

HOW INEQUALITY, FAIRNESS CONCERNS,
AND NORMS AFFECT INDIVIDUAL BEHAVIOUR

ANNA HOCHLEITNER

PhD advisors: Abigail Barr & Silvia Sonderegger



*Thesis submitted to The University of Nottingham
for the Degree of Doctor of Philosophy*

SCHOOL OF ECONOMICS
UNIVERSITY OF NOTTINGHAM

March 2023

ACKNOWLEDGEMENTS

Writing this thesis would not have been possible without all the great people I met along the way and I have a deep gratitude and appreciation for everyone who supported me throughout this journey. First and foremost, I would like to thank my supervisors Abigail Barr and Silvia Sonderegger. Their guidance, expertise, and mentoring has been invaluable to me over the last years. They supported me at all steps of the PhD, always with an open door to discuss new ideas, to give feedback, or to offer advice. Both of them have been role models for me and played a big role in shaping my understanding of what it means to conduct research.

I also want to thank Eugen Dimant and Michele Gelfand, both co-authors on one of my projects. Working together with researchers from diverse institutions and disciplines was a great and highly instructive experience and I learned a lot from their thought processes and the discussions we had for the project. Ultimately, the experience of my PhD would not have been the same without the stimulating environment in the School of Economics, and in particular the Centre for Decision Research and Experimental Economics (CeDEx). Participating and presenting in seminars had a huge impact on me and I am grateful for all the support and feedback that I received over the last years.

This PhD would not have been possible without funding. I am very grateful for the scholarships provided by the School of Economics and the ESRC that allowed me to attend interdisciplinary events and exchange ideas with people from different scientific backgrounds. I would also like to express my sincere gratitude to the ESRC, the School of Economics, CeDEx, Stanford University, and the German Research Foundation (DFG) for providing the resources to conduct the experimental work in this thesis.

In addition to all the scientific support I received, I am also incredibly grateful for the community with fellow PhD students that I was able to experience in Nottingham and all the new friends I made along the way. In particular, I want to thank Sarah Bowen, Candice Ford, Richard Mills, Jose Guinot Saporta, Lara Suraci, and Yuliet Verbel Bustamante. Thanks for all the great memories, the shared laughter, and little adventures. The PhD would have never been the same without you.

I also want to thank my parents Birgit and Robert, my grandparents Waltraud and Engelbert, and my sister Melanie for always having my back. Last but not least, I want to thank my partner Malte Baader for his continuous support and feedback, and for understanding me when technology did not. But most importantly, for making my PhD such a happy and memorable time.

CONTENTS

GENERAL INTRODUCTION	1
I DOES INCREASING INEQUALITY THREATEN SOCIAL STABILITY? EVIDENCE FROM THE LAB	4
1 Introduction	4
2 Experimental design	8
2.1 The battle of the sexes game	8
2.2 Feedback structure	9
2.3 Treatments	9
3 Theoretical framework	11
3.1 Basic model and comparative statics	11
3.2 The effect of inequality on stability	12
3.3 Hypotheses	14
4 Sample and data collection	15
5 Results	16
5.1 Emergence of an unequal convention	17
5.2 Does inequality threaten stability?	18
5.3 Treatment differences	21
5.3.1 The effect of the poor getting poorer	21
5.3.2 Gradual versus sudden increases in inequality	22
5.3.3 Past experiences of (in)stability	23
5.3.4 Reversal of initial conventions	24
6 Discussion and theory revisited	26
7 Conclusion	29
II FAIRNESS IN TIMES OF CRISIS: NEGATIVE SHOCKS, RELATIVE INCOME, AND PREFERENCES FOR REDISTRIBUTION	37
1 Introduction	37
2 Theoretical framework	40
2.1 Basic set-up	40
2.2 The effect of relative income	42
2.3 Experience of shocks	43
3 Experimental design and procedures	45
3.1 Player types and realisation of shocks	45
3.2 Matching and allocation decisions	47
3.3 Hypotheses	47
3.4 Sample and data collection	48

4	Results	48
4.1	Relative income and fairness views	48
4.2	Effect of negative income shocks	49
4.2.1	Constant relative shocks	51
4.3	The role of incomplete information	53
5	Study 2: Income shocks caused by Covid-19	54
5.1	Design and procedures	54
5.2	Results	55
5.2.1	Relative income and fairness views	55
5.2.2	Effect of negative income shocks	56
5.2.3	The role of incomplete information	58
6	Conclusion	59
III	STRATEGIC BEHAVIOUR WITH TIGHT, LOOSE, AND POLARISED NORMS	66
1	Introduction	66
2	Theoretical framework	70
2.1	Case (i): Individuals react only to the mean	71
2.2	Case (ii): Individuals react to the whole distribution	71
2.3	Hypotheses	74
3	Experimental design	75
3.1	Basic set-up and treatment conditions	75
3.2	Two-player PGG and beliefs elicitation	76
3.3	Sample and data collection	78
4	Results	78
4.1	Behaviour in Part I	78
4.2	Effect of variance and shape of the descriptive norm	79
4.3	Personal traits	81
4.4	Effect of high and low means	83
5	Discussion and conclusion	84
	APPENDICES	93
A	Appendix for Chapter 1	93
A.1	Additional analysis	93
A.1.1	Emergence of an unequal convention	93
A.1.2	Does inequality threaten stability?	95
A.1.3	Treatment differences	99
A.2	Theory revisited – simulation results	102
A.3	Instructions	106
B	Appendix for Chapter 2	113
B.1	Theoretical appendix	113
B.1.1	Proofs	113

CONTENTS

B.1.2	Model variation: Shares as reference points	114
B.1.3	Application to experimental setting	115
B.1.4	Extension: self-serving belief formation	116
B.2	Additional analysis	118
B.2.1	Sample and procedures	118
B.2.2	Effect of negative income shocks	119
B.2.3	Constant relative shocks	123
B.2.4	The role of incomplete information	125
B.2.5	Relationship closeness and affect	126
B.2.6	Study 2: Sample and data collection	129
B.2.7	Study 2: Effect of negative income shocks	129
B.2.8	Study 2: The role of incomplete information	133
B.2.9	Study 2: Relationship closeness and affect	134
B.3	Introducing uncertainty about entitlements	136
B.3.1	Motivation and Design	136
B.3.2	Data collection and results	136
B.4	Supplementary materials	140
B.4.1	Instructions	140
B.4.2	Screenshots experiment	142
B.4.3	Study 2: Instructions	149
B.4.4	Study 2: Screenshots experiment	150
C	Appendix for Chapter 3	153
C.1	Theoretical appendix	153
C.1.1	Proofs	153
C.1.2	The disutility of deviating from x_i^a takes the absolute value form	155
C.2	Additional analysis	156
C.2.1	Theoretical framework	156
C.2.2	Experimental Design	157
C.2.3	Sample and data collection	157
C.2.4	Behaviour in Part I	158
C.2.5	Effect of variance and shape of the descriptive norm	160
C.2.6	Personal traits	163
C.3	Quadratic scoring rule	167
C.4	Instructions	168

LIST OF FIGURES

Figure 1.1	The BoS stage game	8
Figure 1.2	Experimental treatments	10
Figure 1.3	SSI for $t \leq 50$	17
Figure 1.4	Stability over time per treatment	19
Figure 1.5	Deviations in treatments with increasing inequality (T1–T3) for $t \leq 50$	21
Figure 1.6	SSI for high inequality in $t > 50$ under different histories	23
Figure 1.7	SSI for low inequality under different histories	24
Figure 1.8	Behaviour in treatments with increasing inequality (by session)	25
Figure 1.9	Simulation of the <i>Shock</i> (T3) treatment if θ_i decreases with H and L	28
Figure 1.10	Simulation of the <i>Shock</i> (T3) treatment if θ_i decreases with Z	29
Figure 2.1	Formation of fairness views and allocation decisions	42
Figure 2.2	Session structure	45
Figure 2.3	Average allocation to self (LHS) and distribution (RHS) by player type.	49
Figure 2.4	Players' reaction to own shock (LHS) and shock to other (RHS)	50
Figure 2.5	Share of players believing that the other suffered a shock by type	53
Figure 2.6	Average allocation to self (LHS) and distribution (RHS) by player type (Study 2)	55
Figure 2.7	Distribution of allocations to self in the absence of shocks to others	56
Figure 2.8	Reactions to own (LHS) and other's (RHS) Covid shocks	57
Figure 2.9	Share of players believing that the other suffered a Covid shock	58
Figure 3.1	$F(x)$ for single peaked (LHS) and polarised (RHS) distributions	72
Figure 3.2	Change in optimal contributions for single-peaked (LHS) and polarised (RHS) distributions	73
Figure 3.3	Experimental conditions: distribution of co-player's contribution	76
Figure 3.4	Overview of the experimental design	77
Figure 3.5	Contributions, aggregate beliefs about the co-player's contribution, and personal values in Part I	79
Figure 3.6	Aggregate beliefs about co-player's contribution in Part II by treatment	80
Figure 3.7	Distribution of participants' contributions in Part II by treatment	80
Figure 3.8	Effect of a high or low mean on contributions	83
Figure A.1	Emergence under low inequality per session	93
Figure A.2	Emergence under high inequality per session	94
Figure A.3	Decrease in switches between A and B over the first 50 periods	94
Figure A.4	Affect responses for $t \leq 50$	98
Figure A.5	Affect responses across treatments	98
Figure A.6	Behaviour in treatments with non-increasing inequality (by session)	101
Figure A.7	Simulation of the <i>Incremental</i> (T1) treatment if θ_i decreases with H and L	102
Figure A.8	Simulation of the <i>Pure Inequality</i> (T2) treatment if θ_i decreases with H and L	102
Figure A.9	Simulation of the <i>Decreasing Inequality</i> (T4) treatment if θ_i decreases with H and L	103
Figure A.10	Simulation of the <i>Control</i> (T5) treatment if θ_i decreases with H and L	103
Figure A.11	Simulation of the <i>Incremental</i> (T1) treatment if θ_i decreases with Z	104

LIST OF FIGURES

Figure A.12	Simulation of the <i>Pure Inequality</i> (T2) treatment if θ_i decreases with Z	104
Figure A.13	Simulation of the <i>Decreasing Inequality</i> (T4) treatment if θ_i decreases with Z	105
Figure A.14	Simulation of the <i>Control</i> (T5) treatment if θ_i decreases with Z	105
Figure B.1	Affect reactions to finding out about a shock/no shock to self	119
Figure B.2	Average allocation to self (LHS) and distribution (RHS) under constant relative shocks	124
Figure B.3	Perception of shocks within the experiment	124
Figure B.4	Share of players believing that the other suffered a shock under constant relative shocks	125
Figure B.5	Reported closeness between players conditional on shocks	126
Figure B.6	Affect responses to other's shock	128
Figure B.7	Share of players believing that the other suffered a Covid shock by player type	133
Figure B.8	Reported closeness between players conditional on Covid shocks	134
Figure B.9	Affect responses to other's Covid shock	135
Figure B.10	Average allocation to self (LHS) and distribution (RHS) by player type (Study 3)	137
Figure B.11	Reaction to a common shock	138
Figure C.1	Cumulative distributions of descriptive norm treatments	156
Figure C.2	Belief elicitation screen	157
Figure C.3	Manipulation checks	158
Figure C.4	Variation in standard deviations of beliefs in Part I	158
Figure C.5	Examples of individual distributions of beliefs in Part I	159
Figure C.6	EEs and NEs in Part I	160
Figure C.7	Distribution of participants' contribution in Part II by variance	160
Figure C.8	Distribution of participants' contribution in Part II (conditional cooperators only)	161
Figure C.9	Different variances and their effect on the high/low end of the distribution	161
Figure C.10	Changes in beliefs about the co-player's distribution between parts	162
Figure C.11	Representation of QSR	168

LIST OF TABLES

Table 1.1	Session Overview	16
Table 1.2	The effect of inequality on coordination over time ($t \leq 50$)	18
Table 1.3	Changes in stability after $t = 50$ by treatment	20
Table 1.4	The effect of the poor getting poorer ($t > 50$)	22
Table 2.1	Pre- and post-shock earnings	46
Table 2.2	The effect of shocks on allocation to self (y_i)	51
Table 2.3	The effect of constant relative shocks on allocation to self (y_i)	52
Table 2.4	The effect of Covid shocks on allocation to self (y_i)	57
Table 3.1	Experimental conditions	75
Table 3.2	Effect of personal traits on contributions in Part II	82
Table 3.3	Effect of high and low mean conditions on contributions in Part II	84
Table A.1	Changes in stability after $t = 50$ – excluding first 10 periods	95
Table A.2	Changes in stability after $t = 50$ – full sample	95
Table A.3	Changes in the probability of choosing A after $t = 50$	96
Table A.4	Probability of deviating from a convention ($t > 50$)	97
Table A.5	The effect of different histories on coordination under high inequality ($\Delta = 75$)	99
Table A.6	The effect of different histories on coordination under low inequality ($\Delta = 25$)	100
Table B.1	Fair allocations to self (m_{i2}^*)	116
Table B.2	Sample characteristics and balance test across shocks	118
Table B.3	Sample characteristics and balance test across player types	119
Table B.4	The effect of shocks on allocation to self (y_i). Testing for order effects.	120
Table B.5	The effect of shocks on allocation to self (y_i). Sub-sample of first decisions.	121
Table B.6	Difference between player types in allocations to self (y_i)	121
Table B.7	The effect of shocks on individuals with different fairness views	122
Table B.8	The effect of shocks on allocation to self (y_i) after controlling for relative perception of shocks	123
Table B.9	The effect of shocks and information on allocation to self (y_i)	126
Table B.10	The effect of shocks on allocation to self (y_i) after controlling for perceived closeness	127
Table B.11	Sample characteristics and balance test across the experience of Covid shocks	129
Table B.12	The difference between player types in allocations to self (y_i) after Covid shocks	130
Table B.13	The effect of Covid shocks on allocation to self (y_i). Testing for order effects.	130
Table B.14	The effect of Covid shocks on allocation to self (y_i). Interaction with order effects.	131
Table B.15	The effect of Covid shocks on allocations to self (y_i). Sub-sample of first decisions.	132
Table B.16	The effect of Covid shocks on allocation to self (y_i) in Study 1	132
Table B.17	The effect of Covid shocks and information on allocation to self (y_i)	134

Table B.18	The effect of Covid shocks on allocations to self (y_i) after controlling for perceived closeness	135
Table B.19	The effect of common shocks on allocations to self (y_i)	139
Table C.1	Effect of variance on contributions in Part II for different sub-samples	162
Table C.2	Effect of personal traits on contributions in Part II — robustness checks	163
Table C.3	Effect of personal traits on contributions in Part II — conditional cooperators only	164
Table C.4	Effect of personal traits on contributions in Part II — controlling for changes in PVs	165
Table C.5	Correlates of personal values, sucker and free-riding aversion	166

GENERAL INTRODUCTION

A recent report commissioned by the French government on future structural challenges for the global economy singled out inequality as one of three major fields of action for economic policy (Blanchard & Tirole, 2021). Given the high societal relevance of the topic, it is not surprising that studying questions of fairness and inequality has a long tradition in the social sciences. A large body of work has shown that people are not behaving fully selfishly, only caring about own material self-interest, but that they are also concerned about fairness and the outcomes of others (see e.g. Hoffman et al., 1994; Fehr & Schmidt, 1999; Camerer, 2011). Differences in these concerns have been shown to matter for real political outcomes and have been linked to the acceptance of inequality and differences in redistributive policies across countries (Alesina et al., 2012; Almås et al., 2020). Investigating how and why individual views on inequality and fairness differ is thus not only important to advance our understanding of economic and political decision-making, but can also directly inform policy makers (see e.g. Cappelen et al., 2007; Mengel & Weidenholzer, 2022; Hvidberg et al., 2023).

In line with this, the first two chapters of this thesis focus on different aspects related to inequality and fairness. Chapter I investigates the implications of increasing inequality for social stability and is based on joint work with Abigail Barr and Silvia Sonderegger. Abstracting from normative arguments, social inequality is often argued to be problematic in itself as it can lead to instability and even social unrest if it reaches a critical magnitude. While the relationship seems intuitive, there are numerous historical counterexamples for this hypothesis and the current empirical evidence is rather mixed (see e.g. Lichbach, 1989; Cramer, 2005). In this chapter, we use a different approach from previous literature and isolate the concepts of inequality and instability in an experimental setting. This allows us to test when and how inequality matters whilst controlling for other environmental factors. We use a repeated battle-of-the-sexes game between two groups, where both groups have an incentive to coordinate on different actions. However, as one action results in higher earnings, the game also encapsulates an inherent element of conflict. Despite the entailed inequality, we observe the endogenous emergence of an unequal convention, where over time each group specialises on a different action. We then exogenously vary the extent of inequality between both actions. Our results show that increasing inequality has a destabilising effect on an established convention. The group that is disadvantaged by the current status quo is thereby the driver of the destabilisation. We show that this behaviour is consistent with a theoretical framework based on disadvantageous inequality aversion and myopic best response. While deviations from the convention can lead into chaos, we also observe the (re-)emergence of stability. Moreover, we find that certain characteristics, such as lowering the payoff for the disadvantaged group or a sudden jump rather than a gradual increase in inequality aggravate the destabilising effect of inequality. Finally, we find that history matters, with people responding differently to the same level of inequality depending on their past experiences.

Chapter II shifts the focus from inequality to fairness views and redistributive preferences. In particular, I investigate whether and how negative economic shocks affect redistributive

preferences. Informed by a theoretical framework, I design an experiment that exogenously varies an individual's experience of negative income shocks before they decide how to redistribute resources. Moreover, the setting introduces differences in relative income, distinguishing between poorer individuals who benefit from redistribution and richer individuals who benefit from the status quo. The results provide causal evidence that poorer and richer individuals hold different fairness views, with the poor favouring higher levels of redistribution. Moreover, these differences in fairness views translate into different responses to negative income shocks. While poorer individuals do not respond to shocks, richer individuals become more opposed to redistribution when they experience a shock themselves, but more supportive of redistribution if poorer individuals are affected. A follow-up study extends and largely confirms these findings by bringing real world income shocks caused by Covid-19 into the experiment.

The third and last chapter of this thesis moves away from the topic of inequality and fairness. Based on joint work with Eugen Dimant, Michele Gelfand, and Silvia Sonderegger, it instead explores another crucial social force: the impact of norms on individual behaviour. A large literature has shown that descriptive norms — the behaviour of other individuals in one's reference group — play a key role in shaping individual decisions (see e.g. [Cialdini & Trost, 1998](#); [Bicchieri, 2005](#)). When characterising the behaviour of others, a standard approach in the literature is to focus on average behaviour. In [Chapter III](#), we argue both theoretically and empirically that not only average behaviour but also the entire distribution of behaviour plays a crucial role in how people react to descriptive norms. Using a representative sample of the U.S. population, we experimentally investigate how individuals react to strategic environments characterised by different distributions of behaviour. Inspired by the literature on tight and loose cultures ([Gelfand et al., 2011](#); [Gelfand, 2021](#)), we thereby focus on the distinction between tight (i.e., characterised by low behavioural variance), loose (i.e., characterised by high behavioural variance), and polarised (i.e., characterised by u-shaped behaviour) environments. We find that individuals indeed strongly respond to differences in the variance and shape of the descriptive norm they are facing. Loose norms generate greater behavioural variance, while polarisation generates polarised responses. Importantly, we also find that relative to tight norms, in polarised and loose environments, personal traits and values play a significantly larger role in determining actual behaviour. This provides important insights into how individuals navigate environments that contain strategic uncertainty.

Even though the three chapters shed light on different facets of individual decision-making and preferences, this thesis thus shows that human behaviour is strongly influenced by the social and economic environment we are interacting within and consequently also responds to changes in the latter. Each of the three chapters is fully self-contained, providing a motivation, study details, results and a conclusion.

References

- Alesina, A., Cozzi, G., & Mantovan, N. (2012). The evolution of ideology, fairness and redistribution. *The Economic Journal*, 122(565), 1244–1261.
- Almås, I., Cappelen, A. W., & Tungodden, B. (2020). Cutthroat capitalism versus cuddly socialism: Are americans more meritocratic and efficiency-seeking than scandinavians? *Journal of Political Economy*, 128(5), 1753–1788.
- Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.
- Blanchard, O. & Tirole, J. (2021). Major future economic challenges. *France Stratégie*.
- Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.
- Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., & Tungodden, B. (2007). The pluralism of fairness ideals: an experimental approach. *American Economic Review*, 97(3), 818–827.
- Cialdini, R. B. & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology* (pp. 151–192). New York: McGraw-Hill.
- Cramer, C. (2005). *Inequality and conflict: A review of an age-old concern*. United Nations Research Institute for Social Development Geneva.
- Fehr, E. & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3), 817–868.
- Gelfand, M. J. (2021). Cultural evolutionary mismatches in response to collective threat. *Current Directions in Psychological Science*, 30(5), 401–409.
- Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., et al. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, 332(6033), 1100–1104.
- Hoffman, E., McCabe, K., Shachat, K., & Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, 7(3), 346–380.
- Hvidberg, K. B., Kreiner, C., & Stantcheva, S. (2023). Social position and fairness views on inequality. *Review of Economic Studies*, forthcoming.
- Lichbach, M. I. (1989). An evaluation of "does economic inequality breed political conflict?" studies. *World Politics*, 41(4), 431–470.
- Mengel, F. & Weidenholzer, E. (2022). Preferences for redistribution. *Journal of Economic Surveys*, (pp. 1–18).

