avoid audience effects on behavior. This may be hard to guarantee, especially because multiple parents may be arriving with their children around the same time. We combated this by identifying the more secluded streets in the school neighborhoods and be present well before the beginning or end of class (30-40 minutes before). The setup of the scene ensured that the scene was staged primarily in the view of the targeted parent, with him or her being the closest to the interaction. In case someone else was approaching while the scene was about to start, the actor waited until the approaching person had passed before starting the scene. This ensured that the parent realized that he or she was the prime candidate to respond to the scene. As a final measure, RA1 recorded whether there were any witnesses who could have possibly intervened with the scene, such that we can control for this in the data analysis.

A third concern was whether the targeted adult was indeed the parent of the child. Children might also be picked up from and dropped off at school by their nannies, grandparents or other caretakers. Insofar as these caretakers still play an important role in the child's education, studying their behavior remains relevant. We might however need to be careful with labeling observed behavior as a tendency of parents *per se*. Therefore, it is important that the share of actual parents is similar across conditions. In case parents exhibit the strongest educative motives, the presence of non-parents would work against our conjectures and make it harder to find an effect. We combated this potential issue by means of the quick survey that

was conducted by RA2 directly after the scene and that allowed us to evaluate the proportion of parents involved in the experiment. A related difficulty pertained to identifying parents arriving at school in the Alone condition, especially when they did not respond to our survey request. Parents approaching the school may simply be random pedestrians not going to the school. For this reason, RA1 tracked the target after the scene to determine whether he or she actually went to school and recorded this.

Power Analysis and Number of Observations We determined an objective of 100 observations per cell and 600 observations in total. Due to the limited literature on this topic, we had no well-established priors with respect to the treatment effect. In comparing V_A and V_C , we drew from the punishment rate in the BasePun treatment of Balafoutas et al. (2014) (17%) and hypothesized a medium effect size due to the presence of the child ($w = 0.3$, *i.e.*, $\sim +11\%$ -point from the baseline). With the specified sample size, we would then achieve a power 98.9% when employing a χ^2 test at a 5% significance level. The same applied to comparing H_A and H_C (a medium effect size implies $\sim +15\%$ -point from the Balafoutas et al. (2014) BaseHelp rate of 39.7%). Moreover, the helping rate in the HelpViolator treatment of Balafoutas et al. (2014) was 18.6%, which implied a power of 99.99% with our sample size when comparing it to their BaseHelp rate to identify indirect punishment.

We stopped the data collection after having met our objective in each of the cells. A misconception of the state of affairs during the last session accidentally led to a 101st observation in H_A instead of a 100th observation in VH_A . Dropping this 601st observation from the data does not affect our results.

Locations We randomly selected 30 out of the 80 public elementary schools Lyon, the third city in France in terms of population size. The vast majority of schools included in the sample host both a kindergarten and a primary school. As a result, the children involved in our experiment are between 3 and 12 years old. In order to make sure that we obtained a representative sample of elementary schools in Lyon, we collected basic information on all public elementary schools in the city, including name and address. We matched each school with the median disposable income and poverty rate of the IRIS area it is located in, and classified each IRIS area as above or below the city-wide median.⁸ We picked schools such that for each socio-economic

⁸IRIS are infra-municipal areas comprising between 1800 and 5000 residents. IRIS is an acronym of ‘aggregated units for statistical information’. France is composed of around 16,100 IRIS. We extracted data from the 2014 edition of the INSEE survey “Revenus, pauvreté et niveau de vie”, available at <https://www.insee.fr/fr/statistiques/3288151>. The poverty rate is measured as the

measure roughly half of the selected schools are classified as above the median, and half are classified as below the median.

We excluded private schools to avoid unobservable selection effects that would possibly interact with our research question. For example, parents using private schools may have different income levels than parents using public schools. Moreover, eligibility for a private school does not depend on the parents' address, while the assignment of a child to a given public school is determined strictly by the parents' address. This ensures that the median income and poverty rate of the school's IRIS give us indirect information about the wealth of parents whose children are assigned to the public school. The resulting diversity in neighborhoods allows us to assess whether the intensity of socialization efforts are affected by the socio-economic environment, as suggested by previous research (Benenson et al., 2007; Bauer et al., 2014; Angerer et al., 2015; Deckers et al., 2017; Kosse et al., 2020). Admittedly, the strength of our socio-economic measures as a proxy may be diminished when differences between IRIS areas are driven by wealthy households that send their children to a private school. This concern is partially alleviated by the fact that we use median income, rather than average income which may be more sensitive to wealthy outliers, and merely classify IRIS areas as below or above the median. Moreover, around 82% of elementary school pupils in the *département* to which Lyon belongs are enrolled in a public school, suggesting that the vast majority of households send their children to a public school.⁹ Together, these two features make it less likely that our classification of the schools' IRIS areas is driven predominantly by households that send their children to a private school. Still, in order for our proxy to be fully convincing, we would need detailed information regarding private-school enrolment on the IRIS level, which is unavailable.

In order to avoid being identified and raise suspicion, we visited each school during one morning and one afternoon on the same day. As a result, the experiment was run on 30 days in total.¹⁰ This also made avoiding previously-targeted parents easier. As a final measure to avoid being identified, we selected multiple suitable spots around the same school by inspecting the school's surroundings beforehand on Google Maps *and* in-person. We made sure to move to a new spot at least once during the same morning or afternoon session. If we deemed the surroundings not suitable for the scenes (open terrain, steep hills, construction work, dead-end street, etc.), we dropped

share of households with a disposable income below 60 percent of the median income in the city.

⁹Own calculations based on statistics available at <http://www.ac-lyon.fr/cid87007/geographie-chiffres-cles.html>

¹⁰We only ran the experiment in the morning on Wednesdays, due to the fact that schools finish around noon on that day of the week.

the school and randomly selected another one to replace it.

3.4 Data Description

The experiment was run in May and June, 2019. We obtained 601 observations: 301 in the Alone condition and 300 in the Child condition. Our primary outcome variables are Punishment and Helping. (Direct) Punishment is a dummy that takes on value 1 if the parent verbally and explicitly punishes the actor for the violation of the non-littering norm, and 0 otherwise. Analogously, Helping is a dummy that takes on value 1 if the parent picks up at least one item (or asks the child to pick one), and zero otherwise.

Our main interest is in the effect of the scene played (Violation, Help, or Violation + Help) and the condition (Child or Alone). In addition, we control for a set of observables. Male Target is a dummy indicating that a father was targeted, while Male Actor indicates that the male actor played the scene. Furthermore, Morning, Rain, and Hot are dummies indicating that the scene was played during the morning, during rainy conditions, and on a hot day, respectively. Witness is a dummy that takes on value 1 if RA1 deemed another non-targeted adult to be observing the scene *and* to be able to intervene. Finally, we created Rich IRIS which takes value 1 if the school's IRIS area median income is above the city-wide median. Table 3.2 shows that our sample is balanced on most controls, with two notable exceptions. That is, the scenes of the Child condition are somewhat more likely to be played in the morning (significant at a 1%-level) and with a witness around (significant only at a 10%-level). In order to control for potentially confounding effects stemming from these differences, we include them in our regression models below.

Regarding the survey, 47% of 504 approached parents respond to it. Here, we exclude 97 parents who could not be reached for various reasons (e.g., they were talking to another person or responded to a phone call). The fairly low response rate should be taken into account when interpreting the survey responses. Parents in the Child condition were more likely to respond to the survey than parents in the Alone condition. This may already hint at parents behaving more pro-socially in the presence of their child.¹¹ Importantly, of those who answered, the vast majority (88%) reported being the parent of the child, rather than a guardian. These rates do not differ between conditions (see Table 3.6 in Appendix 3.B.1). The survey also reveals that the interviewed parents in the Child condition tend to have slightly more kids than the interviewed parents alone.

¹¹However, it may also indicate that parents are more often in a hurry when they are alone.

Table 3.2: Summary Statistics

	CONDITIONS			
	(1) All	(2) Alone	(3) Child	(4) Δ
Male Target	0.35 (0.48)	0.36 (0.48)	0.34 (0.47)	0.02 (0.04)
Witness	0.15 (0.35)	0.12 (0.33)	0.17 (0.38)	-0.05* (0.03)
Male Actor	0.51 (0.50)	0.52 (0.50)	0.50 (0.50)	0.02 (0.04)
Rich IRIS	0.51 (0.50)	0.50 (0.50)	0.51 (0.50)	-0.01 (0.04)
Morning	0.52 (0.50)	0.46 (0.50)	0.58 (0.49)	-0.13*** (0.04)
Rain	0.08 (0.27)	0.09 (0.29)	0.07 (0.25)	0.02 (0.02)
Hot	0.28 (0.45)	0.31 (0.46)	0.26 (0.44)	0.05 (0.04)
Survey response ^a	0.47 (0.50)	0.42 (0.49)	0.51 (0.50)	-0.09** (0.04)
Observations	601	301	300	601

Note: Columns (1)-(3) contain standard deviations in parentheses. Column (4) contains standard errors in parentheses. *a:* Parents who could not be reached are excluded. Hence, the statistics are computed based on 504, 252, 252 observations in All, Alone, and Child, respectively.

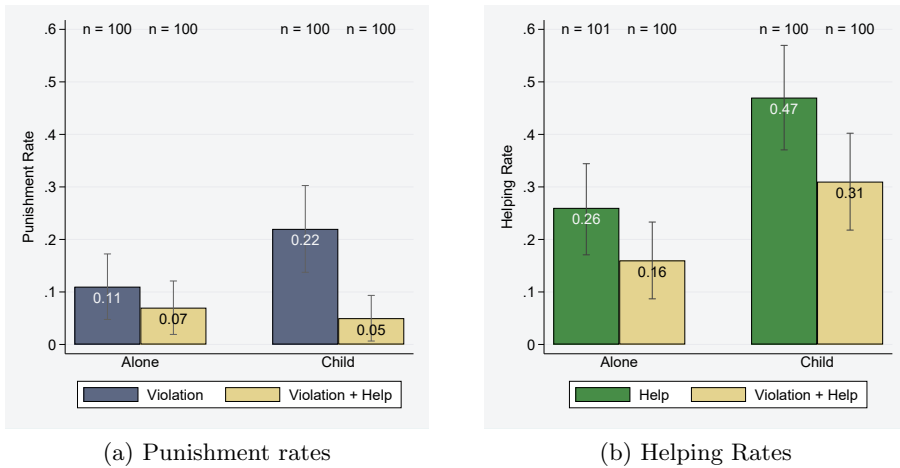
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.5 Results

3.5.1 Main Results

Figure 3.1 displays parents' direct punishment rate (Panel a) and helping rate (Panel b), by treatment (Violation or Help, and Violation + Help) and by condition (Alone or with Child). To test our three conjectures and compare the two conditions we employ χ^2 tests. In addition, Table 3.3 reports coefficient estimates of three linear probability models for each of the dependent variables Punishment and Helping. Models (1) and (4) estimate a simple model including a dummy for the Child condition, a dummy for the Violation + Help scene (VH), and an interaction of these two dummies. To allow for observations at the same school to be correlated, standard errors are clustered at the school-level. Models (2) and (5) also include Male Target, Male Actor, Witness,

Figure 3.1: Behavior of Parents, by Treatment and Condition



Note: Rates of Punishment (Panel a) and Helping (Panel b) across treatments and conditions. The error bars represent a 95% confidence interval around the mean.

Rich IRIS, Morning, Hot, and Rain as controls. Models (3) and (6) add school fixed-effects. Note that we ignore the Help (Violation) scene when examining Punishment (Helping, respectively), thus analyzing roughly 400 observations.

Regarding Conjecture 3.1, we find the direct punishment rate in the Violation treatment to be 22 percent in the Child condition and 11 percent in the Alone condition (compare the dark bars in Figure 3.1a). We can reject the null of no differences in punishment rates between conditions ($\chi_1^2 = 4.39, p = 0.036$) and thus, find evidence in line with Conjecture 3.1. This result is backed up by the linear probability model, as shown by the positive and significant coefficient on *Child* in the left panel of Table 3.3. This analysis supports our first result:

Result 3.1 (Socializing Direct Punishment) *Parents accompanying children are significantly more likely to engage in direct punishment following the violation of a social norm.*

This result is consistent with the willingness of parents to punish more when the child is around in order to teach him or her about the importance of norm compliance and the risk of being sanctioned in case of a violation. We attempt to reject three alternative explanations. To begin with, the negative and insignificant coefficients on the *Witness* dummy in models (2) and (3) of Table 3.3 suggest that we are not estimating a simple social image effect, *i.e.*, parents being more likely to punish

Table 3.3: Regression Analyses of Punishment Rate (Left) and Helping Rate (Right)

	PUNISHMENT			HELPING		
	(1)	(2)	(3)	(4)	(5)	(6)
Child	0.11** (0.05)	0.12** (0.04)	0.12*** (0.05)	0.21*** (0.05)	0.22*** (0.06)	0.25*** (0.06)
VH	-0.04 (0.03)	-0.04 (0.03)	-0.04 (0.05)	-0.10** (0.05)	-0.10** (0.04)	-0.12* (0.06)
VH × Child	-0.13*** (0.04)	-0.12*** (0.04)	-0.13* (0.06)	-0.06 (0.07)	-0.07 (0.08)	-0.10 (0.09)
Male Target		0.06** (0.03)	0.07** (0.03)		0.03 (0.04)	0.02 (0.05)
Male Actor		-0.06 (0.04)	0.15 (0.10)		-0.22*** (0.05)	-0.12 (0.12)
Morning		-0.04 (0.03)	-0.04 (0.04)		0.01 (0.04)	0.02 (0.05)
Witness		-0.04 (0.04)	-0.03 (0.05)		-0.13 (0.08)	-0.14** (0.06)
Rich IRIS		0.02 (0.04)			0.01 (0.05)	
Rain		0.01 (0.04)	-0.01 (0.08)		-0.08 (0.07)	-0.18 (0.12)
Hot		-0.04 (0.05)	-0.08 (0.06)		0.06 (0.05)	0.06 (0.09)
Constant	0.11*** (0.03)	0.14*** (0.05)	0.04 (0.07)	0.26*** (0.04)	0.36*** (0.05)	0.32*** (0.09)
School FE			✓			✓
Observations	400	399	399	401	400	400
Clusters	30	30	30	30	30	30
R^2	0.043	0.068	0.072	0.060	0.131	0.096
F	7.745	5.171	3.086	7.359	6.450	4.257
df	29	29	360	29	29	361

Note: The table contains results from pooled OLS (columns 1, 2, 4, and 5) and linear fixed-effects (columns 3 and 6) regressions. The dependent variable is a dummy for Punishment (columns 1 to 3) or Helping (columns 4 and 6). For the pooled models, standard errors in parentheses are clustered at the school level (30 clusters). One observation is dropped due to missing data on the target's gender.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

purely because they know that they are being observed. Second, we believe that it is unlikely that higher punishment in the presence of the child is driven by the parents' willingness to enhance their image in the eyes of their own child. The drop in direct

punishment rates when comparing the Violation and Violation + Help treatments in the Child condition, as shown in Figure 3.1a and Table 3.3, casts doubt on this argument. However, we cannot exclude the possibility that image concerns also induce parents *not* to punish someone who just dropped his belongings and thereby is already punished by “nature”. Finally, we can also reject an explanation in terms of lower retaliation fear in the presence of the child, *i.e.*, parents being more likely to punish because they believe that a violator is less inclined to retaliate in front of a child. Indeed, punishment decreases between the Violation and Violation + Help treatments in the Child condition *only*, showing that only parents with children seem to decrease direct punishment when indirect punishment opportunities are available. In fact, punishment in the Violation + Help scenes is the only outcome that is not higher in the Child condition as compared to the Alone condition. This argument suggests that parents with children fear retaliation *more*. Moreover, when we analyze punishment in the Violation treatment, we find that only fathers punish significantly more in the presence of the child (see Table 3.7 in Appendix 3.B.2). Assuming that fathers fear retaliation less overall, the presence of the child should decrease the gap in punishment rates between mothers and fathers if children would indeed reduce the fear of retaliation. If anything, we find the opposite. We interpret this as further indirect evidence against the interpretation that parents fear retaliation less when the child is present.¹² Taken together, we thus adopt the teaching motive as the dominant explanation of our results.

Before we move to the analysis of indirect punishment, we note that helping rates in the Help treatments are significantly higher in the Child condition as compared to the Alone condition (compare the dark bars in Figure 3.1b): 47 versus 26 percent ($\chi^2_1 = 9.82, p = 0.002$). This is confirmed by the positive and significant coefficients on *Child* in models (4) to (6) of Table 3.3, and it is consistent with the willingness to teach norm compliance to children when the norm is about helping a stranger in need.¹³ Result 3.2 is thus in line with Conjecture 3.2.

Result 3.2 (Socializing Helping) *Parents accompanying children are significantly more likely to provide help to a stranger in need.*

In both conditions, parents decrease their willingness to help following a norm vi-

¹²We performed a similar exercise with the parent’s estimated height as a proxy for the fear of retaliation, for which we did not find any effect.

¹³Including school fixed effects leads to a significant negative effect of the presence of witnesses on the willingness to help. This could be interpreted as a bystander effect: people are less likely to help when someone else is also able to do it. The introduction of such fixed effects also increases the coefficient associated with the presence of the child, suggesting that the bystander effect might be less active when a child is present.

olation, thereby replicating the pattern of Balafoutas et al. (2014). In particular, the helping rate decreases by 16 percentage points in the Child condition and by 10 percentage points in the Alone condition. Still, the helping rate remains significantly higher in the Violation + Help treatment in the Child condition: 31 versus 16 percent ($\chi_1^2 = 6.26, p = 0.012$). In order to test Conjecture 3.3 formally, we examine the variable VH and the interaction term $VH \times Child$ in our regressions. In models (4) to (6) of Table 3.3 the coefficient estimates on VH suggest that the helping rate decreases significantly when the helping opportunity is preceded by a violation in the Alone and the Child condition. However, as exemplified by the insignificant coefficient on the interaction term, the helping rate does not decrease *more* in the Child condition as compared to the Alone condition, even though the negative coefficient estimate is in the predicted direction. Reassuringly, adding controls leaves the significance of our main variables of interest unchanged. In sum, we find no statistical evidence for Conjecture 3.3. This leads to Result 3.3:

Result 3.3 (Socializing Indirect Punishment) *Parents accompanying children are not significantly more likely to engage in indirect punishment.*

Figure 3.1a suggests that parents withhold help as a substitute for direct punishment. When a helping opportunity is presented to the subject, direct punishment rates decrease in both conditions (again, this is in line with Balafoutas et al., 2014), but substantially more so in the Child condition. What is more, direct punishment rates are even slightly lower in the Child condition. The drop in direct punishment from 22 percent in Violation to 5 percent in Violation + Help is significant at the 1%-level for the Child condition ($\chi_1^2 = 12.37, p < 0.001$). The drop from 11 to 7 percent in the Alone condition is insignificant ($\chi_1^2 = 0.98, p = 0.323$). This suggests that parents may indeed fear retaliation more when they are with their child, as they seem more eager to resort to indirect, rather than direct, punishment. Hence, because of this likely higher fear of retaliation, our result on direct punishment may underestimate parents' true tendency to punish more in the presence of the child.

3.5.2 Robustness Tests

The results presented above are robust to employing proportion tests instead of χ^2 -tests and to using logit models instead of linear probability models. A table containing marginal effect estimates and an accompanying discussion can be found in Appendix 3.B.3. Moreover, we show our results to be robust to a number of sample restrictions and alternative definitions. We summarize the results of this endeavor in Table

3.4, building upon the pooled linear probability models (2) and (5) of Table 3.3. Column (1) and (4) of Table 3.4 show that our results on punishment and helping, respectively, are unaffected by excluding targets who were identified as guardians and not as parents through the survey. In columns (2) and (5), we only include parents accompanying at most one child for whom helping could have been easier compared to parents accompanying several children. Notably, the coefficient on $VH \times Child$ becomes larger for Helping, but remains insignificant. Then, in columns (3) and (6), we discard observations for which a witness was recorded by the RA. Again, this does not change the previous results. Finally, in column (7), we recoded the Helping dummy so that a child encouraged by the parent to provide help is now coded as *no* Helping. Since this only happened in 8 instances (5 in Help and 3 in Violation + Help), it does not change the estimates substantially.

Table 3.4: Robustness Checks for Punishment (Left) and Helping (Right)

	PUNISHMENT			HELPING			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Child	0.13*** (0.05)	0.13** (0.06)	0.11** (0.04)	0.19*** (0.06)	0.18** (0.07)	0.19*** (0.06)	0.17*** (0.05)
VH	-0.05* (0.03)	-0.04 (0.03)	-0.04 (0.03)	-0.11** (0.04)	-0.10** (0.04)	-0.13** (0.06)	-0.10** (0.04)
VH \times Child	-0.14*** (0.05)	-0.13** (0.06)	-0.14*** (0.05)	-0.05 (0.08)	-0.11 (0.10)	-0.01 (0.08)	-0.05 (0.07)
Constant	0.14*** (0.05)	0.13** (0.05)	0.16*** (0.05)	0.39*** (0.05)	0.37*** (0.05)	0.37*** (0.06)	0.36*** (0.05)
Controls	✓	✓	✓	✓	✓	✓	✓
Observations	384	319	347	382	318	334	400
Clusters	30	30	30	30	30	30	30
R^2	0.086	0.069	0.067	0.127	0.110	0.130	0.097
F	4.635	2.776	3.919	6.146	7.175	7.410	5.610
df	29	29	29	29	29	29	29

Note: Dependent variable is a dummy for Punishment (columns (1) to (3)) or Helping (columns (4) to (7)). Standard errors in parentheses are clustered at the school level. Columns (1) and (4) present regression results when identified guardians are excluded from the analysis. In columns (2) and (5), we exclude all parents accompanying 2 or more children. Columns (3) and (6) include only observations for which no witness was recorded. Finally, in column (7) we code the child(ren) helping as *not* helping, rather than helping. These regressions control for the same variables as models (2) and (5) in Table 3.3, but the coefficients are omitted here for the sake of concision. One observation is dropped due to missing data on the target's gender.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Furthermore, we performed some exploratory regressions on heterogeneous treatment effects. Since we have not pre-registered any hypotheses related to these, we discuss the results of this exercise in length in Appendix 3.B.4. In particular, we

examine the effect of the number of children and their gender and age, the gender of the actor and parent, and the relative income in the school's IRIS area on helping and punishment rates. Finally, we show that the timing of the scene does not affect Punishment or Helping rates (see Appendix 3.B.5).

3.6 A Vignette Study for Norm Elicitation

Our results show that parents exhibit a higher tendency to enforce and comply with a social norm when in the presence of a child. We aim to distinguish between two competing explanations for these results by means of a norm-elicitation survey. On the one hand, the presence of the child motivates the parent to enforce and comply with the social norm more than when the parent would be on his or her own, in order to teach the child. On the other hand, the presence of the child could *change* the social norm in itself, meaning that littering becomes a more serious violation or helping a stranger in need more appropriate behavior. If this is the case, teaching may not be the only motive raising parents' tendency to punish. To study this, we conducted a vignette study eliciting the social appropriateness of the violation, direct punishment, *not* helping, and *not* helping after a violation (*i.e.*, indirect punishment) in the presence and in the absence of a child. This survey was conducted in two waves three months after the field experiment and it was not pre-registered.¹⁴

Accompanied by a supportive document from the regional school authority inspection, we sent a letter and a poster to the principals of all the public elementary schools in Lyon. In this letter, we asked them to send all parents in their school a link with an invitation to participate in an online survey, and to place the poster on the information boards next to the school entrance. We also contacted the principals through e-mail asking them to forward our attached message to the parents. Because of this approach, we do not know how many parents actually received or saw the invitation, leaving us in the dark about the response rate. In total, 506 parents responded to our survey. Admittedly, some of these parents may have been targeted during our field experiment. However, we have several arguments to ease this concern. First, with 601 and 506 parents involved in the experiment and vignette study, respectively, and 58000 school-going children in Lyon¹⁵, the probability that a parent participated in both is negligibly small, even if selection into either of the two is non-random and

¹⁴The idea for this vignette study was conceived when the authors attended the presentation of Erin Krupka at the 2019 edition of EWEBE in Lyon, which took place around the end of the field experiment. The delay between the field experiment and survey was caused by the Summer holiday break.

¹⁵<https://www.insee.fr/fr/statistiques/2011101?geo=COM-69123#chiffre-cle-1>

if parents have multiple school-going children. Second, the invitation to participate in the vignette study was sent to all 80 schools in Lyon, while the experiment was run at only 30 of those. Hence, the majority of invited parents had not been targeted in the experiment with certainty. Third, the vignette study was conducted after the Summer break, while the field experiment was run before it. Hence, even if a parent had already been targeted in the field experiment, of which (s)he was not aware, it is unlikely that a parent connected the dots and became suspicious as a result.

In the survey, the respondents had to read vignettes presenting all three treatments of our study in one of the two conditions, Alone or Child (see details in Appendix 3.A.2). The order of the Violation and Help vignettes was randomized, while the Violation + Help vignette was shown last. Thus, the scenes were varied within-subjects, while the condition was varied between-subjects. Since we aim to examine differences in social appropriateness across the Child and Alone conditions, we varied this dimension in a manner identical to the field experiment. At the same time, by varying the scenes within-respondents, we were able to obtain a larger number of observations for each scene. Respondents were asked to rate the social appropriateness of the described behavior on a six-point scale ranging from “very socially inappropriate” to “very socially appropriate” (a neutral option was omitted) and were told that they would have the chance to win a tablet if they chose the option that was chosen by the majority of other respondents. This way, respondents were incentivized to choose the option that they perceived as the social norm (Krupka and Weber, 2013).

Results are presented in Table 3.5. We coded the ratings as equidistant values on a range from -1 to 1, with the former indicating “very socially inappropriate” and the latter “very socially appropriate”. From Panel A, it seems that parents shown the Child condition deemed the littering violation slightly more inappropriate, with an average appropriateness rating of -0.83, as compared to the Alone condition with an average rating of -0.77. However, on further inspection, we discovered a potential confound in the description of the violation scenario that was not present in the actual treatments. In particular, the banana peel was made very salient and not said to be contained in a plastic bag, as was the case in the experiment. This may have raised the perceived risk of the child slipping on the banana peel and, as a result, increased the severity of the violation. We therefore adapted the description of the vignettes mid-way to exclude this confound. When we split the results according to the two different versions of the vignettes, we obtain the results in Panels B (“Scenario w/ Risk”) and C (“Scenario w/o Risk”). As can be seen, the significant difference is driven entirely by the first version of the vignettes, suggesting that the perceived risk for the child played an important role in the vignette study. In either case, parents did not deem

direct punishment more appropriate, which allows us to exclude the possibility that parents in our field study punished more in the presence of a child because they believed this is the social norm. Instead, they did so because they felt more inclined to enforce the same social norm. Similarly, not providing help to a stranger was deemed equally socially inappropriate in the presence and the absence of a child. We see a slight divergence for not providing help to a violator, but this difference fails to be significant. Hence, respondents do not believe that it is more appropriate to help a stranger in need in the presence of a child, as compared to being alone. Finally, note that in both conditions not helping a stranger was deemed significantly less inappropriate when this stranger littered before the helping opportunity presented itself (all respondents: $z_C = 10.113, p < 0.001$; $z_A = 9.742, p < 0.001$; Wilcoxon Signed-Rank tests).¹⁶

¹⁶We acknowledge that the respondents to our vignette study are probably not representative of the population of parents in Lyon, because they self-selected. These respondents may be more involved at school or in the education of their child, or care more about social norms. However, if this is the case, we suspect that these parents may be more sensitive to the presence of the child in the description of the different scenarios. Therefore, not finding significant differences across scenarios with and without the child in this population of respondents suggests that this might be a relatively robust finding. Moreover, we tested whether the parents' responses varied according to socio-demographic variables and the order of scenarios. These variables (age, income, number of children, location, and vignette order) are largely insignificant. There are two exceptions (both significant at the 5%-level): higher educated respondents deem punishment less appropriate (but they do not perceive the violation differently) and males deem not helping less inappropriate than females in the absence of the violation.

Table 3.5: Social Appropriateness of Passerby Behavior

	CHILD CONDITION					ALONE CONDITION					Rank-sum test (z)		
	Mean	-1	-0.6	-0.2	+0.6	+1	Mean	-1	-0.6	-0.2		+0.6	+1
Panel A: All respondents													
Violation	-0.83	0.63	0.32	0.04	0.00	0.00	-0.77	0.56	0.34	0.09	0.00	0.01	2.000**
Direct Punishment	0.52	0.01	0.02	0.06	0.23	0.43	0.51	0.01	0.03	0.04	0.25	0.42	0.388
No Help	-0.50	0.24	0.41	0.28	0.04	0.02	-0.52	0.20	0.45	0.30	0.04	0.00	0.158
No Help (Violation)	-0.13	0.08	0.24	0.31	0.22	0.09	-0.19	0.07	0.25	0.35	0.24	0.07	0.978
Observations	255												
NH vs. NH(V)	10.113***												
Panel B: Scenario w/ Risk													
Violation	-0.81	0.60	0.35	0.04	0.00	0.00	-0.74	0.48	0.41	0.10	0.00	0.01	2.340**
Direct Punishment	0.49	0.02	0.01	0.06	0.26	0.40	0.48	0.01	0.04	0.05	0.28	0.39	0.321
No Help	-0.51	0.24	0.41	0.30	0.03	0.02	-0.51	0.20	0.45	0.29	0.04	0.01	0.247
No Help (Violation)	-0.15	0.07	0.25	0.34	0.21	0.08	-0.21	0.07	0.29	0.35	0.23	0.05	0.935
Observations	156												
NH vs. NH(V)	5.807***												
Panel C: Scenario w/o Risk													
Violation	-0.86	0.68	0.28	0.03	0.00	0.00	-0.82	0.68	0.24	0.08	0.00	0.00	0.381
Direct Punishment	0.57	0.00	0.02	0.05	0.18	0.46	0.56	0.02	0.01	0.02	0.22	0.47	0.229
No Help	-0.48	0.24	0.40	0.25	0.05	0.03	-0.53	0.21	0.45	0.30	0.04	0.00	0.155
No Help (Violation)	-0.11	0.11	0.23	0.25	0.23	0.11	-0.18	0.09	0.21	0.35	0.27	0.09	0.564
Observations	95												
NH vs. NH(V)	8.320***												
Panel D: Scenario w/ Risk													
Violation	-0.86	0.68	0.28	0.03	0.00	0.00	-0.82	0.68	0.24	0.08	0.00	0.00	0.381
Direct Punishment	0.57	0.00	0.02	0.05	0.18	0.46	0.56	0.02	0.01	0.02	0.22	0.47	0.229
No Help	-0.48	0.24	0.40	0.25	0.05	0.03	-0.53	0.21	0.45	0.30	0.04	0.00	0.155
No Help (Violation)	-0.11	0.11	0.23	0.25	0.23	0.11	-0.18	0.09	0.21	0.35	0.27	0.09	0.564
Observations	102												
NH vs. NH(V)	7.236***												

Note: Social appropriateness of each action for the Child and Alone conditions are reported separately. Options range from Very Socially Inappropriate (-1) to Very Socially Appropriate (+1). Entries denote the share of respondents choosing the option belonging to the corresponding column. The final column displays the absolute value of the test statistics of Wilcoxon Rank-Sum (or, Mann-Whitney U) tests comparing the distributions under both conditions. The rows "NH vs. NH(V)" display the test statistics of Wilcoxon Signed-Rank tests examining whether the social appropriateness of not helping is higher following a norm violation. All tests are two-sided.