CHAPTER 1

Does increasing inequality threaten social stability? Evidence from the lab[†]

1 Introduction

The idea that inequality can be a threat for the stability of society has a long tradition in political philosophy. Plato warns in his last big dialogue *The Laws* that inequality bears the danger of civil disintegration and even war, which is why both “*extreme poverty and wealth must not be allowed to arise in any section of the citizen-body*” (Plato, cited in [Cooper et al., 1997](#), 744d). A similar idea is laid out in the *Republic*, where Plato attacks the idea of an oligarchic society that systematically excludes the poor, stating that “*of necessity it isn’t one city but two — one of the poor and one of the rich — living in the same place and always plotting against one another*” (Plato, cited in [Cooper et al., 1997](#), 551d). Although Plato here extends the concept of social inequality to the political domain, he again argues that in the presence of inequality, social stability is menaced by a constant struggle among disparate factions. In the course of history this line of argumentation has been taken up repeatedly in different contexts and nuances, for example in the works of Machiavelli, Montesquieu, or Marx (for a discussion see [Lichbach, 1989](#)). At the core of their arguments is that distributional imbalances can give rise to the polarisation of social groups. The fear is a division of society into the disadvantaged who seek to challenge the existing status quo and the advantaged who seek to defend it. It is therefore not surprising that given the increasing gap between rich and poor in many countries today ([Piketty & Saez, 2014](#)), inequality continues to be a major topic in public and political debates. In fact, it has again been put forward as an explanation for disparate events related to political instability such as the Arab Spring ([Roubini, 2011](#)), the widespread rise of populism ([Inglehart & Norris, 2016](#)), or large scale protests like the international “occupy” movement in 2011 or the “yellow vest” protests 2018 in France, in which millions protest against a system that seems to leave them behind ([Satz, 2019](#)).¹

[†]This chapter is based on joint work with Abigail Barr and Silvia Sonderegger. I am grateful to the ESRC (Research Training Support Grant) and CeDEx for funding this research. I also thank CeDEx members and participants at the CCC conference for comments and feedback. The study received ethical approval from the Nottingham School of Economics Research Ethics Committee on 05/04/2019.

¹We fully acknowledge that all examples are highly complex and that there are many factors that contributed to each movement, as for instance the discussion about fuel taxes in the case of the yellow vests ([Grossman, 2019](#)). However, it is noteworthy that inequality keeps reappearing as an ex-post explanation for such events.

While the linkage between inequality and instability seems to be very intuitive, there exists a vast number of historical counter-examples in which we fail to see any form of unrest despite significant inequalities (Cramer, 2005). A possible counterforce might be that societies with high inequality are characterised by the emergence of influential elites, who can use their power to oppress resistance and to ultimately increase the stability of the prevailing system (Lichbach, 1989). Moreover, the idea of the disadvantaged turning against the established order requires an enormous degree of coordination to overcome problems of collective action (Olson, 1965).² It is thus not surprising that despite a large empirical literature dedicated to this topic, evidence for the “inequality favours instability” hypothesis is rather mixed, with many studies suffering from problems of reversed causality, lack of high quality data and employing varying definitions of inequality and stability. In addition, it is basically impossible to isolate the effect of social inequality from its wider political and historical-ideological context (Cramer, 2003, 2005).

In this paper we test the “inequality favours instability” hypothesis using a very different approach from the previous literature and isolate the core concepts of inequality and stability in a laboratory experiment. This allows us to strip off confounding elements and explore the effects of changes in inequality within a controlled environment. In our study, two groups interact with each other in a repeated battle of the sexes game over 100 periods. Total gains are maximised if groups successfully coordinate on choosing distinct actions, reflecting the idea that society benefits from its members specialising in different areas and then engaging in social exchange. However, while such a specialisation is beneficial for society as a whole, it often gives one group an advantage over another if actions are not remunerated equally (see e.g. Henrich & Boyd, 2008). Mirroring this notion in our experiment, one action results in a higher payoff than the other, creating an inherent conflict about who chooses which action. Over the first 50 periods we hold the payoffs for each action constant and find that despite the inherent inequality, most sessions coordinate on an equilibria in which one group specialises on the high payoff and one on the low payoff action. We refer to such a coordination as the emergence of a convention.³

While this endogenous emergence of inequality is a striking result in its own right, our key focus is to investigate the stability of such equilibria in the context of changing inequalities. To do so, we exogenously vary the payoff difference between actions after a convention has emerged. To inform our design and generate predictions for the experiment, we develop a theoretical framework based on disadvantageous inequality aversion and myopic best response. The model predicts that as inequality increases, for some individuals in the disadvantaged group, the disutility from receiving less than others starts to outweigh the gains from coordination. As more individuals deviate from the convention, it becomes less risky for others to follow and the equilibrium unravels. In line with our model and the philosophical argu-

²Problems of collective action arise as it is preferable for the individual to have others fight for change and benefit from any achieved improvements without risking negative repercussions in case of failure. As a consequence, mobilisation for protest movements can be impeded (Collier, 1999; Blattman & Miguel, 2010).

³See Section 3.2 for a formal definition.

ments outlined above, our experiment shows that increasing inequality has a destabilising effect. The group that is disadvantaged by the current status quo is thereby the driver of the destabilisation and initialises deviations from the established convention.

To develop a more nuanced understanding of how inequality matters, we vary the underlying inequality dynamics across different treatments. The results show that certain factors accelerate instability. First, we observe more instability after a sudden increase in inequality than if the same level of inequality is reached via gradual changes. Second, if inequality is driven by the poor getting poorer responses are stronger than if the same level of inequality is a result of the rich getting richer. Finally, our results show that the history of inequality matters. Having experienced instability in the past makes the emergence of a convention significantly harder even if inequality has decreased. Likewise, having a history of stability helps to maintain a convention even in environments of high inequality. Taken together our results thus confirm the importance of inequality for stability, but also provide a strong argument for taking past experiences as well as the precise inequality dynamics into account.

Our findings contribute to the literature in several ways. First, we provide evidence for the “inequality favours instability hypothesis” using a novel methodology. As discussed above, previous evidence has been mixed with some studies finding a positive relationship between income inequality and socio-political instability or crime (see e.g. [Hsieh & Pugh, 1993](#); [Alesina & Perotti, 1996](#); [Fajnzylber et al., 1998](#); [Kennedy et al., 1998](#); [Stewart, 2000](#); [Østby, 2008](#)), while others argue social conflict is driven by other factors such as absolute deprivation and not income inequality itself ([Collier, 1999](#); [Fearon & Laitin, 2003](#); [Somanthan, 2020](#)). Both of these views are supported by different conceptual underpinnings. *Deprived actor theories*, on the one hand, argue that inequality — or unmet economic and social expectations — lead to frustration, emotional outrage, weaker social cohesion, and ultimately to conflict ([Gurr, 1971](#); [Conteh-Morgan, 2004](#); [Blattman & Miguel, 2010](#)).⁴ This notion is also backed by behavioural economic research on preferences for equity, suggesting that individuals experience a disutility as inequality increases (see e.g. [Fehr & Schmidt, 1999](#); [Charness & Rabin, 2002](#)). On the other hand, *rational actor theories* have emphasised the role of opportunity structures and resources ([Jakobsen, 2010](#)), describing conflict as the consequence of a deliberate cost-benefit analysis ([Eckstein, 2004](#)). Under these theories, social unrest is not expected if costs are too high or the disadvantaged do not possess the necessary resources, implying that inequality is not directly necessary nor sufficient for conflict ([Lichbach, 1989](#); [Cramer, 2005](#)).⁵ Our findings show that inequality indeed leads to instability in interactions between groups. Moreover, we find that the disadvantaged group displays more and more negative emotions as inequality increases. This provides support for the emotion-based argument of deprived actor theories.

Additionally, our paper contributes to theoretical and experimental work on the emergence of unequal conventions. [Young \(1993\)](#) describes conventions as a “*pattern of behaviour*

⁴Note that not only inequality but subjective evaluations of inequality play a crucial role ([Stewart, 2000](#); [Gimpelson & Treisman, 2018](#)). Moreover, some degree of inequality may even be seen as a positive sign for social mobility ([Hirschman & Rothschild, 1973](#)).

⁵Particularly relevant factors in this context are problems of collective action and repression by current elites.

that is customary, expected and self-enforcing” (p.57). Several papers have modelled conventions as the result of an evolutionary coordination process (Lewis, 1967; Baronchelli, 2018; Young, 1996) that can lead to persistent inequalities between groups (Axtell et al., 2001; Binmore et al., 2003; Henrich & Boyd, 2008; Bowles et al., 2014).⁶ In line with these theories, a number of experiments show that groups do in fact coordinate on efficient, but unequal equilibria (Holm, 2000; Dale et al., 2002; Hargreaves-Heap & Varoufakis, 2002; Oprea et al., 2011; Benndorf et al., 2016; Berger et al., 2022).⁷ Our study adds to this literature by providing further evidence for the emergence of unequal conventions, while also showing that environmental factors such as the extent of the involved inequality play a crucial role for the likelihood of successful coordination.

Finally, our paper speaks to the inertia versus the possibility for change of established conventions. Despite their often arbitrary origin, theoretical work shows that unequal conventions can perpetuate over long periods and can even prevail despite being inefficient (Hwang et al., 2018; Belloc & Bowles, 2013). Andreoni et al. (2021) show that even if a convention is undesirable for most people, individuals are still facing major difficulties to initiate behavioural changes. While this inertia can even carry over from one generation to the next (Schotter & Sopher, 2003), changes are not impossible and can occur through external shocks, errors, or conscious deviations (Belloc & Bowles, 2013; Acemoglu & Jackson, 2014; Hwang et al., 2018; Baronchelli, 2018). The change dynamics thereby often exhibit social tipping points, meaning that once a crucial threshold is reached, change is sudden rather than incremental (Young, 2015). Previous experimental studies have shown support for both the notion of social tipping points (Andreoni et al., 2021; Centola et al., 2018), as well as the idea that external shocks can facilitate changes in established equilibria (Brandts & Cooper, 2006). While we share the interest of how and whether a convention can be broken, a crucial difference to our work is that these papers focus on situations of common interest, i.e. scenarios where a deviation from the convention is beneficial for most people but they are trapped in an inefficient equilibrium. In our experiment, by contrast, we intentionally introduce an element of conflict, where some individuals benefit and others lose from the status quo. We show that even in situations without common interest, established conventions can be broken if there is sufficient pressure from the disadvantaged group.

The remainder of this paper is structured as follows. Section 2 describes our experimental design, while Section 3 outlines the theoretical framework and hypotheses. In Section 4 and 5 we describe the data collection and present our findings. In Section 6 we discuss potential extensions of the theoretical framework, before Section 7 concludes.

⁶In fact, some theories assume that inequality is almost inevitable and comes as a natural consequence of specialising on different tasks even without initial differences between individuals (Mookherjee & Ray, 2002).

⁷See also Gächter et al. (2017) on the endogenous emergence of inequality in public good games.

2 Experimental design

2.1 The battle of the sexes game

To test whether inequality threatens stability we design a dynamic experimental environment in which two groups interact repeatedly over 100 periods. Individuals are randomly assigned to a group at the beginning of the experiment ($N=7$) using the minimal group paradigm (Billig & Tajfel, 1973). This means group affiliation has no deeper meaning and is only determined by an arbitrary factor, in our case the random draw of a coloured ball (green or yellow). Group affiliation (green/yellow) is fixed for the duration of the whole experiment.

In each period, a member of the green group interacts with a different, randomly chosen member of the yellow group and plays a battle of the sexes (BoS) game. Each chooses between two actions (A or B) and if both players choose the same action, their payoff is zero, while if they coordinate on different actions each player receives a positive payoff. However, in the case of coordination the player choosing A receives a high payoff h , while the one choosing B receives a low payoff l . Figure 1.1 shows a generic version of the game.

In the first 50 periods, we hold the payoffs for action A and B constant, allowing groups to coordinate on different actions. After $t = 50$, we then introduce different exogenous payoff changes in order to explore the effects of changing inequalities on the stability of coordination. Participants are aware that payoffs can change at any time during the experiment, but do not know how and when. How payoffs and inequality dynamics differ between treatments is explained in Section 2.3.

The BoS game is particularly well suited to study the impact of inequality on stability due to two key characteristics. First, it provides strong incentives for individuals to coordinate on different actions — and thereby accepting inequality —, as the cost of miscoordination is very high (i.e. a payoff of zero). Second, in contrast to a pure coordination game, the BoS game encompasses an inherent element of conflict, as one player has to give in and accept the lower payoff l from choosing B.

As can be seen from Figure 1.1, the stage game has two pure asymmetric Nash equilibria (A, B) and (B, A), as well as one symmetric mixed Nash equilibrium (NE). In a repeated game,

Figure 1.1: The BoS stage game

		Yellow	
		A	B
Green	A	0, 0	h , l
	B	l , h	0, 0

Note: $h > l \geq 0$.

there thus exist many possible strategy profiles, for instance playing a mixed NE over all periods, alternating between pure NE, or repeating the same pure NE in each period. While all these outcomes are theoretically possible, we are particularly interested in strategy profiles in which each group specialises on a distinct action, playing the same pure NE repeatedly. As mentioned above, we refer to such a profile as a convention (see also [Young, 1993](#)). This is not only the most relevant case for the purpose of investigating inter-group inequality, but also requires the least amount of cognitive effort for participants and has been shown to be the one most likely to arise (see e.g. [Hargreaves-Heap & Varoufakis, 2002](#); [Berger et al., 2022](#)).⁸ Once established, a convention eliminates the insecurity about the other group’s behaviour and avoids inefficient miscoordination. This might be preferable even for the group specialising on the low payoff option B, despite the resulting inter-group inequality. In the following we refer to the group predominantly choosing A as the *dominant* group and the group predominantly choosing B as the *submissive* group.

2.2 Feedback structure

In order to allow participants to learn and update their expectations, they receive feedback on their individual outcome and the outcomes of their own group members after each period. More precisely, they are informed about their payoff, as well as the average payoff from action A and B within their own group. Intuitively, our feedback structure captures the idea that individuals can both obtain information from their own personal experience but also from the experience of others. While the latter is often modelled in evolutionary game theory as a random sample of decisions and outcomes of x other players ([Young, 1996, 1993](#)), it is reasonable to assume that this sampling process is non-random in the presence of different groups. Previous research shows that there is a higher probability to learn from ingroup-members and that segregated information networks are widespread ([McPherson et al., 2001](#); [Henrich & Boyd, 2008](#); [DiPrete et al., 2011](#)). This interplay of individual learning and social environment is considered crucial for the perpetuation of inter-group inequalities ([Bowles et al., 2014](#)). The design choice of providing only feedback on own group outcomes can be seen as an extreme case of such segregation, capturing the idea that information is more likely to be passed on within than across social groups.

2.3 Treatments

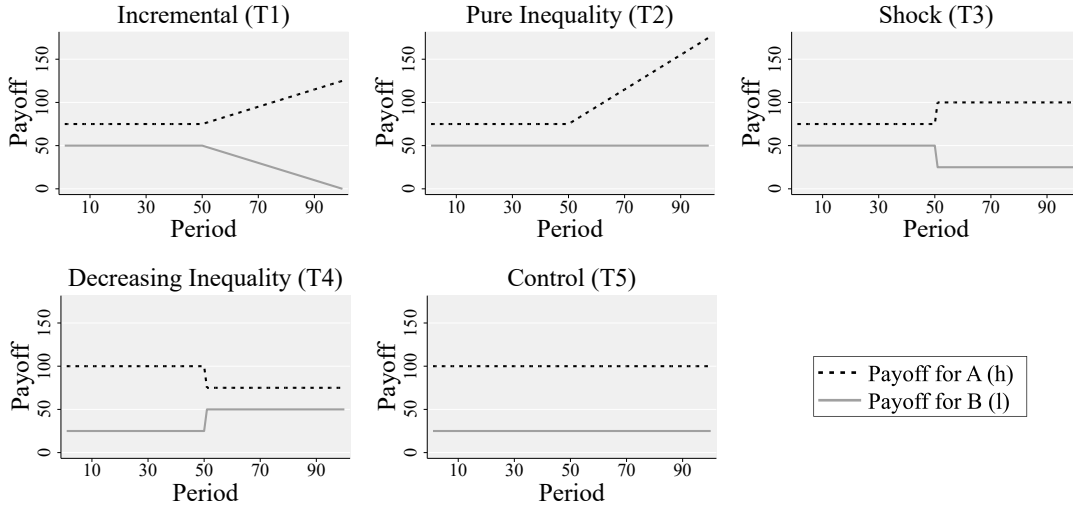
We administer five different treatments that differ both in terms of initial payoff inequality for $t \leq 50$ and the type of change after $t = 50$. Figure 1.2 provides an overview of the different treatments and shows how payoffs for action A and B develop over time.

Let us first focus on the three treatments in the first row (T1–T3). They all start with low levels of inequality in the first 50 periods ($\Delta_t = h_t - l_t = 25$)⁹, which is then increased in the

⁸For instance, while alternating between action A and B avoids inequalities between the two groups, it is very difficult to establish such a rule if participants are randomly re-matched in each period.

⁹The numbers were calibrated using previous work by [Berger et al. \(2022\)](#) and results from a pilot in which we tested whether participants coordinated on an unequal convention under these payoffs.

Figure 1.2: Experimental treatments



second 50 periods, allowing us to assess whether growing inequality destabilises an existing convention. Nevertheless, the treatments differ with respect to the nature of the increase. In the *Incremental* (T1) treatment the gap between h and l widens gradually, both increasing the returns to action A and lowering those to action B. The increase in inequality is thus driven both by the rich getting richer and the poor getting poorer. This implies that, over time, deviations from the convention become less costly for the submissive group. To see the intuition consider the very last period ($t=100$). In that period l_t is equal to zero, meaning that — abstracting from altruistic motives — members of the submissive group will have no incentive whatsoever to uphold the established convention.

To account for this “having less to lose” effect, we also implement the *Pure Inequality* (T2) treatment, in which we increase inequality, but hold l constant. Note that, in this treatment, the extent of inequality is identical to that in the *Incremental* (T1) treatment. Any difference to T1 must therefore be caused by the fact that the payoff for the submissive group remains unchanged.

In both T1 and T2 inequality increases gradually over time. However, external shocks like an economic crisis, may lead to sudden jumps in inequality rather than gradual changes. The *Shock* (T3) treatment reflects this idea and introduces a very sudden and pronounced change in inequality (to $\Delta = 75$) at $t = 51$.¹⁰ This could lead to stronger responses than a gradual increase in inequality. As the change of an established convention requires a critical mass of deviations to be successful, a big one-time jump could serve as a salient signal to change previous behaviour and help to overcome problems of collective action. Moreover, gradual changes might make individuals more accustomed to inequality allowing a convention

¹⁰As we ran the treatments sequentially, the choice of $\Delta = 75$ was informed by results from the *Incremental* (T1) treatment. This is equivalent to the average value of inequality at which the convention broke down in T1.

to be sustained at higher levels of inequality (Andreoni & Samuelson, 2006; Weber, 2006). By contrast, if individuals only care about absolute inequalities there should be no differences between the *Shock* (T3) and the gradual treatments (T1, T2) for $\Delta = 75$.

In contrast to the first three treatments, T4 and T5 exhibit high levels of inequality from the beginning ($\Delta = 75$) and no increase after $t = 50$. In the *Decreasing Inequality* (T4) treatment inequality is even reduced after $t = 50$, while it remains at a constant level in the *Control* (T5) treatment. Focussing on the first 50 periods, these treatments allow us to explore whether unequal conventions also emerge in a high-inequality environment. Moreover, by comparing them to T1–T3, we are able to test in how far different histories of inequality affect behaviour.

3 Theoretical framework

3.1 Basic model and comparative statics

To understand how individuals respond to changes in inequality, we develop a theoretical framework based on inequality aversion and myopic best response. Following the seminal work by Fehr & Schmidt (1999), we assume that, despite individual heterogeneities, most people dislike inequality. To simplify the model, we only focus on disadvantageous inequality aversion, i.e. an aversion of being behind. In our framework, inequality aversion is captured by θ_i that is drawn for each individual at the beginning of the game from the distribution suggested by Fehr & Schmidt (1999), such that $0 \leq \theta_i \leq 4$.¹¹ The second key element of the framework — myopic best responses — implies that individuals react in each period to the given environment by choosing the action that maximises their expected utility in the following period, but without any strategic considerations of future play. An individual would thus choose A if $E(u_{A,t}) \geq E(u_{B,t})$.