** $p < 0.05$, *** $p < 0.01$

3.7 Concluding Discussion

Our study provides evidence of parents' involvement in childhood education of social norms through the enforcement of such norms. By conducting a field experiment in the vicinity of French elementary schools, we have shown that parents are more willing to enforce the social norm of non-littering in the presence of a child. Parents accompanying one or more children, as compared to parents alone, are more likely to engage in direct punishment of the norm violator through verbal confrontation. These parents also punish indirectly through withholding help, but they are not more likely to do so than when they are alone. We argue that indirect punishment may be too subtle a form of norm enforcement. Indeed, not helping someone in need, be that person a norm violator, also violates another social norm: helping others. As a result, parents may believe that the educative signal of indirect punishment is unlikely to be grasped by children and that it does more harm than good: children fail to learn that helping others constitutes a valuable social norm. In line with this, we find that parents accompanying children are more likely to help a stranger in need than when alone, even when this stranger violated before. This suggests that parents prioritize teaching their children to help a stranger in need. Although not the initial aim of this study, we deem the increased willingness of parents to help a stranger in need in front of their child(ren) an important contribution of this study, which further solidifies the importance of modeling desired behavior to children in the social learning process.

By showing evidence of such inter-generational teaching of norms, these results contribute to our understanding of the parental involvement in the maintenance of social norms in the society. We are indeed able to reject explanations of the observed behavior alternative to teaching. First, our norm elicitation vignette study indicates that the presence of the child changes neither the parents' perception of the social norm violation, nor their perception of the appropriateness of punishment and helping. Instead, it strengthens parents' motivation to enforce and comply with the same social norm, which is consistent with an increased willingness to teach.

Second, we argue that our results are not confounded by social image concerns. The presence of witnesses could indeed put a pressure on the parents to enforce the norm more, especially in the presence of their child. But in fact, we find a negative and usually insignificant coefficient on the Witness dummy throughout specifications. Thus, if anything, a regular bystander effect leads to lower helping rate, which is opposite to a social image effect. It remains that it would be interesting to investigate how the same parents would behave if accompanied by another adult instead of their child. This is not the purpose of this study but it would be interesting to explore and

it would further assess the validity of our interpretation as inter-generational teaching of norms. Social image could also matter for parents in the eyes of their own child. However, if parents were more likely to confront verbally the violator in order to show their strength to their child, they should not substitute direct punishment for indirect punishment when this option is available. On the other hand, we acknowledge that parents may not punish in the Violation + Help scene because they do not want to be seen “kicking someone when (s)he is down already.” Still, we found no evidence that image concerns *do* drive parents’ behavior.

The aforementioned substitution effect allows us to reject a third possible confound for Result 3.1: parents might enforce more the littering norm in the presence of their child because they fear retaliation less. In fact, we show that direct punishment significantly decreases in the Violation + Help treatment *only* for the Child condition. Moreover, the presence of the child raises punishment rates for fathers *only*. These two facts together suggest the opposite, if anything: the presence of the child increases the fear of retaliation. Hence, we may be under-estimating the effect of the child on norm enforcement. Taken together, we interpret our results as lending support to the notion that parents punish more in order to teach their children the importance of norm compliance and enforcement.

With this study we contribute to the understanding of the inter-generational transmission of norms from parents to children. By focusing on normative preferences and by assessing parents’ enforcement behavior in a natural setting, we complement studies showing the importance of the cultural transmission of preferences (Bisin and Verdier, 2001) and how preferences evolve during childhood. Economists have only recently started to study empirically parental socialization efforts in this field (Berner et al., 2017; Houser et al., 2016; Cappelen et al., 2020; Sutter and Untertrifaller, 2020). In line with the results from these studies, we show that parents exhibit more socially responsible behavior in the presence of their child. In addition to this, we show that parents not only teach through modeling said behavior, but also that they teach through punishing socially irresponsible behavior. This teaching motive may change the nature of punishment of norm violations: such punishment may no longer be completely altruistic if a future benefit is expected, such as the transmission of values to one’s offspring. This expected benefit may compensate the cost associated with the threat of retaliation.

Incidentally, our study may also contribute to the understanding of the heterogeneity of preferences across social groups. Indeed, we observe that not all parents use the observed norm violation as an opportunity to teach or remind the importance of norm compliance to their children. Some do, but the majority of parents in

our study do not intervene. Our secondary analyses (see Appendix 3.B.4) show that the diversity of parents' reactions to the norm violation cannot be explained by the socio-economic status of the district, the age or gender of the child. This calls for further investigations of the heterogeneity in the inter-generational transmission of values and norms, by examining the individual and institutional determinants of the degree of involvement of parents in teaching normative preferences to their children. For example, our study was conducted in a very anonymous and clean urban setting; would teaching be more likely in a less anonymous environment, and less likely in a less well maintained neighborhood? Our study already suggests that heterogeneity in teaching, and not necessarily only in parents' preferences, may play a role in producing diversity in the formation of normative preferences during childhood. Other interesting extensions of our study would be to connect teaching and learning, to test whether children who have just been taught through example by an adult are themselves more likely to exhibit stronger normative preferences, and whether parents and other adults' teaching make a difference in such endeavor.

Appendices

3.A Experimental Material

3.A.1 Instructions of the Experiment

[Translated from French]

3.A.1.1 Instructions for the RAs and the Actor

You are helping us to collect data for an ongoing research project. None of you are aware of the research goal and topic. We work in research teams of four: two RAs, one actor and one supervisor (who is one of the researchers).

We are asking you to stage a number of scenes in the streets around different elementary schools in Lyon. We are interested in the response of the witness of the scene. This witness is an unaware passerby who is targeted by you. We want you to target two types of witnesses: a parent accompanied by one or multiple children, and single adults (who can also be presumed to be parents). You play these scenes in the morning and the afternoon. In each of these time slots you should target both types of adults. This basically means that you target parents with a child going to school, parents leaving school without a child, parents approaching school without their child, and parents leaving school with their child. You should aim for roughly equal numbers in each of these categories.

There are three different scenes to be played, which are further explained below. Below, you find separate instructions for the actor and the RA. Make sure to read each other's instructions, such that you both know what we expect from each of you.

3.A.1.2 Instructions for the Actor

Materials

- Plain clothes
- Cotton shoulder/shopping bag
- Plastic bag with a banana peel inside
- 7 folders and binders
- Colored pens and markers

- Two tablets for RAs

Before the scene is played, together with the RAs and supervisor, you determine the location of the scene. Make sure not to be too close to the school entrance, in a street that is not too busy. Your first task is to identify your target. You should target either a single adult or a single adult accompanied by children. Do not target adults with a stroller, a bike, a dog, parents of a disabled child, or parents who are holding their children's hands with both hands. Also make sure to avoid parents visibly in a rush or talking on the phone.

For each targeted adult, you play one of three scenes. Before the scenes, make sure to have one handle of the bag on your shoulder and the other loosely hanging down; this makes it easier to reach into your bag. Below is a detailed script of the scenes.

Scene 1: Violation + Help

1. Actor and supervisor identify targeted parent fulfilling the criteria.
2. Actor approaches target from the front.
3. As the actor is roughly 10 meters away, (s)he pauses and pretends to be searching for something in his/her bag.
4. As the target is 5 meters away: the actor throws away the plastic bag with the banana peel inside. The actor makes sure that no one is approaching from behind.
5. Actor takes out all file folders from the bag and starts moving again. As (s)he continues to walk, the actor accidentally drops the entire content.
6. Actor stops walking, reacts visibly upset, stares at dropped items in dismay. This provides the parent with an opportunity to help.
7. Actor starts picking up items as targeted parent passes him/her.
8. Scene ends; RA and Actor record information and clean up scene.

Scene 2: Violation

1. Actor and supervisor identify targeted parent fulfilling the criteria.
2. Actor approaches target from the front.

3. As the actor is roughly 10 meters away, (s)he pauses and pretends to be searching for something in his/her bag.
4. As the target is 5 meters away: the actor throws away the plastic bag with the banana peel inside. The actor makes sure that no one is approaching from behind.
5. Actor starts moving again, but, before the target has reached him/her, then pauses again, going through the bag again as the parent passes.
6. Scene ends; RA and Actor record information and clean up scene.

Scene 3: Help

1. Actor and supervisor identify targeted parent fulfilling the criteria.
2. Actor approaches target from the front.
3. As the actor is roughly 10 meters away, (s)he pauses and pretends to be searching for something in his/her bag.
4. Actor takes out all file folders from the bag and starts moving again. As (s)he continues to walk, the actor accidentally drops the entire content.
5. Actor stops walking, reacts visibly upset, stares at dropped items in dismay. This provides the parent with an opportunity to help.
6. Actor starts picking up items as targeted parent passes him/her.
7. Scene ends; RA and Actor record information and clean up scene.

After the end of each of the scenes, you leave the location in a direction different from that of the targeted adult. In case that the adult confronts you about throwing away the plastic bag and/or demands you clean it up (in scenes 1 and 2), you quietly comply and pick up the plastic bag. If the parent does not respond, the RA makes sure to clean up after the parent has left the scene. After the scene, you meet with the RA and report the following pieces of information:

- Type of scene played
- Reaction of the parent (multiple could apply):
 - Punishment: the parent explicitly addresses you regarding the littering and/or demands you to clean it up.

- Help: the parent picks up at least one of the dropped item from the floor.
- Address child: the parent talks to the child about the violation in a way that is audible for you.
- “Other” circumstances

3.A.1.3 Instructions for Research Assistant 1

Each team includes two RAs, each with different tasks. RA1 observes the scene played closely and notes down the following characteristics:

- ID: School code + number of observation, e.g., GT11 for the 11th observation at Germaine Tillion.
- Setting: what are the circumstances in which the scene is played?
 - Time of day
 - Witnesses: are there any other people around that could possibly intervene in the scene?
 - Weather: sunny, cloudy, or rainy; cold, mild, hot; windy?
 - Direction: from or to school
 - Cleanliness of environment: scale of 1 (dirty) to 5 (clean)
- Treatment: confirm this with the actor after the scene.
 - Condition: child(ren) or alone
 - Type of scene: Violation+Help, Violation, Help
 - In case of littering: did the target see the plastic bag being thrown away?
- Reaction of the parent: confirm this with the actor after the scene.
 - Punishment: does the parent confront the actor by directly addressing him/her about the violation?
 - Help: does the parent help by picking up at least one item?
 - Address child: does the parent talk to the child about the violation?

It is important that RA1 does not stay too close to the parent, because this may contaminate the outcome of the scene. After the end of the scene, RA1 verifies the scene played with the actor and checks whether the parent said something to the actor.

3.A.1.4 Instructions for Research Assistant 2

The task of RA2 is to approach the target after the scene for a seemingly unrelated survey. You tell the target the following:

[Translated from French]

“Good day sir/madam, I am a Master student in Psychology at the University of Lyon 2 and, as part of my courses, I am conducting a survey on the quality of the environment around schools. The survey comprises 5 questions and takes 2 minutes. Could I take some of your time to respond to my questions?”

1. We are close to the elementary school [name of school]. How would you evaluate the quality of the air around this school on a scale of 1 (for a very poor quality) to 7 (for an excellent quality)?
2. Do you think that the circulation of cars should be forbidden in the streets in front of schools to limit the exposure to pollution for children?
3. Today, are you accompanying or have you accompanied your child / a child that you guard / a child of one of your relatives / or no child to this school?
4. If so, what is the age and gender of this child / these children?
5. Finally, do you take the car to arrive at school?

You should also note down a number of characteristics regarding the parent’s appearance. The main purpose for this is to ensure that no parent is targeted twice.

- ID: School code + number of observation, e.g., GT11 for the 11th observation at Germaine Tillion.
- Gender: male or female
- Estimated age
- Ethnicity: caucasian, Arab, African, Asian, other (Indian, South-American)
- Religious signs
- Estimated height
- Build: lean, medium, overweight, obese.
- Hair colour: blond, light brown, dark brown, black, red, gray, other

- Hair style: bold, short, medium, long, curly, straight, ponytail, afro.
- Facial hair: none, moustache, short beard, long beard
- Colour of outer garment (coat, vest, etc.)
- Other: hat, glasses, tattoos, piercings, birth marks, scars, etc.

3.A.2 Instructions of the Vignette Study

[Translated from French]

3.A.2.1 Participant Information Statement

1. What does the study involve? This study involves a very brief questionnaire.
2. Who is carrying out the study? The study is being conducted by professors Fabio Galeotti and Marie Claire Villeval from CNRS and the University of Lyon, and Thijs Brouwer from Tilburg University.
3. How much time will the study take? Answering this questionnaire will take approximately 4 minutes to complete.
4. Can I withdraw from the study? Participating in this questionnaire is completely voluntary. If you do consent, you can withdraw at any time during the questionnaire. Withdrawal from the questionnaire means that you renounce to the chance of winning an electronic tablet, but it will not affect your relationship with the researchers or staff at the CNRS, the University of Lyon or Tilburg University.
5. Will anyone else know the results? All aspects of the questionnaire will be confidential and only the researchers will have access to the responses. A report of the study may be submitted for publication, but all information will only be used in an aggregated form, no personal information will be made public.
6. Will the study benefit me? Responding to the questionnaire will not lead to any payment. However, it will be proposed to the participants to enter a lottery in which one participant will be randomly selected to earn an electronic tablet.

7. Can I tell other people about the study? The researchers request, that for the purpose of maintaining study integrity, you do not share with anybody the nature of the questionnaire.
8. What if I require further information about the study or my involvement in it? If you have specific questions regarding the study, please feel free to contact Marie Claire Villeval by email at villeval@gate.cnrs.fr

3.A.2.2 Scenarios

Below, you will read three short scenarios. In each of the scenarios, you are asked to evaluate the described behavior, choosing between six options ranging from “Very Socially Inappropriate” to “Very Socially Appropriate”. By “socially appropriate” we mean a behavior judged correct and ethical by the majority of people. The objective is to choose, for each scenario, the most common option selected by all other respondents to this questionnaire (all parents with at least one child registered in an elementary school in Lyon).

If you are randomly selected at the end of the study, you will win an electronic tablet (model iPad 32 Go) if your response to one randomly selected question in these scenarios matches the most common response given by all other respondents to the same question. For example, if the most common answer is “Very Socially Inappropriate”, you would receive the tablet if you also answered “Very Socially Inappropriate”. If the most common answer is “Very Socially Appropriate”, you would receive the tablet if you also answered “Very Socially Appropriate”. You will be informed by email if you have won the electronic tablet after all responses have been collected. Please press “Next” to continue.

Vignette 1: Littering + Child /Alone/

A passerby is walking on the street in proximity of an elementary school. This passerby carelessly throws a plastic bag containing food waste on the sidewalk at the sight of a parent with a 6-year-old child / *a parent who has just dropped his/her child at the school* / and no one else around. How would you evaluate the behavior of the passerby? If you give the same response as the majority of the other respondents, you may win a tablet.¹⁷

¹⁷The first version of the vignette emphasized the banana peel more. The introduction of the vignette read: “A passerby is walking on the street in proximity of an elementary school *while eating a banana*. This passerby carelessly throws *the banana peel* on the sidewalk...” We decided to change this because it did not portray the scene accurately and because the perceived risk of

Please choose one option below:

- Very Socially Inappropriate
- Socially Inappropriate
- Somewhat Socially Inappropriate
- Somewhat Socially Appropriate
- Socially Appropriate
- Very Socially Appropriate

Please press “Next” to continue.

The parent addresses the passerby and asks this passerby to pick up the plastic bag. How would you evaluate the behavior of the parent? If you give the same response as the majority of the other respondents, you may win a tablet.

Please choose one option below:

- Very Socially Inappropriate
- Socially Inappropriate
- Somewhat Socially Inappropriate
- Somewhat Socially Appropriate
- Socially Appropriate
- Very Socially Appropriate

Please press “Next” to continue.

Vignette 2: Help + Child /Alone/

A passerby is walking on the street in proximity of an elementary school, while carrying a bag containing folders. The passerby accidentally drops all the folders on the ground at the sight of a parent with a 6-year-old child */a parent who has just dropped his/her child from school/* and no one else around. The parent **does not** go to help

slipping might confound the parents' perception of the severity of the violation.

the passerby with picking up the folders.

How would you evaluate the behavior of the parent? If you give the same response as the majority of the other respondents, you may win a tablet.

Please choose one option below:

- Very Socially Inappropriate
- Socially Inappropriate
- Somewhat Socially Inappropriate
- Somewhat Socially Appropriate
- Socially Appropriate
- Very Socially Appropriate

Please press “Next” to continue.

Vignette 3: Help + Littering + Child /*Alone*/

A passerby is walking on the street in proximity of an elementary school, while carrying a bag containing folders. This passerby carelessly throws a plastic bag containing food waste on the sidewalk at the sight of a parent with a 6-year-old child /*a parent who has just dropped his/her child from the school/* and no one else around. Few instants afterwards, this passerby accidentally drops all his/her folders on the ground. The parent **does not** go to help the passerby with picking up the folders.

How would you evaluate the behavior of the parent? If you give the same response as the majority of the other respondents, you may win a tablet.

Please choose one option below:

- Very Socially Inappropriate
- Socially Inappropriate
- Somewhat Socially Inappropriate
- Somewhat Socially Appropriate
- Socially Appropriate

- Very Socially Appropriate

Please press “Next” to continue.

Before we finish, we would like to ask you a few questions about yourself.

- What is your gender?
 Male Female
- What is your highest educational degree obtained?
 Primary school Less than high school High school diploma or equivalent
 Undergraduate degree Post-graduate degree
- What year were you born (*e.g.*, 1970)?
- How many children do you have?
 0 1 2 3 or more
- What is their gender? How many sons: ---- How many daughters: ----
- What is their age? Your son(s): ---- Your daughter(s): ----
- What is your household monthly earnings category:
 < 2000 Euro 2000-3999 Euro 4000-5999 Euro 6000 Euro and more
- If you live in Lyon, what is your district?
 1 2 3 4 5 6 7 8 9 I don't live in Lyon

Please press “Next” to continue.

Earnings

You may win an electronic tablet if you are randomly selected among all the respondents at the end of our study, and if your response in one randomly selected scenario matches the most common response given by the other respondents. If you are willing to participate in this lottery, please enter your email address below so that we can contact you if you have won the tablet.

Thank you for taking time out of your busy life to participate to this study. If you have any questions concerning this study, you can contact us at villeval@gate.cnrs.fr

3.A.3 Additional Material

Figure 3.2: Materials Used in the Experiment and Scenes



3.B Supplementary Analyses

3.B.1 Summary Statistics of Survey Respondents

Table 3.6 presents summary statistics on survey participants. There are no clear significant differences between the Child and Alone conditions regarding the number of children and their age.

Table 3.6: Summary Statistics of Survey Respondents

	(1) All	(2) Alone	(3) Child	(4) Δ
Own Child	0.88 (0.33)	0.91 (0.29)	0.85 (0.36)	0.06 (0.05)
Son	0.57 (0.50)	0.54 (0.50)	0.59 (0.49)	-0.05 (0.07)
Daughter	0.59 (0.49)	0.55 (0.50)	0.63 (0.49)	-0.07 (0.07)
No. of Children	1.32 (0.72)	1.22 (0.69)	1.40 (0.73)	-0.18* (0.09)
Child Age = 3	0.06 (0.23)	0.06 (0.23)	0.05 (0.23)	0.00 (0.03)
Child Age = 4	0.15 (0.35)	0.12 (0.33)	0.16 (0.37)	-0.04 (0.05)
Child Age = 5	0.23 (0.42)	0.22 (0.42)	0.24 (0.43)	-0.02 (0.06)
Child Age = 6	0.17 (0.38)	0.15 (0.36)	0.19 (0.39)	-0.03 (0.05)
Child Age = 7	0.20 (0.40)	0.17 (0.38)	0.22 (0.42)	-0.05 (0.05)
Child Age = 8	0.19 (0.39)	0.21 (0.41)	0.18 (0.38)	0.03 (0.05)
Child Age = 9	0.17 (0.38)	0.12 (0.33)	0.21 (0.41)	-0.09* (0.05)
Child Age = 10	0.07 (0.26)	0.10 (0.29)	0.05 (0.23)	0.04 (0.03)
Child Age = 11	0.03 (0.16)	0.03 (0.17)	0.02 (0.15)	0.01 (0.02)
Observations	234	105	129	234

Note: For gender and age, totals are not equal to 1 because some parents reported having more than one child at the school.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.B.2 Effect of Parent's Gender on Punishment in Violation

To explore the role of the parent's gender on punishment behavior, Table 3.7 shows an additional analysis of the punishment rate in the Violation treatment only. When the *Child* \times *Male Target* interaction is added, *Child* alone is no longer significant. At the same time, *Child* and *Child* \times *Male Target* are jointly significant ($p = 0.036$). This means that only fathers punish significantly more in the presence of the Child. We take this as indirect evidence that the presence of the child raises the fear of retaliation, but that fathers experience this fear less. This is strengthened by the fact that, in the absence of the Child, fathers only punish insignificantly more, indicating that the fear of retaliation is much more similar between mothers and fathers when the Child is not around.

Table 3.7: Punishment in the Violation Treatment

	(1)	(2)
Child	0.11** (0.04)	0.08 (0.06)
Male Target	0.09** (0.04)	0.05 (0.06)
Child \times Male Target		0.09 (0.15)
Male Actor	-0.08 (0.05)	-0.08 (0.05)
Morning	-0.02 (0.05)	-0.02 (0.05)
Witness	-0.11 (0.08)	-0.10 (0.07)
Rain	-0.00 (0.05)	-0.00 (0.05)
Hot	-0.07 (0.06)	-0.07 (0.06)
Constant	0.15** (0.07)	0.17** (0.08)
Observations	199	199
R^2	0.060	0.063
F	3.684	4.377

Note: Standard errors are clustered on the School level. Only the Violation treatment is included. Male Target and Child \times Male Target are jointly significant ($F(2, 28) = 3.84, p = 0.036$).

** $p < 0.05$

3.B.3 Alternative Estimation Models

Instead of linear probability models, we also estimated logit models. While such models are more suited to analyze binary choice data like ours, they are less suited to study interaction effects. To this end, we took the following approach. We estimated the logit model including the controls and the interaction term. We then estimated marginal effects at $Child = 1$ and $Child = 0$ and used a contrast test to determine whether the marginal effects of VH are significantly different between the conditions. These results are presented in Table 3.8. As can be seen, the marginal effect of VH differs regarding Punishment, but not for Helping.

Table 3.8: Marginal Effects of Logit Estimations

	PUNISHMENT		HELPING	
	Alone (1)	Child (2)	Alone (3)	Child (4)
Child	0.12*** (0.04)	0.12*** (0.04)	0.22*** (0.05)	0.22*** (0.05)
VH	-0.04 (0.03)	-0.16*** (0.04)	-0.10** (0.04)	-0.17** (0.07)
Male Target	0.06** (0.03)	0.12*** (0.04)	0.03 (0.04)	0.04 (0.05)
Male Actor	-0.07 (0.05)	-0.13* (0.07)	-0.21*** (0.04)	-0.27*** (0.05)
Morning	-0.04 (0.03)	-0.08 (0.05)	0.00 (0.04)	0.00 (0.05)
Witness	-0.05 (0.05)	-0.08 (0.08)	-0.13* (0.07)	-0.16* (0.10)
Rich Area	0.03 (0.04)	0.05 (0.07)	0.01 (0.05)	0.02 (0.06)
Rain	0.00 (0.03)	0.00 (0.06)	-0.11 (0.09)	-0.14 (0.11)
Hot	-0.04 (0.05)	-0.07 (0.08)	0.05 (0.05)	0.06 (0.06)
VH _C vs. VH _A	9.88***		0.76	
Observations	399		400	
Clusters	30		30	
(Pseudo) R^2	0.094		0.113	
Wald χ^2	51.98		52.70	

Note: The table contains four sets of marginal effects resulting from two logit estimations: one with Punishment as the dependent variable (columns 1 and 2) and one with Helping as the dependent variable (columns 3 and 4). For each estimation, marginal effects are estimated for Child = 0 and Child = 1. Delta-method standard errors are reported in parentheses. The row “VH_C vs. VH_A” displays the χ^2 -test statistic of a contrast test against the null that the coefficients on VH are the same in the two conditions.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.B.4 Analysis of Heterogeneous Effects

In this part of the Appendix, we report additional regressions in which we allow the presence of the child to have heterogeneous effects with regards to different characteristics of the child, the parent, or the neighborhood. The results of this endeavor

are depicted in Table 3.9 for Punishment and Table 3.10 for Helping. We discuss the tables jointly, as each column in the two tables corresponds to the *same* exercise. Column (1) contains the baseline estimates, as reported in models (2) and (5) in Table 3.3 in the main text with controls included.

Number of Children First, we look at the importance of the number of children accompanying the parent. In column (2) in both tables, we discriminate between parents accompanying one child, two children, and three or more children. It should be noted that only 17 out of 301 parents in the Child condition accompany three or more children (2 in Violation, 9 in Help, 6 in Violation + Help treatments). Parents who accompany one or two children are significantly more likely to punish the actor than parents alone. However, parents accompanying three or more children punish significantly less than parents alone (at the 1% level). This may suggest that parents accompanying three or more children are too occupied paying attention to the children to engage in punishment. A slightly different picture arises when looking at Helping. The helping rate is significantly higher for all numbers of children. However, we see that only parents accompanying one child withhold helping substantially (by 12 pp), even though the coefficient enters insignificantly.

Child's Gender Next, we investigate whether the child's gender matters in the parent's reaction (see column (3) in both tables). To allow for a clean comparison, only single-child observations are classified according to gender. In total, 27 single-child observations have missing gender of the child and are omitted from the analysis. The results show the presence of one girl raises the punishment rate by 19 percentage points, while the presence of one boy raises it by 13 points. This coefficient is borderline significant for the presence of one girl, only. Similarly, the presence of one girl raises the helping rate by 18 percentage points, while the presence of one boy raises it by 17 points. The coefficient is borderline significant for both. For both genders, the additional drop in Helping is insignificant and of similar magnitude. Taken together, parents' educative motive is not really stronger with daughters than with sons.

Child's Age Then, we explore the effect of the age of the child (see column (4)). Again, we do this by focusing on single-child observations for the cleanest comparison. We created dummies for one child aged 5 or younger, one child aged between and including 6 and 8, and one child aged 9 or older. The values of this variable are based either on the parent's response in the survey or on the guess of the research assistants. The results show that the increase in Punishment in the presence of the

child seems to be driven by the middle age category, as the increase in punishment rates is significant only for parents accompanying children aged between 6 and 8 at a magnitude of 20 percentage points. For the youngest category, the coefficient is similar in magnitude to the baseline estimate, but insignificant, while the coefficient is close to zero for the oldest category. A somewhat different picture arises for helping, as parents accompanying the youngest class of children respond strongest to the presence of the child by increasing the helping rate by 27 percentage points (significant at 5%-level). Parents of older children increase their helping rate by less as compared to parents alone, and these increases are insignificant. Moreover, the additional drop in helping rate is significant for the youngest class of children, only. It thus seems that results regarding Helping are driven by the youngest age category.

Targeted Parent and Actor Gender Fathers and mothers may react differently and they may also react to the gender of the actor in the presence of the child. In column (5) and (6) we look at the effects of the gender of the targeted parent. Most importantly, we see no interaction effects between the gender of the parent or actor and the presence of the child, as indicated by the insignificant coefficients on the interaction terms. Interestingly, the coefficient on *Child* now enters insignificantly in the Punishment regression as it now refers to the presence of the child with a female target and the female actor. Furthermore, the interaction term *Male Target* \times *Male Actor* measures the effect of two males interacting. This does not seem to affect outcomes significantly. More generally, most regressions show that punishment does not differ according to the actor's gender, while the male actor receives significantly less help than the actress regardless of the specification.

Income effects Finally, we would like to know whether income influences parents' punishment and helping. To this end, in column (7) we interact the Rich IRIS dummy with the *Child* dummy. We find no effects of this interaction term for both outcomes. To dive deeper into this, in column (8) we classify the IRIS area in which the school is located as Low, Medium Low, Medium High, or High based on the median disposable income. The results show that parents in the highest three income classes punish significantly more than parents in the lowest income class. However, parents do not increase their punishment by more in these neighborhoods in the presence of the child. Regarding helping rates, we do not find effects of income on the parents' tendency to provide help. Additional regressions in which we include the poverty rate in the IRIS area and an interaction term with the *Child* variable lead to similar conclusions. Parents are significantly less likely to punish in poorer areas

(at the 5% level) but do not help less, and the effect of the presence of the child does not differ significantly with the poverty rate (regressions available upon request).

3.B.5 Analysis of Timing Effects

Throughout the experiment, parents arrive at different times at school. It may be that parents who arrive early are different from parents arriving later, either because they are less in a rush or because they are intrinsically different. Similarly, parents who leave the school premises late may be different from parents leaving the premises as soon as possible. If this is the case, this may affect their punishment or helping behavior. In order to test whether this is the case, we take three approaches and report the results in Table 3.11. The first approach uses the timing of the scene in minutes relative to the beginning or end of the school day (8:30 AM, 4:45 PM, or 5:30 PM). The second approach rounds the previous time variable to the nearest ten, in order to discretize the support. The third approach uses the observation number within a session (morning or afternoon) and condition (*i.e.*, Child or Alone). Table 3.11 shows no significant effects of the timing of the scene regardless of the approach retained.

Table 3.9: Secondary Analyses of Punishment Behavior

	PUNISHMENT							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Child	0.12** (0.04)				0.07 (0.05)	0.12** (0.04)	0.15*** (0.05)	0.09*** (0.02)
1 Child		0.13* (0.06)						
1 Boy			0.13 (0.12)					
1 Girl			0.19* (0.09)					
1 Child Age ≤ 5				0.09 (0.14)				
5 < 1 Child Age ≤ 8				0.20** (0.09)				
1 Child Age > 8				0.01 (0.10)				
2 Children		0.11* (0.06)	0.11* (0.06)	0.17*** (0.06)				
3+ Children		-0.14*** (0.05)	-0.16*** (0.05)	-0.17*** (0.06)				
VH	-0.04 (0.03)	-0.04 (0.03)	-0.04 (0.03)	-0.04 (0.03)	-0.04 (0.03)	-0.04 (0.03)	-0.04 (0.03)	-0.05* (0.03)
VH × Child	-0.12*** (0.04)				-0.12*** (0.04)	-0.12*** (0.04)	-0.12** (0.05)	-0.12** (0.05)
VH × 1 Child		-0.13** (0.06)						
VH × 1 Boy			-0.10 (0.13)					
VH × 1 Girl			-0.24** (0.09)					
VH × Child Age ≤ 5				-0.13 (0.14)				
VH × 5 < Child Age ≤ 8				-0.17 (0.10)				
VH × Child Age > 8				-0.07 (0.11)				
VH × 2 Children		-0.11 (0.08)	-0.12 (0.08)	-0.16* (0.08)				
VH × 3+ Children		0.09 (0.06)	0.10 (0.06)	0.11 (0.07)				
Male Target	0.06** (0.03)	0.07** (0.03)	0.06** (0.03)	0.07** (0.03)	0.05 (0.04)	0.08* (0.05)	0.07** (0.03)	0.07** (0.03)
Child × Male Target					0.02 (0.08)			
Male Actor	-0.06 (0.04)	-0.06 (0.04)	-0.07 (0.04)	-0.07 (0.05)	-0.10** (0.05)	-0.05 (0.04)	-0.06 (0.04)	-0.03 (0.03)
Child × Male Actor					0.07 (0.06)			
Male Actor × Male Target						-0.03 (0.05)		
Rich IRIS	0.02 (0.04)	0.03 (0.04)	0.04 (0.04)	0.04 (0.04)	0.02 (0.04)	0.03 (0.04)	0.06 (0.05)	
Rich IRIS × Child							-0.07 (0.06)	
Medium Low Income								0.14*** (0.04)
Medium High Income								0.08** (0.04)
High Income								0.17** (0.06)
Medium Low Income × Child								0.09 (0.07)
Medium High Income × Child								0.04 (0.03)
High Income × Child								-0.03 (0.07)
Constant	0.14*** (0.05)	0.14*** (0.05)	0.13*** (0.05)	0.13*** (0.04)	0.16*** (0.05)	0.13*** (0.04)	0.12** (0.04)	0.03 (0.03)
Observations	399	399	381	381	399	399	399	399
Clusters	30	30	30	30	30	30	30	30
R ²	0.07	0.07	0.09	0.09	0.07	0.07	0.07	0.10
F	5.17	9.87	11.35	8.84	5.41	4.66	5.12	13.36
df	29	29	29	29	29	29	29	29

Note: Standard errors in parentheses are clustered at the School level. Morning, Witness, Rain, and Hot are included in the regressions but omitted from the table for space-saving reasons.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.10: Secondary Analyses of Helping Behavior

	HELPING							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Child	0.22*** (0.06)				0.28*** (0.08)	0.22*** (0.06)	0.16* (0.09)	0.22** (0.09)
1 Child		0.19** (0.07)						
1 Boy			0.17* (0.08)					
1 Girl			0.18* (0.10)					
1 Child Age ≤ 5				0.27** (0.11)				
5 < 1 Child Age ≤ 8				0.11 (0.12)				
1 Child Age > 8				0.07 (0.15)				
2 Children		0.27*** (0.09)	0.28*** (0.09)	0.20** (0.09)				
3+ Children		0.26** (0.10)	0.27** (0.10)	0.24** (0.09)				
VH	-0.10** (0.04)	-0.10** (0.04)	-0.10** (0.04)	-0.10** (0.04)	-0.11** (0.04)	-0.10** (0.04)	-0.10** (0.04)	-0.11** (0.04)
VH × Child	-0.07 (0.08)				-0.06 (0.08)	-0.07 (0.08)	-0.08 (0.07)	-0.07 (0.08)
VH × 1 Child		-0.12 (0.10)						
VH × 1 Boy			-0.15 (0.12)					
VH × 1 Girl			-0.10 (0.11)					
VH × Child Age ≤ 5				-0.31* (0.17)				
VH × 5 < Child Age ≤ 8				-0.06 (0.16)				
VH × Child Age > 8				0.02 (0.17)				
VH × 2 Children		-0.02 (0.14)	-0.03 (0.14)	0.04 (0.16)				
VH × 3+ Children		0.03 (0.24)	0.02 (0.24)	0.05 (0.23)				
Male Target	0.03 (0.04)	0.02 (0.04)	0.02 (0.04)	0.02 (0.04)	0.05 (0.06)	0.05 (0.06)	0.03 (0.04)	0.03 (0.04)
Child × Male Target					-0.06 (0.11)			
Male Actor	-0.22*** (0.05)	-0.23*** (0.05)	-0.22*** (0.05)	-0.22*** (0.05)	-0.19*** (0.06)	-0.21*** (0.05)	-0.22*** (0.05)	-0.20*** (0.05)
Child × Male Actor					-0.07 (0.09)			
Male Actor × Male Target						-0.05 (0.07)		
Rich IRIS	0.01 (0.05)	0.02 (0.05)	0.02 (0.05)	0.03 (0.05)	0.01 (0.05)	0.01 (0.05)	-0.06 (0.06)	
Rich IRIS × Child							0.14 (0.09)	
Medium Low Income								0.08 (0.10)
Medium High Income								-0.00 (0.10)
High Income								0.08 (0.07)
Medium Low Income × Child								-0.02 (0.15)
Medium High Income × Child								0.02 (0.13)
High Income × Child								0.02 (0.09)
Constant	0.36*** (0.05)	0.36*** (0.05)	0.36*** (0.05)	0.35*** (0.04)	0.34*** (0.06)	0.36*** (0.05)	0.40*** (0.06)	0.32*** (0.08)
Observations	400	400	385	385	400	400	400	400
Clusters	30	30	30	30	30	30	30	30
R ²	0.13	0.14	0.14	0.15	0.13	0.13	0.14	0.14
F	6.45	4.81	4.61	4.22	5.76	5.76	15.41	5.98
df	29	29	29	29	29	29	29	29

Note: Standard errors in parentheses are clustered at the School level. Morning, Witness, Rain, and Hot are included in the regressions but omitted from the table for space-saving reasons.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.11: Role of the Timing of Scenes on Punishment and Helping

	PUNISHMENT			HELPING		
	(1)	(2)	(3)	(4)	(5)	(6)
Child	0.215*** (0.054)	0.201*** (0.053)	0.224*** (0.057)	0.116** (0.049)	0.110** (0.047)	0.118*** (0.043)
VH	-0.104** (0.044)	-0.106** (0.044)	-0.092** (0.045)	-0.039 (0.027)	-0.036 (0.028)	-0.035 (0.029)
VH × Child	-0.066 (0.077)	-0.065 (0.078)	-0.072 (0.076)	-0.121*** (0.043)	-0.124*** (0.043)	-0.123*** (0.044)
Time × Arriving	-0.0004 (0.002)			-0.0007 (0.002)		
Time × Leaving	-0.0005 (0.002)			0.001 (0.001)		
T10 × Arriving		-0.0014 (0.002)			-0.0012 (0.002)	
T10 × Leaving		0.002 (0.003)			0.0004 (0.002)	
No. × Arriving			-0.012 (0.009)			-0.002 (0.007)
No. × Leaving			0.001 (0.014)			0.001 (0.010)
Constant	0.365*** (0.047)	0.363*** (0.046)	0.386*** (0.057)	0.145*** (0.050)	0.139*** (0.049)	0.146*** (0.047)
Observations	400	400	400	399	399	399
R^2	0.131	0.133	0.136	0.068	0.068	0.067
F	7.817	7.726	6.004	4.553	4.151	4.287

Note: TIME is a variable that measures the number of minutes from the beginning or end of school. Negative values indicate a time *before* the school bell rings, while positive values correspond to a time *after* the school bell has rung. TIME is truncated at -40 and 40. T10 rounds TIME to the nearest ten, in order to discretize the support. Finally, No. denotes the observation number *within* the same condition and time of day. For example, No. = 2 corresponds to the second observation of a given condition at a given time of day.

** $p < 0.05$, *** $p < 0.01$

An Eye for a Tooth: The Effects of Employer Pressure on Worker Productivity

Treat your employees right, so that they don't use your internet to search for a new job.

MARK ZUCKERBERG

4.1 Introduction

In this chapter, I explore the causal relationship between an employer imposing anti-social incentives on a worker and subsequent productivity of that worker. With anti-social incentives, I refer to incentives that require the worker to hurt a passive outsider in order to preserve the worker's payoff. For example, a second-hand car dealer can pressure a mechanic to conceal a safety issue under the threat of being fired. Recent survey evidence from Europe and the U.S. has highlighted the role of employers in encouraging and sustaining individual employees' malpractices that harm parties outside of the organization, with roughly 1 in every 6 employees being pressured to compromise ethical standards (see Ethics and Compliance Initiative,

2018; Dondé, 2018; Ransijn, 2018). Although this may not seem to be widespread, the consequences of individual cases can be huge, as exemplified by some high-profile cases. For example, Wells Fargo employees reported “extreme pressure to open as many accounts as possible”, which led them, despite some having ethical objections, to open unauthorized bank accounts in the name of unknowing customers who were charged a total of 2.6 million dollars in fees.¹ Whereas consumer harm is an obvious reason why such practices are undesirable, there may also be harmful consequences for the organization itself if it affects worker motivation and productivity. Research by the Ethics and Compliance Initiative (2017) shows that unethical leadership styles are associated with lower levels of worker engagement, which in turn has been negatively associated with various dimensions of job performance (Gallup, 2017). Moreover, the literature examining labor disputes suggests that lower worker satisfaction is associated with decreases in productivity (Krueger and Mas, 2004; Mas, 2006, 2008), although the observational nature of these studies makes it hard to establish the direction of causality. Therefore, I aim to explore the unintended consequences of employer malpractices, by examining whether employer pressures to behave unethically *cause* worker productivity to decrease.

To this end, I conduct a two-stage laboratory experiment in which an Employer (he), a Worker (she), and a passive recipient (a charitable organization) are matched to each other. All three parties start with the same endowment. In Stage 1 of the baseline treatment (called Intentions), the Employer first decides whether to end the stage (Option A) or to force the Worker to make a choice (Option B). In the latter case, the Worker is asked to choose between destroying her own payoff, destroying the payoff of the third party, or engaging in a gamble to have both unaffected against the alternative that both get destroyed. In other words, Option B imposes a trade-off on the Worker which the Worker would prefer to avoid. This feature exemplifies the pressure to behave anti-socially, *i.e.*, to sacrifice the payoff of the passive recipient in order to avoid a payoff reduction. Importantly, Option B yields a higher payoff to the Employer in Stage 1, which may attract him to this option. In Stage 2, the Worker subsequently engages in a coin-identification task (Belot and Schröder, 2013) for which she receives *no additional* payment. In contrast, the Employer’s payoff in Stage 2 depends positively on the Worker’s productivity in the task. I examine whether reciprocal preferences induce the Worker to become less productive in the task following the Employer’s choice of Option B, as compared to Option A.

My hypotheses follow from a model of reciprocal preferences à la Cox et al. (2007).

¹New York Times, *Wells Fargo Fined \$185 Million for Fraudulently Opening Accounts*, Sept. 8, 2016

For negative reciprocity to be activated, the Worker must perceive the Employer's choice of Option B as undesirable *and* intentional. The former implies that the Worker would prefer Option A to be implemented, while the latter means that Option A was intentionally *not* chosen by the Employer. I assess whether Workers indeed prefer Option A by eliciting Workers' mood after the option has been implemented, and find that Workers report a significantly worse mood under Option B. In order to speak to the role of intentions, I employ a No-Intentions Treatment, in which the option is implemented according to a random procedure over which the Employer has no control. This treatment allows me to disentangle the effect of the Employer's intentions from mechanisms related to wealth effects or distributional concerns.

My main contribution is to the literature on the relationship between leadership styles and workplace behavior. By focusing on decreases in worker productivity, this study is related to the concept of counterproductive work behavior (CWB), which constitutes the set of behaviors that employees perform and that hurt their organization. Damages due to CWB are estimated to amount billions of dollars (Bennett and Robinson, 2000). Brown and Treviño (2006a) have conjectured that the quality of leadership plays an important role in affecting CWB, while Blau (1964) and Konovsky and Pugh (1994) emphasize the role of reciprocity in governing the employer-employee relationship. This suggests a negative relationship between leadership quality and the extent to which workers engage in CWB. Indeed, empirical studies in organization theory, largely based on survey data, have shown this (Greenberg, 1990; Brown and Treviño, 2006b; Detert et al., 2007). However, little *causal* evidence on this relationship exists. For example, highly ethical leaders may attract more productive employees, or leaders may become more unethical exactly because workers are slacking. My study aims to contribute in this dimension.

Furthermore, my experiment contains a few novel features compared to traditional experimental papers studying reciprocity in the workplace (Fehr et al., 1993; Gneezy, 2002). First, while most papers examine initial acts of (un)kindness with direct monetary consequences, my design is more subtle: the Employer decides whether or not to impose anti-social incentives on the Worker, who still has control over her own payoffs. This implies that it is not obvious that the Worker will retaliate in the first place; she could just as much blame herself for the outcome obtained. Importantly, the exchange also contains non-monetary elements of (un)kindness. Survey evidence suggests that such elements are likely to be at least as equally relevant as monetary acts of unkindness (*i.e.*, wage decreases) in organizations and employer-employee relationships (Ethics and Compliance Initiative, 2018). As such, the Worker, even though she still holds her monetary destiny in her own hands, could reciprocate

against the induced psychological costs of having to make an undesirable trade-off. This also implies that the Worker reciprocates in a different domain than the initial act. In other words, the Worker trades an eye for a tooth.² Second, trade-offs between one's own wage, promotion, or job, and the interests of a third party are typically created by someone within the organization, rather than exogenously being in place. My experiment mimics this by introducing incentives that are endogenously imposed. As a result, my study adds a layer to traditional studies looking at the relationship between behavior and incentives: it takes into account that workers know that the incentives that they face are imposed by an employer who had other options. Taken together, these two features create a more realistic setting in which to study reciprocal behavior in the workplace.

The results from my experiment show that Worker productivity in the Intentions Treatment is significantly higher when the Employer intentionally abstains from imposing anti-social incentives (Option A) as compared to the Employer imposing anti-social incentives (Option B). In contrast, Worker productivity in the No-Intentions Treatment, where the option is randomly imposed, is comparable for Option A and Option B *and* is similar to Option A in the Intentions Treatment. I interpret this pattern as evidencing that my results are driven by *negative* reciprocity, while positive reciprocity seems absent: Workers decrease their productivity when Option B is intentionally chosen, while they do not increase their productivity when Option A is intentionally chosen. This asymmetric result is in line with Offerman (2002), who concludes that “hurting hurts more than helping helps” [p. 1423], and Kube et al. (2013), who show that only wage *cuts* affect worker moral. Interestingly, Worker productivity under Option B in the Intentions Treatment remains low even if Stage 1 resulted in the same outcome for the Worker and the third party as Option A. Hence, a Worker responds to the Employers' pressure to make an anti-social trade-off *per se*. This has important policy implications for organizations: managers and supervisors should also take into account the psychological costs of the tasks that they bestow upon their employees, in addition to their mere monetary incentives. The relevance of this feature is exemplified by the wage premiums paid to employees in, among others, the weapon or tobacco industry (Schneider et al., 2020), which imply that employees are sensitive to the societal impact of their organization's behavior.

The remainder of this chapter is set up as follows. Section 4.2 discusses related literature. Then, Section 4.3 discusses the design of the experiment and Section 4.4

²Since the Worker reciprocates a psychological cost imposed on her originating from feeling pressured to hurt a third party, it is unclear whether this is an example of *direct* or *indirect* reciprocity. Therefore, in this chapter, I simply refer to “reciprocity” when discussing the Worker's behavior.

presents the theoretical model and the hypotheses. I describe the data in Section 4.5, discuss the results in Section 4.6 and provide a concluding discussion in Section 4.7.

4.2 Related Literature

From the point of view of organization theory, my study relates closely to the concept of counterproductive work behavior (CWB). Counterproductive work behavior (also known as workplace deviance, organizational misbehavior, or worker anti-social behavior) is defined as “voluntary behavior that violates significant organizational norms and in so doing threatens the well-being of an organization, its members, or both” (Robinson and Bennett, 1995, p. 556). It encompasses explicitly harmful employee acts such as employee theft, sabotage and workplace aggression, but also more subtle forms like sloppiness, tardiness, absenteeism or resource wasting. Importantly, it has been argued that reciprocity, leadership styles, and employer treatment play an essential role in governing CWB (Treviño and Brown, 2005; Konovsky and Pugh, 1994). One of the first studies suggesting this, is Greenberg (1990), who studies employee theft rates following temporary pay cuts among a sample of manufacturing workers. Those groups of workers whose pay is temporarily reduced exhibit higher rates of theft. This effect disappears when the reason for the pay cuts was extensively explained to the workers. Zellars et al. (2002) explore the relationship between supervisors’ abusive supervision practices and subordinates’ organizational citizenship behavior (OCB, a virtuous antonym of CWB) by surveying a sample of members of the Air National Guard. They find a negative association, which is stronger for subordinates who perceive OCB as beyond their formal job requirements. Detert et al. (2007) study the relationship between abusive supervision and ethical leadership on the one hand and food loss, as a measure of counterproductive work behavior, using a sample of 265 restaurants. Abusive supervision is shown to correlate positively with food loss, while no relationship is found for ethical leadership, which suggests that workers respond negatively to the former, while they do not respond positively to the latter. Finally, Brown and Treviño (2006b) find deviant behavior to be less prevalent in work groups led by a socialized charismatic leader, where the latter is characterized, among other things, by an ethical leadership style.

An obvious question that arises from these insightful studies concerns the direction of causality: it cannot be established whether Employers cause Workers to decrease their productivity, or whether there is some other reason why Employer behavior and Worker productivity are related. For this reason, experimental economists have started to assess counterproductive behavior in the laboratory. The current chapter

adopts a paradigm introduced by Belot and Schröder (2013). In their experiment, the authors ask subjects to classify a box of Euro coins according to their country of origin. Subjects self-report their productivity and are paid according to this report. The experimenters verify this report afterwards and assess counterproductive behavior under three incentive schemes: fixed pay, piece-rate, and tournament. The authors find that “average counterproductive behavior amounts to 10 percent of average productivity” (Belot and Schröder, 2013, p. 233). Moreover, incentives affect the extent of counterproductive behavior: it is found to be significantly highest under the tournament scheme.

By examining reciprocity in the workplace, this chapter loosely relates to the wage-effort hypothesis (Akerlof, 1982) and the resulting experimental literature on the gift-exchange game. Compared to the seminal paper by Fehr et al. (1993), in which workers choose an effort level following a wage offer and effort costs are induced by a commonly known function, my study differs in two important respects. First, my study relates more closely to the papers that allow workers to reciprocate by actively performing a task that requires physical or cognitive effort. This approach takes into account the psychological cost of exerting effort and is thus closer to an actual workplace setting (Gneezy, 2002). Evidence for a positive wage-effort relation is generally found in the laboratory (Gneezy, 2002; Gächter et al., 2016), while the evidence from field studies is less conclusive (Gneezy and List, 2006; Hennig-Schmidt et al., 2010; Kube et al., 2013; Cohn et al., 2015). In particular, if a positive relationship between wage and effort is found in the field, it is often the result of *negative* reciprocity, while positive reciprocity is virtually absent.

Second, I go beyond wage offers and examine the non-monetary features of the employer’s behavior. In this light, Kube et al. (2012) assess the effect of non-monetary gifts and find even stronger reciprocal responses compared to monetary gifts of the same value. In the domain of monitoring performance, Falk and Kosfeld (2006) show that imposing a minimum effort level, even if it lies well below the average effort choice in the absence of a minimum effort level, causes effort to decrease. The authors show this effect to disappear in case the minimum effort level is exogenously imposed. In the same spirit, Belot and Schröder (2015) find that when mistakes in the coin-identification task of Belot and Schröder (2013) are sanctioned, subjects tend to return the box tardy more often. This shows that imperfect monitoring affects subjects’ counterproductive behavior in dimensions that are not monitored. In a similar spirit, Alempaki et al. (2019) prove the use of dishonesty as a reciprocation device: selfish dictators are more often deceived by their recipients than generous dictators in a subsequent sender-receiver game in which the recipient acts as a sender.

In my study, the non-monetary element entails a psychological cost originating from feeling pressured to hurt someone outside the organization. This element has, to the best of my knowledge, not been studied before. Outside the employer-employee setting, Khadjavi (2017) examine the effect of bestowing a psychological *benefit* upon others and shows that customers of a hair salon tip more following the hairdresser's voluntary efforts to raise money for a charitable organization. By providing customers with an opportunity to obtain a warm glow from donating, this setting could be seen as the opposite of mine.

4.3 Experimental Design

The experiment consists of three parts: a Production Game, a risk-elicitation task and a social value orientation (SVO) task. In addition, at the end of the experiment, subjects answer a questionnaire. Within sessions, the Production Game always comes first, while the order of the risk-elicitation and SVO tasks is counterbalanced across sessions. The full design is illustrated in Figure 4.2. Throughout the experiment, all payoffs are denoted in Tokens, with 10 Tokens equaling 1 Euro.

4.3.1 The Production Game

Subjects are informed that they are randomly matched to another subject in the same session for the two-stage Production Game. Players receive instructions for both stages before the start of Stage 1. In each pair, one subject is assigned the role of Employer (E, neutrally called "Player 1" and assumed male), while the other is assigned the role of Worker (W, "Player 2" and assumed female). Furthermore, each pair is informed that their choices in the Production Game may affect a donation made to a project of the International or Dutch Red Cross, which is randomly-selected from a list of projects shown to the players at the beginning of the experiment.³ The Red Cross represents the passive third party that may be hurt by the combined actions of the Employer and the Worker and can be thought of as the organization's stakeholders like investors, clients, or the general public. I have chosen to include a charitable organization as the third party, rather than an actual human subject, to maximize the use of the student subject pool and avoid the dilution of responsibility

³The projects provide humanitarian aid to people in need in different parts of the world. Subjects are informed that at most one pair per session is matched to any given project. This is done to ease the concern that subjects believe that their project is already receiving a donation from other pairs in the session. In order to avoid pairs becoming more or less motivated for their particular project, each pair's actual project is revealed at the end of the experiment.

that played a role in Chapter 2. Furthermore, the charitable organization is more likely to be perceived as an outsider to the Employer-Worker relationship than another subject in the experiment, which forms a more realistic representation of an actual workplace setting. It is *a priori* unclear how this would affect the results. On the one hand, a charitable organization is more likely to be viewed as an objectively good cause and a Worker may feel more strongly towards an Employer enriching himself at the expense of a charitable organization as compared to another subject, which may intensify the reciprocal response. On the other hand, the more abstract and distant nature of the Red Cross, as compared to an actual subject in the experiment, may dilute the Worker's perception of the consequences of the Employer's and her own decisions, which may in turn abate the reciprocal response. In order to convince the Worker and the Employer that the donation is real, both are informed that the Worker is asked to sign a form authorizing the donation. Moreover, the Worker is given the opportunity to receive a confirmation email of the donation.

Both Players and the Red Cross start the Production Game with an endowment of 50 Tokens (*i.e.*, 5 Euro). I introduce two different treatments of the Production Game, which I vary between-subjects: the Intentions Treatment and the No-intentions Treatment. Below, I discuss their design.

4.3.1.1 Intentions Treatment

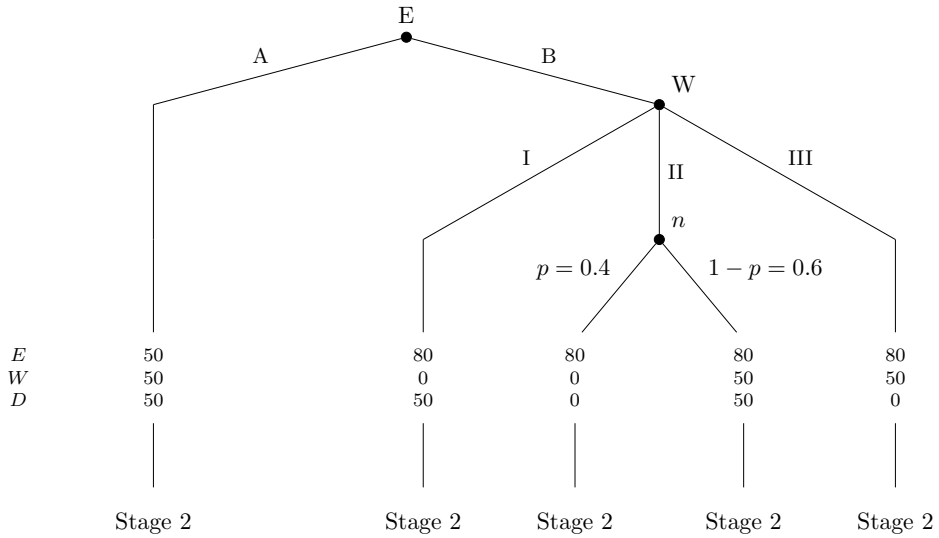
Stage 1 Stage 1 of the Production Game is depicted in extensive form in Figure 4.1. First, the Employer chooses between Option A and Option B. If he chooses Option A, Stage 1 ends immediately and the payoffs of both Players and the Red Cross remain equal to 50 Tokens. If the Employer instead chooses Option B, his payoffs increase to 80 Tokens while the Worker is required to choose between three Alternatives which trade off the Worker's own payoffs with the donation to the Red Cross. Alternative 1 leaves the donation unchanged, but decreases the Worker's payoffs to 0. Alternatively, Alternative 3 leaves the Worker's payoffs unchanged, but decreases the donation to 0. Finally, Alternative 2 entails a lottery that results in one of two outcomes: both the Worker's payoffs and the donation remain unaffected *or* both decrease to 0. The former occurs with a probability of 60 percent and the latter with the complementary probability. A Worker who has chosen Alternative 2 resolves the uncertainty by selecting one box from a field of 100 covered boxes. Underneath the cover, 40 boxes are red and 60 are green. If the selected box is red, both the Worker's payoffs and the donation are reduced to 0, otherwise they remain 50 Tokens.

A few features are worth elaborating upon. The three players start Stage 1 with equal payoffs to avoid any initial inequality between the Worker, the Employer, and the Red Cross that may affect the Employer's and Worker's perception of the options in ways that may be non-trivial to predict. As Option A maintains this egalitarian and fair outcome, I increase the likelihood that both the Employer and the Worker view this as the kind choice. In contrast, the imposed trade-off between the Worker's own payoffs and the donation to the Red Cross under Option B entails the pressure to behave anti-socially: the Worker can ensure a payoff of 50 Tokens *only* by destroying the donation, while she can ensure the preservation of the donation *only* by sacrificing her own payoff. Alternatively, the Worker can try to preserve her payoffs *and* the Donation by choosing Alternative 2, at the risk of losing both. This Alternative allows for the possibility to enter Stage 2 with the exact same outcome as Option A for the Worker and the Red Cross. Furthermore, the Employer's payoff is left independent from the Worker's choice of Alternative so as to exclude any form of reciprocation in Stage 1. Obviously, the difference in Employer payoffs between both options may matter for the Worker's evaluation of the Employer's choice: a lower additional benefit from choosing Option B arguably decreases the degree of understanding that the Worker has for this choice. At the same time, fewer Employers can be expected to choose Option B if the additional benefit is lower. In the end, I determined the parameters in order to obtain a roughly equal split between the two options.⁴ As a result, the (expected) total surplus is higher under Option A, meaning that the Employer, in expectation, destroys part of the surplus when he chooses Option B. Hence, efficiency concerns would make Option B even less desirable compared to Option A, which in turn may intensify the response of an efficiency-minded Worker. Nonetheless, I believe that the destruction of resources is often inherent to organizational malpractices and that it therefore forms a realistic feature of Stage 1.

Stage 2 In Stage 2, the Worker performs a coin-identification task requiring her to classify different Euro coins according to their country of origin and their denomi-

⁴To this end, I conducted two pilot sessions which induced changes to the design. In the first session, I conduct an experiment employing the strategy method with different benefits to the Employer and costs to the Worker, to calibrate these such that roughly half of the Employers can be expected to choose Option B. This resulted in the additional benefit of 30 ECU to the Employer and the probabilities of 60 and 40 percent under Alternative 2. In a second pilot session, I assess whether Option B is indeed perceived as unkind and undesirable from the perspective of the Worker. In order to do so, I ask a sample of student subjects who did *not* participate in the game to rate the social appropriateness of the two options as Very Socially Inappropriate, Socially Inappropriate, Socially Appropriate, or Very Socially Appropriate. To this end, I inform them that they receive 5 Euro if they report the modal social appropriateness rating in the session (Krupka and Weber, 2013). I also let these subjects predict Stage 2 productivity under the two options.

Figure 4.1: Extensive-Form Depiction of Stage 1 in Intentions



Note: E = Employer, W = Worker, D = Donation, n = Nature

nation. This is a computerized version of the task originally designed by Belot and Schröder (2013). Coins from the different countries in the Euro Zone differ in their design of the heads side, which allows one to determine their country of origin. Workers are shown a sequence of Euro coins on the screen and are asked to identify them. To this end, they are provided with a hard-copy catalogue describing the characteristics of the different denominations and displaying the country-specific sides of each coin (see Appendix 4.B.1.3).⁵ Since the Worker needs to process what is shown on the screen and match it to the information in the catalogue, the coin-identification task requires *cognitive* effort and can be performed without having any pre-existing knowledge about the design of Euro coins. As exemplified by Belot and Schröder (2013), the design of this task allows for multiple forms of counterproductive behavior, which are not presented to the Worker explicitly. First, the Worker can work at a slower pace and identify fewer coins. Second, the Worker can work more sloppily and make more mistakes. Third, the Worker could even make *deliberate* mistakes.

The Worker receives no additional payment for this task. In contrast, the Employer is paid according to the Worker's performance. That is, for each successful identification by the Worker, the Employer receives 1 additional Token. At the same

⁵I include 2 Euro, 1 Euro, 50 Euro cents, and 20 Euro cents coins from Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, and Spain. The denominations are chosen to be identifiable without needing to see the tails side of the coin.

time, a mistake by the Worker *decreases* the Employer's earnings by 1 Token. An identification is considered correct if and only if both the denomination and the country of origin are indicated correctly. Workers receive four minutes for this identification task, in which they can identify at most 75 coins.⁶ They do not receive feedback about their performance until time has run out. The Euro coins are shown on their screen in an order that is randomly determined beforehand and identical across treatments. For each coin, Workers first identify the denomination and then identify the country of origin. Images of the interface can be found in Appendix 4.B.4.

Mood After Stage 2, I ask Workers to “describe [their] overall mood *directly after* the implementation of the Option.” I elicit this at the end of Stage 2, rather than at the end of Stage 1, to avoid that the Worker's mood is reinforced before she starts the coin-identification task. Mood is reported on a seven-point scale ranging from *In a very bad mood* to *In a very good mood*. Primarily, I use this variable to check whether Workers indeed dislike Option B being implemented and like Option A being implemented. Moreover, mood could turn out to be a mechanism affecting productivity under both options. As Kirchsteiger et al. (2006) find in a gift-exchange game, individuals in a good mood become more generous, while individuals in a bad mood behave more reciprocally. This mechanism is also captured by my theoretical model, which builds upon Cox et al. (2007)'s notion of reciprocity concerns as an emotional state. In addition to general mood, I also elicit five specific emotions taken from the Positive And Negative Affect Scale (PANAS, Watson et al., 1988): Excited, Upset, Ashamed, Hostile, and Determined.

Other features A few other features of the Production Game deserve mentioning. Before roles are assigned, all players go through a detailed preview of the game. During this preview, they answer comprehension questions about both stages and perform a 30-second trial round of the coin identification task. Subjects can only proceed after they have answered all comprehension questions successfully.

Furthermore, in Stage 1, the Worker is asked which option she expects to be implemented by the Employer. Similarly, while the Worker is performing the coin-identification task, the Employer is asked a few questions. In particular, he is asked why he chose his option in Stage 1, how many *other* Employers in the session he expects to have chosen Option A, how productive he expects a Worker under Option A to be, and how productive he expects a Worker under Option B to be. I use this

⁶Only two Workers make it to the end of the sequence.

information in Subsection 4.6.6 in order to dive deeper into the role of the Worker's and Employer's expectations in determining their behavior in the experiment.

4.3.1.2 No-Intentions Treatment

The No-Intentions Treatment is identical to the Intentions Treatment, with the only difference being the way in which the option in Stage 1 is selected. Instead of being chosen by the matched Employer, the option is now selected through a random procedure identical to the one used for Alternative 2 (see above) and operated by the Worker. The probability of Option B being implemented is 48% and I elaborate upon the selection of this probability below. Thus, in the No-Intentions Treatment, the Worker faces incentives that have not been imposed by the Employer for whom she works. Any feelings of reciprocity should thus be absent in this treatment. The performance under the two options in the No-Intentions Treatment provides a useful benchmark to which I can compare performance levels in the Intentions Treatment. Stage 2 in the No-Intentions Treatment is identical to the Intentions Treatment, including the mood elicitation.

4.3.2 Preference-Elicitation Tasks

Subjects' social and risk preferences may shape their decisions in both stages of the Production Game. I measure social value orientation (SVO) using the procedure developed by Murphy et al. (2011) and elicit only the six primary items. To this end, I use the z-Tree code developed by Crosetto et al. (2019). Furthermore, I elicit risk preferences by means of a bomb-risk elicitation task (Crosetto and Filippin, 2013). Subjects are presented with a field of a 100 boxes on the screen and the software removes one box from the screen every second. Subjects receive 1 Token for each box collected. However, behind one of the boxes, a bomb is hidden which destroys all earnings if it is hidden in one of the boxes that is collected. Subjects decide when to stop the collecting process, with a larger number of boxes indicating a higher degree of risk tolerance.