The expected utility of choosing an action depends on the individual's inequality aversion (θ_i) and the expected payoff of each action ($\pi_t \in \{h_t, l_t, 0\}$). As payoffs depend on both own and other's choices, individual i of group g ($g \in \{Y, G\}$) ultimately needs to assess the share of players in the other group $g' \neq g$ choosing A ($\lambda_t^{g'}$) or B ($1 - \lambda_t^{g'}$) in the following period. In line with the assumption of myopic best response, we formalise $E(\lambda_t^{g'}) = \lambda_t^{g'}$, i.e. individuals expect the same share of players to choose A as in the last period.¹² Since in the very first period individuals cannot turn to past experience, we assume that each individual forms a belief about $\lambda_t^{g'}$ that is drawn from a uniform distribution, $E(\lambda_1^{g'}) \sim U(0, 1)$, and updated in following periods.

Definition: Myopic best response *At time t , an individual of group $g \in \{Y, G\}$ selects the action that maximises their expected utility conditional on $E(\lambda_t^{g'}) = \lambda_t^{g'}$, i.e. on the expectation*

¹¹If in line with Fehr & Schmidt (1999), individuals mind disadvantageous inequality (β_i) more than advantageous inequality (α_i), θ_i can be interpreted as the net difference between β_i and α_i .

¹²We acknowledge that expectations could be modelled in many ways and that the assumption of myopic best response constitutes a simplification, especially in the presence of big changes (as e.g. in the *Shock* (T3) treatment). Nevertheless, the model allows us to focus on the equilibria before and after such changes, as after a period of perturbation agents should return to myopic best response.

that the share of individuals of group $g' \neq g$ selecting action A in the next period equals the share selecting that action in the current period.

Using this definition, we can now formulate an individual decision rule comparing $E(u_{A,t})$ and $E(u_{B,t})$. If individuals are inequality averse, the utility from choosing B depends negatively on both the strength of inequality aversion (θ_i) and the size of the inequality ($\Delta_t = h_t - l_t$). An individual of group g would choose A iff

$$\underbrace{(1 - E(\lambda_t^{g'}))h_t}_{E(u_{A,t})} \geq \underbrace{E(\lambda_t^{g'})(l_t - \theta_i \Delta_t)}_{E(u_{B,t})} \quad (1)$$

or

$$\theta_i \geq \underbrace{\frac{E(\lambda_t^{g'})(l_t + h_t) - h_t}{\Delta_t E(\lambda_t^{g'})}}_{\equiv \theta_{g,t}^*}. \quad (2)$$

Rearranging Equation (1), we derive a group-specific threshold $\theta_{g,t}^*$ that an individual's inequality aversion needs to exceed in order to choose A (see Equation (2)). This allows us to derive a number of comparative statics on the conditions that make it easier or harder to sustain cooperation. Holding everything else constant, $\theta_{g,t}^*$ increases with $E(\lambda_t^{g'})$ and l_t , implying a lower probability of i choosing A.¹³ This captures the notion that with a higher payoff for B or more agents in the other group choosing A, action A becomes less attractive. By contrast, $\theta_{g,t}^*$ decreases with the extent of inequality Δ_t and h_t , making it more likely that an individual chooses A. Intuitively, action A becomes more attractive the higher its payoff is both in absolute terms (h_t) and compared to the payoff for action B (Δ_t). The comparative statics are summarised in Proposition 1. The proof follows directly from Equation 2.

Proposition 1: *Everything else being equal, the probability of an individual of group g choosing A ($\theta_i \geq \theta_{g,t}^*$) decreases with the expected share of A choices in the other group, $E(\lambda_t^{g'})$, as well as the payoff for B (l_t), but increases with the extent of inequality Δ_t and the payoff for A (h_t).*

Note that for the same level of $E(\lambda_t^{g'})$, l_t , h_t , and Δ_t , it is more likely that a person chooses B the lower their individual level of disadvantageous inequality aversion, θ_i . For this reason the group with higher levels of disadvantageous inequality aversion is more likely to become the dominant one.

3.2 The effect of inequality on stability

Before being able to assess stability, we first introduce a formal definition of a convention. As stated above, we speak of a convention if one group specialises on choosing A and the other group on choosing B, i.e. $\lambda_t^g \rightarrow 1$ and $\lambda_t^{g'} \rightarrow 0$. In the dominant group the share of people choosing A must thus at least surpass a certain threshold x . Similarly, in the submissive group the share of individuals choosing A must lie below a certain threshold y . As 50% of

¹³While individuals in the experiment cannot directly observe $\lambda_t^{g'}$, they can derive it from the average payoffs from choosing action A or B within their group.

participants in each group choosing A is equivalent to perfect chaos, the weakest requirement for a convention would be that the majority of members in the submissive group chooses B ($0 \leq y < 0.5$), while the majority of members in the dominant group chooses A ($0.5 < x \leq 1$). Following the description of a convention by Young (1993) as “*customary, expected and self-enforcing*” (p.57), we moreover require participants to expect that the convention will still be followed in the next period.

Definition: (x,y) – convention with g dominance We say that a (x,y) – convention with g dominance holds at t when (i) $\lambda_t^g \geq x > 0.5$, (ii) $\lambda_t^{g'} \leq y < 0.5$, (iii) $E(\lambda_t^g) \geq x$, and (iv) $E(\lambda_t^{g'}) \leq y$.

We can then define stability as the share of individuals following a convention. From the comparative statics, we see that an increase in inequality will translate into a higher share of individuals choosing A. If inequality becomes sufficiently large, more and more members of the submissive group will prefer A over B and the convention starts to attenuate. To see that deviations must always be initialised by the submissive group note that $\theta_{g,t}^*$ is group-specific due to its dependence on expectations. Assume we have a (x,y) – convention with g dominance such that $\lambda_t^g, E(\lambda_t^g) \geq x, \lambda_t^{g'},$ and $E(\lambda_t^{g'}) \leq y$. With most members of g' choosing B, an agent of the dominant group g would only deviate from the convention and choose B iff

$$\theta_i \leq \theta_{g,t}^* = \frac{E(\lambda_t^{g'})(l_t + h_t) - h_t}{\underbrace{E(\lambda_t^{g'})\Delta_t}_{< 0 \text{ as } E(\lambda_t^{g'}) \leq y < 0.5}} \quad (3)$$

As Equation (3) shows, $\theta_{g,t}^*$ is negative for members of the dominant group independent of the degree of inequality Δ_t , as $E(\lambda_t^{g'}) \leq y < 0.5$. As by definition $\theta_i \geq 0$, the condition for choosing B will never be met for a member of the dominant group, not even for those with no inequality aversion. Consequently, if there is a deviation from an established convention it can never be initialised by the dominant group.¹⁴ Intuitively, for a member of a dominant group, with more than 50% of individuals in g' choosing B ($E(\lambda_t^{g'}) \leq y < 0.5$), a deviation would mean giving up getting h with a high probability for a higher chance of receiving zero.

Let us now turn to members of the submissive group g' and explore whether they would deviate from the convention and choose A. From Equation (2), we see that as Δ_t increases, the threshold for choosing A ($\theta_{g,t}^*$) is lowered. This means that the condition for choosing A will be eventually met for individuals with a sufficiently high inequality aversion θ_i . By deviating to A, individuals knowingly accept a high risk of receiving zero ($E(\lambda_t^g) > 0.5$), because inequality has passed the threshold that they are willing to accept. As inequality continues to increase, this is true for a growing number of individuals in g' , further destabilising the existing convention.

Moreover, the direct consequence of an individual in g' deviating to A is that $\lambda_t^{g'}$ increases.

¹⁴If we allow for advantageous inequality aversion and assume that the latter is larger than disadvantageous inequality aversion, deviations could be theoretically initialised by the dominant group.

This can in turn induce reactions from the dominant group. The first individuals in g who will react are the ones with the lowest degree of inequality aversion θ_i , as deviating implies forgoing h_t for l_t . As the lowest possible level of inequality aversion in our model is $\theta_i = 0$, the minimum share of agents in g' required to induce a deviation for a member of the dominant group is given by

$$0 \geq \frac{E(\lambda_t^{g'}) (l_t + h_t) - h_t}{E(\lambda_t^{g'}) \Delta_t} \longrightarrow E(\lambda_t^{g'}) \geq \frac{h_t}{l_t + h_t}. \quad (4)$$

As by definition $\Delta_t > 0$, it follows that $\frac{h_t}{l_t + h_t} > 0.5$. In other words, as long as the expected share of individuals in g' who deviate to A is not above 50%, there will not be any reaction from the dominant group. However, once the share of individuals in g' choosing A surpasses the threshold defined in Equation (4), individuals in the dominant group will start deviating to B , further destabilising the former convention and accelerating additional deviations from the submissive group.

Proposition 2: Inequality favours instability *Increases in inequality will destabilise an existing convention. Deviations are always initialised by the submissive group. This in turn can induce reactions from the dominant group once $E(\lambda_t^{g'}) \geq \frac{h_t}{l_t + h_t} > 0.5$.*

3.3 Hypotheses

Using our theoretical framework, we can derive predictions for behaviour across our experimental treatments. First, let us focus on the first 50 periods of the game, where we keep payoffs constant in order to let an unequal convention emerge. In three treatments (*Incremental (T1)*, *Pure Inequality (T2)*, *Shock (T3)*) individuals are initially exposed to low levels of inequality ($\Delta_t = 25$), whereas in the other two (*Decreasing Inequality (T4)*, *Control (T5)*) they face high levels of inequality ($\Delta_t = 75$). For a convention to emerge, inequality aversion must lie above the critical threshold for choosing A for most people in one group ($\theta_i \geq \theta_{g,t}^*$), but below for most people in the other group ($\theta_i \leq \theta_{g,t}^*$). If inequality is larger, $\theta_{g,t}^*$ will be lower, making it less likely that a majority of individuals chooses B in one group (see Proposition 1). This means that both the probability that a convention emerges, as well as the strength of the convention depend negatively on Δ_t .

Hypothesis 1: *A higher payoff inequality between actions A and B (Δ_t) impedes coordination on different actions. An unequal convention is less likely to emerge in the first 50 periods in treatments with higher payoff inequality ($T4$, $T5$) and if it does emerge, compliance will be lower.*

Next, we turn to what happens under changing inequality after $t = 50$. From Proposition 2 we expect that an increase in inequality will destabilise an established convention, with deviations being initialised by the submissive group.

Hypothesis 2: *i) In treatments where inequality increases ($T1$, $T2$, $T3$), we expect lower levels of stability in the second 50 than in the first 50 periods of the experiment. ii) Deviations from the established convention will be initialised by the submissive group.*

Finally, we turn to how different inequality dynamics affect behaviour. Both the *Incremental* (T1) and *Pure Inequality* (T2) treatment show the same gradual increase in inequality, but only T1 is characterised by a decreasing payoff for action B (l_t). Proposition 1 shows that a lower l_t increases the probability of choosing A, implying that deviations from the convention will be faster in T1 (holding everything else constant).

Hypothesis 3: *Holding θ_i and $E(\lambda_t^{g'})$ constant, a convention will destabilise no earlier in the Pure Inequality (T2) than in the Incremental (T1) treatment. If an individual decides to deviate in a given period in T2, they would have also deviated in T1.*

The assumption of myopic best response implies that individuals' choices depend only on the conditions in the current period, but not on the history of the game, i.e. past experiences of inequality or stability. Holding $E(\lambda_t^{g'})$ and θ_i constant, we thus expect the same level of stability in treatments that show the same level of inequality and payoffs. For instance, the *Shock* (T3) treatment at $t = 51$ and the *Incremental* (T1) and *Pure Inequality* (T2) treatments at $t = 75$ share the same level of inequality, but differ with respect to the nature of change (sudden/gradual). In addition, T3 has the same level of inequality for $t > 50$ as the *Control* (T5) treatment, but a history of low inequality for $t \leq 50$. Similarly, the *Decreasing Inequality* (T4) treatment exhibits the same amount of inequality for $t > 50$ as T1–T3 for $t \leq 50$, but has a history of high inequality. The model predicts that these differences do not matter.

Hypothesis 4: *The effect of inequality on stability is independent of the history of the game. The level of stability will be the same for i) the Shock (T3) at $t = 51$ and the Incremental (T1) and Pure Inequality treatments (T2) at $t = 75$, ii) the Shock (T3) and Control (T5) treatment for $t > 50$ and iii) the Decreasing Inequality (T4) treatment for $t > 50$ and T1–T3 for $t \leq 50$.*

4 Sample and data collection

All experimental sessions took place between May and September 2019 in the CeDEx laboratory at the University of Nottingham. The experiment was programmed using z-Tree (Fischbacher, 2007). Students from the University of Nottingham were recruited via the online recruitment system for economic experiments (ORSEE) (Greiner, 2015). Interactions were anonymous and communication was not allowed during the experiment. Upon their arrival participants were randomly assigned to computer terminals and informed about the procedure. In order to establish common knowledge among all participants, instructions were read aloud (see Appendix A.3 for instructions). After answering control questions, we randomly divided participants into a yellow and a green group by the draw of a coloured ball, with seven participants in each group. In each period participants were then randomly rematched with a player of the other group (either green or yellow) and played the BoS game.

In addition to the individual choice data, we elicited emotional affect, focusing on the dimension of valence — i.e. positive versus negative feelings (Russell, 2003) — at repeated points in the experiment. We thereby used a simple and fast variant of a pictorial assessment scale developed by Desmet et al. (2001). The reason why affect might be an interesting variable in

Table 1.1: Session Overview

Treatment	Participants	Sessions	h,l at $t \leq 50$	h,l at $t > 50$
Incremental (T1)	84	6	75,50	75+x,50-x
Pure Inequality (T2)	84	6	75,50	75+2x,50
Shock (T3)	80	6	75,50	100,25
Decreasing Inequality (T4)	84	6	100,25	75,50
Control (T5)	84	6	100,25	100,25

Note: $x \in \{1, 50\}$ for periods 51 to 100. Due to recruitment difficulties we ran two sessions with six instead of seven participants per group in T3. In T4, we ran one session with six and one session with eight participants per group. Controlling for group size shows that the latter does not affect results.

our set-up is the proposed link between inequality and instability via the emotional response to relative grievances (Cramer, 2005; Blattman & Miguel, 2010). After the final period, participants received feedback on their accumulated payoffs and a summary of the performance across groups. We then elicited risk aversion, using a version of Holt & Laury (2002) adjusted to our context, personality via the Big Five (Gerlitz & Schupp, 2005) and social value orientation (SVO) to measure social preferences (Murphy et al., 2011; Murphy & Ackermann, 2014).¹⁵ These measures serve as additional controls for understanding individual choices in the BoS game, as the latter involves both elements of risk (i.e. the risk of receiving zero) and pro-sociality (i.e. the willingness to give in and choose B). Finally, we collected information on demographics.

We ran six sessions for each treatment, resulting in 30 sessions in total (see Table 1.1). Each session consisted of 100 periods and lasted between 60 and 75 minutes. The final payoff consisted of the aggregate profits across all periods. In addition, for the risk elicitation and the SVO task, one decision was chosen randomly to be payoff relevant, resulting in average earnings of £9.56 per hour (SD 2.44).

5 Results

To investigate the emergence and stability of a convention, we construct an aggregate index that measures stability within a session, the social stability index (SSI). It compares the share of individuals choosing action A in the green ($\lambda_{s,t}^G$) versus the yellow group ($\lambda_{s,t}^Y$) in a given session s and period t (see Benndorf et al., 2016, for a similar approach):

$$SSI_{s,t} = |\lambda_{s,t}^G - \lambda_{s,t}^Y|, \text{ with } 0 \leq SSI_{s,t} \leq 1 \quad (5)$$

Intuitively, if everyone chooses the same action, $SSI = 0$, indicating complete chaos.¹⁶

¹⁵The Big Five measure key dimensions of personality, namely openness, conscientiousness, extroversion, agreeableness and neuroticism. The SVO task was implemented using the z-Tree code by Crosetto et al. (2012).

¹⁶While the social stability index (SSI) is calculated for each session and period, we drop the subscripts s and t for ease of notation.

By contrast, $SSI = 1$ describes a situation of perfect coordination and the strongest possible convention, with all members in one group choosing A and all members in the other group choosing B. We use the SSI throughout the remainder of this chapter to test our hypotheses.

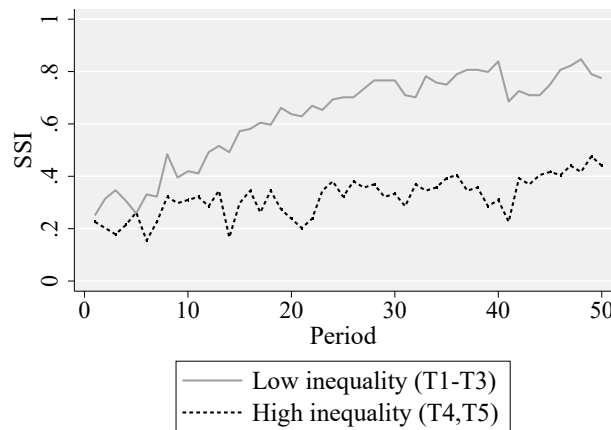
5.1 Emergence of an unequal convention

The necessary basis for all further analysis in this chapter is that participants coordinate on an unequal convention in the first 50 periods of the experiment. During these periods, inequality is fixed at either a low level with $\Delta = 25$ (T1, T2, T3) or a high level with $\Delta = 75$ (T4, T5). We can thus compare the emergence of a convention in treatments with low and high inequality.

Figure 1.3 shows the average SSI for the first 50 periods under different levels of inequality. While we see a clear upward trend in treatments with low inequality (solid gray), stability only increases marginally over time under high inequality (dashed black). Focusing on the last 5 periods as an approximation of equilibrium play, we see that stability is substantially lower under high inequality ($SSI = 0.81$ for $\Delta = 25$ versus $SSI = 0.45$ for $\Delta = 75$). Across all periods, the stability in high inequality treatments reaches only 52% of the stability in low inequality treatments (t-test, $p < 0.001$).

Table 1.2 presents the results of a random effects panel regression of the SSI as the dependent variable on a time trend, a high inequality dummy and their interaction. The results confirm that the difference between high and low inequality treatments is statistically significant.¹⁷ Table 1.2 shows that the time trend for the SSI is positive and significant in the low inequality treatments, confirming the emergence of an unequal convention. For the high inequality treatments, by contrast, stability is not only already lower in the first period, but also shows a significantly weaker increase over time. These results are robust to the inclusion of controls in column (2) and fixed effects in column (3). We thus find evidence in support

Figure 1.3: SSI for $t \leq 50$



¹⁷A Levin-Lin-Chu unit root test shows that the data is stationary ($p < 0.001$), supporting the use of a static panel data model. This is true for $t \leq 50$, as well as for all 100 periods.

Table 1.2: The effect of inequality on coordination over time ($t \leq 50$)

	(1)	(2)	(3)
High inequality	-0.132*** (0.049)	-0.156*** (0.056)	
Time trend	0.011*** (0.001)	0.011*** (0.001)	0.011*** (0.001)
High inequality x time trend	-0.007*** (0.002)	-0.007*** (0.002)	-0.007*** (0.002)
Constant	0.355*** (0.040)	0.378*** (0.137)	0.302*** (0.024)
Demographic Controls	No	Yes	FE
N observations	1500	1500	1500
N clusters	30	30	30
N periods	50	50	50
R ²	0.36	0.41	0.33

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the session level.

Note: Results of three panel regressions for $t \leq 50$. The dependent variable is the SSI in a given session and period and varies between 0 and 1. *High inequality* is a binary variable that takes the value 0 if $\Delta = 25$ (T1–T3) and 1 if $\Delta = 75$ (T4, T5). Demographic controls include the difference in the share of female participants as well as in average measures of Big 5, risk aversion, prosociality and age between the groups in a session. Columns (1) and (2) are random effects models, while column (3) reports results of a fixed effects estimation.

of Hypothesis 1 that coordinating on a convention is harder if the convention is synonymous with high inequality.

While a high SSI could in principle also reflect an equilibrium where groups coordinate within each period but switch who chooses A between periods, examining the share of A choices over time shows that in most low inequality sessions groups converge on repeatedly playing the same pure NE. This means that one group increasingly specialises on action A and the other on action B (see Appendix A.1, Figure A.1). This is also reflected in the reduction of switches between actions for individuals over time (see Appendix A.1, Figure A.3).

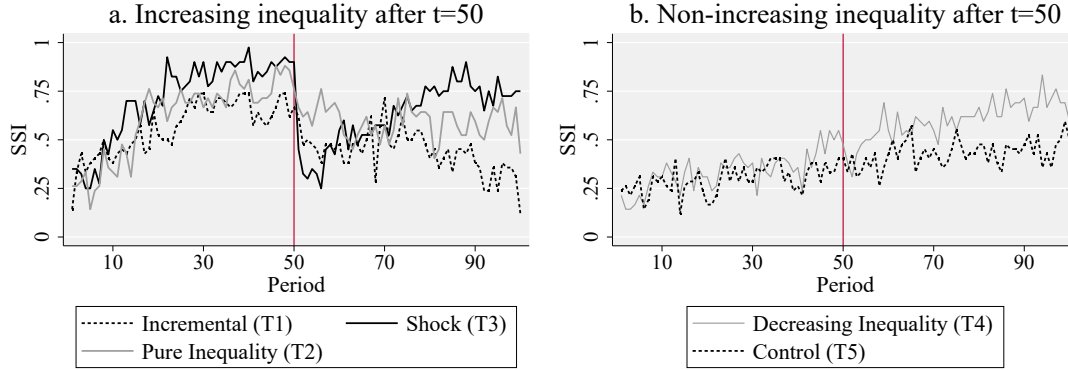
Result 1: *Under low inequality most sessions coordinate on an unequal convention by the end of the first 50 periods. Higher inequality leads to significantly lower levels of stability.*

Finally, it is interesting to note that the higher stability of an unequal convention in the low inequality treatments (T1–T3) leads to substantial differences in earnings. Individuals under high inequality earn on average 30% less than individuals under low inequality (t-test, $p < 0.001$). However, the higher stability comes at the cost of inequality. At $t = 50$, a member of a submissive group in T1–T3 earned on average only 71% of what a member of a dominant group earned (t-test, $p < 0.001$), underlining the unequal nature of the convention.