4.3.3 Survey and Procedures

Survey The experiment is concluded with a survey in order to gather more information on the subjects in the experiment. In addition to checking whether the sample is balanced across treatments, I use this information in order to measure some of the preference parameters in the theoretical model and assess heterogeneous treatment

(see below). First, I ask for the subjects' sex, age, nationality, study program, and experience with Euro coins. Second, subjects self-report the extent to which they exhibit Positive Reciprocity, Negative Reciprocity, Indirect Reciprocity, and Altruism on a seven-point Likert scale using the questions developed by Falk et al. (2018). For negative reciprocity, I include two slightly differently formulated versions. Subjects are also asked to report their monthly amount of charitable donations. Third, subjects are asked whether they thought the Euro Identification task was difficult, exciting, and effortful. Finally, I measure subjects' Big Five personality traits using a 15-item questionnaire developed by Hahn et al. (2012), with each item being evaluated on a seven-point Likert scale. For the precise formulation of the survey questions, I refer to Appendix 4.B.2.

Subjects and Sessions I conducted 18 sessions of the experiment between March and October 2020 in the CentERlab, the experimental laboratory of Tilburg University, the Netherlands. All subjects were students who were recruited using Sona Systems.⁷ The first four sessions of the experiment featured the Intentions Treatment and were used to determine the exogenous probability of Option B being implemented in the No-Intentions Treatment (*i.e.*, 48%). I decided to only use the first four sessions in order to be able to run both treatments in the first week of the study (which contained eight sessions) and avoid unobservable selection of subjects into treatments due to subjects enrolling in the first week being different from subjects enrolling at a later point. Coincidentally, CoViD-19 measures caused the university, and the laboratory, to close down for a duration of three months after this first week. Later sessions were conducted in June, September, and October under a strict protocol approved by Tilburg University's health and safety advisor. Among other things, the protocol featured a reduced laboratory capacity to facilitate social distancing and a health check to be performed by subjects and experimenters. Subjects were still recruited via Sona Systems, with some additional effort expended on invitation emails and reaching out to students attending the few lectures taking place on the university campus.

Procedures I received approval from Tilburg University's institutional review board (IRB-EXE 2020-001) in February, 2020. The experiment was computerized using the z-Tree software developed by Fischbacher (2007). All instructions were read aloud by the experimenter in order to establish common knowledge and any questions were an-

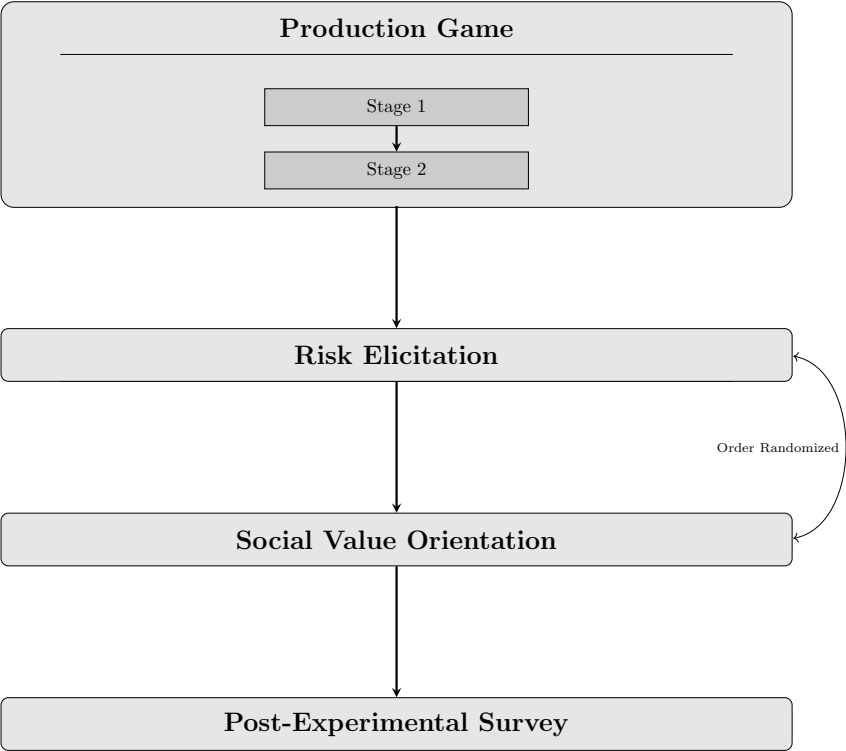
⁷Sona Systems provides universities with software to manage and recruit subjects for research purposes. More information can be found on <https://www.sona-systems.com/about.aspx>.

swered in private. The instructions used in the experiment can be found in Appendix 4.B.1. In line with the CoViD-19 protocol, no paper instructions were distributed in the sessions taking place after the outbreak. Instead, the instructions were shown on the screen at the beginning of each Part. At any point in time, subjects could access these instructions through a button on their interface. Moreover, the coin identification catalogues were laminated and attached to the inside of the subjects' cubicles.

Payments Subjects received a 4 Euro show-up fee. In addition, they received their earnings from the Production Game and their earnings from either the SVO or the risk-elicitation task. The experimenter performed the random selection of this task by flipping a coin in public at the end of the session, such the same task is selected for all subjects in the same session. Sessions lasted at most 60 minutes and subjects earned 12.86 Euro on average. Again, due to CoViD-19 measures, payments were executed electronically during sessions after the outbreak.

Donations Each session featured 10 to 12 different Red Cross projects, which were taken from either the Dutch or International Red Cross page. I selected those projects with their own distinct webpage, and replaced a project when donations to the particular project were no longer possible. Over the course of the experiment, subjects made 64 donations to 13 Red Cross projects for a total of 320 Euro. At the end of each session, Workers whose project received a donation authorized this donation and indicated whether they wanted to receive a confirmation email once the donation was made (33 out of 64 Workers request a confirmation). I list all Red Cross projects and their donated amounts in Appendix 4.B.3.

Figure 4.2: Experimental Design Summary



4.4 Model and Hypotheses

Below, I present my theoretical model and resulting hypotheses. The purpose of the theoretical model is twofold. First, it allows me to provide a theoretical foundation of my two main hypotheses building upon previous scholars' models of reciprocal preferences. The model not only differentiates in terms of net productivity between Option A and Option B, but also predicts how behavior in the No-Intentions Treatment compares to this. Second, the theoretical model provides testable predictions with respect to heterogeneous treatment effects and predicts which Workers respond more strongly to the Employers' intentions.

4.4.1 Set-Up

Workers, indexed $i = 1, 2, 3, \dots, N/2$, and Employers, indexed $j = 1, 2, 3, \dots, N/2$, are drawn from the same population consisting of N individuals and matched to each other at random. Each pair participates in treatment $t \in \{NI, I\}$, with NI for the No-Intentions Treatment and I for the Intentions Treatment. The game's structure is as follows. Employer j (or nature in the No-Intentions Treatment) first chooses an option $o_j \in \{A, B\}$. If Option B is chosen, Worker i chooses an Alternative $a_i \in \{1, 2, 3\}$. Subsequently, the Worker engages in the Identification Task, exerts effort, and achieves net productivity $e_i \in [-75, 75]$. The Employer, the Worker and the charitable organization earn expected monetary payoffs $\pi_E(o_j, e_i)$, $\pi_W(o_j, a_i)$, and $D(o_j, a_i)$, respectively. Using the parameters from Section 4.3, these monetary payoffs (in Tokens) are as follows:

$\pi_E(o_j, e_i)$	$\pi_W(o_j, a_i)$	$D(o_j, a_i)$	Condition
$50 + e_i$	50	50	if $o_j = A$
$80 + e_i$	0	50	if $o_j = B$ & $a_i = 1$
$80 + e_i$	50	50	if $o_j = B$ & $a_i = 2$ & lottery won
$80 + e_i$	0	0	if $o_j = B$ & $a_i = 2$ & lottery lost
$80 + e_i$	50	0	if $o_j = B$ & $a_i = 3$

4.4.2 Utility

The players' utility functions are based on Cox et al. (2007), who design a tractable model to capture reciprocity concerns. In their model, each player maximizes a utility function which includes the own payoff and another player's payoff. The weight θ_i

attached to the other's payoff is assumed to depend on the emotional state felt towards the other player, which in turn is determined by reciprocity concerns r .⁸ This results in the following linear version of the utility function:

$$U_i(\cdot) = \pi_i(\cdot) + \theta_i(r)\pi_j(\cdot)$$

In comparison to Cox et al. (2007), I adjust the model in three ways. First, I add the donation as a third payoff term in the utility function, which has a potentially different weight attached to it. Second, I add an effort cost function in order to model the Worker's disutility from performing the coin-identification task. Third, I introduce a specific functional form for the reciprocity term r on which I elaborate below. For simplicity, I assume here that both players are risk-neutral. In Appendix 4.A.4, I discuss how risk aversion affects behavior in the model.

4.4.2.1 Worker Behavior

Worker i 's utility is given by:

$$U_i^W(o_j, a_i, e_i) = \pi_W(o_j, a_i) + \theta_i^D D(o_j, a_i) + (\theta_i^E + \rho_i r_i(t, o_j))\pi_E(o_j, e_i) - c(e_i) \quad (4.1)$$

As explained above, the Worker attaches a weight to the monetary payoffs of the other players in the game. In line with Cox et al. (2007), the weight on $\pi_E(\cdot)$ is a linear combination of a "residual altruism" parameter $0 \leq \theta_i^E \leq \bar{\theta}$ and a reciprocity term $\rho_i r_i(t, o_j)$. In the latter, $r_i(t, o_j)$ captures the (un)kindness of the Employer's action *as perceived by Worker i* , while the slope coefficient $0 \leq \rho_i \leq 1$ represents the *intensity* of reciprocity concerns. This parameter allows some Workers to respond more strongly than others to the same action of the Employer. The weight attached to the donation to the Red Cross only consists of the residual altruism term θ_i^D , which may differ from θ_i^E , and reciprocity concerns are naturally absent. I assume that θ_i^E and θ_i^D are private information, positively correlated, and drawn from the same distribution with some commonly known probability density function $f(\theta)$ and cumulative distribution function $F(\theta)$. Similarly, ρ_i is private information and distributed according to some commonly known distribution with probability density function $g(\rho)$ and cumulative distribution function $G(\rho)$. In the experiment, I elicit θ_i^E and ρ_i using the SVO task and the survey questions on reciprocity (see Appendix 4.B.2), respectively. Due to

⁸Cox et al. (2007) also include status concerns to affect the weight. However, since players in my experiment are *ex ante* equal in the status dimension, I do not consider status concerns.

its positive correlation with θ_i^E , θ_i^D can be proxied by the SVO task, as well.

Reciprocity In order to determine the reciprocity term $\rho_i r_i(t, o_j)$, I assume that the Worker cares about the consequences to her own monetary payoff $\pi_W(\cdot)$ and the consequences to the donation $D(\cdot)$ following the Employer's choice. The Worker derives the kindness of the Employer's choice by comparing the actual payoff to herself and the charity to what she expected *a priori* and deems the Employer's action kind (unkind) if the actual payoff is higher (lower, respectively) than expected.⁹ In Appendix 4.A.1, I elaborate on the construction of the kindness term $r_i(t, o_j)$, building upon the specification by Cox et al. (2007). For the current discussion, the resulting reduced form, in which $\eta_i \in [0, 1]$ captures the Worker's prior belief that the Employer chooses Option A, suffices:

$$r_i(t, o_j) = \begin{cases} 1 - \eta_i & \text{if } t = I \text{ \& } o_j = A \\ 0 & \text{if } t = NI \\ -\eta_i & \text{if } t = I \text{ \& } o_j = B \end{cases} \quad (4.2)$$

Thus, choosing Option B ($o_j = B$) is deemed unkind by a magnitude of $-\eta_i$, while choosing Option A ($o_j = A$) is deemed kind by a magnitude of $1 - \eta_i$. As a result, the parameter η_i affects the (a)symmetry of the reciprocal response. For the knife-edge case of $\eta_i = 0.5$, $r(I, A) = -r(I, B)$, so that the extent to which Option A is deemed kind is equal to the extent to which Option B is deemed unkind. As Option A becomes the more expected choice ($\eta_i \rightarrow 1$), actually picking Option A becomes less kind, while picking Option B becomes more unkind. The opposite reasoning applies to the case where $\eta_i \rightarrow 0$. In the experiment, Workers report their prior expectation about the option chosen, which results in a crude measure of η_i . Naturally, the Worker does not evaluate the kindness of the Employer in the No-Intentions Treatment and $r_i(NI, o_j) = 0$, irrespective of o_j .

Cost of Effort The final element of the utility function consists of the Worker's effort cost from performing the coin-identification task. Remember that the Worker chooses $e_i \in [-75, 75]$, where $e_i < 0$ represents the Worker making more mistakes than successful identifications. For simplicity, I assume that the marginal cost of identifying an additional coin is independent from whether or not the coin is iden-

⁹It should be noted that the Worker looks at the *expected* payoff to herself and the donation at the time of the Employer's choice. This implies that the Worker only blames the Employer for forcing the choice upon her and not, for example, for having lost $\pi_W(\cdot)$ and $D(\cdot)$ after choosing Alternative 2 and losing the lottery.

tified correctly. Moreover, I assume that the Worker only makes *deliberate*, and no accidental, mistakes. Together, these assumptions imply that when a Worker intends to achieve a net productivity of, say, $-x$ Tokens, she does so by making exactly x mistakes, instead of, for example, identifying y coins correctly and making $y + x$ mistakes to compensate for the correct identifications. This equates net productivity accrued by the Employer to the effort exerted by the Worker. In order to facilitate an internal solution, effort costs are convex. This captures that the repetitive nature of the coin-identification task increases the marginal disutility from identifying an additional coin as the Worker has already identified more coins. This yields the following effort cost function, in which the parameter μ scales the marginal cost from identifying an additional coin.

$$c(e_i) = \frac{\mu}{2} e_i^2 \quad (4.3)$$

Optimal Productivity Note that $r_i(t, o_j)$ does not depend on the Worker's choice of Alternative a_i in Stage 1. Since the Worker's Stage 1 choice is not the primary focus of this model, I discuss the analysis of the Worker's choice of Alternative in more length in Appendix 4.A.2. Following Stage 1, the Worker decides how many coins to identify correctly or incorrectly in Stage 2. To this end, she maximizes Equation (4.1) with respect to net productivity e_i . This yields the following optimal productivity level:

$$e_i^*(t, o_j) = \frac{\theta_i^E + \rho_i r_i(t, o_j)}{\mu} \quad (4.4)$$

Note that Equation (4.2) states that $r_i(NI, B) = r_i(NI, A) = 0$, which in turn implies that the optimal productivity in the No-Intentions Treatment is independent from the option implemented: $e_i^*(NI, B) = e_i^*(NI, A) = e_i^*(NI)$.

Then, I can draft the following proposition:

Proposition 4.1 *Define $e_i^{NI} \equiv e_i^*(NI)$, $e_i^A \equiv e_i^*(I, A)$, and $e_i^B \equiv e_i^*(I, B)$ as the Worker's optimal net productivity in the No-Intentions Treatment, under Option A in the Intentions Treatment, and under Option B in the Intentions Treatment, re-*

spectively. Then:

$$e_i^{NI} = \frac{\theta_i^E}{\mu} \quad (4.5)$$

$$e_i^A = e_i^0 + \frac{\rho_i(1 - \eta_i)}{\mu} \quad (4.6)$$

$$e_i^B = e_i^0 - \frac{\rho_i\eta_i}{\mu} \quad (4.7)$$

As a result, $e_i^A > e_i^{NI} > e_i^B$ for $\rho_i > 0, \eta_i \in (0, 1)$.

In words, in the Intentions Treatment, Option A induces a positively reciprocal response and increases the Worker's net productivity, relative to the No-Intentions benchmark. After all, the Worker considers this as a kind action by the Employer. Analogously, the Worker considers choosing Option B as an unkind action, which induces her to reduce her net productivity, relative to the No-Intentions benchmark. Finally, note that, in case the Worker expects one of the options with certainty (*i.e.*, $\eta_i = 1$ or $\eta_i = 0$), there is no reciprocal response to one of the two options: $e_i^{NI} = e_i^A$ or $e_i^{NI} = e_i^B$.

Testable Hypotheses Based on Proposition 4.1, I formulate two testable hypotheses, which have been pre-registered with AsPredicted (#36388). Since I look at Workers in the aggregate in a between-subjects design, I examine *average* net productivity e^{NI} , e^A , and e^B .

Hypothesis 4.1 *In the Intentions Treatment, net productivity by the Worker is higher when the Employer chooses Option A as compared to when he chooses Option B.*

Furthermore, e^{NI} is the same under both options and weakly in between e^A and e^B . I refer to the difference between e^A and e^{NI} as positive reciprocity, as it represents the increase in productivity from *intentionally* choosing Option A. Analogously, the difference between e^0 and e^B captures the extent of negative reciprocity. In the aggregate, the average expectation η (without the subscript) determines whether e^0 is relatively closer to e^A or e^B , and whether positive or negative reciprocity is relatively larger: negative reciprocity is predicted to be the dominant element when $\eta > 0.5$, while positive reciprocity is predicted to be the dominant force when $\eta < 0.5$. For η strictly between 0 and 1, I can formulate the following hypothesis:

Hypothesis 4.2 *Net productivity by the Worker under Option B (Option A) is higher (lower) in the No-Intentions Treatment as compared to the Intentions Treatment.*

In the extreme case that $\eta = 1$ ($\eta = 0$), the model predicts that I only observe negative (positive) reciprocity, simply because choosing Option A (B) is the universally expected choice among Workers. In these extreme cases, the model predicts no differences in productivity between treatments for one of the two options.

4.4.2.2 Employer Behavior

I close the model by examining the Employer j 's optimal behavior. Employers are drawn from the same population and therefore maximize a similar utility function as the Worker. Obviously, Employers exhibit no reciprocity concerns, as they move first, and do not incur effort costs from the coin identification task:

$$U_j^E(o_j, e_i) = \pi_E(o_j, e_i) + \theta_j^D D(o_j, a_i) + \theta_j^W \pi_W(o_j, a_i) \quad (4.8)$$

Similar to θ_j^E for the Worker, θ_j^W is drawn from the same distribution as θ_j^D . Since Employer j cares about the monetary payoff to Worker i and the Red Cross, the Employer's choice also depends on the Alternative chosen by the Worker under Option B and the Employer needs to form a belief about this. I define F_1 , F_2 and F_3 as the perceived probability that the Worker chooses Alternative 1, 2 and 3, respectively, with $F_2 = 1 - F_1 - F_3$.¹⁰ In addition, the Employer needs to form a belief about the net productivity of the Worker under both options. I call the belief of Employer j about the net productivity of Worker i $\hat{e}_{j,i}^A$ and $\hat{e}_{j,i}^B$ for Option A and Option B, respectively. Importantly, I do not require these beliefs to be correct. Then, the Employer chooses Option A if (also see Appendix 4.A.3):

$$\hat{e}_{j,i}^A - \hat{e}_{j,i}^B > 30 - (20 - 10(2F_1 - 3F_3))\theta_j^D - (20 - 10(2F_3 - 3F_1))\theta_j^W \quad (4.9)$$

$$\hat{e}_{j,i}^A - \hat{e}_{j,i}^B > 30 - \theta_j(40 + 10(F_1 + F_3)) \quad (4.10)$$

Here, Equation (4.10) denotes the case in which $\theta_j^D = \theta_j^E = \theta_j$. Hence, if the Employer believes that the productivity gap between Option A and Option B is sufficiently high, he finds it worthwhile to choose Option A. For sufficiently high values of θ_j , the Employer would choose Option A even if he believes that both options yield the

¹⁰Under risk-neutrality and correct beliefs about the Worker's cut-off strategy, $F_1 = \Pr(\theta_i^D > \frac{3}{2})$ and $F_3 = \Pr(\theta_i^D \leq \frac{2}{3})$. Also see Appendix 4.A.2.

Table 4.1: Number of Workers in Each Cell

Treatment	Option		Total
	A	B	
NI	21	24	45
I	25	36	61
Total	46	60	106

same productivity (*i.e.*, the left hand side of (4.10) is zero). I examine the consistency between the beliefs of Employers and their choices in Subsection 4.6.6.

4.5 Data Description

In total, 228 subjects participated in the experiment. The data from the first No-Intentions session needed to be dropped since I erroneously included the wrong exogenous probability for the two options being implemented (16 observations). As a result, I end up with a sample of 212 observations. As I am interested in Workers, I examine the behavior of the 106 subjects in this role in relation to the treatment and the option implemented. A total of 61 pairs participated in the Intentions Treatment, while 45 pairs participated in the No-Intentions Treatment. Of the 61 Employers in the Intentions Treatment, 36 chose Option B (59%). In the No-Intentions Treatment, the exogenous probability of 48% for Option B resulted in 24 out of 45 Workers ending up with Option B (53.3%). Table 4.1 summarizes the number of Workers in each treatment-option combination.

Tables 4.2 and 4.3 report summary statistics for Workers across treatments and options, respectively. It should be noted that, although the option arises endogenously during the experiment, it is by construction exogenous to the Worker as it is imposed by a randomly-matched Employer (in Intentions) or by a random draw (in No-Intentions). In Column (4) of Table 4.2 and Columns (4) and (7) in Table 4.3, I report the results of *t*-tests of no difference between treatments and options, respectively, with the standard error of the difference in parentheses, in order to check whether the sample is balanced. As discussed below, the sample is unbalanced on a few dimensions, thus motivating the need for the linear regressions in Section 4.6.

Panel A shows information on the background of the Workers. Across the entire sample, 42 percent of the subjects are male, 74 percent of the subjects are European, and 65 percent of the subjects participate in an Economics or Business program. Moreover, subjects are 21.6 years old and have about 11.7 years experience with the

Euro. The t -tests indicate that more subjects in the No-Intentions Treatment are male (significant at a 10%-level). Regarding the comparison across options, I observe some significant differences within the No-Intentions Treatment *only*. In particular, compared to subjects under Option A, subjects under Option B in the No-Intentions Treatment are significantly younger (at a 10%-level), more likely to be European (at a 5%-level), and have more experience with the Euro (at a 5%-level). Obviously, the latter two characteristics are likely to be highly correlated. Having more experience with the Euro may make Workers under Option B in the No-Intentions Treatment more productive, which should be taken into account when evaluating the difference between options in this treatment.

Panel B includes information on the preferences of the subjects. The variables Reciprocity and Altruism both measure the subjects' self-reported inclination for the corresponding preference on a seven-point scale, where Reciprocity is the average report for Positive Reciprocity, Negative Reciprocity, Negative Reciprocity 2, and Indirect Reciprocity. A higher score on these items indicates a higher degree of reciprocity and altruism, respectively. Risk Tolerance indicates the number of boxes collected in the risk-elicitation task and is a measure of subjects' risk preferences (risk-neutral subjects should collect 50 boxes). On average, subjects collect 38 boxes, which is somewhat lower than the average reported by Crosetto and Filippin (2013) themselves and implies that subjects are on average risk averse. Finally, SVO Angle measures the subjects' social value orientation elicited by the corresponding incentivized task, with a larger angle indicating a higher degree of prosociality. The average angle is 20 degrees, which means that, according to the classification by Murphy et al. (2011), the average subject is classified as an individualist.¹¹ The t -tests indicate that subjects under Option A in the No-Intentions Treatment self-report being more altruistic (significant at the 1%-level). However, this does not translate into a higher angle in the incentivized SVO task for these subjects. This could reflect that Option A made Workers feel more altruistic, while they do not act on this when push comes to shove.

Finally, Panel C contains information on the subjects' Big Five personality traits. For each domain (Conscientiousness, Extraversion, Agreeableness, Openness, and Neuroticism), the scores of the three items covering that domain are averaged to

¹¹Note that this classification is rather arbitrary. As Murphy et al. (2011) note, a *consistent* individualist decision maker has an angle between -7.8 and +7.8, while a consistent prosocialist has an angle between 37.09 and 52.91 degrees. By bisecting the range in between, the authors arrive at the cut-off of 22.45 degrees. Translating this to the model of Section 4.4 implies an average θ_i of roughly $20/45 = 0.42$ (an angle of 0 implies $\theta_i = 0$ and an angle of 45 implies $\theta_i = 1$), which is substantially different from zero.

obtain a single score between 1 and 7. In the process, the scores of items that negatively relate to the particular domain are reversed to make averaging feasible. Across treatments and options, the sample is balanced on all Big Five dimensions, with the exception of conscientiousness across options in the Intentions Treatment (significant at a 5%-level). Note that the higher degree of conscientiousness in Option B, in combination with the reported negative relationship between counterproductive work behavior and conscientiousness, would work against my main hypothesis and make it harder to find an affirmative result.

As explained above, I conducted the experiment before and after the outbreak of the CoViD-19 virus. In Appendix 4.C.6, I show that subjects participating in a session after the outbreak are somewhat less risk-averse, which may point towards self selection of more risk-loving subjects into the experiment. Reassuringly, I also show that results before and after the outbreak are similar.

4.6 Results

In the current section, I present the results. I begin by showing that the manipulation was successful in the sense that Workers dislike Option B being implemented, while they like Option A being implemented (Subsection 4.6.1). I then turn to testing my two hypotheses using Mann-Whitney U tests in Subsections 4.6.2 and 4.6.3. Subsequently, I perform a linear regression in order to control for observables (Subsection 4.6.4), I analyze productivity according to Stage 1 outcomes (Subsection 4.6.5), and I look at Employer behavior (Subsection 4.6.6).

4.6.1 Manipulation Check: Workers under Option B Report a Worse Mood

I first check whether Workers indeed dislike Option B being implemented, while they like Option A being implemented. To this end, I examine Workers' self-reported mood following the implementation of the option.¹² Mood is reported on a seven-point scale ranging from *In a very bad mood* (coded as -3) to *In a very good mood* (coded as +3). I present a comparison across options for both treatment separately in Figure 4.3. Interestingly, I find that Workers report a significantly better mood under Option A as compared to Option B in *both* treatments (Intentions: 1.4 vs. -0.5, Mann-Whitney U test: $z = 4.717$, $p < 0.001$; No-Intentions: 1.1 vs. -0.5, Mann-Whitney

¹²This manipulation check and the hypothesis that mood is worse in Option B was also pre-registered.

Table 4.2: Summary Statistics of Workers and Balance across Treatments

	TREATMENTS			
	(1) All	(2) I	(3) NI	(4) Δ
A. Personal Characteristics				
Male	0.40 (0.49)	0.33 (0.47)	0.50 (0.51)	-0.17* (0.10)
Age	21.68 (3.42)	21.38 (3.17)	22.09 (3.73)	-0.71 (0.67)
European	0.72 (0.45)	0.75 (0.43)	0.67 (0.48)	0.09 (0.09)
Economics & Business	0.63 (0.48)	0.66 (0.48)	0.60 (0.50)	0.06 (0.10)
Euro Years	11.76 (8.09)	11.90 (8.11)	11.58 (8.14)	0.32 (1.60)
B. Preferences				
Reciprocity	4.32 (0.91)	4.23 (0.87)	4.44 (0.96)	-0.21 (0.18)
Altruism	4.89 (1.53)	4.92 (1.69)	4.84 (1.30)	0.07 (0.30)
Risk Tolerance	39.30 (17.31)	37.92 (19.27)	41.18 (14.25)	-3.26 (3.40)
SVO Angle	20.09 (13.10)	20.80 (13.65)	19.14 (12.40)	1.66 (2.58)
C. Big Five				
Conscientiousness	4.84 (1.01)	4.76 (1.01)	4.96 (1.01)	-0.20 (0.20)
Extraversion	4.56 (1.28)	4.51 (1.39)	4.62 (1.14)	-0.11 (0.25)
Agreeableness	5.14 (1.05)	5.09 (1.07)	5.20 (1.02)	-0.11 (0.21)
Openness	4.94 (1.13)	5.02 (1.09)	4.83 (1.18)	0.19 (0.22)
Neuroticism	4.26 (1.34)	4.34 (1.28)	4.14 (1.43)	0.20 (0.26)
Observations	106	61	45	106

Note: Balance of Workers' characteristics across treatments. Column (4) displays the difference between the treatments using a t -test, with the standard error of the difference displayed between parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

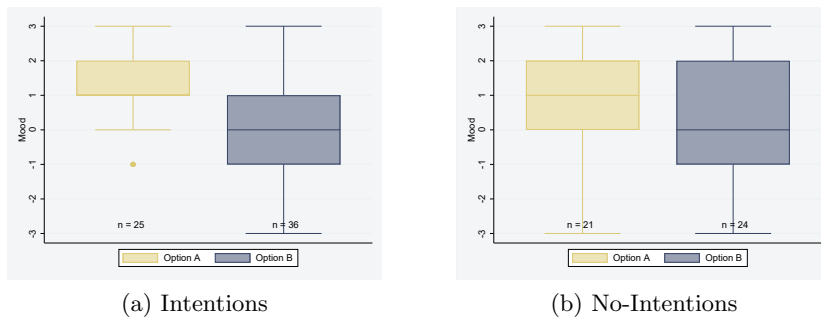
Table 4.3: Summary Statistics of Workers and Balance across Options

	INTENTIONS				NO-INTENTIONS		
	(1) All	(2) A	(3) B	(4) Δ	(5) A	(6) B	(7) Δ
A. Background							
Male	0.40 (0.49)	0.40 (0.50)	0.28 (0.45)	0.12 (0.12)	0.38 (0.50)	0.61 (0.50)	-0.23 (0.15)
Age	21.68 (3.42)	21.12 (3.63)	21.56 (2.84)	-0.44 (0.83)	23.10 (4.44)	21.21 (2.78)	1.89* (1.09)
European	0.72 (0.45)	0.76 (0.44)	0.75 (0.44)	0.01 (0.11)	0.48 (0.51)	0.83 (0.38)	-0.36** (0.13)
Economics & Business	0.63 (0.48)	0.68 (0.48)	0.64 (0.49)	0.04 (0.13)	0.62 (0.50)	0.58 (0.50)	0.04 (0.15)
Euro Years	11.76 (8.09)	10.20 (8.40)	13.08 (7.80)	-2.88 (2.10)	8.86 (8.53)	13.96 (7.12)	-5.10** (2.33)
B. Preferences							
Reciprocity	4.32 (0.91)	4.17 (0.87)	4.26 (0.88)	-0.09 (0.23)	4.60 (1.01)	4.30 (0.91)	0.29 (0.29)
Altruism	4.89 (1.53)	5.28 (1.84)	4.67 (1.55)	0.61 (0.44)	5.38 (1.07)	4.38 (1.31)	1.01*** (0.36)
Risk Tolerance	39.30 (17.31)	37.28 (20.75)	38.36 (18.46)	-1.08 (5.06)	37.52 (11.63)	44.38 (15.74)	-6.85 (4.18)
SVO Angle	20.09 (13.10)	21.07 (15.25)	20.61 (12.66)	0.46 (3.58)	19.92 (10.84)	18.46 (13.82)	1.46 (3.74)
C. Big Five							
Conscientiousness	4.84 (1.01)	4.41 (1.25)	5.00 (0.73)	-0.59** (0.25)	5.19 (0.92)	4.75 (1.06)	0.44 (0.30)
Extraversion	4.56 (1.28)	4.84 (1.18)	4.29 (1.49)	0.55 (0.36)	4.51 (1.22)	4.72 (1.08)	-0.21 (0.34)
Agreeableness	5.14 (1.05)	4.93 (1.10)	5.19 (1.05)	-0.26 (0.28)	5.22 (1.02)	5.18 (1.05)	0.04 (0.31)
Openness	4.94 (1.13)	4.89 (1.24)	5.11 (0.98)	-0.22 (0.28)	4.84 (1.23)	4.82 (1.17)	0.02 (0.36)
Neuroticism	4.26 (1.34)	4.16 (1.30)	4.47 (1.27)	-0.31 (0.33)	4.16 (1.45)	4.13 (1.44)	0.03 (0.43)
Observations	106	25	36	61	21	24	45

Note: Balance of Workers' characteristics across options in both treatments. options are implemented by the matched Employer (in Intentions) or by a random procedure performed by the Worker (in No-Intentions). Columns (4) and (7) display the difference between the options using a *t*-test, with the standard error of the difference displayed between parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 4.3: Average Mood under Option A and Option B



Note: The Figure displays boxplots of self-reported mood of Workers under the two options, ranging from -3 (very bad mood) to +3 (very good mood).

Table 4.4: Average Performance in Each Cell

A. Net Productivity					B. Mistakes								
Treatment	Option		Avg.		Treatment	Option		Avg.					
	A	B				A	B						
NI	20.9	(2.7)	23.4	(5.1)	22.2	(3.0)	NI	3.5	(0.7)	6.8	(3.0)	5.2	(1.7)
I	23.4	(2.8)	10.1	(4.4)	15.5	(2.9)	I	4.3	(0.8)	11.0	(2.6)	8.2	(1.7)
Avg.	22.3	(2.0)	15.4	(3.4)	18.4	(2.2)	Avg.	3.9	(0.6)	9.3	(2.0)	7.0	(1.2)

Note: The table contains average net productivity (Panel A) and mistakes (Panel B) in each treatment-option combination, in each treatment and option overall, and of all Workers together. The standard deviation is denoted in parentheses.

U test: $z = 3.400$, $p < 0.001$). The pattern seems to be somewhat more pronounced in the Intentions Treatment, but the difference in differences is insignificant across treatments. This suggests that the Worker's mood is to a large extent affected by the option itself, irrespective of *who* implemented it, which could merely reflect the Worker's disappointment of not having Option A implemented. However, only in the Intentions Treatment can the Worker blame the Employer for this disappointment. Therefore, only in the Intentions Treatment would we expect differences in mood to translate into a reciprocal response of the Worker towards the Employer. Taken together, the analysis of the Worker's mood validates the presumption that she would prefer Option A, rather than Option B, to be implemented. I now turn to assessing the consequences that this has for Worker productivity in Stage 2.

4.6.2 Result 4.1: Workers in Intentions Are Less Productive under Option B

I begin with a comparison of Worker behavior between Option A and Option B in the Intentions Treatment. First, I examine Workers' net productivity, which I define

as the difference between successful and unsuccessful identifications (*i.e.*, mistakes). Net productivity thus equals the number of Tokens by which the Employer's payoff increases. On average, Workers identify 32.3 coins, of which 25.3 are identified correctly and 7 are identified incorrectly. As a result, the average productivity across the entire experiment equals 18.4 Tokens. In line with Hypothesis 4.1, Figure 4.4a and Panel A of Table 4.4 show that the average net productivity is indeed lower under Option B (10.1 Tokens), as compared to Option A (23.4). A Mann-Whitney U test confirms the null of no differences (MWU: $z = 2.789, p = 0.005$.) Thus, Workers in the Intentions Treatment are less productive after the Employer has chosen Option B, as compared to Option A. It should be noted that Workers attain a positive net productivity on average even under Option B, which they report to dislike. I discuss potential explanations for this in Section 4.7.

Next, I examine *how* Workers under Option B retaliate against the Employer. This exercise could inform me whether counterproductive behavior is explicit or more subtle. After all, Workers could decrease productivity by identifying fewer coins in total or by making more mistakes. Importantly, Employers observe the number of correct *and* incorrect identifications, meaning that Workers could use this to signal their dissatisfaction with the Employers' choice. Arguably, making (deliberate) mistakes forms a stronger signal than simply being idle for four minutes.¹³ Figure 4.4b and Panel B of Table 4.4 show the number of mistakes made in each treatment-option combination. On average, Workers make 4.3 mistakes under Option A and 11 mistakes under Option B. A Mann-Whitney U test shows this difference to be significant at a 5%-level (MWU: $z = -2.319, p = 0.020$). At the same time, Workers do not work more slowly under Option B: they identify an identical number of coins under both Options (A: 32; B: 32).

Result 4.1 *The Worker achieves a higher productivity under Option A as compared to Option B. This difference seems to be driven by a higher number of mistakes made under Option B, rather than a lower number of identifications.*

Figures 4.20a and 4.20b in Appendix 4.C.1 display the distributions of net productivity and mistakes across options and treatments.

¹³As anecdotal evidence for this, one subject remarked that he/she "was mean to chose one-wrong-one-correct alternatively in the euro identification test."

4.6.3 Result 4.2: Workers Are Less Productive when Option B Is Intentionally Chosen

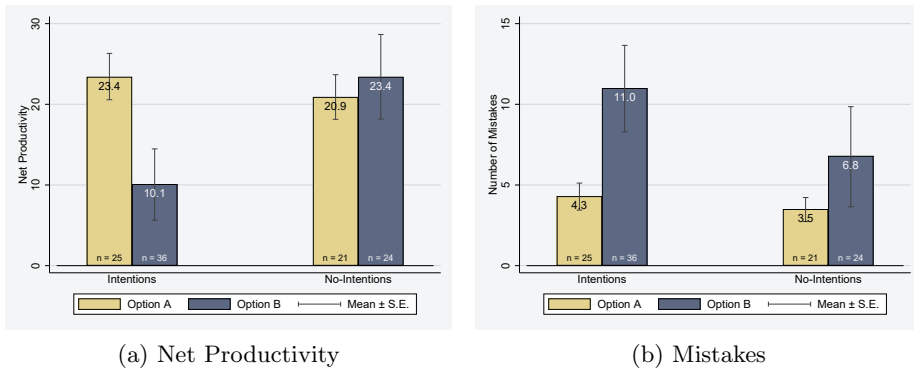
Next, I examine the role of the Employer's intentions by comparing the Intentions Treatment to the No-Intentions Treatment, in which one of the two options is implemented randomly. I do so by comparing the Worker's performance in each option across treatments. If the difference between Option A and Option B in the Intentions Treatment is caused solely by the benevolent or malicious intentions of the Employer, I should observe no differences between Option A and Option B in the No-Intentions Treatment. This appears to be the case: Workers achieve a Net Productivity of 20.9 Tokens under Option A and 23.4 Tokens under Option B, with the difference between the two going in the opposite direction and being insignificant (MWU: $z = -1.525$, $p = 0.130$). Moreover, I find that net productivity under Option B is significantly higher in the No-Intentions Treatment as compared to the Intentions Treatment (MWU: $z = 3.254$, $p = 0.001$), while there is no difference in net productivity under Option A across treatments (MWU: $z = -0.408$, $p = 0.690$). It thus appears that choosing Option B is punished by Workers, while choosing Option A is not rewarded. This suggests that, while Workers deem choosing Option B as a testament of the Employer's *bad* intentions, they do not think that choosing Option A reflects *good* intentions. Instead, they seem to think of Option A as the obvious choice for any Employer. In sum, I find mixed evidence in line with Hypothesis 4.2.

Result 4.2 *Compared to the Intentions Treatment, productivity in the No-Intentions Treatment, where Employer intentions are absent, is higher under Option B, while it is not lower under Option A.*

4.6.4 Regression Analysis

In order to control for background variables and test for heterogeneous treatment effects, I estimate an ordinary least-squares (OLS) model with standard errors clustered on the session level. Results are reported in Table 4.5. Model (1) only includes dummies for Option B and the No-Intentions Treatment, and an interaction between the two. As a result, Option A in the Intentions Treatment is the omitted category, meaning that *Option B* measures the effect of Option B being chosen in the Intentions Treatment, while *No-Intentions* measures the effect of Option A being implemented in the No-Intentions Treatment. Then, I include controls for being male, studying an Economics or Business program and years of experience with the Euro as a means of payment in Model (2). In Model (3), I add standardized measures of the SVO, risk

Figure 4.4: Worker Performance across Options and Treatments



Note: Average net productivity (Panel A) and number of mistakes (Panel B) across options and treatments. The error bars span the mean plus and minus one standard error of the mean.

tolerance, and reciprocity variables, and introduce interaction terms with Reciprocity in Model (4). Then, I examine the role of the Worker's mood in Model (5). Finally, instead of the preference variables, I include standardized big-five personality traits in Model (6).

Throughout all specifications, with the exception of Model (5), I find a significantly negative effect of Option B in the Intentions Treatment of roughly 13 Tokens. This result provides further evidence for Hypothesis 4.1: Workers in the Intentions Treatment attain a lower productivity under Option B as compared to Option A. The small and insignificant coefficient on *No-Intentions* shows that the difference in average productivity between treatments under Option A can be considered negligible. Hence, whether or not Option A is implemented intentionally does not affect a Worker's productivity significantly. At the same time, the coefficient on *Option B* \times *No-Intentions* is larger in absolute magnitude and significant at a 10%-level in some specifications. This resonates the earlier findings and provides partial evidence in support of Hypothesis 4.2: Workers seem to decrease their productivity following the intentional choice of Option B, while they do not increase their productivity following the intentional choice of Option A.

Adding control variables does not alter my results and none of them enter significantly. This applies to the personal characteristics in Panel A, the preferences in Panel B and the Big Five personality traits in Panel E. Interestingly, the significantly negative coefficient on *Reciprocity (std.)* \times *Option B* in Model (4) shows that subjects with a stronger reciprocal inclination decrease their productivity *more* when

Option B is intentionally chosen. Note that this is in line with the theoretical model of Section 4.4. To be precise, a Worker under Option B in the Intentions Treatment decreases her productivity by 13 Tokens more per standard deviation increase in the reciprocity variable. The insignificant coefficient on *Reciprocity (std.)* indicates that a Worker's degree of reciprocity does not affect her response to Option A being intentionally implemented. Moreover, the insignificant coefficient for *Option B* in Model (5) shows that Option B does not alter the Worker's net productivity if the Worker's mood is controlled for. Furthermore, the significantly positive coefficient on *Mood*, the significantly negative coefficient on *Mood × No-Intentions*, and the insignificant coefficients on the remaining two interaction terms together suggest that Workers with a better mood achieve a higher net productivity in the Intentions Treatment only. It should be noted that the direction of causality between mood and net productivity cannot be established with certainty, as a low productivity could reinforce an already bad mood. However, one would then expect to observe the same correlation within the No-Intentions Treatment, which is not the case. Hence, the exercises in Models (4) and (5) suggest that, even though negative mood is present in both treatments following Option B, a more negative mood and a more intense reciprocal inclination only translate into a stronger negative response in the Intentions Treatment, where the Employer can actually be blamed for having chosen Option B.

In Table 4.9 in Appendix 4.C.3, I perform the same analysis with mistakes as the dependent variable and obtain similar results. Furthermore, I show in Appendix 4.C.4 that the results become more pronounced when I winsorize net productivity and mistakes at a 5%-level. This approach takes care of the outliers that make more than 70 mistakes and drag down performance.

4.6.5 Stage 1 Outcomes following Option B

Workers who were faced with Option B differ in the outcome obtained in Stage 1: those who chose Alternative 1 retained the Donation and lost their own payoff, those who chose Alternative 3 retained their own payoff and lost their donation, and those who chose Alternative 2 either retained both or lost both. Across both treatments, 6 Workers (10%) choose Alternative 1, 26 Workers (43.3%) choose Alternative 2, and 28 Workers (46.7%) choose Alternative 3.¹⁴ In Appendix 4.C.5, I show that the Worker's choice of Alternative is consistent with the predictions of the model (laid down in Appendix 4.A.2): Workers with a higher SVO Angle are more likely to choose Alternative 1 or 2, and Workers with a higher Risk Tolerance are more likely

¹⁴A χ^2 test indicates no difference in the distribution across treatments.

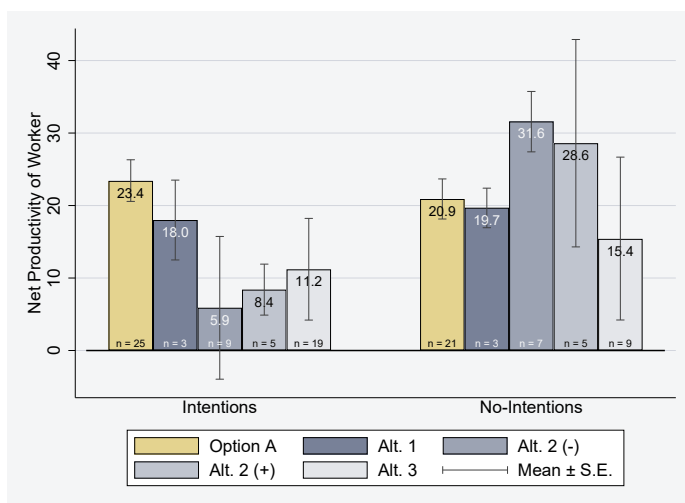
Table 4.5: Regression Analysis of Net Productivity

	DEPENDENT VARIABLE: NET PRODUCTIVITY					
	(1)	(2)	(3)	(4)	(5)	(6)
A. Treatment-Option						
Option B	-13.384** (5.425)	-13.131** (5.786)	-12.830* (6.541)	-13.774** (5.479)	-2.092 (6.629)	-12.424* (6.123)
No-Intentions	-2.535 (3.668)	-1.991 (3.776)	-0.723 (4.027)	-3.773 (3.627)	6.989 (5.041)	-3.243 (4.169)
Option B \times No-Intentions	15.896** (6.698)	14.474* (7.580)	12.777 (8.537)	16.280** (7.547)	1.023 (7.825)	14.796* (7.993)
B. Personal Characteristics						
Male		3.850 (4.873)	4.595 (5.563)	4.011 (5.722)	5.687 (5.792)	3.023 (4.849)
Economics & Business		4.623 (4.061)	3.582 (4.142)	2.084 (4.265)	4.538 (3.609)	3.375 (3.192)
Euro Years		0.141 (0.201)	0.147 (0.204)	0.142 (0.225)	0.237 (0.217)	0.045 (0.190)
C. Preferences						
SVO Angle (std.)			-0.042 (1.555)	1.252 (1.559)	0.390 (1.298)	
Risk Tolerance (std.)			0.313 (2.246)	-0.042 (2.214)	-0.186 (2.183)	
Reciprocity (std.)			-2.791 (2.003)	4.520 (3.003)	-1.574 (2.531)	
Rec. (std.) \times No-Intentions				-1.104 (3.660)		
Rec. (std.) \times Option B				-13.642*** (2.951)		
Rec. (std.) \times Option B \times No-Intentions				2.009 (8.632)		
D. Mood						
Mood					4.482* (2.426)	
Mood \times No-Intentions					-5.931* (2.876)	
Mood \times Option B					4.421 (4.799)	
Mood \times Option B \times No-Intentions					-1.571 (5.874)	
E. Big Five						
Conscientiousness (std.)						2.336 (1.797)
Openness (std.)						-1.631 (2.632)
Extraversion (std.)						1.868 (2.434)
Neuroticism (std.)						-2.234 (1.923)
Agreeableness (std.)						-1.260 (2.131)
Constant	23.440*** (3.101)	17.316*** (5.056)	17.363*** (5.605)	19.377*** (4.784)	8.987 (5.301)	19.957*** (4.301)
Observations	106	105	105	105	105	105
Clusters	17	17	17	17	17	17
R^2	0.07	0.10	0.11	0.19	0.20	0.14
F	2.23	3.66	2.96	24.54	11.93	11.50
df	16	16	16	16	16	16

Note: OLS model with net productivity as the dependent variable. Standard errors are clustered on the session level. One observation is dropped due to missing information on that subject's gender.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 4.5: Net Productivity per Stage 1 Outcome



Note: Average net productivity across treatments and outcomes. Alt. 2 (+) represents the Worker choosing Alternative 2 and winning the gamble, while Alt. 2 (-) represents the Worker choosing Alternative 2 and losing the gamble. The error bars span the mean plus and minus one standard error of the mean.

to choose Alternative 2.

Next, I examine whether differences in productivity in the Intentions Treatment can be traced back to the outcomes in Stage 1, or whether productivity is universally lower across outcomes. An important caveat of this exercise is the small sample size in each of the outcomes, which hampers a powerful and reliable statistical analysis. Therefore, the claims in this Subsection should be interpreted with caution. Moreover, it should be noted that Workers to a large extent self-select into the different outcomes, which implies that I cannot treat the relationship as causal: factors that might have driven the Worker to a particular Alternative may also have affected her productivity in Stage 2.

Figure 4.5 shows average net productivity under Option A and under each of the four outcomes following Option B in both treatments. The number at the base of each bar shows how Workers are distributed across Stage 1 outcomes in both treatments. As can be seen, net productivity in the Intentions Treatment is (slightly) lower in each Option B outcome as compared to Option A. The differences are significant for all outcomes (at a 10%-level, all $p < 0.060$), except Alternative 1. Most interestingly, the Worker's net productivity in the Intentions Treatment is low even if Stage 1 results in the most beneficial scenario for the Worker (*i.e.*, a successful Alternative 2

gamble). Moreover, this productivity level is similar to the least beneficial scenario (*i.e.*, an unsuccessful Alternative 2 gamble). This suggests that the Worker responds to the Employer's intentions, even though eventual payoff consequences are identical to the Employer opting for Option A. In the absence of intentions, net productivity is much closer to Option A for all outcomes, with the outcomes related to Alternative 2 even resulting in a significantly higher net productivity (at a 10%-level). Despite the small sample sizes, the observation that net productivity is consistently low across options reassures that the Result 4.1 is not driven by one particular outcome under Option B.

4.6.6 Other Results

Worker Expectations While the Employer was choosing his option, I asked the Worker whether she expected Option A or Option B to be chosen. Section 4.4 predicts that the extent to which the Worker interprets Option B as unkind and responds to this in the coin-identification task is increasing in the expectation that the Employer chooses Option A. As can be seen in Panel A of Table 4.6, only 17 out of 61 (28%) Workers expect Option A to be implemented, which suggests an average η of 0.28. In terms of my theoretical model, this would imply that, on average, Option A should be rewarded relatively *more* than Option B is punished. As stated above, this is not the case.

On an individual basis, the model predicts that those Workers expecting Option B would not punish Option B and reward Option A, while those Workers expecting Option A would punish Option B and not reward Option A. Again, I do not find this to be the case as can be seen in Panels B and C of Table 4.6. To begin with, Workers who expected Option A and Workers who expected Option B report similar moods, for both options *actually* chosen by the Employer. Moreover, when the Employer chooses Option B, the 9 Workers expecting Option A achieve a net productivity of 22.3 Tokens, while the 27 Workers expecting Option B attain a net productivity of only 6.0 Tokens. The direction of this difference is inconsistent with the theoretical model. This could suggest that Workers expecting Option A differ from Workers expecting Option B in other dimensions, for example, because they are more optimistic or altruistic in general.

Employer Expectations In a similar spirit, I asked Employers about the predicted productivity under both options (within-subjects). Interestingly, they respond in a manner that is consistent with their own choice. This can be seen in Panel A of Table

Table 4.6: Worker Behavior in Intentions Treatment by Expectations

A. Distribution				B. Mood						C. Net Productivity							
Exp.	Option			Exp.	Option				Exp.	Option							
	A	B	Total		A	B	Avg.			A	B	Avg.					
A	8	9	17	A	1.5	(0.4)	-0.8	(0.4)	0.3	(0.4)	A	23.0	(4.3)	22.3	(5.5)	22.6	(3.6)
B	17	27	44	B	1.4	(0.3)	-0.4	(0.2)	0.3	(0.2)	B	23.6	(3.6)	6.0	(5.3)	12.8	(3.8)
Total	25	36	61	Avg.	1.4	(0.3)	-0.5	(0.2)	0.3	(0.2)	Avg.	23.4	(2.8)	10.1	(4.4)	15.5	(3.0)

Note: The table contains the distribution of Workers across expectations and options (Panel A), self-reported Mood (Panel B) and Net Productivity (Panel C) across each expectation-option combination for Workers in the Intentions Treatment. The standard deviation is denoted in parentheses.

4.7, where I summarize the difference in beliefs about productivity under both options. In the Intentions Treatment, Employers who chose Option A predict productivity to be about 38 Tokens higher on average under Option A as compared to Option B, which would offset the gain from choosing Option B. At the same time, Employers who chose Option B predict a difference of only 19 Tokens, which would result in a net gain from choosing Option B. In other words, Employers choosing either option on average believe that their chosen option yields the highest payoff. Since this exercise was not incentivized, I cannot determine whether this reflects Employers' true beliefs or whether this is merely a manifestation of ex-post rationalization of the option chosen. In any case, it seems that Employers do understand the potential consequences of their actions for Worker productivity in Stage 2 and anticipate this when making their choice. Reassuringly, I observe only a slight insignificant difference in beliefs across Employers under Option A and Option B in the No-Intentions Treatment: 3 Tokens vs. 6 Tokens, respectively ($z = -0.726, p = 0.475$).

Employer Profits Finally, I briefly examine Employer profits, in order to see whether choosing Option B pays off on average. Panel B of Table 4.7 summarizes the Employers' total profit from the game. Remember that choosing Option B yields the Employer an immediate benefit of 30 Tokens, which may be offset by a substantial drop in the Worker's subsequent productivity. However, due to the modest productivity decrease (23.4 to 10.1 Tokens), Employers still earn more on average when choosing Option B in the Intentions Treatment: 90.1 vs. 73.4 Tokens. A Mann-Whitney U test of no differences is able to reject the null that earnings are the same for the two options at a 1%-level ($z = -3.794, p < 0.001$). This implies that the beliefs of Employers choosing Option A are too optimistic about the productivity gap between both Options.

Table 4.7: Employer Outcomes in Each Cell

A. Belief Difference							B. Total Profit						
Treatment	Option						Treatment	Option					
	A		B		Avg.			A		B		Avg.	
NI	2.6	(5.9)	6.0	(4.1)	4.4	(3.5)	NI	70.9	(2.7)	103.4	(5.1)	88.2	(3.9)
I	37.8	(9.0)	19.0	(7.0)	26.7	(5.6)	I	73.4	(2.8)	90.1	(4.4)	83.2	(3.0)
Avg.	21.7	(6.2)	13.8	(4.6)	17.2	(3.7)	Avg.	72.3	(2.0)	95.4	(3.4)	85.4	(2.4)

Note: The table contains belief differences (Panel A) and total profit (Panel B) in each treatment-option combination. Belief Difference is defined as the expected net productivity under Option A minus the expected net productivity under Option B. The standard deviation is denoted in parentheses.

4.7 Concluding Discussion

The current paper examines the effect of imposing an anti-social trade-off on a Worker on the subsequent productivity of said Worker. The results show that Workers achieve a significantly lower productivity when the Employer chooses to impose anti-social incentives (Option B), as compared to Employer abstaining from doing so (Option A). Perhaps surprisingly, net productivity of the Worker under Option B in the Intentions Treatment is still positive and equal to 10.1 Tokens. This could be a manifestation of residual altruism: even after adjusting her weight on the Employer's payoff, the Worker still cares enough about him to attain a positive net productivity. Alternatively, this could reflect that Workers enjoy performing the coin-identification task and derive intrinsic utility from performing the task well. This seems a plausible claim especially for those Workers who are apprehensive to make deliberate mistakes, since the alternative would be to remain idle for four minutes. At the same time, I observe that the difference is driven by mistakes, rather than fewer identifications. This could be a manifestation of Workers signaling their dissatisfaction to the Employer. The extent to which Workers make mistakes may also correlate with their personality and image concerns. When I check this by regressing the number of mistakes made on observable characteristics and Big Five traits, I do not find any of these to predict mistakes.

By means of the No-Intentions Treatment, I show that intentions play an important role in explaining the productivity gap between both options, with the gap disappearing when the Employer has no control over the option implemented. Comparing productivity under the same option across treatments shows that the Worker decreases her productivity when Option B is intentionally chosen, as opposed to being randomly implemented, while she does not increase productivity when Option A is intentional chosen. This suggests that negative reciprocity is at work, while positive reciprocity is absent, which is in line with the results from previous studies. For example, Kube et al. (2013) show that wage cuts damage worker morale, while wage

raises have no positive effect, and Offerman (2002) shows that subjects respond much more strongly to unkind acts than to kind acts. The absence of positive reciprocity in my experiment suggests that the Worker feels entitled to keeping their endowment and deems Option A as the only acceptable option. On the one hand, this is not too surprising, given that the Worker's payoff is framed as an endowment that may be destroyed and Option A basically comprises maintaining the status quo. In terms of my theoretical model, this would imply that η_i lies close to one on average.¹⁵ On the other hand, most Workers expect Option B to be chosen and those Workers expecting Option B also achieve the lowest net productivity when it is actually implemented. This pattern is inconsistent with the theoretical model, as the model predicts the reciprocal response to Option B to be strongest among those Workers not expecting it, although it must be acknowledged that my measure of the Worker's expectation is rather crude. Taken together, this suggests that initial property rights play an important role in determining the reciprocal response. Therefore, future efforts could be invested in tweaking the framing of the two options and further examining the role of Worker beliefs in affecting the reciprocal response.

Another interesting extension would be to vary the immediate gain from choosing Option B. With the current constellation of parameters, the Employer is better off choosing Option B, as the gain from doing so is higher than the loss in productivity. A higher gain would not only make Option B more attractive; it may also make it more acceptable from the Worker's point of view. Analogously, choosing Option A may be deemed more praiseworthy by the Worker, as the Employer foregoes a larger gain. In other words, an increased immediate gain could affect lower the parameter η_i measuring the expectation of Option A being chosen, and the reciprocal response following Option B (Option A) may be weaker (stronger, respectively) as a result. Furthermore, a higher immediate gain may remove the destruction of resources which characterizes Option B. Currently, this waste of resources may form an additional motive to punish for an efficiency-minded Worker. Thus, part of my result could be driven by efficiency concerns, and, although a realistic feature of organizational malpractices, it would be interesting to see whether I find the same results when the two options are identical in terms of (expected) total surplus in Stage 1.

My results exemplify that Worker productivity is affected by how she is treated by her Employer. This stresses the importance of leadership within organizations. Obviously, an actual work environment is richer than the abstract setting of my

¹⁵Average net productivity in the No-Intentions Treatment of 22.2 is 1.2 Tokens lower than Option A in the Intentions Treatment and 12.1 Tokens higher than Option B in the Intentions Treatment. A back-of-the-envelope calculation using Proposition 4.1 then shows that $\eta = \frac{12.1}{12.1+1.2} = 0.91$.

experiment. On the one hand, Workers in actual firms may be held accountable for their drop in productivity, which is why they would refrain from counterproductive behavior. On the other hand, in an actual work environment, workers have many more ways in which they can retaliate against organization, and it is likely that they will somehow find a way to do so unpunished. In this sense, my study fits in with previous studies showing that workers may reciprocate in whichever dimension possible, even if it is a different domain that the initial act of (un)kindness (Alempaki et al., 2019; Belot and Schröder, 2015).

Another conceptual contribution of this chapter concerns the fact that the Worker still holds her destiny in her own hands: after the Employer has chosen Option B, the Worker still can ensure a payoff to herself and, with some luck, can even end up in a situation in which the outcome is the same as under Option A. I have shown that Worker productivity remains low when this occurs, which suggests that even if actions do not result in actual consequences, the Employer is still blamed for his intentions. This is in line with experiments showing that subjects respond differently to identical outcomes depending on the intentions of the counterparty (Charness and Rabin, 2002; Falk et al., 2003; Cox, 2004; Charness and Levine, 2007; Sebald, 2010). For example, Charness and Levine (2007) show that agents may react differently to outcome-equivalent situations depending on the history of past choices and random draws: an agent's effort is lower when the outcome was the result of the principal's malicious intentions, rather than of bad luck. As explained above, my experiment contains an additional step where the agent (*i.e.*, the Worker) makes a choice and she might blame the employer for putting her on the spot in the first place.

In sum, the current study exemplifies the scope for studying reciprocity originating from other elements in the worker-employer relationship than monetary benefits.

Appendices

4.A Theoretical Derivations

4.A.1 Derivation of Kindness Term

Here, I derive the kindness term $r_i(t, o_j)$, as displayed in Equation (4.2). For ease of exposition, I define the linear combination of own monetary payoffs and the donation as:

$$m_i(o_j, a_i) \equiv \pi_W(o_j, a_i) + \theta_i^D D(o_j, a_i) \quad (4.11)$$

In case the Employer chooses Option A, the Worker does not need to make a choice, which implies:

$$m_i(A, a_i) = m_i(A) = 50 + 50\theta_i^D \quad (4.12)$$

I then capture Employer kindness in the following formula, taken from Cox et al. (2007):

$$r_i(t, o_j) = \mathbb{1}_{t=I} \cdot \frac{\max_{a_i} m_i(o_j, a_i) - m_i(o_i^0, a_i)}{m_i(A) - \max_{a_i} m_i(B, a_i)} \quad (4.13)$$

To begin with, the indicator function $\mathbb{1}_{t=I}$ ensures that the kindness term only enters the utility function in the Intentions Treatment ($t = I$) and not in the No-Intentions Treatment ($t = NI$). As a result, $r_i(NI, o_j) = 0$ if intentions are absent. Then, the numerator $\max_{a_i} m_i(o_j, a_i) - m_i(o_i^0, a_i)$ represents the kindness of the Employer's choice, which is the difference between the highest payoff that the Worker can ensure for herself under the Employer's actual choice and the payoff attained under some neutral choice by the Employer ($m_i(o_i^0, a_i)$). The denominator $m_i(A) - \max_{a_i} m_i(B, a_i)$ measures the distance in payoff under Option A and the highest attainable $m_i(\cdot)$ for the Worker under Option B, and ensures that $r_i(t, o_j)$ is between -1 and 1.