5.2 Does inequality threaten stability?

From the previous section, we already got a first indication that inequality impedes the emergence of a stable convention. But how do changes in inequality affect behaviour if there is already a convention in place? To answer this question, we now examine behaviour across all 100 periods. Figure 1.4 plots the SSI over all periods. The left graph shows the dynamics for treatments where inequality increases after $t = 50$ (T1–T3). We can see that behaviour

Figure 1.4: Stability over time per treatment



changes substantially, with average stability decreasing after $t = 50$. To analyse this change in behaviour more formally, we test for a structural break in the series for each treatment at $t = 50$, following the approach by [Bai & Perron \(1998, 2003\)](#).¹⁸ In line with the visual evidence, the tests confirm a significant break in all three treatments ($p < 0.001$). In contrast, the right graph in Figure 1.4 shows the SSI for treatments with no increase in inequality (T4, T5). In both these treatments stability increases over time continuing the positive trend from the first 50 periods. When testing for structural breaks at $t = 50$, we cannot reject the null that there are no breaks in the data ($p = 0.84$ and $p = 0.86$).

Further examining the effects of inequality on conventions, Table 1.3 presents estimated coefficients for a single random effects panel regression taking the SSI as the dependent variable and a dummy variable that is equal to 1 if $t > 50$ as the explanatory variable of interest. We interact this dummy variable with treatment indicators and take a different treatment as the base category in each column, allowing us to present the change in stability by treatment. As we are interested in the effect of increasing inequalities on an established convention, we exclude the first 20 periods from the analysis as they still show relatively low levels of stability (average SSI = 0.37).¹⁹ On average, stability decreases significantly in the second half of the experiment for treatments involving an increase in inequality, as shown in columns (1) – (3). The coefficients for T1–T3 are not statistically different from each other. When looking at individual choice data, we see that the decline in stability is driven by more individuals choosing action A, which becomes more attractive as inequality increases (see Appendix C.2, Table A.3).

Result 2i: *An increase in inequality destabilises an existing convention, as more individuals in the Incremental (T1), Pure Inequality (T2) and Shock (T3) treatment choose action A after $t = 50$.*

In contrast, in treatments where inequality is not increasing we observe a significantly different pattern, as shown in columns (4) and (5). Here stability increases after $t = 50$,

¹⁸The test is implemented in Stata following [Ditzen et al. \(2021\)](#).

¹⁹Tables A.1 and A.2 in Appendix C.2 report how results change when only excluding the first 10 or no periods. While the sign of all coefficients remains the same, some become insignificant.

Table 1.3: Changes in stability after $t = 50$ by treatment

	(T1) Incremental	(T2) Pure Inequality	(T3) Shock	(T4) Decreasing Inequality	(T5) Control
After $t=50$	-0.22*** (0.07)	-0.15* (0.09)	-0.23*** (0.08)	0.21*** (0.06)	0.10** (0.05)
Constant	0.65*** (0.14)	0.74*** (0.05)	0.86*** (0.05)	0.40*** (0.10)	0.33*** (0.06)
N observations	2400				
N clusters	30				
N periods	80				
R ²	0.20				

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the session level.

Note: Results of a single random effects panel regression. The dependent variable is the *SSI* in a given session and period and varies between 0 and 1. *After* $t = 50$ is a dummy variable that takes the value 0 for the periods before $t = 50$ and 1 for periods 51–100. The regression controls for treatment indicators and their interaction with *After* $t = 50$. Each column shows the results for a different treatment as the base category. Treatment differences are omitted from the table for conciseness. The first 20 periods are excluded from the analysis, as it takes time for the convention to emerge and the first periods still exhibit a large degree of switching behaviour (average *SSI* = 0.37).

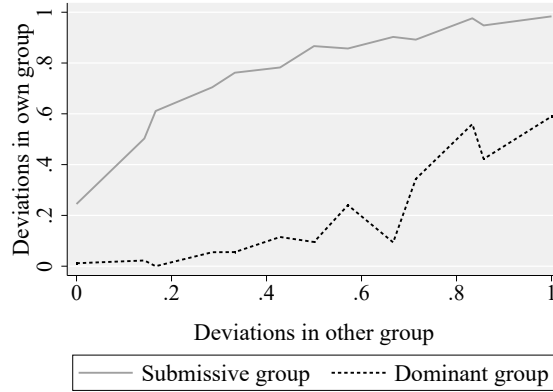
with the coefficients for both T4 and T5 being significantly larger than the ones for T1–T3 ($p \leq 0.001$).²⁰ Intuitively, coordination in the *Decreasing Inequality* (T4) treatment becomes easier as inequality decreases. The larger stability in the *Shock* (T3) treatment after $t = 50$ suggests that even under high levels of inequality individuals achieve coordination over a sufficiently long time horizon.

Finally, let us examine who initialises deviations from the established convention. To do this, we focus on treatments where inequality increases after $t = 50$ (T1–T3), as those are the ones where we expect to see a breakdown of the convention. An individual deviates from a convention if a member of a submissive group switches to choosing A or a member of a dominant group switches to choosing B in a given period.²¹ Figure 1.5 plots the average share of deviations after $t = 50$ in the own group (y-axis) against the average share of deviations in the other group (x-axis). For instance, for a dominant group member, deviations in the own group are defined as the share of individuals in the dominant group choosing B, while deviations in the other group are defined as the share of individuals in the submissive group choosing A. Figure 1.5 shows that in line with Hypothesis 2.1, submissive groups are the ones initialising the deviations and show more deviations, irrespective of the deviations in the dominant group. Even if the share of individuals who deviate in the dominant group is zero, the share is already above 20% in the submissive groups.²² Moreover, dominant groups only start deviating from the convention once 50 – 70% of submissive group members are

²⁰The effects for T4 and T5 are not statistically different from each other.

²¹As a measure for which group becomes dominant and which submissive, we look at the share of A choices in the last 5 periods before $t = 50$ as an approximation of equilibrium play.

²²A t-test indicates that the first period in which an individual deviates is significantly earlier for submissive compared to dominant groups ($p < 0.001$).

Figure 1.5: Deviations in treatments with increasing inequality (T1–T3) for $t \leq 50$ 

Note: Dominant (submissive) group refers to the group specialising in action A (B) in the last 5 periods of the first part of the experiment as an approximation of equilibrium play.

deviating.²³ After that point deviations increase more rapidly. This behaviour is in line with the concept of tipping points, stating that a critical threshold of deviations needs to be reached before we see a change in the current status quo.

Result 2ii: *Deviations from a convention are initialised by the submissive group. Members of the dominant group only start changing their behaviour once a critical threshold is reached.*

In addition to being a member of the submissive group, we find that some individual characteristics predict deviations. Individuals with lower levels of risk aversion, higher levels of conscientiousness, and more negative affect responses are more likely to be in the vanguard of deviators (see Appendix A.1, Table A.4). In fact, already over the first 50 periods we see that submissive groups show more negative affect responses (see Appendix A.1, Figure A.4).²⁴ Together with the link between negative affect and the likelihood to deviate, this provides support for the argument that relative grievances can lead to discontent and that such emotional factors, in turn, drive instability (Cramer, 2005; Blattman & Miguel, 2010). Finally, we find that the stronger a convention is, the less likely individuals deviate in both groups.

5.3 Treatment differences

5.3.1 The effect of the poor getting poorer

In this section, we explore whether certain inequality dynamics increase the negative effect on stability. First, we test Hypothesis 3 which states that deviations from a convention are more likely if increased inequality is not only a result of the rich getting richer (*Pure Inequality*

²³The theoretical model predicts that members of the dominant group should not deviate from a convention before $E(\lambda_t^{g'}) \leq \frac{h_t}{l_t + h_t} > 0.5$ is reached. Our data is thus in line with this prediction.

²⁴In line with the difference in inequality dynamics, affect responses vary across treatments in the second half of the experiment (see Appendix A.1, Figure A.5).

Table 1.4: The effect of the poor getting poorer ($t > 50$)

	(1) SSI	(2) A choices	(3) Deviations
<i>Baseline = Incremental treatment (T1)</i>			
Pure inequality (T2)	0.16 (0.11)	-0.05** (0.03)	-0.21** (0.09)
Constant	0.44*** (0.09)	0.68*** (0.02)	0.48*** (0.07)
N observations	1500	20800	20800
N panels	30	416	416
N clusters	30	30	30
N periods	50	50	50
R ²	0.07	0.01	0.03

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the session level.

Note: Results of three random effects panel regressions for $t > 50$. The dependent variable in model (1) is the SSI in a given session and period and varies between 0 and 1. The dependent variable in model 2 is an individual's choice in a given period and takes the value 0 if an individual chooses B and 1 if they choose A. The dependent variable in model (3) is also binary, with a value of 0 if an individual does not deviate from an established convention in a given period and 1 if they do. All regressions include treatment indicators for T3, T4, and T5. They are omitted from the table for conciseness.

treatment, T2), but also of the poor getting poorer (*Incremental* treatment, T1).

Table 1.4 compares stability in T1 and T2 for $t > 50$. Column (1) uses the SSI as the dependent variable, column (2) the probability of choosing A for each individual, and column (3) the individual probability to deviate from an established convention. These different measures for stability are regressed on treatment indicators with T1 as the baseline. As this is our comparison of interest, only the coefficient for the difference between the *Incremental* (T1) and the *Pure Inequality* (T2) treatment is reported in Table 1.4, the coefficients for the difference with T3–T5 are omitted for conciseness. As can be seen, the *Pure Inequality* (T2) treatment is characterised by higher levels of stability compared to the *Incremental* (T1) treatment across all three measures. Even though the coefficient for the SSI is statistically insignificant ($p = 0.25$), we conclude that overall lowering the payoff for the submissive group exacerbates the destabilising effect of inequality. This finding indicates a “having less to lose” effect that makes miscoordination less costly.

Result 3: *If an increase in inequality is combined with the poor getting poorer, this aggravates the destabilising effect of inequality.*

5.3.2 Gradual versus sudden increases in inequality

Next, we explore Hypothesis 4i that predicts no difference between gradual and sudden changes in inequality, but only a reaction to the absolute level of inequality. By contrast, if gradual increases lead to a habituation effect or if a sudden shock helps people to coordinate their deviations, stability will be higher for the same level of inequality under gradual changes. To test Hypothesis 4i, we compare the level of stability between the *Shock* (T3) treatment at $t = 51$ with the *Incremental* (T1) and *Pure Inequality* (T2) treatment at $t = 75$, as in those periods inequality is identical ($\Delta = 75$). We find that even though stability — as measured by the

SSI — is 0.16 points higher in T1 and T2 than in T3 for that period, this difference is statistically not significant (Mann-Whitney test, $p = 0.20$). However, when examining a five period interval (periods 51–55 in T3 and 75–79 in T1 and T2) to enhance statistical power ($N=18$ for 1 period), the difference in average stability grows to 0.24 points and becomes significant (Mann-Whitney test, $p < 0.001$). The same result holds when looking at the probability of choosing action A as an individual level measure for stability (t-tests, $p = 0.20$ and $p = 0.04$ for a one and five year window). The data thus provides weak evidence against Hypothesis 4i and in favour of the idea that a gradual increase in inequality will be more tolerated.

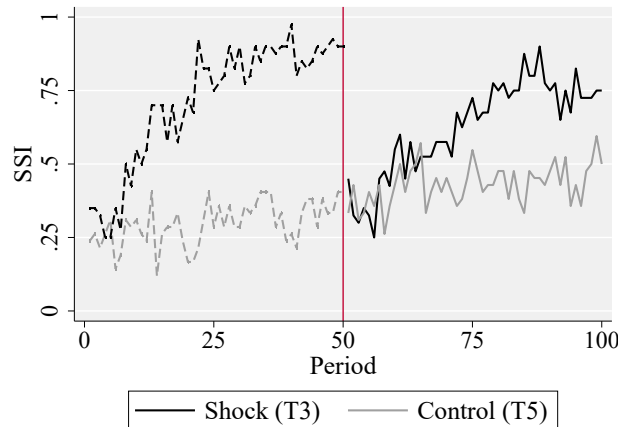
Result 4i: *If there is a sudden jump in inequality, this leads to higher levels of instability than if the same level of inequality is reached gradually.*

5.3.3 Past experiences of (in)stability

Hypothesis 4ii and 4iii state that after controlling for $E(\lambda_t^{g'})$, past experiences of (in)stability should not affect how individuals respond to a given level of inequality. As a first test of this notion, we compare the *Shock* (T3) and the *Control* (T5) treatment for $t > 50$, where $\Delta = 75$ for both treatments. However, T3 has a history of low and T5 a history of high inequality. Figure 1.6 shows that contrary to Hypothesis 4ii, stability is significantly higher in T3 for $t > 50$ (t-test, $p < 0.001$). This difference also holds in a regression model, showing that although the initial difference in stability between treatments is insignificant between the two treatments, there is a significantly stronger increase in stability over time for T3 (see Appendix C.2, Table A.5). We conclude that the experience of stability in the past also increases the likelihood of establishing stability in a more unequal environment.

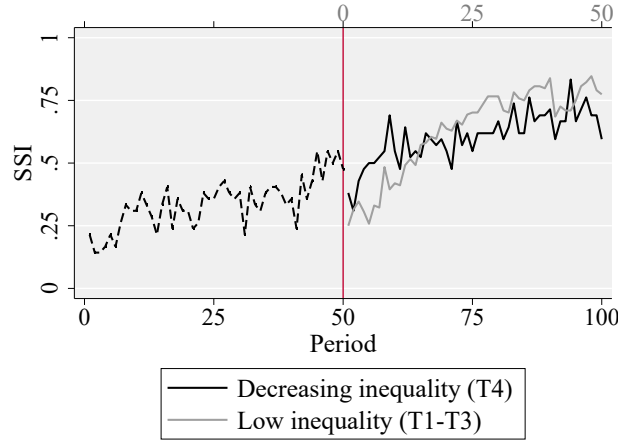
With the same objective in mind, we compare the first 50 periods in treatments with initially low inequality (T1–T3) and the last 50 periods in the *Decreasing Inequality* (T4) treatment. While in both cases participants experience low levels of inequality ($\Delta = 25$), in T1–T3 they

Figure 1.6: SSI for high inequality in $t > 50$ under different histories



Note: Solid lines represent stability across treatments under the same level of inequality ($\Delta = 75$) and are thus our focus of interest. The dashed lines represent past experiences of stability in each treatment.

Figure 1.7: SSI for low inequality under different histories



Note: Solid lines represent stability across treatments under the same level of inequality ($\Delta = 25$) and are thus our focus of interest. The dashed line represents past experiences of stability. As we compare the first 50 periods in the treatments with low inequality (T1–T3) to the last 50 periods in the *Decreasing Inequality* (T4) treatment, only the latter has a previous history for our comparison window.

start without any previous experiences, whereas in T4 they have a history of high inequality. Figure 1.7 shows that even though average stability is not different (t-test, $p = 0.49$), the positive time trend is significantly stronger in T1–T3 that lack a history of high inequality. This is confirmed in a regression analysis, showing a significant difference in time trends (see Appendix C.2, Table A.6). In line with the last result, this shows that history matters. The emergence of a convention can be slowed down by past experiences of instability or accelerated by a history of stability.

Result 4ii–iii: *Past experiences affect the current level of stability. ii) A history of stability helps to maintain or re-establish stability even under higher inequality, while iii) a history of instability slows down the emergence of a convention even when inequality has decreased.*

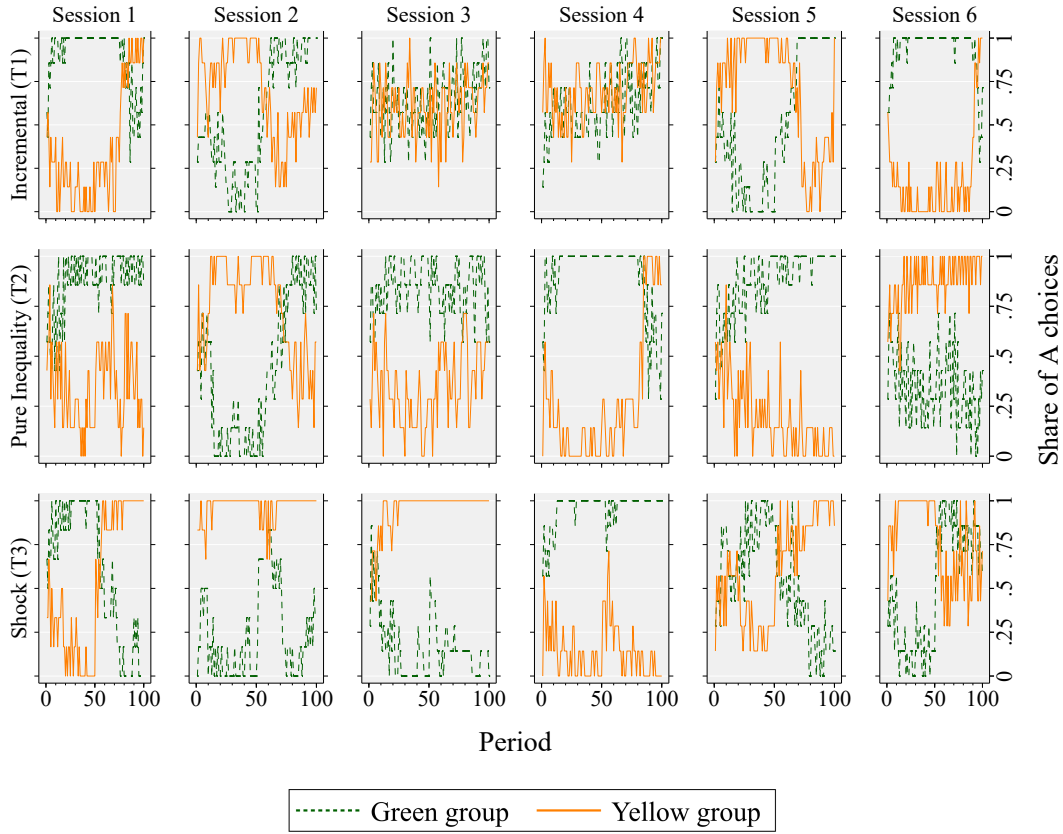
5.3.4 Reversal of initial conventions

While looking at the SSI provides a useful aggregate measure for stability, it hides substantial heterogeneity in the dynamics within each session. To account for this, Figure 1.8 plots the share of A choices in T1–T3 for yellow and green groups per session.²⁵ From the previous analysis, we know that on average stability decreases in the second half of the experiment for T1–T3. However, Figure 1.8 shows that across sessions, deviations from the convention can lead to quite different outcomes.

Focussing first on the *Shock* (T3) treatment, we can see that the increase in inequality at $t = 51$ is either followed by a period of chaos (see Session 6), a reversal of the previous

²⁵In line with the random assignment of individuals to groups, Figure 1.8 also shows that there is no tendency for whether the green or yellow group becomes dominant. See Appendix C.2, Figure A.6 for the behaviour in treatments with non-increasing inequality (T4 and T5).

Figure 1.8: Behaviour in treatments with increasing inequality (by session)



convention (see Sessions 1, 5) or a re-emergence of the previous convention after a failed attempt to change it (see Sessions 2, 3, 4). While in the *Incremental* (T1) and *Pure Inequality* (T2) treatment we also see reversals (e.g. Session 2 in both treatments), the strength of the new convention is much lower, and as inequality continues to increase most sessions that reversed end up in chaos with both groups choosing A with $p > 0.5$. Note also that reversals are less likely in T2 compared to T1 (2 vs. 4 sessions). This underlines once more that it is especially the effect of the poor getting poorer that drives destabilisation.

Unfortunately, we cannot analyse which individual characteristics predicts reversals (or failed reversals) as the data boils down to one observation per session (reversal yes or no). However, we find that reversals have a significant effect on final earnings and earnings inequality. While overall participants earn 13% more in sessions without a reversal (t-test, $p = 0.005$), the final inequality in earnings between dominant and submissive groups is 43% higher in these sessions (t-test, $p < 0.001$). This further highlights the trade-off between stability and inequality in our setting. While higher levels of stability lead to fewer cases of miscoordination and thus higher earnings, adherence to an unequal convention by definition leads to higher inequalities between groups.