The neutral choice o_i^0 merits discussing further. As shown above, I employ the individual-specific parameter $\eta_i \in [0, 1]$ capturing the Worker's belief that the Employer chooses Option A, so the linear combination $m_i(o_i^0, a_i) = \eta_i m(A) + (1 - \eta_i) \max_{a_i} m_i(B, a_i)$ can then be interpreted as the Worker's expectation *a priori* about Stage 1 to which she compares the actual outcome. When η_i is close to one, the Worker expects Option A to be chosen with a high probability, while

she expects Option B to be implemented when η_i is close to zero. As a result, a higher η_i raises the expectation $m_i(o_i^0, a_i)$, which in turn makes choosing Option A less kind and makes Option B more unkind. This more flexible approach differs from previous literature, which has often taken the average payoff under the most and least favourable alternative ($\eta_i = 0.5$) or the original property rights ($\eta_i = 1$) as $m(o_i^0)$ (see *e.g.*, Dufwenberg and Kirchsteiger, 2004; Cox et al., 2007). Since I can write $m_i(o_i^0, a_i) = \eta_i m_i(A) + (1 - \eta_i) \max_{a_i} m_i(B, a_i)$ and $m_i(o_j, a_i) = \mathbb{1}_{o_j=B} \cdot \max_a m_i(B, a_i) + (1 - \mathbb{1}_{o_j=B}) \cdot m_i(A)$, we can then simplify Equation (4.13) in the following way, where the indicator function $\mathbb{1}_{o_j=B}$ takes on value 1 if Option B is implemented, and 0 otherwise.

$$r_i(t, o_j) = \mathbb{1}_{t=I} \cdot \left(\frac{\mathbb{1}_{o_j=B} \cdot \max_{a_i} m_i(B, a_i) + (1 - \mathbb{1}_{o_j=B}) \cdot m_i(A)}{m_i(0) - \max_{a_i} m_i(B, a_i)} - \frac{\eta_i \cdot m_i(A) + (1 - \eta_i) \cdot m_i(B, a_i)}{m_i(0) - \max_{a_i} m_i(B, a_i)} \right) \quad (4.14)$$

$$= \mathbb{1}_{t=I} \cdot \frac{(1 - \mathbb{1}_{o_j=B} - \eta_i) (m_i(A) - \max_{a_i} m_i(B, a_i))}{m_i(A) - \max_{a_i} m_i(B, a_i)} \quad (4.15)$$

$$= \mathbb{1}_{t=I} (1 - \mathbb{1}_{o_j=B} - \eta_i) \quad (4.16)$$

$$= \begin{cases} 1 - \eta_i & \text{if } t = I \text{ \& } o_j = A \\ 0 & \text{if } t = NI \\ -\eta_i & \text{if } t = I \text{ \& } o_j = B \end{cases} \quad (4.17)$$

4.A.2 Worker Choice of Alternative

Below, I analyze Worker i 's choice of Alternative a_i under Option B. From Section 4.4, remember that utility of the Worker takes the following form:

$$U_i^W(o_j, a_i, e_i) = \pi_W(o_j, a_i) + \theta_i^D D(o_j, a_i) + (\theta_i^E + \rho_i r_i(t, o_j)) \pi_E(o_j, e_i) - c(e_i) \quad (4.1)$$

The Alternatives have the following consequences for the Worker and the Red Cross:

Alt. 1 $\pi_W(B, 1) = 0$, $D(B, 1) = 50$

Alt. 2 $\pi_W(B, 2) = 0$, $D(B, 2) = 0$ with probability $p = 0.4$

$\pi_W(B, 2) = 50$, $D(B, 2) = 50$ with probability $p = 0.6$

Alt. 3 $\pi_W(B, 3) = 50$, $D(B, 3) = 0$

The Employer's payoff in Stage 1 always equals 80 when Option B is implemented, irrespective of the Alternative chosen by the Worker. The Worker chooses the Alternative with the highest expected payoff. Risk-neutrality implies that the expected utility from choosing Alternative 2 is $0.6(50 + \theta_i^D \cdot 50) + 0.4(0 + \theta_i^D \cdot 0) = 30(1 + \theta_i^D)$. Then, it can be shown that the Worker chooses:

$$a_i^* = \begin{cases} 1 & \text{if } \theta_i^D > \frac{3}{2} \\ 2 & \text{if } \frac{3}{2} \geq \theta_i^D > \frac{2}{3} \\ 3 & \text{if } \theta_i^D \leq \frac{2}{3} \end{cases} \quad (4.18)$$

Thus, the Worker chooses for her own payoff if and only if she cares insufficiently for the charity. Similarly, she chooses the donation if and only if she cares strongly for the charity. For intermediate values of θ_i^D , she is willing to gamble on keeping both.

4.A.3 Employer Behavior

Employer j chooses between Option A and Option B. Since Worker i 's θ_i^E and θ_i^D are private information to Worker i , the Employer forms belief $\widehat{\theta}_{j,i}^E$ and $\widehat{\theta}_{j,i}^D$, and $\widehat{e}_{j,i}^0 = \frac{\widehat{\theta}_{j,i}^E}{\mu}$. Furthermore, the Employer anticipates a net productivity of $\widehat{e}_{j,i}^A$ and $\widehat{e}_{j,i}^B$ following Option A and Option B, respectively. Regarding the Alternative, the Employer is aware of the decision rule from Equation (4.18), and expects the Worker to have $\theta_i^D < \frac{2}{3}$ and choose Alternative 3 with probability $F(2/3) \equiv F_3$. Similarly, he expects the Worker to have $\theta_i^D \geq \frac{3}{2}$ and choose Alternative 1 with probability $1 - F(3/2) \equiv F_1$. Alternative 2 is chosen with probability $1 - F(2/3) - (1 - F(3/2)) \equiv 1 - F_1 - F_3$. The Employer's utility is:

$$U_j^E(o_j, e_i) = \begin{cases} 50 + e_i^A + 50\theta_j^D + 50\theta_j^W \\ 80 + e_i^B + F_1 \cdot 50\theta_j^D + (1 - F_1 - F_3) \cdot (30\theta_j^D + 30\theta_j^W) + F_3 \cdot 50\theta_j^W \end{cases} \quad (4.19)$$

$$= \begin{cases} 50(1 + \theta_j^D + \theta_j^W) + e_i^A \\ 80 + e_i^B + 10((3 + 2F_1 - 3F_3)\theta_j^D + (3 + 2F_3 - 3F_1)\theta_j^W) \end{cases} \quad (4.20)$$

Hence, the Employer chooses Option A if and only if the utility from doing so is higher than the utility from Option B. In doing so, he forms beliefs $\widehat{e}_{j,i}^A$ and $\widehat{e}_{j,i}^B$.

$$50(1 + \theta_j^D + \theta_j^W) + e_i^A > 80 + e_i^B + 10((3 + 2F_1 - 3F_3)\theta_j^D + (3 + 2F_3 - 3F_1)\theta_j^W) \quad (4.21)$$

$$\widehat{e}_{j,i}^A - \widehat{e}_{j,i}^B > 30 - (20 - 10(2F_1 - 3F_3))\theta_j^D - (20 - 10(2F_3 - 3F_1))\theta_j^W \quad (4.22)$$

4.A.4 Risk Preferences

I now add risk aversion to the model. I do this in the following quasi-linear way. I subject the linear combination of the Worker's payoff and the Donation ($m_i(o_j, a_i)$) to a constant relative risk aversion transformation $m_i(o_j, a_i)^\beta$, with $\beta > 0$. Note that $\beta = 1$ implies risk-neutrality, while $\beta < 1$ implies risk-aversion, and $\beta > 1$ implies risk-lovingness. In the bomb-risk elicitation, 15 percent of the subjects display risk-loving preferences (*i.e.*, collect more than 50 boxes), which is why I allow for values of β larger than one. This yields the following utility function:

$$U_i^W(o_j, a_i, e_i) = [\pi_W(o_j, a_i) + \theta_i^D D(o_j, a_i)]^\beta + (\theta_i^E + \rho_i r_i(t, o_j))\pi_E(o_j, e_i) - \frac{\mu}{2} e_i^2 \quad (4.23)$$

Since the reciprocal response is independent from the Alternative chosen, I can restrict attention to the term $m_i(B, a_i)^\beta \equiv M_i^W(B, a_i)$. As a result, I can write:

$$M_i^W(B, a_i) = \begin{cases} (50\theta_i^D)^\beta & \text{if } a_i = 1 \\ \frac{3}{5}(50(1 + \theta_i^D))^\beta & \text{if } a_i = 2 \\ 50^\beta & \text{if } a_i = 3 \end{cases} \quad (4.24)$$

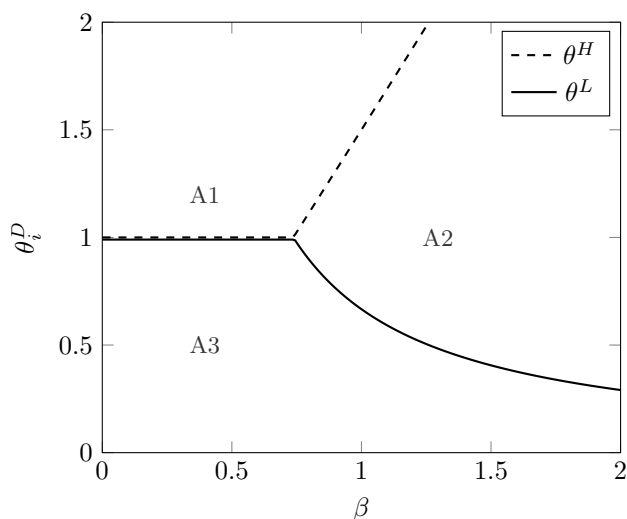
Standard algebra implies that the Worker chooses:

$$a_i = 1 \text{ if } \theta_i^D > \max \left\{ 1, \frac{3^{\frac{1}{\beta}}}{5^{\frac{1}{\beta}} - 3^{\frac{1}{\beta}}} \right\} \equiv \theta^H \quad (4.25)$$

$$a_i = 2 \text{ if } \theta^L < \theta_i^D \leq \theta^H \quad (4.26)$$

$$a_i = 3 \text{ if } \theta_i^D \leq \min \left\{ 1, \frac{5^{\frac{1}{\beta}} - 3^{\frac{1}{\beta}}}{3^{\frac{1}{\beta}}} \right\} \equiv \theta^L \quad (4.27)$$

On the interval $\beta \in (0, 2]$ and $\theta_i^D \in (0, 2]$, θ^H and θ^L follow the pattern below:



As can be seen from the graph, sufficiently risk-averse Workers will never choose the lottery, because they prefer to opt for one of the certain outcomes. In this case, they opt for Alternative 1 if and only if they weigh the donation more heavily than their own payoff: $\theta_i^D > 1$. For low degrees of risk-aversion, the Worker chooses Alternative 2 for values of θ_i^D that are large enough to stay away from destroying the donation with certainty (Alternative 3) and small enough to not prefer to keep the donation with certainty (Alternative 1).

Stage 2 Conveniently, this specification does not affect Stage 2 behavior of the Worker. Note that a utility function in which the Employer's payoff would also be subject to a constant relative risk aversion parameter would inhibit an internal solution, due to the convex effort costs.

4.B Experimental Materials

4.B.1 Instructions

4.B.1.1 General Instructions

Instructions

Thank you for participating in this decision-making experiment. All decisions that you make during the experiment are completely anonymous. Your final earnings depend on the choices made by you and others in this experiment, as well as on chance. We denote earnings in Tokens. The exchange rate is **10 Tokens = 1 Euro (or 1 Token = 10 Eurocents)**.

Today's experiment consists of three Parts. You receive a show-up fee of 4 Euro. On top of this, you receive your Earnings from Part 1 and your Earnings from **either Part 2 OR Part 3** (so, only one of the two). To determine this, the experimenter tosses a coin at the end of the session. Please be reminded that any form of communication with other participants is prohibited and leads you to be removed from the experiment.

You receive instructions for each Part before the start of the Part. The Parts are independent from each other; decisions and outcomes in one Part are unrelated to any of the other Parts.

We now continue with the instructions of Part 1.

While the experiment is running, all computers are connected to the same Zoom session with audio muted and camera off. If you have a question, you can use the chat function to talk to the experimenter. This way, the experimenter does not need to approach you to answer your question. Please do only use this option when you are sure that you cannot find the answer in the instructions.

4.B.1.2 Instructions and Materials Production Game

Part 1 – Production Game

Part 1 consists of two stages. You are paired with the **same participant** for both stages. Your earnings for Part 1 equal your total earnings over the two stages. One of you is assigned the role of **Player 1** (we will assume he is male in the instructions), the other that of **Player 2** (female). Each pair of participants is also matched to a project of the International Red Cross. All Red Cross projects provide humanitarian aid in different parts of the world. **Only in Stage 1**, your choices may have actual consequences for a Donation made to your project. Importantly, each project has **at most** one pair matched to it. We reveal your pair's project at the end of Part 1, but only Player 2 knows the Donation.

Stage 1: Both Players and the Red Cross project start Stage 1 with **Earnings of 50 Tokens**. First, Player 1 chooses between Option A and Option B. If he chooses Option A, Stage 1 ends immediately and the Earnings of both Players and the Red Cross remain equal to 50 Tokens. If Player 1 chooses Option B, **his Earnings** increase to 80 Tokens and Player 2 needs to choose one of three Alternatives:

- Alt. 1:** Player 2's Earnings decrease from 50 to 0 Tokens, the Donation remains 50 Tokens.
- Alt. 2:** Player 2's Earnings **and** the Donation decrease from 50 to 0 Tokens with a 40% percent probability. They both remain equal to 50 Tokens with a 60% probability. A random procedure determines which of the two outcomes occurs.
- Alt. 3:** Player 2's Earnings remain 50 Tokens, the Donation decreases from 50 to 0 Tokens.

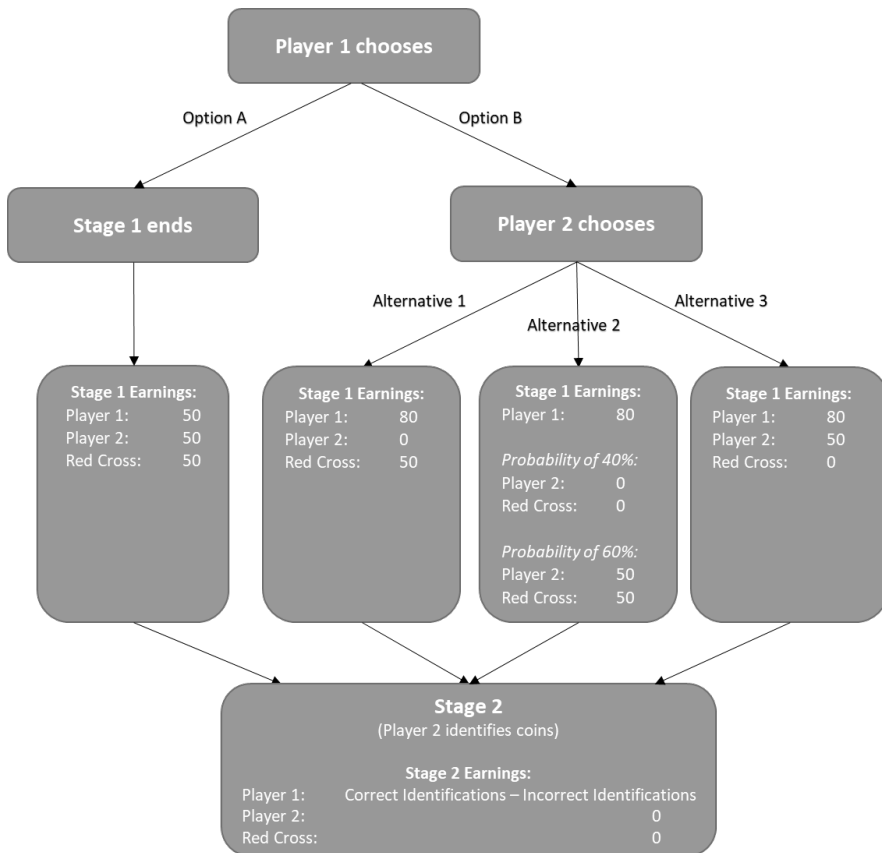
Player 1 **never** observes Player 2's chosen Alternative

Stage 2: Player 2 is shown the Heads side of different Euro coins on the screen and she is asked to identify the coins according to their value and country of origin. To do so, Players receive a hard-copy catalogue detailing the characteristics of the different Euro coins. Included are 2 Euro, 1 Euro, 50 Eurocents and 20 Eurocents coins from Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal and Spain. Player 2 does **not** receive any additional Earnings for this task. However, for every correct identification, the Earnings of Player 1 **increase** by 1 Token. An identification is correct when both the value **and** the country of origin are indicated correctly. For every incorrect identification, the

Earnings of Player 1 **decrease** by 1 Token. Player 2 receives 4 minutes to identify as many Euro coins as she wants (the maximum is 75 coins). During the task, Player 2 does not receive feedback; both Players are shown the number of correct and incorrect identifications by Player 2 after time has run out.

The Figure on the next page illustrates Part 1 graphically. On the screen, we first show a detailed preview of Part 1, including comprehension questions and a trial round of the coin identification task, before assigning the roles of Player 1 and Player 2.

Figure 4.6: Production Game



Note: Earnings in Tokens (1 Tokens is 10 Eurocents)

4.B.1.3 Coin Identification Catalogue

Figure 4.7: Page 1 of Catalogue (Discolored)



Figure 4.8: Page 2 of Catalogue (Discolored)

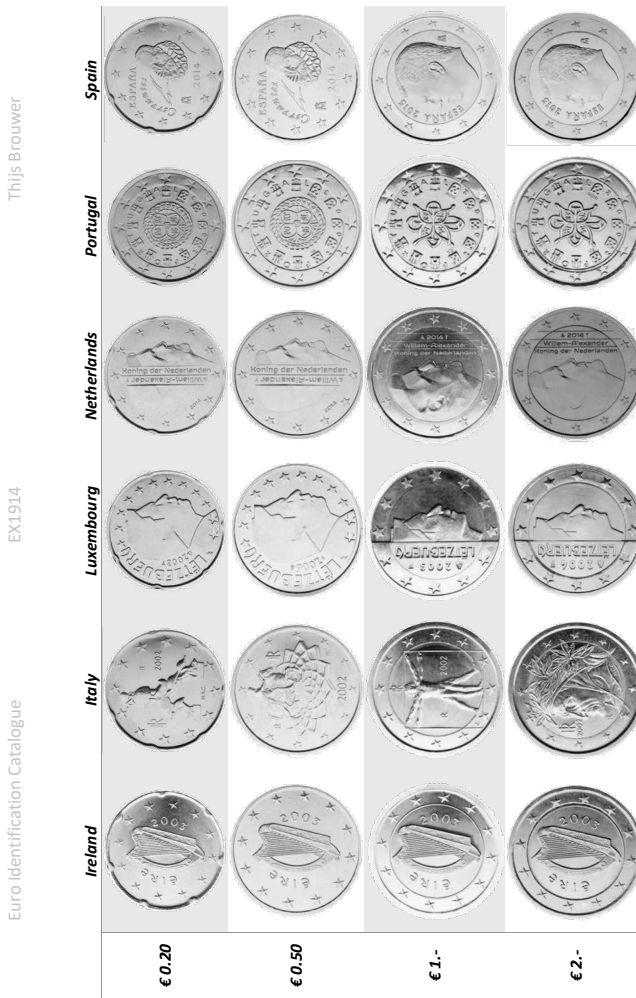






Figure 4.9: Page 3 of Catalogue (Discolored)

Thijs Brouwer

EX1914

Euro Identification Catalogue

Denominations

Denomination	Image	Description
€ 0.20		<p>Shape: Spanish flower shape Colour: Gold Composition: Nordic gold</p>
€ 0.50		<p>Shape: Round Colour: Gold Composition: Nordic gold</p>
€ 1.-		<p>Shape: Round Colour: Outer part: gold; inner part: silver Composition: Outer part: nickel brass; inner part: three layers: copper-nickel, nickel, copper-nickel</p>
€ 2.-		<p>Shape: Round Colour: Outer part: silver; inner part: gold Composition: Outer part: copper-nickel; inner part: three layers: nickel brass, nickel, nickel brass</p>

4.B.1.5 Instructions SVO Task

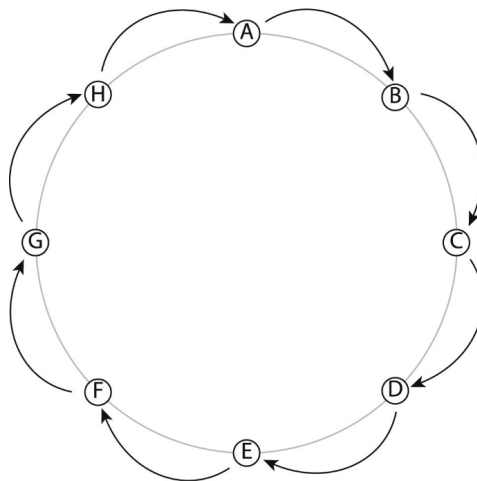
Part 3 – Choice Task

In Part 3, we ask you to make six choices. Each choice requires you to distribute Tokens between yourself and a randomly selected participant in this experiment. For each choice, you choose between nine possible allocations of Tokens. The matching of participants is shown in the Figure below: Participant A allocates Tokens between him-/herself and Participant B, Participant B allocates Tokens between him-/herself and Participant C, and so on. The order of participants is determined randomly. Please note that there are more Participants in this experiment than shown in the Figure.

For your payment of this Part, we select **one of the six** choices at random and implement the allocation chosen. For each participant, a different choice may be selected. Suppose that you are **Participant B** in the Figure and that Choice 3 is selected for you. Then, you earn the Tokens allocated to yourself and **Participant C** earns the Tokens that you allocated to him/her in Choice 3. In addition, you earn the Tokens that **Participant A** allocated to you in the choice that was selected for him/her.

We randomly select the choice and announce your Earnings for Part 3 **at the end of the experiment**. Please be reminded that each Token is worth 10 Eurocents.

Figure 4.11: Matching of Participants in Part 3



4.B.2 Survey Questions

PAGE 1

- What is your birth year (please write your birth year as “YYYY”)?
- What is your gender?
 - Female
 - Male
 - Other
 - I prefer not to say
- At which school are you studying?
 - Tilburg School of Economics and Management (TISEM)
 - Tilburg Law School (TLS)
 - Tilburg School of Social and Behavioral Sciences
 - Tilburg School of Humanities and Digital Sciences (THSD)
 - Tilburg School of Catholic Theology
 - University College Tilburg
 - TIAS School for Business and Society
 - Not at any school
 - I am on exchange
- What is your major/study program?
- What is your nationality?

PAGE 2 How well do the following statements describe you as a person? Please indicate your answer on a scale from 1 to 7, where 1 means “does not describe me at all” and a 7 means “describes me perfectly”.

- **(Positive Reciprocity)** When someone does me a favor I am willing to return it.
- **(Negative Reciprocity)** If I am treated very unjustly, I will take revenge at the first occasion, even if there is a cost to do so.

How willing are you to act in the following ways? Please again indicate your answer on a scale from 1 to 7, where 1 means you are “completely unwilling to do so” and 7 means you are “very willing to do so”.

- **(Negative Reciprocity 2)** How willing are you to punish someone who treats you unfairly, even if there may be costs for you?
- **(Indirect Reciprocity)** How willing are you to punish who treats others unfairly, even if there may be costs for you?

PAGE 3

- **(Altruism)** How willing are you to give to good causes without expecting anything in return? Please again indicate your answer on a scale from 1 to 7, where 1 means you are “completely unwilling to do so” and 7 means you are “very willing to do so”.
- **(Giving)** How much do you donate to charity each month (approximately in Euro)?
 - Nothing
 - Between 1 and 10 Euro
 - Between 11 and 20 Euro
 - Between 21 and 50 Euro
 - More than 50 Euro

PAGE 4 The Euro was first introduced as a means of payment in 2002 in Belgium, France, Germany, Ireland, Italy, Luxembourg, Netherlands, Greece, Spain, Portugal, Austria and Finland. Slovenia joined in 2007, Cyprus and Malta in 2008, Slovakia in 2009, Estonia in 2011, Latvia in 2014, and Lithuania in 2015.

- **(Euro Years)** For how many years have you been living in a Euro country in total during your life (only count the years during which the Euro was actually used as a means of payment)?
- Please complete the following sentence for each of the scales: “I thought that the Euro Identification Task was...”
 - Very easy - - - Very difficult (seven-point scale)
 - Not effortful at all - - - Very effortful (seven-point scale)
 - Very boring - - - Very exciting (seven-point scale)

PAGE 5 For each of the descriptions below, please indicate to what extent they apply to you on a scale from 1 to 7, where 1 means “does not apply to me at all” and 7 means “applies to me perfectly”.

“I see myself as someone who...

C ... does a thorough job.”

E ... is communicative, talkative.”

A- ... is sometimes somewhat rude to others.”

O ... is original, comes up with new ideas.”

N ... worries a lot.”

A ... has a forgiving nature.”

C- ... tends to be lazy.”

- E ... is outgoing, sociable.”
- O ... values artistic experience.”
- N ... gets nervous easily.”
- C ... does things effectively and efficiently.”
- E- ... is reserved.”
- A ... is considerate and kind to others.”
- O ... has an active imagination.”
- N- ... is relaxed, handles stress well.”

C = Conscientiousness, E = Extraversion, A = Agreeableness, O = Openness, N = Neuroticism. Items that reversely predict the trait are indicated with a “-”

PAGE 6 Open remark field

PAGE 7 Payment details

4.B.3 List of Red Cross Projects in the Experiment

Table 4.8: Red Cross Projects

Project	Amount	Description
Humanitarian Aid and Famine Prevention in Yemen	€ 55	<i>“As the brutal conflict continues to rage in Yemen, the need for emergency aid grows. Years of conflict, drought and instability have left thousands of Yemenis struggling to get food, water and medicine. The continued airstrikes and fighting on the ground are putting many innocent lives in peril. Despite the danger and obstacles, we are supplying hospitals and clinics with medicines and emergency supplies to treat the wounded and sick. But there is so much more to do. With your help we can continue delivering life-saving aid.”</i> (Source: International Committee of the Red Cross, Yemen Crisis Appeal)
Humanitarian Aid after Armed Conflict in Syria	€ 40	<i>“Eight years of violence have brought death and destruction to Syria. Millions of people have been forced from their homes or have fled the country. Many of them are children. In spite of the dangers, the International Committee of the Red Cross has been providing food and life-saving support to Syrians since the beginning. We, alongside the Syrian Arab Red Crescent, are one of the few humanitarian organizations that can help people in hard-to-reach areas. But the scale of the crisis is greater than anything we have faced in the last 15 years. We must do more.”</i> (Source: International Committee of the Red Cross, Syria Crisis Appeal)

Project	Amount	Description
Rebuilding Post-Conflict Iraq	€ 15	<i>“The devastating conflict in Iraq has left over 3 million people homeless and in desperate need. Although the fighting is now over in Mosul, the people there still desperately need our help. Many cannot return to their neighbourhoods as their homes are no longer standing. Water stations, schools and hospitals have also been destroyed. We are ready to help people rebuild their lives in Mosul with aid, including food and medical supplies. The people of Iraq need our help now more than ever.”</i> (Source: International Committee of the Red Cross, Iraq Crisis Appeal)
Helping Vulnerable Regions Battling the Outbreak of CoViD-19	€ 5	<i>“COVID-19 represents a dramatic threat to life in war-torn countries. The impact of an outbreak of the virus could be nothing short of catastrophic. Health systems in areas of conflict are already severely strained, where millions do not have access to basic health care. Fleeing violence, people seek shelter in crowded camps with inadequate sanitation. For people in such fragile situations, the virus is yet another threat to their lives. They need your urgent support.”</i> (Source: International Committee of the Red Cross, CoViD-19 Emergency Appeal)

Project	Amount	Description
Helping People Affected by the Explosion in Beirut (Lebanon)	€5	<i>“The horrific explosion in Beirut has caused dramatic loss of life, countless injuries, widespread destruction across the city and has left people in need of medical help, shelter and assistance in locating their loved ones. The city and people are shaken to the core. The ICRC is on the ground, working alongside the Lebanese Red Cross (LRC) and other Red Cross Red Crescent (RCRC) partners. We’re supporting hospitals with much-needed emergency medical supplies and treating the wounded, we’re helping families trace their loved ones, we’re making sure that the dead are identified and treated with dignity, and we’re fixing damaged water infrastructure. We are following the situation closely and will be providing more support as needed, working closely with the LRC and other RCRC partners to ensure a coordinated response.”</i> (Source: International Red Cross, Donate to Lebanon)
Natural Disaster Prevention in Developing Countries (worldwide)	€10	<i>“Why wait with providing help until disaster strikes? Simple measures such as installing warning systems or planting mangrove forests may save lives and prevent or reduce damage to homes, facilities and infrastructure. The Prinses Margriet Fonds (Princess Margriet Fund), together with the Netherlands Red Cross, invests in preventive measures in order to make sure that natural phenomena do not become natural disasters.”</i> (Source: The Netherlands Red Cross, Prinses Margriet Fonds)

Project	Amount	Description
Providing Life-Saving Aid to Venezuelan Families	€ 25	<p><i>“The situation in Venezuela is becoming more complex with every passing day, where ordinary families are left feeling the brunt of the increase in violence. Many have fled the country, but those who have stayed are in a vulnerable position and face an uncertain future. We are on the ground in Venezuela and we are scaling up our response to provide more life-saving aid and support. Every donation towards our operations brings some hope to the men, women and children who truly need our help.”</i> (Source: International Committee of the Red Cross, Venezuela Crisis Appeal)</p>
Supporting Physical Rehabilitation Centers in War-Torn Regions (world-wide)	€ 50	<p><i>“At our physical rehabilitation centres, we have been providing artificial limbs and physiotherapy to people with disabilities in areas affected by conflict for 30 years. But these are just the first steps towards social reintegration. Alongside these services, we have introduced sport as a way of combining physical rehabilitation, social inclusion and fun, as well as programmes for education, employment and vocational training. Every year, more and more people come to our physical rehabilitation centres looking for help. You can give adults and children with disabilities a chance to rebuild their lives. With your donation, we will support these patients for as long as it takes for them to regain their independence.”</i> (Source: International Committee of the Red Cross, Physical Rehabilitation Appeal)</p>

Project	Amount	Description
Family Re-union in Conflict-Affected South Sudan	€ 40	<i>“Since the conflict broke out in South Sudan, hundreds of thousands of men, women and children have been forced to flee their homes. Millions of people are in dire straits. The ICRC and Red Cross volunteers are working hard every day to bring them life-saving aid, such as food, water and shelter. Many face starvation. Your donation will help us deliver this urgently needed aid in South Sudan.”</i> (Source: International Committee of the Red Cross, South Sudan Crisis Appeal)
Providing Aid amidst the Violence around Lake Chad	€ 15	<i>“The violence in the region now affects all four countries (Nigeria, Chad, Niger, Cameroon) around Lake Chad. Civilians have been targeted and killed in the crisis. Over 2.4 million have been forced to flee their homes, while millions are in need of food, water, shelter and access to health care. The ICRC is building shelters, distributing food and essential household items, providing access to health care and helping families separated by the fighting to get back in touch.”</i> (Source: International Committee of the Red Cross, Lake Chad Crisis Appeal)

Project	Amount	Description
Providing Refuge to Families from the Rakhine area (Myan- mar)	€ 25	<i>“The conflict in Rakhine has forced many men, women and children to abandon their homes and flee the violence. In the rush to reach safety, they have had to leave everything they own behind. With scarce food, water and shelter, the already harsh conditions are becoming unbearable. We are working around the clock to help everyone affected by the fighting. Families are suffering, and we must be there to support them all. With your help we can continue delivering life-saving aid.”</i> (Source: International Committee of the Red Cross, Myanmar Crisis Appeal)
Assisting with Basic of Life in the Cen- tral African Republic	€ 30	<i>“Central African Republic is one of the poorest and most unstable countries in the world. The 2013 crisis has turned into an inter-community conflict, leading to the total collapse of the already weak socio-economic infrastructure. Basic social services are non-existent. We provide aid, run livelihood-support projects and repair water and sanitation systems. We visit detainees, restore contact between relatives separated by conflict and promote international humanitarian law.”</i> (Source: International Committee of the Red Cross)

Project	Amount	Description
Humanitarian Aid following the Conflict in Libya	€5	<i>“The continuing conflict in Libya has left devastation in its wake. Only a fraction of hospitals and health-centres are functional. Roads and schools have been destroyed. Fleeing the fighting, many families take only what they can carry, and many lose touch with their loved ones along the way. Despite the danger and obstacles, our work brings medicine and supplies to treat the wounded and sick, and delivers food and household items to people scarred by years of fighting. We also help migrants reconnect with their families and provide them with the support they need. With your help we can continue delivering life-saving aid.”</i> (Source: International Committee of the Red Cross, Libya Crisis Appeal)

4.B.4 Selection of z-Tree Screens

Figure 4.12: Stage 1 Decision of Employer (Discolored)

Remaining time (sec): 117

FIGURE INSTRUCTIONS

You have been selected as **Player 1**.

At the beginning of Stage 1, your Earnings are 50 Tokens, Player 2 Earnings are 50 Tokens, and the Donation is 50 Tokens.
 As Player 1, you choose between Option A and B. The details of both Options are in the instructions.
 Please study both Options carefully and indicate below which Option you would like to choose.

YOUR DECISION AS PLAYER 1
 Which Option do you choose?
 Option A
 Option B

SUBMIT

Figure 4.13: Stage 1 Expectation of Worker (Discolored)

Remaining time (sec): 114

FIGURE INSTRUCTIONS

You have been selected as Player 2.

At the beginning of Stage 1, your Earnings are **50** Tokens, Player 1 Earnings are **50** Tokens, and the Donation is **50** Tokens.
Player 1 is now choosing between Option A and Option B. The details of both Options are in the instructions.

As Player 2, you need to choose between three Alternatives: **if and only if** Player 1 chooses Option B. Otherwise, Stage 1 ends immediately.

While Player 1 makes his choice, we ask you to predict Player 1's choice.

YOUR EXPECTATION:
Which Option do you expect Player 1 to choose?
 Option A
 Option B

SUBMIT

Figure 4.14: Alternatives of Worker (Discolored)

Remaining time (sec): 115

FIGURE **INSTRUCTIONS**

You have been selected as Player 2.

Player 1 has chosen **Option B**. As a result, you need to choose between the three Alternatives below. Please make your choice by clicking on one of the SELECT buttons.

Alternative 1

Earnings Player 1: 80 Tokens
Earnings Player 2: 0 Tokens
Donation: 50 Tokens

SELECT

Alternative 2

Earnings Player 1: 80 Tokens
With a probability of 40 percent:
Earnings Player 2: 0 Tokens
Donation: 0 Tokens
With a probability of 60 percent:
Earnings Player 2: 50 Tokens
Donation: 50 Tokens

SELECT

Alternative 3

Earnings Player 1: 80 Tokens
Earnings Player 2: 50 Tokens
Donation: 0 Tokens

SELECT

Figure 4.15: Random Draw in Alternative 2 (Discolored)

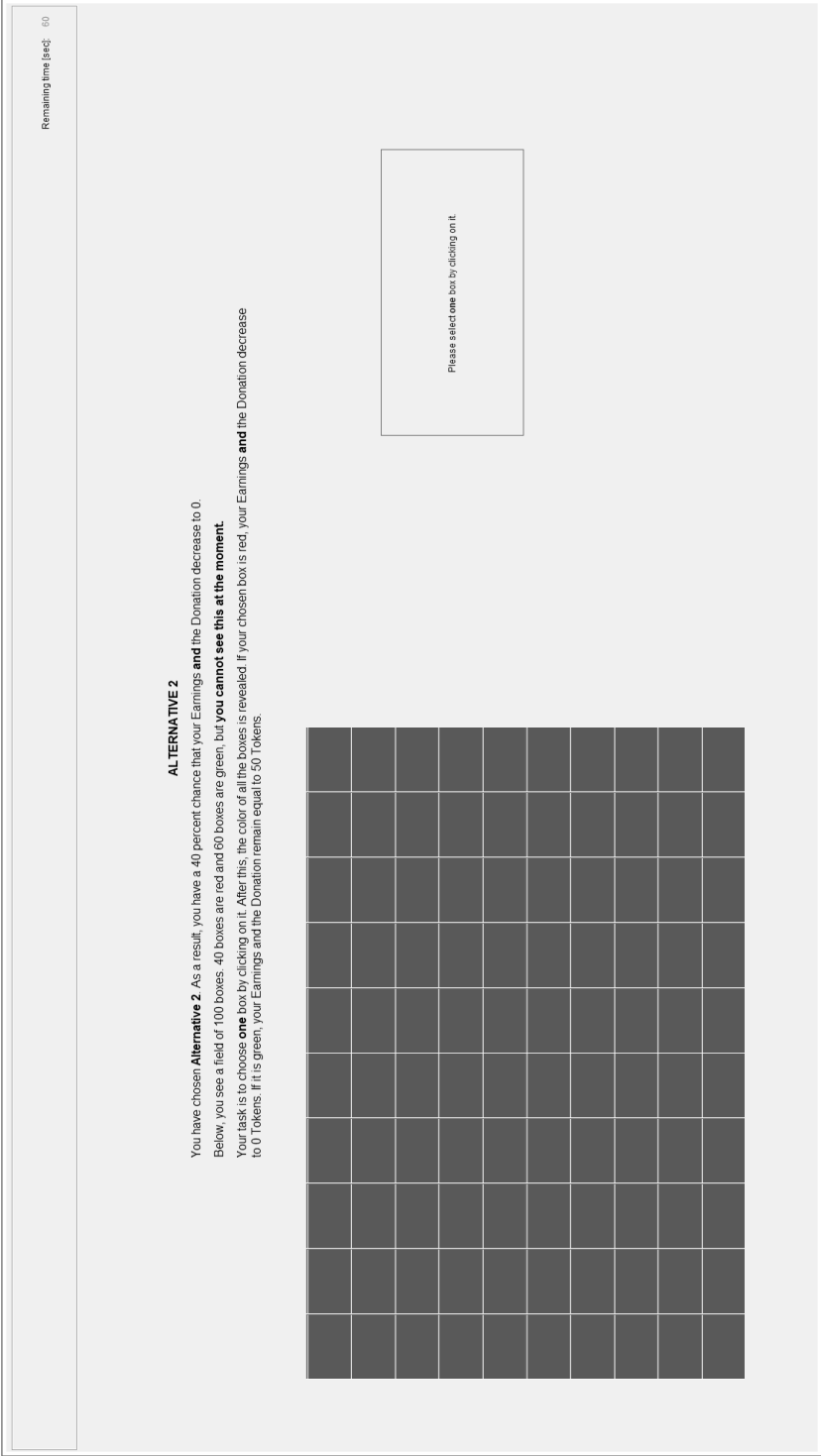
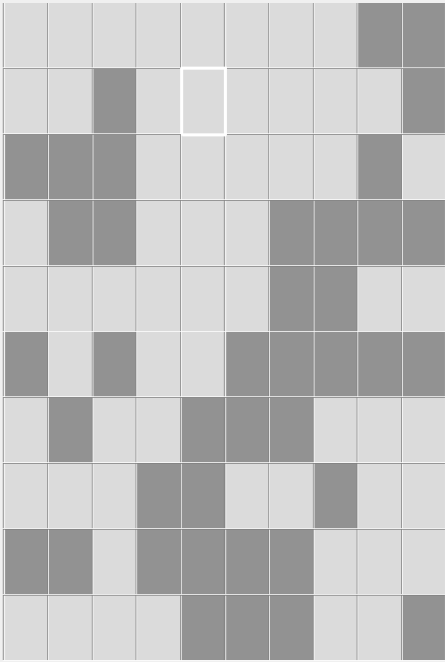


Figure 4.16: Random Draw in Alternative 2 - Outcome (Discolored)

Remaining time (sec): 56

ALTERNATIVE 2

You have chosen **Alternative 2**. As a result, you have a 40 percent chance that your Earnings and the Donation decrease to 0. Below, you see a field of 100 boxes. 40 boxes are red and 60 boxes are green, but **you cannot see this at the moment**. Your task is to choose **one** box by clicking on it. After this, the color of all the boxes is revealed. If your chosen box is red, your Earnings and the Donation remain equal to 50 Tokens. If it is green, your Earnings and the Donation remain equal to 30 Tokens.



Please select **one** box by clicking on it.


You selected box: 49
This box is: Green
As a result, your Earnings remain 50 Tokens
As a result, the Donation remains 50 Tokens

CONTINUE

Figure 4.17: Coin Identification Task - Denomination (Discolored)

Remaining time (sec): 239

Current coin number (out of 75): 1



20 cents 50 cents 1 Euro 2 Euro

The image shows a coin identification task interface. At the top, there is a timer showing 'Remaining time (sec): 239' and a counter showing 'Current coin number (out of 75): 1'. In the center, a 20-cent Euro coin is displayed. Below the coin, there are five buttons representing different denominations: '20 cents', '50 cents', '1 Euro', and '2 Euro'. The '20 cents' button is highlighted, indicating the correct answer.

Figure 4.18: Coin Identification Task - Country-of-Origin (Discolored)

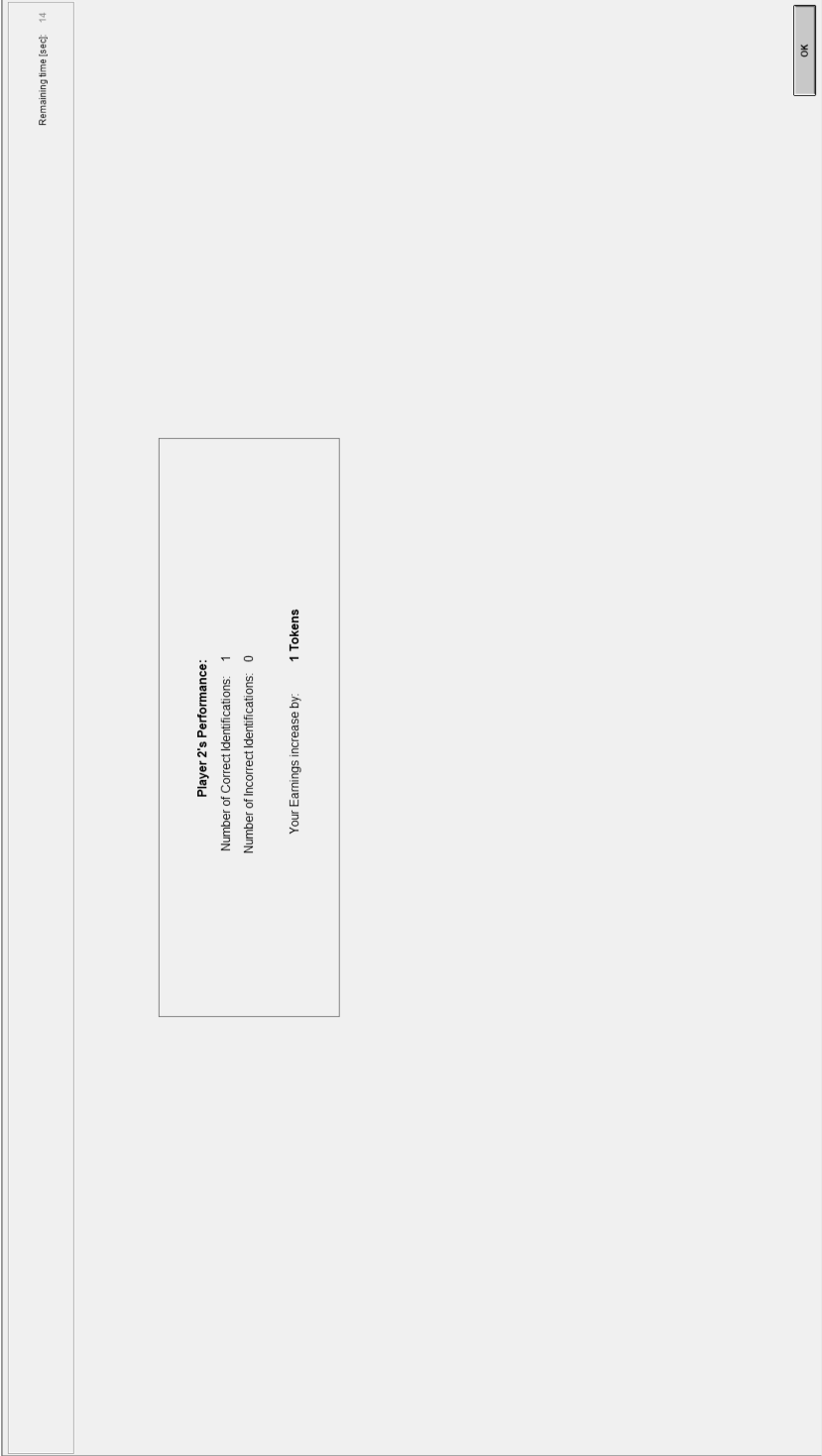
Remaining time (sec): 232

Current coin number (out of 75): 1



Austria Belgium Finland France Germany Greece Ireland Italy Luxembourg Netherlands Portugal Spain

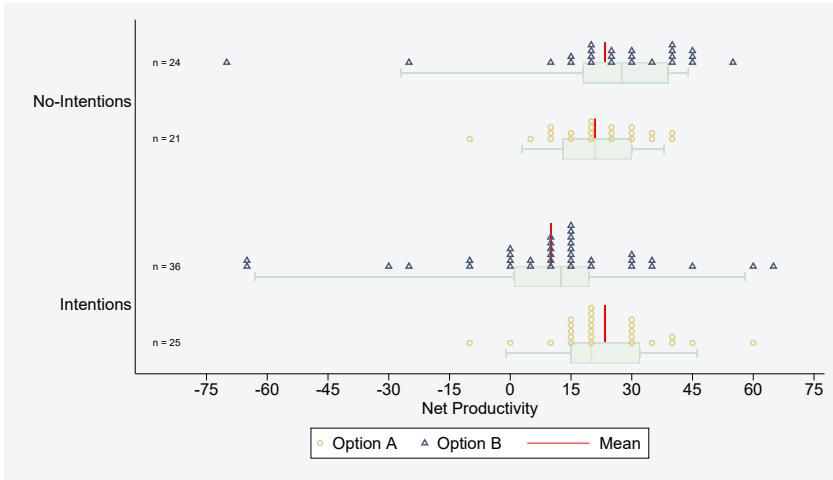
Figure 4.19: Coin Identification Task - Performance (Discolored)



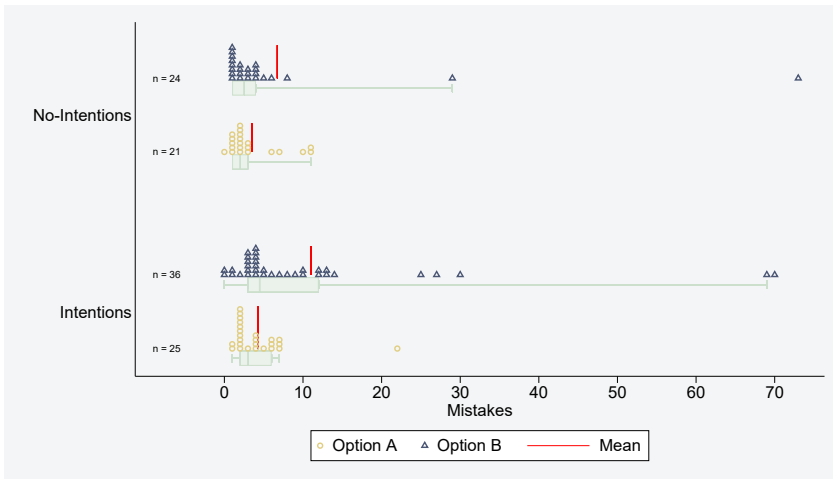
4.C Supplementary Analyses

4.C.1 Distribution of Performance

Figure 4.20: Distribution of Performance across Options and Treatments



(a) Net Productivity



(b) Mistakes

Note: The Figure displays net productivity levels (Panel a) and mistakes (Panel b) across treatments and options. Each marker represents one subject. For ease of exposition, net productivity levels are grouped in bins of width 5, while the bin width is 1 for mistakes. The boxes span the interquartile range with the median in between and the whiskers spanning the 5th and 95th percentile. The red markers denote the average net productivity for each treatment-option combination.

4.C.2 Mood

Figure 4.21: Excitement of Workers in Stage 1

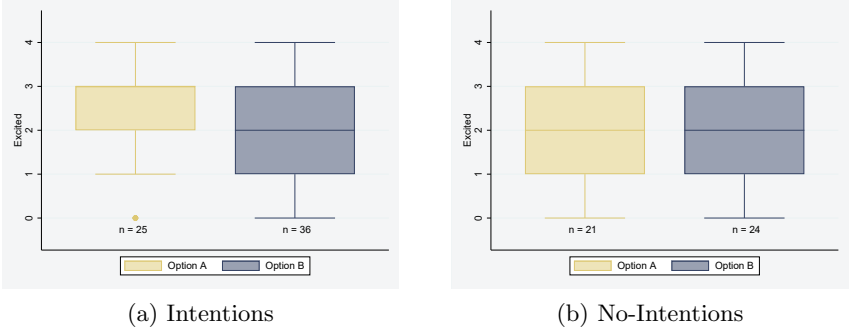


Figure 4.22: Degree of Upset of Workers in Stage 1

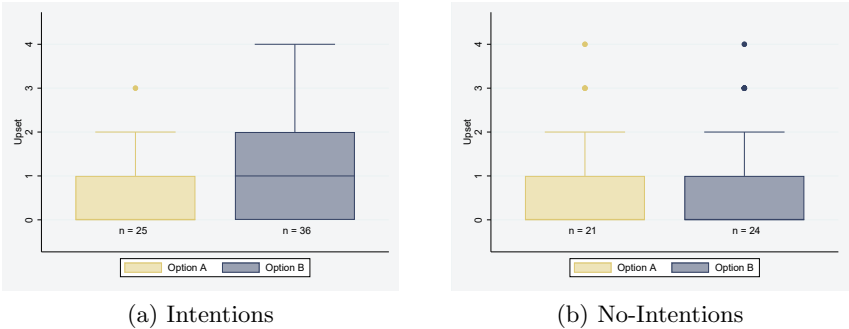


Figure 4.23: Shame of Workers in Stage 1

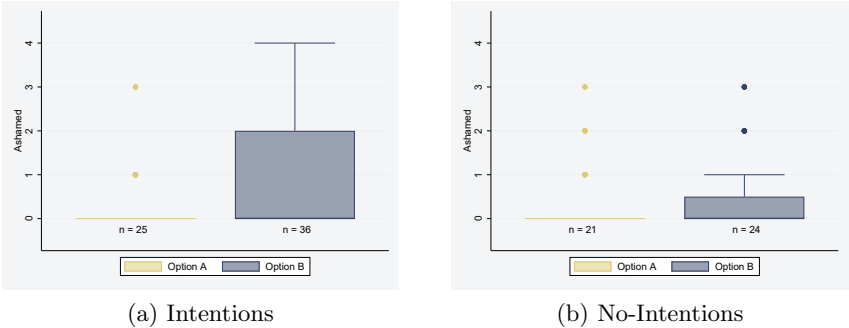


Figure 4.24: Hostility of Workers in Stage 1

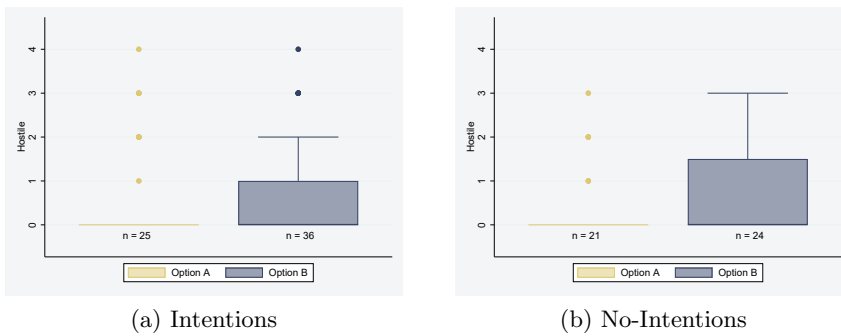
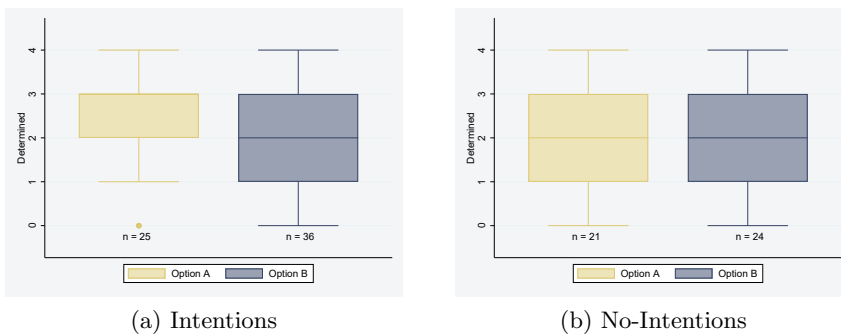


Figure 4.25: Determination of Workers in Stage 1



4.C.3 Mistakes

Table 4.9 presents the results of OLS regressions with the number of mistakes as the dependent variable. These results mimic those reported in Subsection 4.6.4. Option B induces Workers to make more mistakes only in the Intentions Treatment, as exemplified by the positive and significant coefficient on *Option B*. As opposed to Subsection 4.6.4, the insignificant coefficients on *No-Intentions* and *Option B × No-Intentions* imply that I cannot identify whether Workers respond positively to Option A by making fewer mistakes, or negatively to Option B by making more mistakes, or both, by comparing the number of mistakes made to those in the No-Intentions Treatment. Furthermore, Workers who report to have stronger reciprocal tendencies seem to respond more strongly to Option B being intentionally chosen by making more mistakes.

4.C.4 Outliers

Below, I perform regressions with net productivity (Table 4.10) and mistakes (Table 4.11) winsorized at a 5%-level. This means that productivity levels below the 2.5th and 97.5th percentile are replaced by the value of the 2.5th and 97.5th percentile, respectively. In contrast, mistakes are winsorized in the upper tail only, meaning that mistake levels above the 95th percentile are replaced by the 95th percentile. I then perform the same regressions as above.

4.C.5 Worker's Stage 1 Choice

In Table 4.12, I analyze the Worker's choice of Alternative under Option B. I pool the Workers across treatments and implement a multinomial logit model twice. First, I regress the Alternative on SVO Angle, Risk Tolerance, Male, European, and Economics & Business (Columns 1 to 3). Then, I replace SVO Angle with a dummy Pro-Social, which takes on value 1 if Worker is classified as being pro-social according to the SVO task (Columns 4 to 6).

As can be seen, Workers with a higher SVO Angle are more likely to choose Alternative 1 (which preserves the donation), while Workers with a higher Risk Tolerance are more likely to select the lottery. In addition, the dummy Pro-Social shows that pro-social Workers shift away from Alternative 3 (which preserves the own payoff). With SVO Angle and Pro-Social as proxies for θ_i^D in the model, this is in line with the prediction that Workers choose Alternative 1 when their θ_i^D is sufficiently high.

Table 4.9: Regression Analysis of Mistakes

	DEPENDENT VARIABLE: MISTAKES					
	(1)	(2)	(3)	(4)	(5)	(6)
A. Treatment-Option						
Option B	6.692** (3.111)	5.762* (2.793)	5.427* (3.104)	5.920** (2.714)	0.416 (3.148)	5.962** (2.749)
No-Intentions	-0.804 (1.236)	-0.587 (1.432)	-2.117 (1.631)	-0.400 (1.471)	-4.392 (2.677)	-0.095 (1.603)
Option B × No-Intentions	-3.418 (3.472)	-4.142 (4.091)	-2.338 (4.533)	-4.274 (4.159)	3.024 (4.748)	-4.523 (4.326)
B. Personal Characteristics						
Male		0.361 (2.627)	-0.536 (3.245)	-0.292 (3.204)	-1.416 (3.365)	0.718 (2.515)
Economics & Business		-3.049 (2.148)	-2.214 (1.940)	-1.421 (2.036)	-2.526 (1.507)	-2.637 (1.632)
Euro Years		0.295** (0.103)	0.265** (0.103)	0.267** (0.116)	0.199* (0.095)	0.283** (0.099)
C. Preferences						
SVO Angle (std.)			-0.857 (0.954)	-1.600* (0.861)	-1.067 (0.819)	
Risk Tolerance (std.)			-0.078 (1.183)	0.097 (1.212)	0.236 (1.144)	
Reciprocity (std.)			3.078** (1.357)	-0.819 (1.505)	2.319 (1.479)	
Rec. (std.) × No-Intentions				0.318 (1.643)		
Rec. (std.) × Option B				7.115*** (1.746)		
Rec. (std.) × Option B × No-Intentions				-0.079 (4.427)		
D. Mood						
Mood					-1.633 (1.284)	
Mood × No-Intentions					1.756 (1.452)	
Mood × Option B					-3.930 (3.774)	
Mood × Option B × No-Intentions					4.101 (3.874)	
E. Big Five						
Conscientiousness (std.)						-0.741 (0.885)
Openness (std.)						-0.754 (1.423)
Extraversion (std.)						-0.664 (1.121)
Neuroticism (std.)						1.125 (1.207)
Agreeableness (std.)						-0.363 (1.507)
Constant	4.280*** (0.974)	3.204 (2.437)	3.714 (2.625)	2.675 (2.434)	7.245** (3.305)	2.608* (1.396)
Observations	106	105	105	105	105	105
Clusters	17	17	17	17	17	17
R^2	0.06	0.11	0.17	0.24	0.26	0.13
F	6.11	2.28	1.71	7.56	4.61	3.89
df	16	16	16	16	16	16

Note: OLS model with Mistakes as the dependent variable. Standard errors are clustered on the session level. One observation is dropped due to missing information on that subject's gender.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4.10: Regression Analysis of Winsorized Net Productivity

	DEPENDENT VARIABLE: WINSORIZED NET PRODUCTIVITY					
	(1)	(2)	(3)	(4)	(5)	(6)
A. Treatment-Option						
Option B	-11.633** (4.637)	-11.577** (5.087)	-11.496* (5.472)	-12.276** (4.571)	-1.425 (4.970)	-11.291** (5.219)
No-Intentions	-1.895 (3.408)	-1.371 (3.578)	-0.909 (3.786)	-3.132 (3.442)	7.285* (4.160)	-2.990 (3.738)
Option B × No-Intentions	15.562** (5.334)	13.788** (6.100)	13.099* (6.729)	15.689** (5.908)	2.223 (5.406)	14.764** (6.225)
B. Personal Characteristics						
Male		4.513 (3.634)	4.619 (3.914)	4.125 (3.959)	5.479 (3.924)	4.017 (3.531)
Economics & Business		2.598 (3.427)	2.293 (3.719)	1.090 (3.821)	3.140 (3.244)	1.705 (2.923)
Euro Years		0.209 (0.162)	0.208 (0.171)	0.205 (0.189)	0.268 (0.186)	0.114 (0.164)
C. Preferences						
SVO Angle (std.)			0.015 (1.220)	0.979 (1.142)	0.405 (1.060)	
Risk Tolerance (std.)			0.475 (1.814)	0.169 (1.775)	0.050 (1.653)	
Reciprocity (std.)			-1.028 (1.601)	4.967** (2.223)	-0.143 (1.903)	
Rec. (std.) × No-Intentions				-1.696 (3.041)		
Rec. (std.) × Option B				-10.951*** (3.049)		
Rec. (std.) × Option B × No-Intentions				2.780 (5.972)		
D. Mood						
Mood					4.726** (1.874)	
Mood × No-Intentions					-6.203** (2.399)	
Mood × Option B					1.981 (2.547)	
Mood × Option B × No-Intentions					1.397 (3.731)	
E. Big Five						
Conscientiousness (std.)						2.844** (1.328)
Openness (std.)						-1.775 (1.859)
Extraversion (std.)						1.640 (1.816)
Neuroticism (std.)						-1.568 (1.617)
Agreeableness (std.)						-1.479 (1.471)
Constant	22.800*** (2.788)	17.100*** (4.404)	17.172*** (4.542)	18.804*** (3.672)	8.884** (3.958)	19.568*** (3.978)
Observations	106	105	105	105	105	105
Clusters	17	17	17	17	17	17
R^2	0.11	0.14	0.15	0.22	0.25	0.20
F	3.00	3.82	3.78	9.93	10.38	23.80
df	16	16	16	16	16	16

Note: OLS model with Mistakes as the dependent variable. Standard errors are clustered on the session level. One observation is dropped due to missing information on that subject's gender.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4.11: Regression Analysis of Winsorized Mistakes

	DEPENDENT VARIABLE: WINSORIZED MISTAKES					
	(1)	(2)	(3)	(4)	(5)	(6)
A. Treatment-Option						
Option B	4.248** (1.746)	3.738** (1.590)	3.625* (1.716)	3.866** (1.544)	0.582 (2.280)	4.023** (1.396)
No-Intentions	-0.804 (1.236)	-0.667 (1.276)	-1.284 (1.245)	-0.539 (1.238)	-3.387 (2.353)	-0.059 (1.091)
Option B × No-Intentions	-2.974 (1.851)	-3.383 (2.043)	-2.641 (2.404)	-3.481 (2.097)	0.687 (2.582)	-3.979** (1.820)
B. Personal Characteristics						
Male		0.140 (1.375)	-0.119 (1.503)	-0.025 (1.393)	-0.504 (1.552)	0.390 (1.340)
Economics & Business		-1.383 (1.119)	-1.147 (1.240)	-0.771 (1.364)	-1.323 (1.145)	-1.110 (0.969)
Euro Years		0.163** (0.056)	0.150** (0.053)	0.150** (0.059)	0.126* (0.060)	0.159** (0.064)
C. Preferences						
SVO Angle (std.)			-0.480 (0.569)	-0.824* (0.461)	-0.594 (0.537)	
Risk Tolerance (std.)			-0.219 (0.606)	-0.133 (0.618)	-0.055 (0.510)	
Reciprocity (std.)			1.213* (0.593)	-0.716 (0.998)	0.898 (0.753)	
Rec. (std.) × No-Intentions				0.447 (1.147)		
Rec. (std.) × Option B				3.308** (1.536)		
Rec. (std.) × Option B × No-Intentions				-0.128 (1.979)		
D. Mood						
Mood					-1.309 (1.090)	
Mood × No-Intentions					1.586 (1.205)	
Mood × Option B					-1.074 (1.612)	
Mood × Option B × No-Intentions					0.725 (1.708)	
E. Big Five						
Conscientiousness (std.)						-0.899 (0.549)
Openness (std.)						-0.345 (0.725)
Extraversion (std.)						-0.428 (0.763)
Neuroticism (std.)						0.786 (0.606)
Agreeableness (std.)						-0.067 (0.680)
Constant	4.280*** (0.974)	3.502** (1.503)	3.746** (1.387)	3.260** (1.198)	6.131** (2.275)	2.920** (1.054)
Observations	106	105	105	105	105	105
Clusters	17	17	17	17	17	17
R^2	0.10	0.14	0.18	0.23	0.25	0.19
F	4.78	6.31	7.79	48.89	13.83	23.41
df	16	16	16	16	16	16

Note: OLS model with Mistakes as the dependent variable. Standard errors are clustered on the session level. One observation is dropped due to missing information on that subject's gender.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4.12: Multinomial Logit Analysis of Alternative Chosen

	DEPENDENT VARIABLE: ALTERNATIVE					
	(1)	(2)	(3)	(4)	(5)	(6)
	Alt. 1	Alt. 2	Alt. 3	Alt. 1	Alt. 2	Alt. 3
SVO Angle (std.)	0.056*	0.064	-0.121			
	(0.032)	(0.081)	(0.083)			
Pro-Social				0.111*	0.226	-0.336*
				(0.064)	(0.165)	(0.172)
Risk Tolerance (std.)	0.033	0.175**	-0.208**	0.039	0.192**	-0.231**
	(0.029)	(0.088)	(0.093)	(0.031)	(0.091)	(0.096)
Male	-0.109	-0.090	0.200	-0.111	-0.095	0.206
	(0.083)	(0.178)	(0.182)	(0.086)	(0.180)	(0.185)
Economics & Business	-0.024	0.102	-0.078	-0.026	0.138	-0.112
	(0.063)	(0.161)	(0.163)	(0.066)	(0.166)	(0.169)
European	-0.011	0.082	-0.071	-0.003	0.078	-0.075
	(0.078)	(0.177)	(0.178)	(0.081)	(0.178)	(0.180)
Observations	59	59	59	59	59	59
R^2	0.12	0.12	0.12	0.13	0.13	0.13
Log L	-49.37	-49.37	-49.37	-48.87	-48.87	-48.87
df	10	10	10	10	10	10

Note: Multinomial logit model with Alternative as the dependent variable. Marginal effects on the probability of each Alternative are displayed, with delta-method standard errors between parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4.C.6 Effects of CoViD-19 Measures on Sample and Results

Exactly half of the subjects participated in an experimental session before the Dutch government imposed measures in relation to the CoViD-19 virus. In this part of the Appendix, I examine (i) whether subjects in the experiment after measures were imposed differ from those who participated before measures were imposed, and (ii) whether this affects our results. Table 4.13 shows the characteristics of *all* subjects (both roles) before and after the outbreak. As can be seen, a significantly larger share of subjects after the outbreak is enrolled in an economics or business program (significant at a 1%-level). This probably reflects the additional recruiting efforts during lectures in economics- or business-related programs. Furthermore, subjects after the outbreak are significantly more risk-tolerant (at a 10%-level), which may reflect that risk-averse subjects are not willing to travel to the laboratory and run the risk of contracting the virus.