From a theoretical perspective, the most common outcome after an increase in inequality

is the destabilisation of the initial convention. While reversals are theoretically possible they are not very likely. To see why, recall that within the theoretical framework individuals choose A if their level of inequality aversion exceeds a certain threshold (θ^*). Most individuals in the submissive group would thus need a relatively high level of inequality aversion to deviate to A ($\theta_i \geq \theta^*$), while most individuals in the dominant group would need a relatively low level of inequality aversion to start choosing B ($\theta_i \leq \theta^*$). While not impossible it is unlikely that inequality aversion is systematically lower in the dominant group, especially as they were the ones specialising on A in the first place.

What is not in line with our theoretical framework is a re-emergence of the initial convention, like we observe in the cases of failed reversals in the *Shock* (T3) treatment. If individuals in the submissive group deviate to A, they do so despite knowing that they will receive zero with a very high probability. This means that if inequality remains the same, they should never revert back to B. One possible explanation for the observed behaviour is that the assumption of myopic best response fails in the presence of large changes. The jump in inequality at $t = 51$ may also lead to a jump in expectations about the behaviour of others ($E(\lambda_t^{g'})$). If the dominant group keeps choosing A, members of the submissive group may then correct their expectations and revert back to choosing B. This explanation is supported by the fact that we do not observe failed reversals in any other treatment. While changes in expectations can thus explain the existence of failed reversals in the *Shock* (T3) treatment, they cannot rationalise the overall high number of reversals and the finding that reactions to the same level of inequality depend on past experiences. For this reason, we outline possible extensions of the theoretical framework in the next section.

6 Discussion and theory revisited

As discussed above, while the theoretical framework correctly predicts the destabilising effect of inequality (Hypotheses 1-3), it does not accommodate the large degree of history dependence that we observe in our data (contrary to Hypothesis 4). Such history dependence is in line with models of endogenous preferences (see e.g. Bowles, 1998; Henrich et al., 2010; Fehr & Hoff, 2011), which in our context means that inequality aversion (θ_i) may not be fixed but change with experiences. Our goal in this section is not to formalise a model of dynamic inequality aversion. Instead, informed by previous literature, we want to look at two potential factors that could affect θ_i : the experience of inequality and the experience of miscoordination or coordination.

When it comes to the experience of inequality, there exist both arguments that the latter could increase or decrease inequality aversion. For individuals in the submissive group, the repeated experience of being behind might lead to more envy and feelings of unfairness (Hong & Bohnet, 2007; Buttrick & Oishi, 2017). This can over time increase inequality aversion. By contrast, other studies show that sometimes disadvantaged groups defend the existing system (Lerner, 1978; Jost et al., 2003; Malahy et al., 2009) and that the current level of inequality significantly influences public opinion on tolerable inequality (Trump, 2018). Being exposed to inequality could thus also decrease inequality aversion within the submissive group. Similarly,

the experience of belonging to the dominant group could affect inequality aversion in both directions. While some studies have shown that higher levels of inequality can lead to more selfish behaviour among the rich and a larger feeling of entitlement (Kraus & Keltner, 2013; Nishi et al., 2015; Côté et al., 2015), the lack of justification for their position could give rise to guilt (Beranek et al., 2015) and make them more likely to accept a reversal of roles.

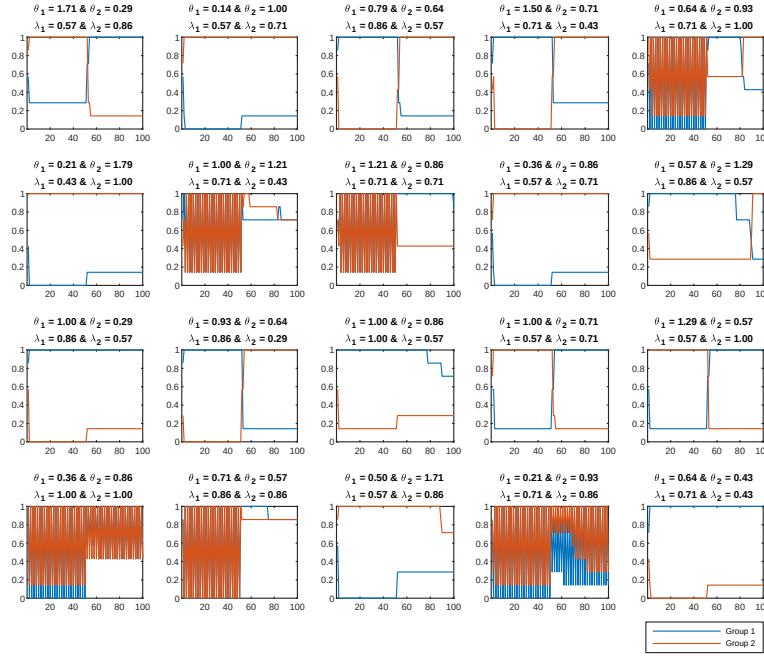
Within the context of the experiment, we can measure the experience of being dominant or submissive by the number of times an individual experiences a high payoff or a low payoff, such that $H = \sum_0^t \pi = h$, $L = \sum_0^t \pi = l$. Inequality aversion can then be expressed as a function of H and L ($\theta_i(H, L)$).²⁶ Due to the contrasting arguments in previous work, we refrain from imposing a specific direction of H and L on θ_i . Instead we take an explorative approach, testing which specification can explain the patterns in our data.

Another dimension that can influence inequality aversion is the experience of miscoordination or coordination. Brandts & Cooper (2006) show experimentally that a history of successful coordination can also help to maintain it in a more adverse environment. On the other hand, the repeated experience of miscoordination may decrease inequality aversion, as participants want to avoid receiving zero again. In line with this notion, papers have shown that growing up in a recession or experiencing economic shocks can lead to more conservative investment decisions and an avoidance of risk (Malmendier & Nagel, 2011; Shigeoka, 2019). Within our set-up, the experience of miscoordination can be expressed as the number of times an individual receives zero: $Z = \sum_0^t \pi = 0$. We would then expect that θ_i decreases with Z .

In order to examine which specifications of history dependence are consistent with our data, we execute a series of simulations, adjusting how θ_i depends on H , L , and Z . To do so, we follow our theoretical framework and first draw an inequality aversion parameter for each individual from the distribution specified in Fehr & Schmidt (1999) and randomly draw the initial expectation about others' behaviour from a uniform distribution, such that $E(\lambda_t^{g'}) \sim U(0, 1)$. Next, we simulate the choice for A or B for each individual using Equation (2), with beliefs being adjusted using $E(\lambda_t^{g'}) = \lambda_t^{g'}$ for $t > 1$. For each specification we run 100 simulations over all 100 periods. The key facts that come out from the experiment and should be captured by the simulations are first of all that increasing inequality destabilises an established convention (i). While this can lead to reversals in some sessions (ii), in other sessions a less stable version of the initial convention can be maintained (iii).

Our simulations show that the data cannot be rationalised with a model where θ_i increases with H or L . Intuitively, (iii) is not possible if θ_i increases with L or is independent of L . We can only recreate the persistence of a convention under increased inequality if θ_i decreases with L , i.e. if individuals mind disadvantageous inequality less the more often they experience it. Similarly, the existence of reversals (ii) cannot be simulated if θ_i increases with H or is independent of H , as in that case participants in the dominant group would never deviate to choosing B . In fact, for a successful reversal to happen θ_i must not only decrease over time,

²⁶We also implement non-linear versions of history dependence but do not find that they describe the data better than the simple linear specification.

Figure 1.9: Simulation of the *Shock* (T3) treatment if θ_i decreases with H and L

Note: Behaviour was simulated using the decision rule derived in the theoretical framework (see Equation (2)).

The values for λ and θ above each panel show the randomly drawn average starting parameters for each group per simulation.

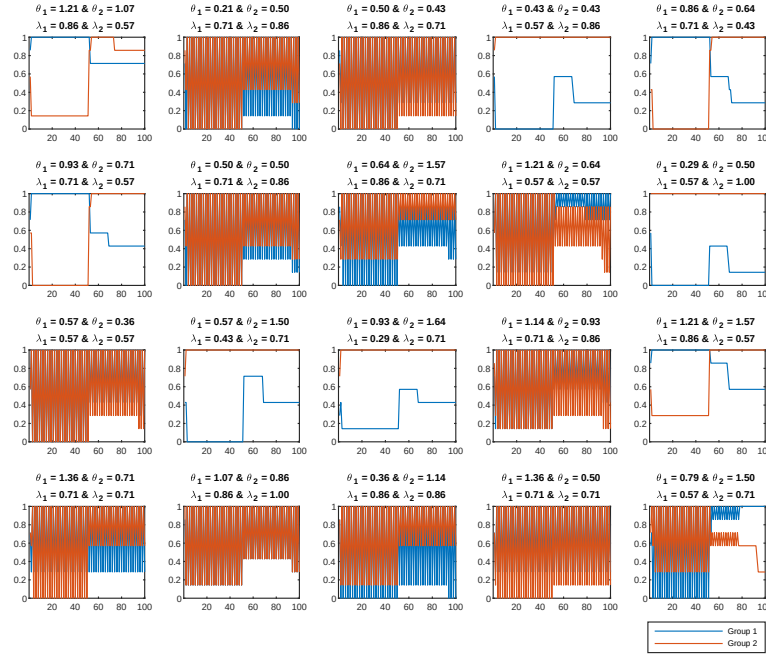
but do so at a faster rate for the dominant than the submissive group.²⁷ Figure 1.9 shows a sub-sample of 20 simulations for the *Shock* (T3) treatment for the case where θ_i decreases with both H and L.²⁸ As we can see, if we let θ_i decrease with both H and L, we can replicate both reversals and the persistence of initial conventions, with an overall lower level of stability after $t = 50$.

Note that one feature that cannot be explained by a model where θ_i decreases with H and L is failed reversals. As noted above, this could be accommodated by allowing expectations to change in the presence of large jumps in inequality. Alternatively, Figure 1.10 shows that the failed reversals in T3 can be accounted for by a model where θ_i decreases with the experience of miscoordination (Z).²⁹ Moreover, this version of history dependence provides a possible explanation for the surprisingly high level of stability at the end of the experiment in the *Control* (T5) treatment. While initially inequality is too high to achieve coordination, the repeated experience of receiving zero decreases inequality aversion and therefore allows for the emergence of a convention over time.

²⁷Note that the dynamic nature is crucial. If we just introduce an additional fixed parameter that accounts for advantageous inequality aversion, the patterns in our data cannot be replicated

²⁸The *Shock* (T3) treatment is particularly interesting, as we observe both reversals and failed reversals. See Appendix A.2, Figures A.7–A.10 for simulations of the other treatments. A sub-sample of 20 simulations per treatment is chosen as results are repetitive.

²⁹See Appendix A.2, Figures A.11–A.14 for simulations of the other treatments.

Figure 1.10: Simulation of the *Shock* (T3) treatment if θ_i decreases with Z 

Note: Behaviour was simulated using the decision rule derived in the theoretical framework (see Equation (2)). The values for λ and θ above each panel show the randomly drawn average starting parameters for each group per simulation.

7 Conclusion

In this chapter, we examine whether inequality has a destabilising effect on coordination between groups. While arguments that inequality threatens social stability can be traced back to ancient philosophers, the empirical evidence on the matter is mixed. In this project, we use a novel approach and manipulate the extent of the inequality individuals are facing in a dynamic laboratory environment, allowing us to isolate its effect on stability. In particular, we look at behaviour in a repeated battle of the sexes game between two groups. In the game, individuals have both an incentive to coordinate on different actions but at the same time face an inherent conflict as one of the two possible actions results in a higher payoff. During the first half of the experiment, we keep inequality constant and find that if inequality is not too pronounced, groups coordinate on an unequal convention, where each group specialises on one action. While this helps to avoid miscoordination, over time, it also creates substantial inequality between groups.

When we start to exogenously manipulate inequality between actions in the second half of the experiment, we find that increasing inequality has a strongly destabilising effect and deviations from the convention are initialised by the submissive group. While this change of behaviour can lead to chaos, we observe the (re-)emergence of a convention in some sessions. We show that the negative effect of inequality on stability can be rationalised by a theoretical framework based on disadvantageous inequality aversion and myopic best response. In

each period, individuals decide which action to take based on current payoff differences, expectations about the behaviour of the other group, and their individual inequality aversion. Once inequality passes a threshold which they are not willing to tolerate, individuals from the submissive group start to deviate from the convention.

By looking at different inequality dynamics across treatments, we show that certain characteristics of changing inequality exacerbate its destabilising effect. If inequality is driven by the poor getting poorer, deviations from the submissive group are faster and more pronounced than if it is driven by the rich getting richer. This underlines the additional importance of absolute grievances (see e.g. [Collier, 1999](#); [Fearon & Laitin, 2003](#); [Somanthan, 2020](#)). Moreover, we find evidence that a large one-time jump in inequality leads to higher instability in the short run than if the same level of inequality is reached by gradual changes. Finally, we find that responses to a given level of inequality are not only affected by how it was reached but also by the history of the game. Past experiences of stability can help to maintain a convention even under higher levels of inequality. Similarly, past experiences of instability can compromise the emergence of a convention even if inequality is reduced.

As history dependence cannot be rationalised with a fixed inequality aversion, we further explore how the predictions of our theoretical framework change when allowing inequality aversion to depend on previous outcomes. Conducting a simulation exercise, we show that both a model where inequality aversion decreases with the experience of inequality (i.e. experiencing being ahead/behind) or the experience of miscoordination perform best in explaining the behavioural patterns observed in our experiment.

There are several interesting extensions that can be tested using our experimental framework. For instance, it would be interesting to explore what happens to stability under increasing inequality if we include stabilising counter forces or obstacles to change, such as giving the dominant group means to repress attempts to revolt against the status quo or allow them to endogenously lower inequality if they observe deviations. This could enable the dominant group to maintain the highest possible level of inequality that still results in a stable convention. We thus hope that this project can encourage further work and applications for inequality research.

Overall, our findings shine a new light on the debate around inequality and instability, showing that in a controlled environment increasing inequality has adverse effects on stability. This is of direct applied relevance as according to our findings it is crucial to understand and monitor inequality dynamics in order to achieve a stable and functioning society.

References

- Acemoglu, D. & Jackson, M. O. (2014). History, expectations, and leadership in the evolution of social norms. *The Review of Economic Studies*, 82(2), 423–456.
- Alesina, A. & Perotti, R. (1996). Income distribution, political instability, and investment. *European Economic Review*, 40(6), 1203–1228.
- Andreoni, J., Nikiforakis, N., & Siegenthaler, S. (2021). Predicting social tipping and norm change in controlled experiments. *Proceedings of the National Academy of Sciences*, 118(16), e2014893118.
- Andreoni, J. & Samuelson, L. (2006). Building rational cooperation. *Journal of Economic Theory*, 127(1), 117–154.
- Axtell, R. L., Epstein, J. M., & Young, H. P. (2001). The emergence of classes in a multiagent bargaining model. *Social Dynamics*, 27, 191–211.
- Bai, J. & Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1), 47–78.
- Bai, J. & Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1), 1–22.
- Baronchelli, A. (2018). The emergence of consensus: a primer. *Royal Society Open Science*, 5(2), 172189.
- Belloc, M. & Bowles, S. (2013). The persistence of inferior cultural-institutional conventions. *American Economic Review*, 103(3), 93–98.
- Benndorf, V., Martinez-Martinez, I., & Normann, H.-T. (2016). Equilibrium selection with coupled populations in hawk–dove games: Theory and experiment in continuous time. *Journal of Economic Theory*, 165(2016), 472–486.
- Beranek, B., Cubitt, R., & Gächter, S. (2015). Stated and revealed inequality aversion in three subject pools. *Journal of the Economic Science Association*, 1(1), 43–58.
- Berger, J., Vogt, S., & Efferson, C. (2022). Pre-existing fairness concerns restrict the cultural evolution and generalization of inequitable norms in children. *Evolution and human behavior*, 43(1), 1–15.
- Billig, M. & Tajfel, H. (1973). Social categorization and similarity in intergroup behaviour. *European Journal of Social Psychology*, 3(1), 27–52.
- Binmore, K., Samuelson, L., & Young, P. (2003). Equilibrium selection in bargaining models. *Games and Economic Behavior*, 45(2), 296–328.
- Blattman, C. & Miguel, E. (2010). Civil war. *Journal of Economic Literature*, 48(1), 3–57.

- Bowles, S. (1998). Endogenous preferences: The cultural consequences of markets and other economic institutions. *Journal of Economic Literature*, 36(1), 75–111.
- Bowles, S., Loury, G. C., & Sethi, R. (2014). Group inequality. *Journal of the European Economic Association*, 12(1), 129–152.
- Brandts, J. & Cooper, D. J. (2006). A change would do you good.... an experimental study on how to overcome coordination failure in organizations. *American Economic Review*, 96(3), 669–693.
- Buttrick, N. R. & Oishi, S. (2017). The psychological consequences of income inequality. *Social and Personality Psychology Compass*, 11(3), e12304.
- Centola, D., Becker, J., Brackbill, D., & Baronchelli, A. (2018). Experimental evidence for tipping points in social convention. *Science*, 360(6393), 1116–1119.
- Charness, G. & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3), 817–869.
- Collier, P. (1999). Doing well out of war. *World Bank Report*, 28137.
- Conteh-Morgan, E. (2004). *Collective political violence: An introduction to the theories and cases of violent conflicts*. Psychology Press.
- Cooper, J. M., Hutchinson, D. S., et al. (1997). *Plato: complete works*. Hackett Publishing.
- Côté, S., House, J., & Willer, R. (2015). High economic inequality leads higher-income individuals to be less generous. *Proceedings of the National Academy of Sciences*, 112(52), 15838–15843.
- Cramer, C. (2003). Does inequality cause conflict? *Journal of International Development: The Journal of the Development Studies Association*, 15(4), 397–412.
- Cramer, C. (2005). Inequality and conflict: A review of an age-old concern. *United Nations Research Institute for Social Development. Programme Paper*, (11).
- Crosetto, P., Weisel, O., & Winter, F. (2012). A flexible z-tree implementation of the social value orientation slider measure (murphy et al. 2011)—manual. *Jena Economic Research Paper*, 62.
- Dale, D. J., Morgan, J., Rosenthal, R. W., et al. (2002). Coordination through reputations: A laboratory experiment. *Games and Economic Behavior*, 38(1), 52–88.
- Desmet, P., Overbeeke, K., & Tax, S. (2001). Designing products with added emotional value: Development and application of an approach for research through design. *The Design Journal*, 4(1), 32–47.
- DiPrete, T. A., Gelman, A., McCormick, T., Teitler, J., & Zheng, T. (2011). Segregation in social networks based on acquaintanceship and trust. *American Journal of Sociology*, 116(4), 1234–83.

- Ditzen, J., Karavias, Y., & Westerlund, J. (2021). Testing and estimating structural breaks in time series and panel data in stata. *arXiv preprint arXiv:2110.14550*.
- Eckstein, H. (2004). Theoretical approaches to explaining collective political violence. In *Political Psychology* (pp. 432–448). Psychology Press.
- Fajnzylber, P., Lederman, D., & Loayza, N. (1998). Determinants of crime rates in latin america and the world: an empirical assessment. *The World Bank Report*, ISBN: 978-0-8213-4240-4.
- Fearon, J. D. & Laitin, D. D. (2003). Ethnicity, insurgency, and civil war. *American Political Science Review*, 97(1), 75–90.
- Fehr, E. & Hoff, K. (2011). Introduction: Tastes, castes and culture: The influence of society on preferences. *The Economic Journal*, 121(556), 396–412.
- Fehr, E. & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3), 817–868.
- Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2), 171–178.
- Gächter, S., Mengel, F., Tsakas, E., & Vostroknutov, A. (2017). Growth and inequality in public good provision. *Journal of Public Economics*, 150(2017), 1–13.
- Gerlitz, J.-Y. & Schupp, J. (2005). Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. *DIW Research Notes*, 4, 2005.
- Gimpelson, V. & Treisman, D. (2018). Misperceiving inequality. *Economics & Politics*, 30(1), 27–54.
- Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association*, 1(1), 114–125.
- Grossman, E. (2019). France’s yellow vests - symptom of a chronic disease. *Political Insight*, 10(1), 30–34.
- Gurr, T. R. (1971). *Why Men Rebel*. Princeton and Oxford: Princeton University Press.
- Hargreaves-Heap, S. & Varoufakis, Y. (2002). Some experimental evidence on the evolution of discrimination, co-operation and perceptions of fairness. *The Economic Journal*, 112(481), 679–703.
- Henrich, J. & Boyd, R. (2008). Division of labor, economic specialization, and the evolution of social stratification. *Current Anthropology*, 49(4), 715–724.
- Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., & Henrich, N. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science*, 327(5972), 1480–1484.