Table 4.13: Summary Statistics Before and After CoViD-19 Outbreak

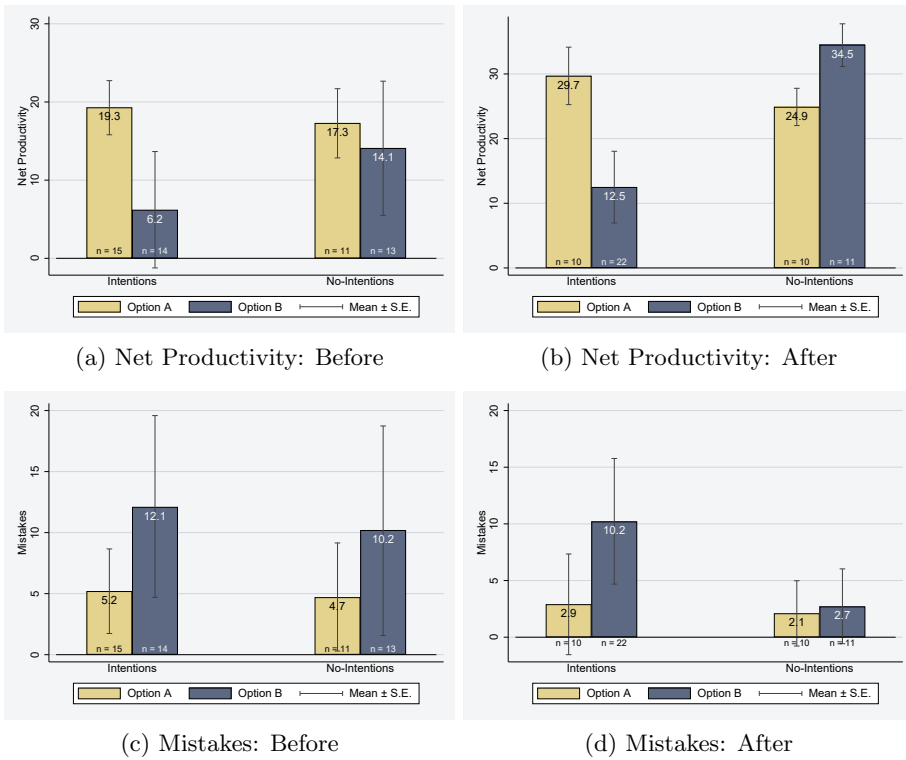
	CoViD-19			
	(1) All	(2) Before	(3) After	(4) Δ
A. Background				
Male	0.42 (0.50)	0.38 (0.49)	0.46 (0.50)	-0.08 (0.07)
Age	21.64 (3.25)	21.45 (3.20)	21.83 (3.30)	-0.38 (0.45)
European	0.74 (0.44)	0.75 (0.44)	0.73 (0.45)	0.02 (0.06)
Economics & Business	0.65 (0.48)	0.57 (0.50)	0.74 (0.44)	-0.17*** (0.06)
Euro Years	11.66 (8.02)	11.40 (8.08)	11.92 (7.99)	-0.53 (1.10)
B. Preferences				
Reciprocity	4.27 (0.90)	4.33 (0.93)	4.21 (0.87)	0.12 (0.12)
Altruism	4.78 (1.56)	4.80 (1.42)	4.75 (1.69)	0.05 (0.21)
Risk Tolerance	37.84 (16.85)	35.93 (15.40)	39.75 (18.05)	-3.82* (2.30)
SVO Angle	19.96 (13.75)	20.01 (13.02)	19.91 (14.51)	0.10 (1.89)
C. Big Five				
Conscientiousness	4.87 (1.00)	4.82 (1.03)	4.92 (0.97)	-0.09 (0.14)
Extraversion	4.68 (1.27)	4.62 (1.31)	4.75 (1.23)	-0.13 (0.17)
Agreeableness	5.24 (1.00)	5.20 (1.05)	5.27 (0.96)	-0.07 (0.14)
Openness	4.82 (1.25)	4.75 (1.23)	4.89 (1.27)	-0.13 (0.17)
Neuroticism	4.19 (1.33)	4.22 (1.27)	4.17 (1.39)	0.05 (0.18)
Observations	212	106	106	212

Note: Balance across subjects in sessions before and after the outbreak of CoViD-19. Columns (4) displays the difference between the two, with the standard error of the difference between parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Re-assessing my main results, I show in Figure 4.26 that my main results still

Figure 4.26: Worker Performance Before and After CoViD-19 Outbreak



stand: before and after the outbreak, Workers in Intentions are more productive under Option A as compared to Option B, and this results seems to be driven by negative reciprocity. In addition, three observations can be made regarding performance before and after the outbreak. First, a larger share of Employers seem to choose Option B after the outbreak. Second, productivity is higher in each cell after the outbreak, potentially caused by the more easily accessible laminated catalogues. Third, productivity in No-Intentions is significantly higher (at a 10%-level) under Option B as compared to Option A.

Bibliography

- Abbink, K., Irlenbusch, B., and Renner, E. (2000). The moonlighting game: An experimental study on reciprocity and retribution. *Journal of Economic Behavior & Organization*, 42(2):265–277.
- Adriani, F. and Sonderegger, S. (2009). Why do parents socialize their children to behave pro-socially? An information-based theory. *Journal of Public Economics*, 93(11-12):1119–1124.
- Aina, C., Battigalli, P., and Gamba, A. (2020). Frustration and anger in the ultimatum game: An experiment. *Games and Economic Behavior*.
- Akerlof, G. A. (1982). Labor contracts as partial gift exchange. *The Quarterly Journal of Economics*, 97(4):543–569.
- Alan, S., Baydar, N., Boneva, T., Crossley, T. F., and Ertac, S. (2017). Transmission of risk preferences from mothers to daughters. *Journal of Economic Behavior & Organization*, 134:60–77.
- Alempaki, D., Doğan, G., and Saccardo, S. (2019). Deception and reciprocity. *Experimental Economics*, 22(4):980–1001.
- Andersen, S., Ertac, S., Gneezy, U., List, J. A., and Maximiano, S. (2013). Gender, competitiveness, and socialization at a young age: Evidence from a matrilineal and a patriarchal society. *Review of Economics and Statistics*, 95(4):1438–1443.
- Angerer, S., Glätzle-Rützler, D., Lergetporer, P., and Sutter, M. (2015). Donations, risk attitudes and time preferences: A study on altruism in primary school children. *Journal of Economic Behavior & Organization*, 115:67–74.
- Artavia-Mora, L., Bedi, A. S., and Rieger, M. (2017). Intuitive help and punishment in the field. *European Economic Review*, 92:133–145.

- Attanasio, O., Cattan, S., Fitzsimons, E., Meghir, C., and Rubio-Codina, M. (2020). Estimating the production function for human capital: Results from a randomized controlled trial in Colombia. *American Economic Review*, 110(1):48–85.
- Balafoutas, L. and Nikiforakis, N. (2012). Norm enforcement in the city: A natural field experiment. *European Economic Review*, 56(8):1773–1785.
- Balafoutas, L., Nikiforakis, N., and Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*, 111(45):15924–15927.
- Balafoutas, L., Nikiforakis, N., and Rockenbach, B. (2016). Altruistic punishment does not increase with the severity of norm violations in the field. *Nature Communications*, 7(13327):1–6.
- Bandura, A. (1965). Vicarious processes: A case of no-trial learning. In *Advances in Experimental Social Psychology*, volume 2, pages 1–55. Elsevier.
- Bandura, A. (1977). *Social learning theory*. Oxford: Prentice Hall.
- Bartuli, J., Djawadi, B., and Fahr, R. (2016). Business ethics in organizations: An experimental examination of whistleblowing and personality. IZA Discussion Papers no. 10190, Institute of Labor (IZA).
- Bauer, M., Chytilová, J., and Pertold-Gebicka, B. (2014). Parental background and other-regarding preferences in children. *Experimental Economics*, 17(1):24–46.
- Belot, M. and Schröder, M. (2013). Sloppy work, lies and theft: A novel experimental design to study counterproductive behaviour. *Journal of Economic Behavior & Organization*, 93:233–238.
- Belot, M. and Schröder, M. (2015). The spillover effects of monitoring: A field experiment. *Management Science*, 62(1):37–45.
- Ben-Ner, A., List, J. A., Putterman, L., and Samek, A. (2017). Learned generosity? An artefactual field experiment with parents and their children. *Journal of Economic Behavior & Organization*, 143:28–44.
- Benenson, J. F., Pascoe, J., and Radmore, N. (2007). Children’s altruistic behavior in the dictator game. *Evolution and Human Behavior*, 28(3):168–175.
- Bennett, R. J. and Robinson, S. L. (2000). Development of a measure of workplace deviance. *Journal of Applied Psychology*, 85(3):349.

- Berger, J. and Hevenstone, D. (2016). Norm enforcement in the city revisited: An international field experiment of altruistic punishment, norm maintenance, and broken windows. *Rationality and Society*, 28(3):299–319.
- Berkowitz, M. W. and Grych, J. H. (1998). Fostering goodness: Teaching parents to facilitate children’s moral development. *Journal of Moral Education*, 27(3):371–391.
- Bettinger, E. and Slonim, R. (2006). Using experimental economics to measure the effects of a natural educational experiment on altruism. *Journal of Public Economics*, 90(8-9):1625–1648.
- Binswanger, H. P. (1981). Attitudes toward risk: Theoretical implications of an experiment in rural India. *The Economic Journal*, 91(364):867–890.
- Bisin, A. and Verdier, T. (2001). The economics of cultural transmission and the dynamics of preferences. *Journal of Economic Theory*, 97(2):298–319.
- Biziou-van Pol, L., Haenen, J., Novaro, A., Liberman, A. O., and Capraro, V. (2015). Does telling white lies signal pro-social preferences? *Judgment & Decision Making*, 10(6).
- Blau, P. M. (1964). *Exchange and Power in Social Life*. Transaction Publishers.
- Brenøe, A. A. and Epper, T. (2019). Parenting values moderate the intergenerational transmission of time preferences. Working Paper no. 333, University of Zurich, Department of Economics.
- Brown, M. E. and Treviño, L. K. (2006a). Ethical leadership: A review and future directions. *The Leadership Quarterly*, 17(6):595–616.
- Brown, M. E. and Treviño, L. K. (2006b). Socialized charismatic leadership, values congruence, and deviance in work groups. *Journal of Applied Psychology*, 91(4):954.
- Cappelen, A. W., List, J. A., Samek, A., and Tungodden, B. (2020). The effect of early education on social preferences. *Journal of Political Economy*, in press.
- Cappelen, A. W., Sørensen, E. Ø., and Tungodden, B. (2013). When do we lie? *Journal of Economic Behavior & Organization*, 93:258–265.
- Capra, M. C. (2004). Mood-driven behavior in strategic interactions. *American Economic Review*, 94(2):367–372.

- Casari, M. (2012). Weak reciprocity alone cannot explain peer punishment. *Behavioral and Brain Sciences*, 35(1):21–22.
- Castillo, M. (2020). Negative childhood experiences and risk aversion: Evidence from children exposed to domestic violence. IZA Discussion Papers no. 13320, Institute of Labor Economics (IZA), Bonn.
- Charness, G. and Levine, D. I. (2007). Intention and stochastic outcomes: An experimental study. *The Economic Journal*, 117(522):1051–1072.
- Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869.
- Chowdhury, S., Sutter, M., and Zimmermann, K. (2018). Evaluating intergenerational persistence of economic preferences: A large scale experiment with families in Bangladesh. ZEF Discussion Papers on Development Policy no. 250, Center for Development Research (ZEF).
- Chowdhury, S., Sutter, M., and Zimmermann, K. F. (2020). Economic preferences across generations and family clusters: A large-scale experiment. GLO Discussion Papers no. 3642651, Global Labor Organization.
- Cipriani, M., Giuliano, P., and Jeanne, O. (2013). Like mother like son? Experimental evidence on the transmission of values from parents to children. *Journal of Economic Behavior & Organization*, 90:100–111.
- Cohn, A., Fehr, E., and Goette, L. (2015). Fair wages and effort provision: Combining evidence from a choice experiment and a field experiment. *Management Science*, 61(8):1777–1794.
- Coleman, J. S. (1994). *Foundations of social theory*. Harvard University Press, Cambridge MA.
- Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, 46(2):260–281.
- Cox, J. C., Friedman, D., and Gjerstad, S. (2007). A tractable model of reciprocity and fairness. *Games and Economic Behavior*, 59(1):17–45.
- Crosetto, P. and Filippin, A. (2013). The “bomb” risk elicitation task. *Journal of Risk and Uncertainty*, 47(1):31–65.

- Crosetto, P., Weisel, O., and Winter, F. (2019). A flexible z-Tree and oTree implementation of the Social Value Orientation slider measure. *Journal of Behavioral and Experimental Finance*, 23:46–53.
- Cunha, F. and Heckman, J. J. (2010). Investing in our young people. NBER Working Paper no. 16201, National Bureau of Economic Research.
- Deckers, T., Falk, A., Kosse, F., Pinger, P. R., and Schildberg-Hörisch, H. (2017). Socio-economic status and inequalities in children’s IQ and economic preferences. IZA Discussion Papers no 11158, Institute of Labor (IZA).
- Degner, J. and Dalege, J. (2013). The apple does not fall far from the tree, or does it? A meta-analysis of parent–child similarity in intergroup attitudes. *Psychological Bulletin*, 139(6):1270.
- Denant-Boemont, L., Masclet, D., and Noussair, C. N. (2007). Punishment, counter-punishment and sanction enforcement in a social dilemma experiment. *Economic Theory*, 33(1):145–167.
- Detert, J. R., Treviño, L. K., Burris, E. R., and Andiappan, M. (2007). Managerial modes of influence and counterproductivity in organizations: A longitudinal business-unit-level investigation. *Journal of Applied Psychology*, 92(4):993.
- Doepke, M. and Zilibotti, F. (2017). Parenting with style: Altruism and paternalism in intergenerational preference transmission. *Econometrica*, 85(5):1331–1371.
- Dohmen, T., Falk, A., Huffman, D., and Sunde, U. (2012). The intergenerational transmission of risk and trust attitude. *Review of Economic Studies*, 79(2):645–677.
- Dondé, G. (2018). *Ethics at Work: 2018 survey of employees – Europe*. Institute of Business Ethics.
- Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2):268–298.
- Dunn, J. R. and Schweitzer, M. E. (2005). Feeling and believing: The influence of emotion on trust. *Journal of Personality and Social Psychology*, 88(5):736.
- Ellingsen, T. and Johannesson, M. (2008). Pride and prejudice: The human side of incentive theory. *American Economic Review*, 98(3):990–1008.
- Erat, S. and Gneezy, U. (2012). White lies. *Management Science*, 58(4):723–733.

- Ethics and Compliance Initiative (2017). Ethical leadership around the World - and why it matters. Global Business Ethics Survey, Ethics and Compliance Initiative.
- Ethics and Compliance Initiative (2018). The state of ethics & compliance in the workplace. Global Business Ethics Survey, Ethics and Compliance Initiative.
- Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692.
- Falk, A., Fehr, E., and Fischbacher, U. (2003). On the nature of fair behavior. *Economic Inquiry*, 41(1):20–26.
- Falk, A., Fehr, E., and Fischbacher, U. (2008). Testing theories of fairness: Intentions matter. *Games and Economic Behavior*, 62(1):287–303.
- Falk, A. and Kosfeld, M. (2006). The hidden costs of control. *American Economic Review*, 96(5):1611–1630.
- Falk, A. and Kosse, F. (2016). Early childhood environment, breastfeeding and the formation of preferences. SOEP Paper no. 882, SOEP.
- Fehr, E. and Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2):63–87.
- Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994.
- Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868):137–140.
- Fehr, E., Kirchsteiger, G., and Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *The Quarterly Journal of Economics*, 108(2):437–459.
- Fischbacher, U. (2007). z-Tree: Zurich Toolbox for Ready-made Economic Experiments. *Experimental Economics*, 10(2):171–178.
- Fischbacher, U. and Föllmi-Heusi, F. (2013). Lies in disguise: An experimental study on cheating. *Journal of the European Economic Association*, 11(3):525–547.
- Gächter, S., Huang, L., and Sefton, M. (2016). Combining “real effort” with induced effort costs: The ball-catching task. *Experimental Economics*, 19(4):687–712.

- Gallup (2017). State of the global workplace. Technical report, Gallup.
- Gino, F., Ayal, S., and Ariely, D. (2013). Self-serving altruism? The lure of unethical actions that benefit others. *Journal of Economic Behavior & Organization*, 93:285–292.
- Gneezy, U. (2002). Does high wage lead to high profits: An experimental study of reciprocity using real efforts. Unpublished.
- 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. and List, J. A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, 74(5):1365–1384.
- Greenberg, J. (1990). Employee theft as a reaction to underpayment inequity: The hidden cost of pay cuts. *Journal of Applied Psychology*, 75(5):561.
- Gross, J., Leib, M., Offerman, T., and Shalvi, S. (2018). Ethical free riding: When honest people find dishonest partners. *Psychological Science*, 29(12):1956–1968.
- Hahn, E., Gottschling, J., and Spinath, F. M. (2012). Short measurements of personality—validity and reliability of the GSOEP Big Five inventory (BFI-S). *Journal of Research in Personality*, 46(3):355–359.
- Hardy, S. A., Padilla-Walker, L. M., and Carlo, G. (2008). Parenting dimensions and adolescents’ internalisation of moral values. *Journal of Moral Education*, 37(2):205–223.
- Heckman, J. and Rubinstein, Y. (2001). The importance of noncognitive skills: Lessons from the ged testing program. *American Economic Review*, 91(2):145–149.
- Heckman, J., Stixrud, J., and Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, 24(3):411–482.
- Heckman, J. J. and Mosso, S. (2014). The economics of human development and social mobility. *Annual Review of Economics*, 6(1):689–733.

- Hennig-Schmidt, H., Sadrieh, A., and Rockenbach, B. (2010). In search of workers' real effort reciprocity – a field and a laboratory experiment. *Journal of the European Economic Association*, 8(4):817–837.
- Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5):1644–1655.
- Houser, D., List, J. A., Piovesan, M., Samek, A., and Winter, J. (2016). Dishonesty: from parents to children. *European Economic Review*, 82:242–254.
- Janssen, M. A. and Bushman, C. (2008). Evolution of cooperation and altruistic punishment when retaliation is possible. *Journal of Theoretical Biology*, 254(3):541–545.
- Kerschbamer, R., Neururer, D., and Gruber, A. (2019). Do altruists lie less? *Journal of Economic Behavior & Organization*, 157:560–579.
- Khadjavi, M. (2017). Indirect reciprocity and charitable giving – evidence from a field experiment. *Management Science*, 63(11):3708–3717.
- Khadjavi, M. and Nicklisch, A. (2018). Parents' ambitions and children's competitiveness. *Journal of Economic Psychology*, 67:87–102.
- Kirchsteiger, G., Rigotti, L., and Rustichini, A. (2006). Your morals might be your moods. *Journal of Economic Behavior & Organization*, 59(2):155–172.
- Kocher, M. G., Schudy, S., and Spantig, L. (2017). I lie? We lie! Why? Experimental evidence on a dishonesty shift in groups. *Management Science*, 64(9):3995–4008.
- Konishi, N. and Ohtsubo, Y. (2015). Does dishonesty really invite third-party punishment? results of a more stringent test. *Biology Letters*, 11(5):20150172.
- Konovsky, M. A. and Pugh, S. D. (1994). Citizenship behavior and social exchange. *Academy of Management Journal*, 37(3):656–669.
- Kosse, F., Deckers, T., Pinger, P., Schildberg-Hörisch, H., and Falk, A. (2020). The formation of prosociality: Causal evidence on the role of social environment. *Journal of Political Economy*, 128(2):434–467.
- Kosse, F. and Pfeiffer, F. (2012). Impatience among preschool children and their mothers. *Economics Letters*, 115(3):493–495.

- Krueger, A. B. and Mas, A. (2004). Strikes, scabs, and tread separations: Labor strife and the production of defective Bridgestone/Firestone tires. *Journal of Political Economy*, 112(2):253–289.
- Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524.
- Kube, S., Maréchal, M. A., and Puppe, C. (2012). The currency of reciprocity: Gift exchange in the workplace. *American Economic Review*, 102(4):1644–62.
- Kube, S., Maréchal, M. A., and Puppe, C. (2013). Do wage cuts damage work morale? Evidence from a natural field experiment. *Journal of the European Economic Association*, 11(4):853–870.
- Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1(3):593–622.
- Levine, E. E. and Schweitzer, M. E. (2014). Are liars ethical? On the tension between benevolence and honesty. *Journal of Experimental Social Psychology*, 53:107–117.
- Levine, E. E. and Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, 126:88–106.
- Maccoby, E. E. (1992). The role of parents in the socialization of children: An historical overview. *Developmental Psychology*, 28(6):1006–1017.
- Mas, A. (2006). Pay, reference points, and police performance. *The Quarterly Journal of Economics*, 121(3):783–821.
- Mas, A. (2008). Labour unrest and the quality of production: Evidence from the construction equipment resale market. *The Review of Economic Studies*, 75(1):229–258.
- Murphy, R. O., Ackermann, K. A., and Handgraaf, M. (2011). Measuring social value orientation. *Judgment and Decision Making*, 6(8):771–781.
- Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics*, 92(1-2):91–112.
- Offerman, T. (2002). Hurting hurts more than helping helps. *European Economic Review*, 46(8):1423–1437.

- Ohtsubo, Y., Masuda, F., Watanabe, E., and Masuchi, A. (2010). Dishonesty invites costly third-party punishment. *Evolution and Human Behavior*, 31(4):259–264.
- Ostrom, E. (2000). Collective action and the evolution of social norms. *The Journal of Economic Perspectives*, 14(3):137–158.
- Piaget, J. (1932). *The moral judgement of the child*. Routledge and Kegan Paul.
- Proto, E., Sgroi, D., and Nazneen, M. (2019). Happiness, cooperation and language. *Journal of Economic Behavior & Organization*, 168:209–228.
- Przepiorka, W. and Berger, J. (2016). The sanctioning dilemma: A quasi-experiment on social norm enforcement in the train. *European Sociological Review*, 32(3):439–451.
- Ransijn, B. (2018). Integriteit bij de rijksoverheid. Technical report, FNV.
- Reuben, E. and Stephenson, M. (2013). Nobody likes a rat: On the willingness to report lies and the consequences thereof. *Journal of Economic Behavior & Organization*, 93:384–391.
- Robinson, S. L. and Bennett, R. J. (1995). A typology of deviant workplace behaviors: A multidimensional scaling study. *Academy of Management Journal*, 38(2):555–572.
- Roest, A., Dubas, J. S., and Gerris, J. R. (2009). Value transmissions between fathers, mothers, and adolescent and emerging adult children: The role of the family climate. *Journal of Family Psychology*, 23(2):146.
- Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, 23(3):393–404.
- Schneider, F., Brun, F., and Weber, R. A. (2020). Sorting and wage premiums in immoral work. Working Paper no. 353, University of Zurich, Department of Economics.
- Sebald, A. (2010). Attribution and reciprocity. *Games and Economic Behavior*, 68(1):339–352.
- Shalvi, S. and De Dreu, C. K. W. (2014). Oxytocin promotes group-serving dishonesty. *Proceedings of the National Academy of Sciences*, 111(15):5503–5507.

- Smetana, J. G. (1999). The role of parents in moral development: A social domain analysis. *Journal of Moral Education*, 28(3):311–321.
- Sugden, R. (1986). *The economics of rights, cooperation and welfare*. Basil Blackwell, Oxford.
- Sutter, M., Angerer, S., Glätzle-Rützler, D., and Lergepporter, P. (2018). Language group differences in time preferences: Evidence from primary school children in a bilingual city. *European Economic Review*, 106:21–34.
- Sutter, M. and Untertrifaller, A. (2020). Children’s heterogeneity in cooperation and parental background: an experimental study. *Journal of Economic Behavior & Organization*, 171:286–296.
- Sutter, M., Zoller, C., and Glätzle-Rützler, D. (2019). Economic behavior of children and adolescents – a first survey of experimental economics results. *European Economic Review*, 111:98–121.
- Tabellini, G. (2008). The scope of cooperation: Values and incentives. *The Quarterly Journal of Economics*, 123(3):905–950.
- Treviño, L. K. and Brown, M. E. (2005). The role of leaders in influencing unethical behavior in the workplace. In *Managing Organizational Deviance*, pages 69–96. SAGE Publications Inc.
- Utikal, V. and Fischbacher, U. (2013). Disadvantageous lies in individual decisions. *Journal of Economic Behavior & Organization*, 85:108–111.
- Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6):1063.
- Weisel, O. and Shalvi, S. (2015). The collaborative roots of corruption. *Proceedings of the National Academy of Sciences*, 112(34):10651–10656.
- White, F. A. and Matawie, K. M. (2004). Parental morality and family processes as predictors of adolescent morality. *Journal of Child and Family Studies*, 13(2):219–233.
- Wiltermuth, S. S. (2011). Cheating more when the spoils are split. *Organizational Behavior and Human Decision Processes*, 115(2):157–168.

-
- Young, H. P. (2008). Social norms. In Durlauf, S. N. and Lawrence, E. B., editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan.
- Zellars, K. L., Tepper, B. J., and Duffy, M. K. (2002). Abusive supervision and subordinates' organizational citizenship behavior. *Journal of Applied Psychology*, 87(6):1068.
- Zumbuehl, M., Dohmen, T. J., and Pfann, G. A. (2013). Parental investment and the intergenerational transmission of economic preferences and attitudes. SOEP Paper no. 570, SOEP.

CENTER DISSERTATION SERIES

CentER for Economic Research, Tilburg University, the Netherlands

No.	Author	Title	ISBN	Published
617	Matjaz Maletic	Essays on international finance and empirical asset pricing	978 90 5668 618 5	January 2020
618	Zilong Niu	Essays on Asset Pricing and International Finance	978 90 5668 619 2	January 2020
619	Bjorn Lous	On free markets, income inequality, happiness and trust	978 90 5668 620 8	January 2020
620	Clemens Fiedler	Innovation in the Digital Age: Competition, Cooperation, and Standardization	978 90 5668 621 5	October 2020
621	Andreea Popescu	Essays in Asset Pricing and Auctions	978 90 5668 622 2	June 2020
622	Miranda Stienstra	The Determinants and Performance Implications of Alliance Partner Acquisition	978 90 5668 623 9	June 2020
623	Lei Lei	Essays on Labor and Family Economics in China	978 90 5668 624 6	May 2020
624	Farah Arshad	Performance Management Systems in Modern Organizations	978 90 5668 625 3	June 2020
625	Yi Zhang	Topics in Economics of Labor, Health, and Education	978 90 5668 626 0	June 2020
626	Emiel Jerphanion	Essays in Economic and Financial decisions of Households	978 90 5668 627 7	July 2020
627	Richard Heuver	Applications of liquidity risk discovery using financial market infrastructures transaction archives	978 90 5668 628 4	September 2020
628	Mohammad Nasir Nasiri	Essays on the Impact of Different Forms of Collaborative R&D on Innovation and Technological Change	978 90 5668 629 1	August 2020
629	Dorothee Hillrichs	On inequality and international trade	978 90 5668 630 7	September 2020
630	Roland van de Kerkhof	It's about time: Managing implementation dynamics of condition-based maintenance	978 90 5668 631 4	October 2020

No.	Author	Title	ISBN	Published
631	Constant Pieters	Process Analysis for Marketing Research	978 90 5668 632 1	December 2020
632	Richard Jaimes	Essays in Macroeconomic Theory and Natural Resources	978 90 5668 633 8	November 2020
633	Olivier David Armand Zerbib	Asset pricing and impact investing with pro-environmental preferences	978 90 5668 634 5	November 2020
634	Laura Capera Romero	Essays on Competition, Regulation and Innovation in the Banking Industry	978 90 5668 635 2	December 2020
635	Elisabeth Beusch	Essays on the Self-Employed in the Netherlands and Europe	978 90 5668 636 9	December 2020
636	Sophie Zhou	Essays on the Self-Employed in the Netherlands and Europe	978 90 5668 637 6	November 2020
637	Vincent Peters	Turning modularity upside down: Patient-centered Down syndrome care from a service modularity perspective	978 90 5668 638 3	December 2020
638	Pranav Desai	Essays in Corporate Finance and Innovation	978 90 5668 639 0	January 2021
639	Kristy Jansen	Essays on Institutional Investors, Asset Allocation Decisions, and Asset Prices	978 90 5668 640 6	January 2021
640	Riley Badenbroek	Interior Point Methods and Simulated Annealing for Nonsymmetric Conic Optimization	978 90 5668 641 3	February 2021
641	Stephanie Koornneef	It's about time: Essays on temporal anchoring devices	978 90 5668 642 0	February 2021
642	Vilma Chila	Knowledge Dynamics in Employee Entrepreneurship: Implications for parents and offspring	978 90 5668 643 7	March 2021
643	Minke Remmerswaal	Essays on Financial Incentives in the Dutch Healthcare System	978 90 5668 644 4	July 2021
644	Tse-Min Wang	Voluntary Contributions to Public Goods: A multi-disciplinary examination of prosocial behavior and its antecedents	978 90 5668 645 1	March 2021

No.	Author	Title	ISBN	Published
645	Manwei Liu	Interdependent individuals: how aggregation, observation, and persuasion affect economic behavior and judgment	978 90 5668 646 8	March 2021
646	Nick Bombajj	Effectiveness of Loyalty Programs	978 90 5668 647 5	April 2021
647	Xiaoyu Wang	Essays in Microeconomics Theory	978 90 5668 648 2	April 2021
648	Thijs Brouwer	Essays on Behavioral Responses to Dishonest and Anti-Social Decision-Making	978 90 5668 649 9	May 2021

This dissertation consists of three essays that discuss the punishment of different types of undesirable behavior, which all have in common that they do not necessarily hurt the punisher directly. Instead, they include (the threat of) harming a passive third party or they constitute behavior of which the punisher may simply disapprove. The first essay examines how people respond to dishonest behavior that benefits them (but hurts a third party) by examining their trust towards the dishonest decision-maker. The results of a laboratory experiment do not provide conclusive evidence, with people trusting honest and dishonest decision-makers to a similar extent. Instead, trust is higher when the moral dilemma is completely avoided by chance and the highest payoff was obtained without having to lie. This implies that moral dilemmas are better prevented than cured. The second essay assesses the extent to which parents engage in norm compliance and norm enforcement in front of their children, with the aim of teaching them the importance of social norms. The field experiment shows that parents are more likely to punish a norm violation by a stranger in front of their children and that they are more likely to help a stranger in need in this case. As such, this essay documents that parents teach their children not simply by modeling desired behavior to them, but also by punishing undesirable behavior. Finally, the third essay examines how anti-social incentives, imposed by an employer, which encourage workers to harm an outside party to the benefit of the employer may backfire and hurt subsequent productivity of workers. In a laboratory experiment, subjects in the role of workers punish employers who imposed anti-social incentives by performing worse in a task that benefits the employer. This effect disappears when the employer had no control over the incentives imposed on the worker. Together, these results show that workers negatively reciprocate the psychological costs imposed on them intentionally and exemplify the importance of organizational leadership.

THIJS BROUWER (Nijmegen, the Netherlands, 1992) obtained his Bachelor degree in Economics and Business Economics in 2014 from Tilburg University. In the same year, he started the Research Master program in Economics at the CentER Graduate School of Tilburg University. After graduating cum laude in 2016, he joined the Department of Economics as a PhD candidate where he was supervised by prof. dr. Jan Potters and prof. dr. Eric van Damme. In 2019, he spent five months at GATE Lyon-St.-Etienne as part of a research visit.

ISBN: 978 90 5668 649 9

DOI: 10.26116/center-lis-2107