- Hirschman, A. O. & Rothschild, M. (1973). The changing tolerance for income inequality in the course of economic development: With a mathematical appendix. *The Quarterly Journal of Economics*, 87(4), 544–566.
- Holm, H. J. (2000). Gender-based focal points. *Games and Economic Behavior*, 32(2), 292–314.
- Holt, C. A. & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5), 1644–1655.
- Hong, K. & Bohnet, I. (2007). Status and distrust: The relevance of inequality and betrayal aversion. *Journal of Economic Psychology*, 28(2), 197–213.
- Hsieh, C.-C. & Pugh, M. D. (1993). Poverty, income inequality, and violent crime: a meta-analysis of recent aggregate data studies. *Criminal Justice Review*, 18(2), 182–202.
- Hwang, S.-H., Naidu, S., & Bowles, S. (2018). Social conflict and the evolution of unequal conventions. *Santa Fe Institute Working Paper*, 2018.
- Inglehart, R. F. & Norris, P. (2016). Trump, brexit, and the rise of populism: Economic have-nots and cultural backlash. *HKS Working Paper*, (RWP16-026).
- Jakobsen, T. (2010). Theories of collective violence: The continuing rational actor versus deprived actor debate. In T. Jakobsen (Ed.), *War: An Introduction to Theories and Research* (pp. 19–34). New York: Nova Science Publishers, Inc.
- Jost, J. T., Pelham, B. W., Sheldon, O., & Ni Sullivan, B. (2003). Social inequality and the reduction of ideological dissonance on behalf of the system: Evidence of enhanced system justification among the disadvantaged. *European journal of Social Psychology*, 33(1), 13–36.
- Kennedy, B. P., Kawachi, I., Prothrow-Stith, D., Lochner, K., & Gupta, V. (1998). Social capital, income inequality, and firearm violent crime. *Social Science & Medicine*, 47(1), 7–17.
- Kraus, M. W. & Keltner, D. (2013). Social class rank, essentialism, and punitive judgment. *Journal of Personality and Social Psychology*, 105(2), 247–261.
- Lerner, M. (1978). *The belief in a just world. A fundamental delusion*. New York: Plenum Press.
- Lewis, D. (1967). *Convention: A philosophical study*. Cambridge: Harvard University Press.
- Lichbach, M. I. (1989). An evaluation of "does economic inequality breed political conflict?" studies. *World Politics*, 41(4), 431–470.
- Malahy, L. W., Rubinlicht, M. A., & Kaiser, C. R. (2009). Justifying inequality: A cross-temporal investigation of us income disparities and just-world beliefs from 1973 to 2006. *Social Justice Research*, 22(4), 369–383.
- Malmendier, U. & Nagel, S. (2011). Depression babies: Do macroeconomic experiences affect risk taking? *The Quarterly Journal of Economics*, 126(1), 373–416.

- McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1), 415–444.
- Mookherjee, D. & Ray, D. (2002). Is equality stable? *American Economic Review*, 92(2), 253–259.
- Murphy, R. O. & Ackermann, K. A. (2014). Social value orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review*, 18(1), 13–41.
- Murphy, R. O., Ackermann, K. A., & Handgraaf, M. (2011). Measuring social value orientation. *Judgment and Decision Making*, 6(8), 771–781.
- Nishi, A., Shirado, H., Rand, D. G., & Christakis, N. A. (2015). Inequality and visibility of wealth in experimental social networks. *Nature*, 526(7573), 426–429.
- Olson, M. (1965). *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge: Harvard University Press.
- Oprea, R., Henwood, K., & Friedman, D. (2011). Separating the hawks from the doves: Evidence from continuous time laboratory games. *Journal of Economic Theory*, 146(6), 2206–2225.
- Østby, G. (2008). Polarization, horizontal inequalities and violent civil conflict. *Journal of Peace Research*, 45(2), 143–162.
- Piketty, T. & Saez, E. (2014). Inequality in the long run. *Science*, 344(6186), 838–843.
- Roubini, N. (2011). The instability of inequality. *RGE Global EconoMonitor*, October 17, 2011.
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145–172.
- Satz, D. (2019). What’s wrong with inequality? *The IFS Deaton Review*, available at: <https://www.ifs.org.uk/inequality/expert-comment/whats-wrong-with-inequality/>.
- Schotter, A. & Sopher, B. (2003). Social learning and coordination conventions in intergenerational games: An experimental study. *Journal of Political Economy*, 111(3), 498–529.
- Shigeoka, H. (2019). Long-term consequences of growing up in a recession on risk preferences. *NBER Working Papers*, 26352.
- Somanthan, R. (2020). Group inequality in democracies: Lessons from cross-national experiences. In J.-M. Baland, F. Bourguignon, J.-P. Platteau, & T. Verdier (Eds.), *The Handbook of Economic Development and Institutions* (pp. 137–152). Princeton: Princeton University Press.
- Stewart, F. (2000). Crisis prevention: Tackling horizontal inequalities. *Oxford Development Studies*, 28(3), 245–262.
- Trump, K.-S. (2018). Income inequality influences perceptions of legitimate income differences. *British Journal of Political Science*, 48(4), 929–952.

- Weber, R. A. (2006). Managing growth to achieve efficient coordination in large groups. *American Economic Review*, 96(1), 114–126.
- Young, H. P. (1993). The evolution of conventions. *Econometrica: Journal of the Econometric Society*, 61(1), 57–84.
- Young, H. P. (1996). The economics of convention. *Journal of Economic Perspectives*, 10(2), 105–122.
- Young, H. P. (2015). The evolution of social norms. *Annual Review of Economics*, 7(1), 359–387.

CHAPTER 2

Fairness in times of crisis: Negative shocks, relative income, and preferences for redistribution[†]

1 Introduction

Economic crises give rise to public and political debates surrounding inequality and redistribution. Examples include concerns regarding aggravated inequalities caused by the Covid-19 pandemic ([Adams-Prassl et al., 2020](#); [Butler, 2021](#); [The Economist, 2021](#); [Ten et al., 2022](#)), or the current cost-of-living crisis and its impact on the most vulnerable ([Siemplenski, 2022](#); [United Nations, 2022](#)). While negative economic shocks caused by such crises increase the salience of redistributive issues, a key question is whether they also affect preferences for redistribution. If yes, this can have significant implications for the acceptance of inequality or support for social policies (see e.g. [Alesina et al., 2012](#); [Almås et al., 2020](#)).

In this chapter, I study how negative economic shocks affect redistributive decisions. Addressing this question, poses several empirical challenges. The degree to which an individual is affected by a shock could correlate with socioeconomic characteristics. Moreover, a crisis will not only impact individuals through economic channels but also lead to wider political and societal changes. To isolate the causal effect of economic shocks on redistributive preferences, I thus design an experiment that — informed by a theoretical framework — allows me to exogenously vary an individual’s loss of income.

To understand the conditions under which economic shocks facilitate or hinder redistributive policies, it is important to consider heterogeneous responses within society. Previous research has argued that redistributive preferences are formed in a self-serving way, with support for redistribution being negatively correlated with an individual’s position within the income distribution ([Piketty, 1995](#); [Alesina & La Ferrara, 2005](#); [Alesina & Giuliano, 2011](#); [Cappelen et al., 2013b](#); [Durante et al., 2014](#); [Hvidberg et al., 2023](#)). For this reason, I randomly assign individuals to different income levels in the experiment, distinguishing between two types: poorer individuals who benefit from redistribution and richer individuals who benefit from the status quo. This does not only allow me to provide causal evidence for the relation-

[†]I thank the School of Economics and CeDEx for the funding that made this project possible. I am also grateful for feedback and comments from CeDEx members during and outside of seminars, as well as for feedback received during different conferences. The study received ethical approval from the Nottingham School of Economics Research Ethics Committee on 20/05/2021.

ship between relative income and redistributive preferences, but also to test whether different socioeconomic groups show distinct responses to shocks.

To conceptualise notions of distributive fairness I focus on two views that stand out from the literature: fairness as equality of outcomes, and fairness as proportionality to individual inputs (see e.g. [Konow, 2003](#); [Cappelen et al., 2007](#)). In the following, I refer to those ideas as egalitarian and contribution-based fairness views. In the theoretical framework, I assume that individuals first form fairness views on which they base their redistributive decisions. If there are differences in relative income, it becomes optimal for different income types to hold distinct fairness views and engage in motivated reasoning to avoid redistribution ([Jia, 2022](#)). Richer individuals will favour contribution-based views, while poorer individuals will favour egalitarian views.¹ This will ultimately translate into differences in redistributive decisions.

These differences in fairness views have direct implications for reactions to negative economic shocks. If, in accordance with egalitarianism, final income is to be shared equally, it does not matter if one party suffers a negative shock. By contrast, under the contribution-based fairness view, shocks to earned income will be compensated, as the unexpectedness of a crisis leaves pre-shock earnings as a natural reference point that affects post-shock allocation decisions.² The model thus predicts that only richer individuals with a contribution-based fairness view will respond to shocks. They will become more opposed to redistribution when they are hit by economic shocks themselves, but more supportive of redistribution if poorer individuals are affected.

The purpose of the theoretical framework is to provide a structure for the empirical analysis. In the experiment, participants work to earn an initial endowment and are later given the opportunity to redistribute total earnings between themselves and another player (following [Cappelen et al., 2007, 2013b](#)). I introduce differences in relative income by randomly assigning participants to either a hard or an easy task for which they earn different initial endowments. A crucial and novel aspect of my study is that, after earning their endowments, but before the redistribution decision, participants may incur a shock that reduces their earned income.³ Similarly, I vary the information individuals have about the experience of the other participant. They either know that the other suffered a shock, that the other did not suffer a shock, or are in a situation where the other's shock is unknown. This allows me to disentangle the effect of own shocks and shocks to others on redistributive decisions. Finally, the experiment is designed in a way that post-shock earnings are held constant across all conditions, which allows to exclude alternative explanations such as differences in inequality or total earnings.

¹In line with this, people tend to judge beneficial outcomes as more fair (see e.g. [Babcock & Loewenstein, 1997](#); [Ubeda, 2014](#); [Neuber, 2021](#)).

²See e.g. [Kahneman & Tversky \(1979\)](#); [Kőszegi & Rabin \(2006\)](#) on the relevance of reference points for decision making. More recently, studies have shown that these concepts also matter for distributional preferences ([Roth & Wohlfart, 2016](#); [Charité et al., 2022](#)).

³Income shocks within the lab have been previously used to study different topics such as solidarity and insurance agreements ([Selten & Ockenfels, 1998](#)), poverty and discount rates ([Haushofer et al., 2019](#)), the intra-personal consistency of fairness views ([Ubeda, 2014](#)), reciprocity between employers and employees in the presence of economic shocks ([Gerhards & Heinz, 2017](#)), or the effect of own failures on redistribution decisions as an unaffected observer ([Cassar & Klein, 2019](#)).

The results indicate that both relative income and the experience of shocks matter for redistributive preferences. Consistent with the theoretical framework, participants with a higher initial endowment are more likely to distribute total post-shock income in line with contribution-based fairness views, while players with a lower initial endowment are more likely to distribute total post-shock income equally. Moreover, as predicted by the theoretical framework, relative income has a moderating effect, with only richer individuals responding to shocks. Finally, in line with the literature on moral wiggle room (see e.g. [Dana et al., 2007](#); [Haisley & Weber, 2010](#); [Regner, 2018](#)) and willful ignorance (see e.g. [Grossman & Van der Weele, 2017](#)), a lack of information about the other’s shock provides an excuse for individuals who would otherwise compensate shocks to ignore them.

To provide more context to the experiment, I run a follow-up study that incorporates shocks caused by a real world crisis into the lab. In particular, I conduct the same experiment with participants who did/did not lose their employment due to Covid-19. The results confirm that participants who lost their employment allocate more resources to themselves, whereas learning that others lost their employment makes people redistribute more towards those in need. While the Covid-19 crisis is a highly relevant case study, my results are based on its economic component and can thus be translated to other large-scale crisis such as mass layoffs, recessions or natural disasters.

This paper contributes to the literature on redistribution and preference formation in several ways.⁴ First, my research complements previous studies that investigate the effects of economic shocks on distributive behaviour, analysing both macro shocks such as a general economic crisis ([Fisman et al., 2015](#); [Cappelen et al., 2021](#); [Lotti & Pethiyagoda, 2022](#)), as well as individual level shocks such as becoming unemployed ([Barr et al., 2016](#); [Naumann et al., 2016](#); [Martén, 2019](#)), exposure to natural disasters ([Stein & Weisser, 2022](#)), or correcting beliefs about the income distribution ([Cruces et al., 2013](#); [Karadja et al., 2017](#); [Hvidberg et al., 2023](#)). While results on macro shocks are mixed, negative individual shocks seem to increase support for redistribution, which is also confirmed in a recent experiment by [Gagnon et al. \(2021\)](#).⁵ An experimental study by [Mérola & Helgason \(2016\)](#) moreover shows that absolute and relative changes in income can have different effects on demand for redistribution. I extend these findings by showing that individuals do not only react to own income shocks, but also compensate the shocks of others. Moreover, distinguishing between different types of individuals allows a more nuanced analysis. Depending on which socioeconomic groups are affected by economic shocks, times of crises can both generate support for and resistance to redistributive policies.

Second, my study provides causal evidence for the notion that the relative position within society matters for redistributive preferences (see e.g. [Piketty, 1995](#); [Alesina & La Ferrara, 2005](#); [Alesina & Giuliano, 2011](#); [Cappelen et al., 2013b](#); [Durante et al., 2014](#); [Hvidberg et al., 2023](#)). In line with their own material self-interest the rich tend to demand lower levels of redistribution

⁴For a recent survey on determinants of redistributive preferences see [Mengel & Weidenholzer \(2022\)](#).

⁵There is also research on positive shocks such as winning the lottery ([Doherty et al., 2006](#); [Powdthavee & Oswald, 2014](#)), showing a reduction in the support for redistribution. See [Margalit \(2019\)](#) for an overview.

than the poor.⁶ This is in line with literature on the *veil of ignorance*, showing that individuals who take distributive decisions before their own position is revealed (behind the veil) take significantly different decisions from those who already know their position within society (see e.g. Sutter & Weck-Hannemann, 2003; Schildberg-Hörisch, 2010; Durante et al., 2014). My results also confirm recent theoretical and empirical work by Gallenstein (2021), who in a simultaneous paper studies exogenous variation in relative income. A key distinction is that in my paper earned income differences interact with exogenous shocks, yielding a rich set of results that inform the formation of redistributive preferences under different environments.

Third, I show that the source of inequality matters for redistributive decisions (see also Preuss et al., 2022). While the initial income inequality in the experiment can be justified by different difficulty levels of the real effort task, the subsequent income shock is unambiguously random. The finding that participants with a contribution-based fairness view compensate shocks but not initial differences, supports previous research showing that inequalities based on merit are more accepted than those produced by luck (see e.g. Alesina & La Ferrara, 2005; Cappelen et al., 2013a; Almås et al., 2020; Nettle & Saxe, 2020; Trump, 2020).⁷

Finally, my paper has interesting implications for understanding the formation and limits of self-serving biases and motivated reasoning in the perception of fairness (see e.g. Bénabou, 2015; Bénabou & Tirole, 2016; Deffains et al., 2016; Amasino et al., 2021). The experiment indicates that participants behave in line with the fairness view that is most beneficial to them. However, once an individual commits to a fairness view, they appear to be bound by it and do not engage in further self-serving behaviour such as only reacting to own but not others' shocks.⁸

The remainder of this paper is structured as follows. Section 2 outlines the theoretical framework, conceptualising how income differences and shocks affect fairness views. Section 3 describes the experimental design and hypotheses, while Section 4 presents results. Section 5 outlines the design and results for the follow-up study using Covid-19 shocks, before Section 6 concludes.

2 Theoretical framework

2.1 Basic set-up

Let us consider two individuals i and j who at time t earn respective endowments x_{it} and x_{jt} . Individuals can then decide to offset potential income differences by redistributing total earnings X_t ($X_t = x_{it} + x_{jt}$) among each other. Previous research has shown that in addition to their own material benefit, most people are guided by fairness considerations and social motives (see e.g. Hoffman et al., 1994; Huck & Oechssler, 1999; Camerer, 2011). Following the

⁶Relative income has moreover been linked to differences in giving (Erkal et al., 2011).

⁷Even though individuals may fail to account for luck when judging merit (see Andre, 2021).

⁸Such a commitment may be explained by an individual's self-image, preferences for consistency (Cialdini et al., 1995; Falk & Zimmermann, 2011) or a desire to think of oneself as a moral person (Brekke et al., 2003; Bénabou & Tirole, 2011, 2016).

seminal work by Cappelen et al. (2007, 2013b), an individual's allocation decision can then be modelled as a trade-off between own material benefit (y_{it}) and the amount they consider as a fair allocation to themselves (m_{it}), yielding the following utility function:

$$U_{it} = y_{it} - \beta_i \frac{(y_{it} - m_{it})^2}{2X_t}, \beta_i \geq 0, \quad (1)$$

where β_i is a sensitivity parameter that describes how much an individual values behaving in accordance with their own fairness judgement. When solving (1) for the optimal allocation to self this results in

$$y_{it}^* = m_{it} + \frac{X_t}{\beta_i}. \quad (2)$$

If individuals do not care about adhering to their own fairness judgement ($\beta_i = 0$), the solution is trivial and individuals always allocate all earnings to themselves. In the following, I thus focus on the more interesting case of $\beta_i > 0$. The higher β_i , the closer the allocation will be to m_{it} .

While individuals' decisions depend on m_{it} , a key question is how they decide on what constitutes a fair allocation. In an early and influential contribution to this literature, Konow (2003) categorises fairness along the dimensions of *equality and need* and *equity and desert*. In the following, I focus on two fairness views that embody these respective categories: the egalitarian and the contribution-based view. The egalitarian view abstracts from individual contributions and always prescribes an equal split of the total pie X_t . According to the contribution-based view, by contrast, each individual should receive exactly what they contributed to the overall endowment (x_{it}, x_{jt}).⁹ This is also in line with Konow's accountability principle, stating that a fair distribution should be proportionate to a person's discretionary inputs (Konow, 2000, 2003).¹⁰ The two fairness views can be summarised as follows:

Egalitarian fairness view: *Fairness means equality of outcomes ($m_{it} = \frac{X_t}{2}$). It is fair to redistribute total earnings such that each individual receives the same amount independent of individual contributions.*

Contribution-based fairness view: *Fairness means that each individual receives an amount that is proportional to their individual contributions ($m_{it} = x_{it}$). This is independent of whether such an allocation causes inequalities between individuals.*

The two fairness views can be seen as endpoints of a continuum, with some individuals being closer to the egalitarian and others closer to the contribution-based view (see also Barr et al., 2015). Formally, the allocation an individual perceives as fair can then be represented as a weighted average between these two notions:

$$m_{it} = \alpha_i x_{it} + (1 - \alpha_i) \frac{X_t}{2}, \quad (3)$$

⁹Cappelen et al. (2007) refer to this as a *libertarian* fairness view in line with the libertarian principle of non-interference.

¹⁰While Konow (2000, 2003) argues that a fair allocation will be a function of inputs and endowments, he also states that participants should not be held accountable for exogenous variables.

where α_i determines where on the continuum between the egalitarian and contribution-based fairness view an individual falls. If m_{it} is the fair allocation to self, the fair allocation to the other player from the point of view of individual i is consequently

$$m_{jt} = X_t - m_{it} = X_t - (\alpha_i x_{it} + (1 - \alpha_i) \frac{X_t}{2}). \quad (4)$$

In the following I refer to the fair allocation to self and other as m_{kt} , with $k \in \{i, j\}$.

While this conceptualisation represents a formal restriction on what can be considered as fair, it captures the most important ideas discussed in the literature and thus provides a useful way to think about the possible range of fairness judgements (see e.g. Cappelen et al., 2007, 2013b; Barr et al., 2015; Almås et al., 2020).

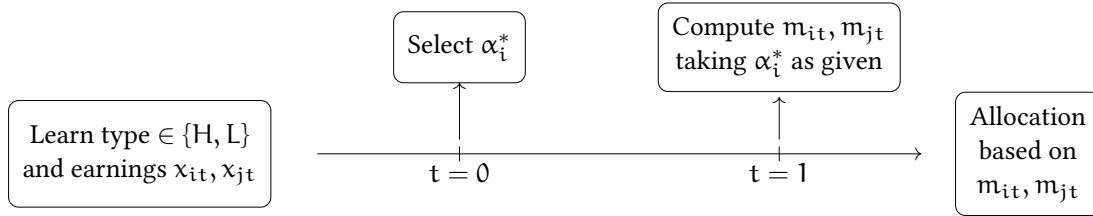
2.2 The effect of relative income

To explore how differences in relative income affect what an individual perceives as fair, consider a situation where initial endowments differ between i and j . Some individuals earn a high and some a low endowment ($i, j \in \{H, L\}$), generating differences in relative income.

As mentioned above, individuals tend to follow the fairness view that is most beneficial to them (Konow, 2000; Cappelen et al., 2013b; Ubéda, 2014; Barr et al., 2015; Gallenstein, 2021). Conceptually, this means that individuals choose α_i such that their utility is maximised. One can think of the allocation decision as a two-period process ($t \in \{0, 1\}$), similar to models of motivated reasoning (see e.g. Bénabou, 2015; Bénabou & Tirole, 2016). At $t = 0$, an individual decides which fairness view (which α_i) they should apply given a specific situation. At $t = 1$ the selected α_i^* is then taken as given and i decides on what they consider a fair allocation m_{kt} , and ultimately how they want to distribute X_t (see Figure 2.1).¹¹

To find the optimal weight between the two fairness views α_i^* , I substitute Equation (3) into the utility function. This yields a corner solution where $\alpha_i^* \in \{0, 1\}$, implying that individuals are either following the contribution-based ($m_{kt} = x_{kt}$) or the egalitarian fairness view ($m_{kt} = \frac{X_t}{2}$). If there are differences in relative income, it is always true that for richer individuals ($i = H$) own contributions are larger than average contributions ($x_{it} > \frac{X_t}{2}$), while the opposite is true for poorer individuals ($i = L$). To maximise their own material benefit,

Figure 2.1: Formation of fairness views and allocation decisions



¹¹Note that fairness views do not translate one-to-one into allocation decisions. If β_i — the importance individuals assign to adhering to their own fairness views — is low, individuals will allocate amounts to themselves that exceed m_{it} .

richer individuals should thus hold a contribution-based, and poorer individuals an egalitarian fairness view. Individuals may justify such self-serving judgements by interpreting the situation in question in a favourable way. For instance, one contextual factor that can justify different fairness views could be whether differences in earnings are deserved (see e.g. [Barr et al., 2015](#); [Jakiela, 2015](#)). Intuitively, if individuals differ in the assessment of the situation — e.g. regarding the deservingness of income differences — they may also differ in their fairness judgement.¹² Proposition 1 summarises the idea that differences in income can lead to differences in fairness views. All proofs can be found in Appendix B.1.1.

Proposition 1: *Fairness views depend on relative income. For richer individuals $\alpha_i^* = 1$, implying that they hold a contribution-based fairness view ($m_{kt} = x_{kt}$). Poorer individuals, with $\alpha_i^* = 0$ hold an egalitarian fairness view ($m_{kt} = \frac{X_t}{2}$).*

2.3 Experience of shocks

Finally, but most importantly, consider the situation where endowments are not fixed but subject to negative shocks. Formally, this means that individuals suffer a reduction in earnings between $t = 1$ and $t = 2$ such that $x_{k2} < x_{k1}$. While shocks change the allocation decision, pre-shock earnings can still affect fairness judgements by generating reference points ([Roth & Wohlfart, 2016](#); [DellaVigna et al., 2017](#); [Charité et al., 2022](#)). Conceptually, the allocation perceived as fair can thus be written as a weighted average of the initial and the post-shock situation:¹³

$$m_{i2}^* = (1 - \rho_i)m_{i2} + \rho_i m_{i1} \quad (5a)$$

$$m_{j2}^* = (1 - \rho_i)m_{j2} + \rho_i m_{j1}, \quad (5b)$$

where $\rho_i \in [0, 1]$ is exogenously given and can be interpreted a measure for the stickiness or saliency of the initial judgement (m_{k1}). If $\rho_i > 0$ the pre-shock situation will influence which allocation individuals consider as fair. m_{k2} describes what an individual would have considered as a fair allocation in the post-shock situation, had they been presented with it from the start. However, given the shock, the final fairness judgement (m_{k2}^*) may be different from the hypothetical judgement (m_{k2}).

Note that Equations (5a) and (5b) are subject to a feasibility constraint, $m_{i1} + m_{j1} = X_2$, as a negative shock implies that the total distributable amount (X_t) is shrinking. It may thus become impossible to stick to the allocation that was originally considered as fair. Individuals need to weigh off the fair reference allocation for themselves against the one for the other player. This trade-off is measured by γ_i .¹⁴

¹²A crucial aspect of the experimental design is that income differences are generated in a way that allows for different perceptions of deservingness (see Section 3), with previous literature showing a self-serving bias in the perception of deservingness ([Deffains et al., 2016](#); [Amasino et al., 2021](#)).

¹³The specifications in Equations (5a) and (5b) are consistent with the standard model for reference dependent preferences (see [Kőszegi & Rabin, 2006](#)).

¹⁴In Appendix B.1.2, I discuss an alternative version of the model where reference points are not expressed in absolute terms but in shares. This makes γ_i redundant. The predicted effect of shocks does not change using this alternative specification.

The feasible m_{i1} can then be written as $\gamma_i m_{i1} + (1 - \gamma_i)(X_2 - m_{j1})$ and analogously for m_{j1} . Substituting this expression and the feasibility constraint ($m_{i1} + m_{j1} = X_2$) into Equation (5a), results in the following expression for the allocation i perceives as fair:

$$m_{i2}^* = m_{i2} + \rho_i \gamma_i (m_{i1} - m_{i2}) - \rho_i (1 - \gamma_i) (m_{j1} - m_{j2}). \quad (6)$$

The allocation perceived as fair for individual j is derived analogously.

How does relative income affect reactions to shocks? Let us first consider how richer individuals respond to shocks. As argued in Section 2.2, richer individuals adhere to a contribution-based fairness view with $m_{kt} = x_{kt}$. The change in m_{i2}^* caused by an own shock is thus given by $\frac{\partial m_{i2}^*}{\partial (x_{i1} - x_{i2})} = \rho_i \gamma_i \geq 0$, while the response to a shock to others is given by $\frac{\partial m_{i2}^*}{\partial (x_{j1} - x_{j2})} = -\rho_i (1 - \gamma_i) \leq 0$.¹⁵

Corollary: *Under a contribution-based fairness view, the effect of an own negative shock on m_{i2}^* is strictly positive if $\rho_i, \gamma_i > 0$. The effect of a shock to other is strictly negative if $\rho_i > 0$ and $\gamma_i < 1$.*

The framework thus predicts that — given certain parameter restrictions — richer individuals ultimately allocate more to themselves after an own shock and less to themselves after the other suffered a shock. In other words, they become less supportive of distributing income to the poor if they receive a shock themselves, but more supportive if poorer individuals are hit by a shock. As Equation (6) shows, the framework assumes that the effects of an own shock and a shock to the other are additive, i.e. independent from each other. If both players suffer a shock, the weight individuals give to their own versus the other's reference point (γ_i) determines which of the two effects prevails.¹⁶ The predictions for richer individuals are summarised in Proposition 2.

Proposition 2: *Richer individuals react to negative income shocks. For own shocks $\frac{\partial m_{i2}^*}{\partial (x_{i1} - x_{i2})} = \rho_i \gamma_i \geq 0$, while for shocks to others $\frac{\partial m_{i2}^*}{\partial (x_{j1} - x_{j2})} = -\rho_i (1 - \gamma_i) \leq 0$.*

For poorer individuals, by contrast, the egalitarian view implies that $m_{kt} = \frac{X_t}{2}$. This means that independent of who suffers a shock, the effect on m_{i2}^* will be identical. The response to a shock is given by $\frac{\partial m_{i2}^*}{\partial (x_{k1} - x_{k2})} = \rho_i \gamma_i - \rho_i (1 - \gamma_i)$. If $\gamma_i = 0.5$, the effect is equal to zero. Intuitively, if poorer individuals believe that the fair distribution is an equal split, it does not matter whether individual earnings change due to a negative shock. As stated in Proposition 3, the framework thus predicts that if $\gamma_i = 0.5$ fairness judgements and ultimately allocation decisions will not be affected by income shocks.

Proposition 3: *Poorer individuals take the same allocation decision independent of who suffers a shock. If $\gamma_i = 0.5$, the effect of a negative shock is equal to zero.*

¹⁵Note that in the experiment x_{k2} is held constant across treatments. Negative shocks are thus driven by changes in x_{k1} .

¹⁶Section B.1.4 extends the model to situations with incomplete information about the other's shock.

3 Experimental design and procedures

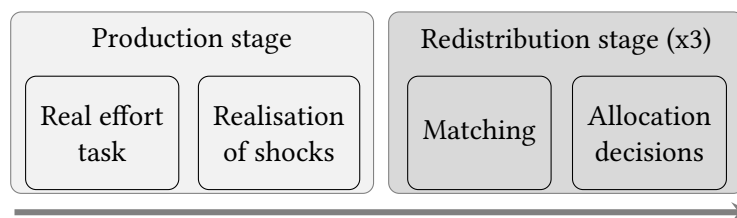
The following section describes the experimental design and the hypotheses derived from the theoretical framework. The experiment was pre-registered and builds on the structure of previous allocation experiments (see e.g. Cappelen et al., 2007, 2013b).¹⁷ It consists of two stages: A *production* and a *redistribution* stage (see Figure 2.2). In the production stage, participants individually generate earnings that depend on i) the difficulty level of a real effort task and ii) whether a negative income shock occurs. In the redistribution stage participants are then matched in pairs and are given the opportunity to freely redistribute the total earnings both players brought into this stage between each other. Each player is thereby matched with three different players who differ with respect to their experience of shocks.

3.1 Player types and realisation of shocks

At the outset of the experiment, participants have to perform a real effort task and receive a flat fee upon completion. They are either completing a *more difficult* task with higher earnings (x_{H1}) or an *easier* task with lower earnings (x_{L1}), introducing differences in relative income.¹⁸ Whether an individual receives high or low initial earnings thus only depends on the random assignment of the real effort task and not on factors like skill or effort. In the following, I refer to participants who receive the high initial earnings as H and to those who receive the low initial earnings as L players. Note that the difference in earnings can be seen as deserved or a consequence of luck. On the one hand, who is assigned to which version of the task is random. On the other hand, knowing that one did a more difficult task could generate a sense of entitlement. This ambiguity allows participants to interpret the income inequality between them in a self-serving way (see also Deffains et al., 2016; Amasino et al., 2021, for self-serving attributions of luck and effort).¹⁹

After participants have completed the real-effort task, a negative income shock may occur. Participants know from the beginning that their earnings depend on external factors that will

Figure 2.2: Session structure



¹⁷AEA RCT Registry. July 06, 2021. <https://doi.org/10.1257/rct.7913-1.0>.

¹⁸In the task, participants have to correctly reverse 10 strings consisting of 6 letters each (see e.g. Zhu et al., 2018). In the easy version of the task, typed letters are visible, in the difficult version they are replaced by asterisks. The median completion time for the hard task is 11 seconds longer (X^2 -test, $p < 0.001$), with the overall median time being 2 minutes 36 seconds.

¹⁹While here individuals know exactly how much each participant contributes to total earnings, in Appendix B.3 I present the results of an additional treatment that introduces uncertainty about individual contributions.

Table 2.1: Pre- and post-shock earnings

	<i>H players</i>		<i>L players</i>	
	No shock	Shock	No shock	Shock
Pre-shock	300	400	100	200
	↓	↓	↓	↓
Post-shock	300	300	100	100

be revealed to them later on (see Appendix B.4.1 for experimental instructions). Half of the participants randomly suffer a shock, while the other half does not. The realisation of a shock is thereby independent of player type and performance in the real effort task. Moreover, shocks do not change the relative income position. Players are immediately informed whether they have been hit by a shock as well as about their new post-shock earnings (x_{H2} , x_{L2}).

A key feature of the design is that post-shock earnings are always identical, independent of own and other's shocks ($x_{H2} = 300$, $x_{L2} = 100$). In other words, the final allocation decision is fixed, but the pathway of how participants get there differs. This is achieved by varying pre-shock earnings across shock and no shock conditions (see Table 2.1). If H players have been assigned to the shock condition they start with pre-shock earnings of 400. They then learn that they suffered a negative shock of -100 and are left with post-shock earnings of 300. If, by contrast, they are assigned to the no shock condition they directly start with 300. The same is true for L players. In the shock condition they start with 200 tokens and are left with 100 after the shock, while they already start with 100 tokens in the no shock condition.²⁰ The treatment manipulation is thus very subtle and any differences in allocation decisions can be attributed to the psychological experience of losing income.

Note that in the case of a shock, both H and L players receive the same *absolute* reduction in earnings (-100). This implies, however, that L players are affected more by shocks in relative terms. While the instructions make absolute losses salient, this does not rule out the possibility that people think in relative terms. If I find that — as predicted by the theoretical framework — H players redistribute more towards L players after the latter suffer a shock, a potential concern is that they only become more generous because L players suffer a relatively larger shock. To exclude this possibility, I run an additional treatment in which relative and not absolute shocks are held constant across player types. The only difference in this variation is that H players start with pre-shock earnings of 600 instead of 400. While absolute shocks are now larger for H players (-300), this implies that H and L players suffer identical shocks in relative terms (-50%). The results for constant relative shock are discussed in Section 4.2.1. In addition, I include questions about the perception of absolute and relative differences in a post-experimental survey that I use as controls in the analysis.

²⁰Differences in pre-shock earnings can be justified, as participants know that there are “different difficulty levels” and people earn more for higher difficulty levels. Moreover, participants never interact with players of the same type (H/L) and thus do not know that some earn more/less for the same task.

3.2 Matching and allocation decisions

In order to investigate how people react to different shocks, I employ a 2x3 design, varying both *own* experience of shocks (shock/no shock) and what is known about the *other's* shock (shock/no shock/no information). In the redistribution stage, H players are always matched with L players. In a random order, each participant faces three scenarios and decides how to distribute total post-shock earnings between themselves and i) a player who has suffered a shock, ii) a player who has not suffered a shock and iii) a player for whom they have no information (while knowing that half of all participants suffered a shock). What is known about the other's shock thus varies within participants, while the own experience of shocks varies between participants. As it is possible that earlier allocation decisions have an impact on later ones, the decision order is randomised. For the incomplete information scenario, I also elicit incentivised beliefs about the shock to the other player.²¹

After being matched with another player, participants can freely redistribute the total post-shock earnings between themselves and the other person in an unbounded dictator game. At the end of the experiment, first one of the three scenarios is selected at random, then the allocation of one of the two players is chosen randomly to be relevant for the bonus.

3.3 Hypotheses

The theoretical framework allows to derive precise hypotheses about how differences in earnings and shock experiences affect distribution decisions in the experiment. See Appendix B.1.3 for a detailed application of the framework to this setting. Proposition 1 shows that fairness views depend on relative income, resulting in the following prediction for H and L players:

Hypothesis 1: *Poorer (L) players allocate less to themselves than richer (H) players. In line with an egalitarian view, L players' allocations will be closer to an egalitarian split (200-200), while H players' allocations will be closer to individual contributions (300-100).*

From Proposition 2 and 3, I can derive hypotheses for the expected reaction to shocks for each player type:

Hypothesis 2: *Richer players react to income shocks. They i) allocate more to themselves after suffering a negative shock and ii) less to themselves after the other suffered a shock.*

Hypothesis 3: *For poorer players, allocations are not affected by the experience of shocks.*

Finally, under incomplete information individuals may downplay the probability that the other suffered shock. Such biased beliefs would be in line with own material interests for H players, as — in line with the contribution-base fairness view — a shock to the other person should be compensated (see Appendix 4.3). Pretending that the other did not suffer a shock, by contrast, can justify a similar allocation as in the situation where it is known that the other did not suffer a shock.

²¹Participants receive 50 tokens if their guess about the other's shock is correct and 0 otherwise. To control for order effects, I randomise whether a participant first takes the allocation decision in iii) or is first asked to state their beliefs about the other player. The results show no significant differences between both orders.

Hypothesis 4: *If information is incomplete, richer players form beliefs in a self-serving way. Allocations under incomplete information will then be similar to a situation where it is known that the other did not suffer a shock.*

3.4 Sample and data collection

The experiment was programmed using LIONESS Lab (Giamattei et al., 2020) and participants were recruited online via Prolific in July 2021. The median completion time was 15.5 minutes and participants earned on average £7.35 per hour. Based on a power analysis, I recruited 536 participants that were equally distributed across player type (H, L) and shock experience (yes, no), resulting in 134 participants in each cell. This allows me to detect an effect size of 0.36 standard deviations at a significance level of 5% with 90% power.²² Participants were stratified with respect to real world shocks to ensure balance across treatment cells (see Table B.3 in Appendix B.2 for balance checks). Using data provided by Prolific, I distinguished between three different sub-samples: 1) participants who became unemployed due to Covid-19 (large shock), 2) participants who were full-time employed and now work part-time (medium shock) and 3) participants who still work full-time (no shock).

After participants completed the experiment, they were asked to fill out a short questionnaire that collected demographic information, perceived closeness between participants (Aron et al., 1992), positive/negative affect reactions to shocks (see Appendix B.2.5), attitudes towards redistribution and solidarity during a crisis, as well as perceptions of inequality and shocks within the experiment.

4 Results

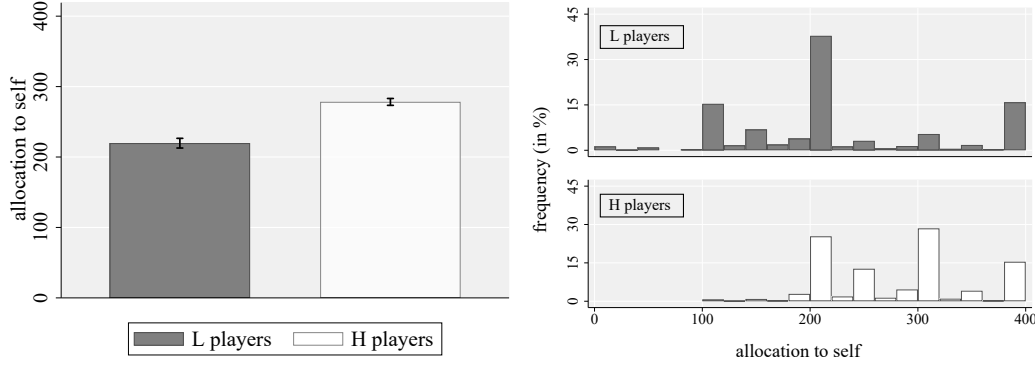
4.1 Relative income and fairness views

The data confirms that different player types take significantly different allocation decisions. The graph on the LHS in Figure 2.3 shows the average allocation to self by player type, with H players allocating significantly more to themselves than L players (t-test, $p < 0.001$). This is in line with previous research showing that initial earnings generate a feeling of entitlement (see e.g. Barr et al., 2015; Jakiela, 2015). Moreover, the graph shows that L players allocate on average 220 tokens to themselves. This exceeds their contribution to total resources (100 tokens), revealing a significant degree of redistribution (t-test, $p < 0.001$). H players, by contrast, allocate on average 278 tokens to themselves, which is only slightly less than their contribution of 300 (t-test, $p < 0.001$). H players thus support a much lower level of redistribution than L players.

The graph on the RHS in Figure 2.3 depicts the distribution of allocations to self for both H and L players. The distribution for H players is not only shifted to the right, but also differs with respect to modal allocations. While the most common allocation to self for H players

²²The power analysis was based on comparisons of means. The effect size was informed by previous research on redistributive preferences (Fisman et al., 2015; Barr et al., 2015; Cassar & Klein, 2019).

Figure 2.3: Average allocation to self (LHS) and distribution (RHS) by player type.



Note: Whiskers represent 95% confidence intervals.

is 300 in line with a contribution-based fairness view, the modal choice for L players is the egalitarian 200-200 split. A Kolmogorov-Smirnov test confirms that the distributions are statistically different from each other ($p < 0.001$). While there is a second mode for H players at 200, indicating some heterogeneity in fairness views, the results confirm that there are substantial differences in both mean and modal allocations between player types, supporting Hypothesis 1.

Result 1: *Relative income differences matter for redistribution decisions. L players are more likely to choose an egalitarian allocation, while H players are more likely to allocate earnings in line with individual contributions.*

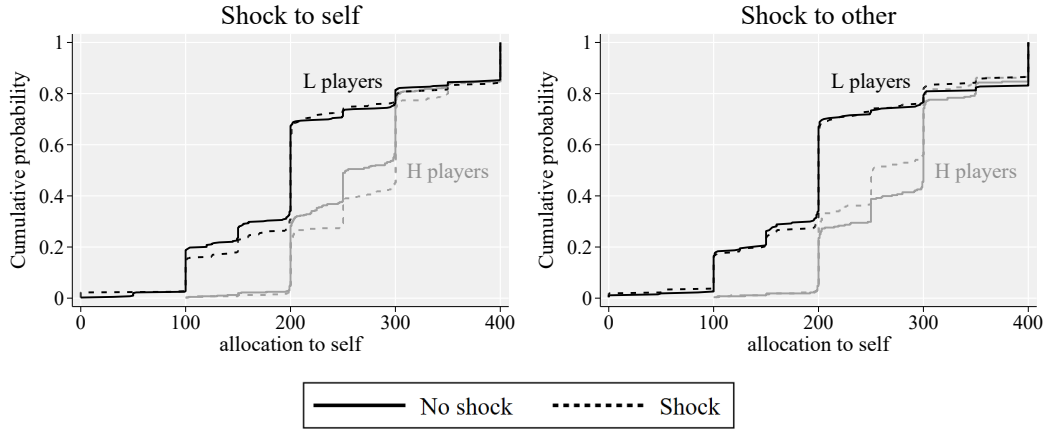
While the results are compatible with H and L players holding different fairness views, an alternative explanation would be that differences in relative income affect the weight individuals put on fairness (β_i). In other words, both richer and poorer individuals hold the same fairness view but care differently about following it. However, this alternative explanation predicts the same response to shocks. The next sub-section shows that this is not compatible with the data, providing support for the interpretation that H and L players hold different fairness views.

4.2 Effect of negative income shocks

Figure 2.4 provides graphical evidence on how allocations to self change after players themselves (left graph) or others (right graph) suffered a shock.²³ As illustrated by the light dotted lines, H players show a significant reaction to shocks with the cumulative distribution of allocations shifting to the right after a shock to self and to the left after a shock to the other (Kolmogorov-Smirnov tests, $p = 0.01$ for own shocks, $p = 0.03$ for shocks to other). H players thus allocate on average more to themselves after an own shock (t-test, $p = 0.02$), but less

²³The affect measure used in the ex-post survey shows that participants perceive shocks as a negative event. They report significantly more negative feelings after having suffered a shock (see Appendix B.2, Figure B.1).

Figure 2.4: Players' reaction to own shock (LHS) and shock to other (RHS)



to themselves if the other suffers a shock (t-test, $p = 0.05$). L players, by contrast, neither react to own nor other's shocks. The cumulative distributions after a shock (dark dotted lines) are statistically not distinguishable from the ones without a shock (dark solid lines).

These findings are also confirmed in a regression analysis. Table 2.2 reports results from regressing the number of tokens individuals allocate to themselves on indicators for own and other's shock.²⁴ In line with the graphical evidence, column (1) shows that H players allocate more to themselves after experiencing a negative shock and less to themselves if the other player suffered a negative shock. These effects are robust to the inclusion of controls in column (2). To test whether reactions to the shock of others depend on own experience of shocks, column (3) includes an interaction term. The latter is insignificant ($p = 0.35$), implying that being affected by a shock does not cause participants to become more or less responsive to the shock of their co-player.²⁵

While shocks have a significant effect on allocation decisions for H players, columns (4) – (6) show that the picture is very different for L players. The coefficients on *shock self* and *shock other* are small and insignificant for all specifications. Different player types thus show very different reactions to shocks (see Appendix B.2, Table B.6 for Chow tests), confirming Hypotheses 2 and 3. This is in line with H and L players holding different fairness views.²⁶

Result 2: *H players react to negative income shocks. They allocate i) more to themselves after experiencing an own shock and ii) less to themselves if the other suffered a negative shock.*

Result 3: *For L players, allocation decisions are independent of income shocks.*

When it comes to the strength of reactions to own and other's shocks for H players, Table

²⁴While the regression controls for order effects, the latter are not significant. Moreover, the results hold when restricting the analysis to the first decision of individuals (see Appendix B.2, Tables B.4 and B.5).

²⁵While this is in line with the theoretical framework where shocks are additive, it contrasts previous research by Cassar & Klein (2019). They find that own prior experiences moderate how people respond to the experience of others. A key difference is that in my experiment shocks happen simultaneously. It would be interesting to test the effect of own shocks on future interactions.

²⁶Independent of player type, I find that individuals with different fairness views show distinct reactions to shocks (see Appendix B.2, Table B.7).

Table 2.2: The effect of shocks on allocation to self (y_i)

	H players			L players		
	(1)	(2)	(3)	(4)	(5)	(6)
Shock self	13.10* (7.87)	14.99* (7.77)	11.87 (8.57)	4.00 (11.63)	3.03 (11.86)	6.68 (12.35)
Shock other	-12.06*** (3.29)	-11.71*** (3.39)	-14.82*** (4.33)	-2.67 (3.65)	-2.20 (3.61)	1.40 (5.50)
Shock self x Shock other			6.24 (6.78)			-7.30 (7.20)
Constant	277.61*** (6.74)	305.43*** (35.65)	306.99*** (35.83)	211.43*** (10.03)	291.91*** (43.38)	290.11*** (43.61)
Demographic controls	No	Yes	Yes	No	Yes	Yes
N observations	536	518	518	534	508	508
N clusters	268	259	259	267	254	254
R ²	0.016	0.117	0.118	0.004	0.133	0.134

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the individual level.

Note: Results of OLS regressions. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Shock self* is a binary variable that takes the value of 1 if an individual i suffered a shock themselves and 0 otherwise. *Shock other* is a binary variable that takes the value of 1 if the player i has been matched with suffered a shock and 0 otherwise. Demographic controls include employment shock due to Covid-19, age, gender, income, household size, education, region, subjective social status, and fairness of task. All regressions control for order effects.

2.2 shows that the coefficient on *shock self* is slightly larger than the one on *shock other* in the models without interactions. However, Wald tests fail to reject the null hypothesis of no difference, implying that H players show an equally strong reaction to own and other's shocks. One caveat is that while H and L players suffer the same shock in absolute terms, L players are affected more in relative terms.²⁷ A more careful interpretation of the results is thus that H players put at least as much weight on their own shocks as on the shocks to others.

4.2.1 Constant relative shocks

As discussed above, relative shocks are not identical for H and L players in the main experiment. One concern is thus that H players only react to the other's shock because it is relatively larger. To address this issue, I run an additional treatment where both H and L players suffer an identical shock in relative terms (−50% of their initial earnings). Data was collected with a sample of $N=536$ Prolific workers in July 2022 using the same procedures as for the main study. The additional data collection was pre-registered.²⁸

Table 2.3 shows how allocation decisions are affected by the experience of constant relative shocks, using the same specifications as in Section 4.2. Columns (1) – (3) confirm that H players show a significant reaction to the other's shock, replicating Result 2ii. The effect in the main experiment is thus not an artefact of the relatively larger shock to other. Moreover, columns (4) – (6) support Result 3 that for L players allocation decisions are independent of shocks.

²⁷The reaction to the other's shock is robust to controlling for relative perception of shocks, see Appendix B.2, Table B.8).

²⁸AEA RCT Registry. July 11, 2022. <https://doi.org/10.1257/rct.7913>.

Table 2.3: The effect of constant relative shocks on allocation to self (y_i)

	H players			L players		
	(1)	(2)	(3)	(4)	(5)	(6)
Shock self	-6.69 (8.60)	-5.88 (9.01)	-7.98 (9.48)	-6.24 (12.73)	-16.82 (12.92)	-16.55 (13.26)
Shock other	-10.82*** (2.96)	-9.67*** (3.13)	-11.80** (4.64)	-2.66 (2.98)	-2.30 (3.27)	-2.03 (5.29)
Shock self x Shock other			4.20 (6.27)			-0.54 (6.56)
Constant	290.03*** (7.29)	311.81*** (35.85)	312.88*** (35.78)	257.89*** (11.82)	310.65*** (45.49)	310.51*** (45.46)
Demographic controls	No	Yes	Yes	No	Yes	Yes
N observations	536	502	502	536	494	494
N clusters	268	251	251	268	247	247
R ²	0.010	0.081	0.081	0.028	0.200	0.200

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the individual level.

Note: Results of OLS regressions. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Shock self* is a binary variable that takes the value of 1 if an individual i suffered a shock themselves and 0 otherwise. *Shock other* is a binary variable that takes the value of 1 if the player i has been matched with suffered a shock and 0 otherwise. Demographic controls include employment shock due to Covid-19, age, gender, income, household size, education, region, subjective social status, and fairness of task. All regressions control for order effects.

Interestingly, the additional data collection does not confirm that H players allocate more to themselves after an own shock (see Result 2i). This is surprising as the shock for H players is both larger in absolute and relative terms. In the main experiment, H players lost 100 tokens (25% of their initial income), while they lose 300 tokens (50% of their initial income) in the constant relative shocks treatment. One explanation for this finding is that the initial income difference between L and H players if the latter suffer a shock may be too large to be justifiable (600 vs 200/100), leading H players to judge the advantageous inequality more negatively (see e.g. [Fehr & Schmidt, 1999](#)). In the post-experimental survey I ask participants how they perceived the initial difference in earnings between H and L players. H players with initial earnings of 600 judge income differences to be significantly less fair than H players with initial earnings of 300 (t-test, $p = 0.09$), providing some support for this explanation.

The additional data collection also allows to test whether the effect of relative income on allocation decisions can be replicated. In line with Result 1, I find that L players are most likely to choose an allocation based on egalitarian fairness views, while H players are most likely to allocate resources proportional to individual contributions (see [Figure B.2](#), [Appendix B.2](#)).

Finally, using the post-experimental survey, I find that 60% of participants perceive shocks in relative terms, while 30% view them in an absolute way (see [Appendix B.2](#), [Figure B.3](#)). The share of individuals thinking in absolute terms is significantly higher in the main experiment than in the constant relative shocks treatment (t-test, $p < 0.001$). This opens interesting questions for future research on the perception of income shocks.

4.3 The role of incomplete information

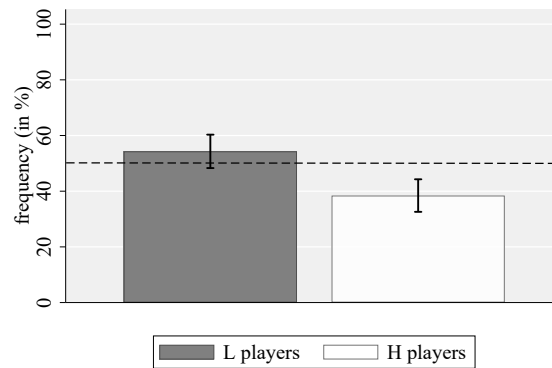
In the incomplete information scenario, participants do not know what happened to the player they are matched with. They only know that the other might have suffered a shock and that overall half of the participants in the experiment suffer a shock and half of them do not. When asked about their beliefs, on average, 50% of participants should thus state that the other did suffer a shock and 50% that the other did not suffer a shock.

Figure 2.5 shows that while L players do not deviate from this prediction (t-test, $p = 0.16$), the beliefs of H players are significantly downward biased with less than 50% stating that the other suffered a shock (t-test, $p < 0.001$). This bias is in line with Hypothesis 4. As H players are the ones who react to shocks, acknowledging that the other player might have suffered a shock implies that they should redistribute more towards the other. If, by contrast, they downplay the probability that the other suffered a shock, H players can justify higher allocations to self. For L players, on the other hand, there is no incentive to distort beliefs, as the egalitarian fairness view translates into an equal split independent of the experience of shocks.²⁹

In addition, a regression analysis shows that incomplete information has a significant effect on allocation decisions for H, but not for L players (see Appendix B.2, Table B.9), with H players behaving as if the other player did not suffer a shock. For L players, by contrast, allocation decisions are independent of the available information.

Result 4: *Under incomplete information H players' beliefs are systematically biased towards the other not suffering a shock. This is also reflected in allocation decisions. L players do not show biased beliefs and allocate earnings in the same way across all scenarios.*

Figure 2.5: Share of players believing that the other suffered a shock by type



Note: The dashed line represents the actual share of players suffering a shock (50%). Whiskers represent 95% confidence intervals.

²⁹These findings are also replicated in the variation with constant relative shocks (see B.2, Figure B.4).

5 Study 2: Income shocks caused by Covid-19

The study described above isolates the effect of income differences and negative shocks through a controlled lab environment. As a next step, I incorporate real world shocks into the experiment. In a follow-up study, I use information on how people's economic situation has been affected by the Covid-19 pandemic and test how this affects allocation decisions in an experiment. This provides additional insights into how people react to shocks in a less abstract environment.

5.1 Design and procedures

Study 2 closely follows the design of the previous study. As before, at the beginning of the experiment, half of the participants complete an easier real effort task, earning a lower flat fee x_L , the other half a more difficult task, earning a higher flat fee x_H . They then decide on how to redistribute total earnings between themselves and another player. However, this time there is no negative shock that reduces earnings from the production stage. Instead, after being matched with another player, participants learn about the *real world shock* of the other person. They either learn that the other i) did suffer a significant loss in livelihood due to Covid-19, ii) did not suffer a significant loss in livelihood due to Covid-19 or iii) are not provided with information about what happened to the other player. Appendix B.4.3 provides instructions for Study 2.

To define whether participants suffered a *significant loss in livelihood*, I use information about changes in their employment status that is accessible via Prolific. Participants who experienced a change from full-time employment to unemployment due to Covid-19 are classified as having suffered a significant shock, while participants who remained in full-time employment are classified as not having suffered a significant shock. I use the formulation a *significant loss in livelihood* to avoid any connotations with being unemployed. While this information is used to inform participants about the shocks of the other player, I directly ask participants whether they suffered a significant loss in livelihood in the ex-post questionnaire and use this question as a proxy for own shock.³⁰ Even if somebody lost their job due to Covid-19, whether this constitutes a significant loss in livelihood will depend on many other factors such as assets, savings or income provided by other family members. I thus use the survey question to make shocks to self and other more comparable. This results in 62% of participants being categorised as having experienced no Covid shock, while 38% are categorised as having suffered a shock.³¹

Study 2 allows me to test the same hypotheses as developed for the first experiment (see Section 3.3). There are however two key differences to the previous design. First, while the shock to the other player remains exogenous in Study 2, the own experience of shocks be-

³⁰The question uses a likert scale from 1-7. I distinguish between people who rather disagree with having suffered a shock (1-4) and those who rather agree with having suffered a shock (5-7).

³¹The correlation between stating to have suffered a significant loss in livelihood and losing employment due to Covid-19 is 0.44 ($p < 0.001$).

comes endogenous. Second, the environment in which I study negative shocks is very different. While exploring income shocks within the experiment provides an abstract and highly controlled environment, Study 2 moves closer to people's real world experiences and explores shocks in a specific economic context. Study 2 thus complements and extends the findings from the main study and shows how the combination of different degrees of control and realism can contribute to a broader understanding of behavioural mechanisms.

The experiment was programmed with LIONESS Lab (Giamattei et al., 2020) and run online via Prolific in July 2021. The median completion time was 14.7 minutes and participants earned on average £7.76 per hour. In line with the power calculations and sample size for the previous study, a total of 536 participants was equally split by player type and loss in livelihood due to Covid-19.

A balance test shows that own shocks are no longer orthogonal to individual socio-economic characteristics (see Appendix B.2, Table B.11). However, generalising from a lab experiment will always be accompanied by a loss of control, while offering a deeper understanding of the problem in a specific context (see Falk & Heckman, 2009, for a methodological discussion on laboratory experiments and causal inference). In the analysis, I control for these characteristics when analysing own shocks.

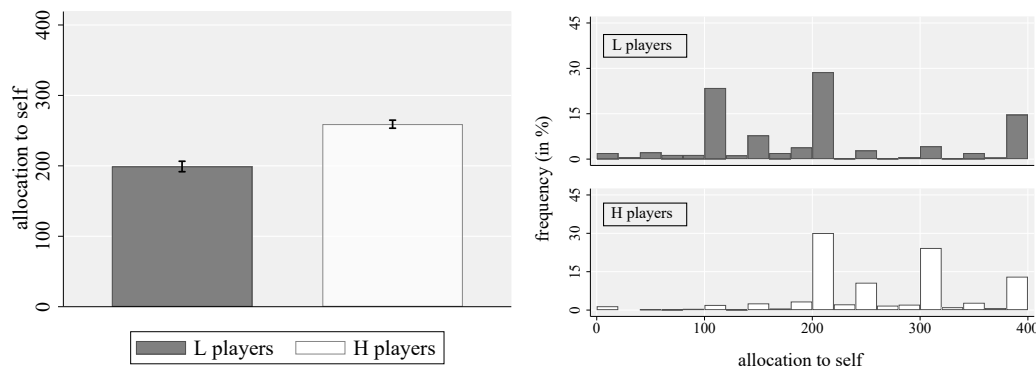
5.2 Results

5.2.1 Relative income and fairness views

The results confirm that differences in relative income translate into different allocation decisions. In line with Result 1, H players allocate significantly more to themselves than L players ($p < 0.001$) and favour lower levels of redistribution (see Figure 2.6, LHS). Even though the distributions of allocations are different across types (Kolmogorov-Smirnov test, $p < 0.001$), the right graph in Figure 2.6 shows that the most common allocation for both H and L players is now 200-200, suggesting an egalitarian fairness view. Also the share of participants who follow a contribution-based fairness view (300-100) is comparable across player types.

When analysing the results for Study 2, it is important to note that the difference between

Figure 2.6: Average allocation to self (LHS) and distribution (RHS) by player type (Study 2)

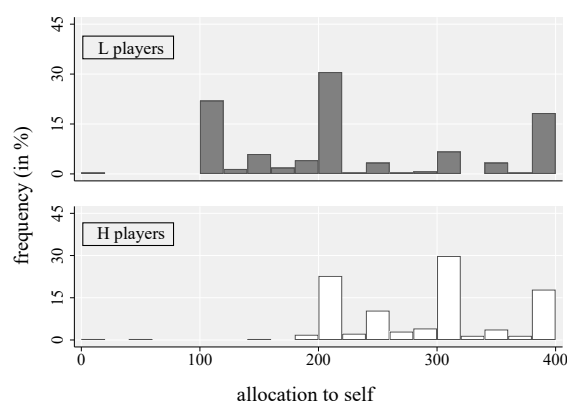


Note: Whiskers represent 95% confidence intervals.

H and L players is generated within the lab, while the information about the other's shock is based on recent and severe real world experiences. It is thus conceivable that the information about real world shocks has such a strong impact that differences between experimental player types become less important when deciding on what is fair. In fact, when restricting the sample to cases where it is known that the other did not suffer a shock, the finding that different player types choose allocations in line with different fairness views re-appears. As Figure 2.7 shows, H players are again most likely to allocate 300 to themselves, while the modal allocation for L players remains an equal split.

Relative income differences in the lab thus matter for allocation decisions in the absence of real world shocks. When introducing large real world shocks this difference becomes less pronounced and an egalitarian split becomes the most common allocation for all players.

Figure 2.7: Distribution of allocations to self in the absence of shocks to others



5.2.2 Effect of negative income shocks

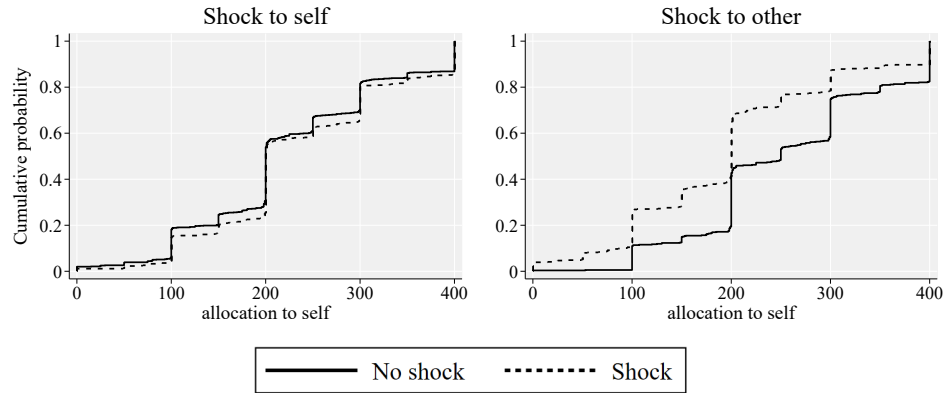
As stated above, I find that player types induced in the lab lose their importance in the face of a large real world shock. When exploring the effect of negative shocks, a Chow test confirms that the reaction to both own and other's shock is statistically indistinguishable between types (see Appendix B.2, Table B.12). For the rest of the analysis, I thus pool H and L players and only control for level differences.

Figure 2.8 shows how real world shocks affect allocation decisions. In the case of a shock to self (LHS), the cumulative distribution is shifted to the right (dashed line). This means that own real world shocks are associated with higher allocations to self (t-test, $p = 0.04$). By contrast, when learning someone else suffered a Covid shock (RHS), participants distribute significantly more to that player and less to themselves (t-test, $p < 0.001$).³² All participants thus respond to shocks in the same way as H players in the previous study (see Result 2).

These results are confirmed by a regression analysis, with allocations to self as the dependent variable and indicators for own and other's shock as the explanatory variables of interest

³²A Kolmogorov-Smirnov test shows no difference in the overall distributions for shock to self, but a significant difference for shock to other ($p < 0.001$).

Figure 2.8: Reactions to own (LHS) and other's (RHS) Covid shocks



(see Table 2.4).³³ All regressions control for differences between H and L players. The results show that own shocks have a positive, significant effect on allocations to self, while shocks to others have a negative one. These findings are robust to the inclusion of demographic controls in column (2). As in the previous study, the interaction between own and other's shock included in column (3) is insignificant. Individuals thus react to the other's shocks in the same way independent of their own experience. Finally, while player types do not matter for the reaction to shocks, they do for absolute allocations to self. Across all specifications, H players

Table 2.4: The effect of Covid shocks on allocation to self (y_i)

	(1)	(2)	(3)
Shock self	16.47** (7.70)	16.85* (9.00)	14.96 (9.49)
Shock other	-56.88*** (3.25)	-57.00*** (3.33)	-58.42*** (4.32)
Shock self x Shock other			3.78 (6.75)
H Player	61.38*** (7.46)	64.00*** (7.77)	64.00*** (7.77)
Constant	212.25*** (7.70)	223.80*** (31.03)	224.51*** (30.94)
Demographic controls	No	Yes	Yes
N observations	1072	1026	1026
N clusters	536	513	513
R ²	0.17	0.21	0.21

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the individual level.

Note: Results of OLS regressions. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Shock self* is a binary variable that takes the value of 1 if an individual i suffered a shock themselves and 0 otherwise. *Shock other* is a binary variable that takes the value of 1 if the player i has been matched with suffered a shock and 0 otherwise. *H Player* is a binary variable that takes the value of 1 for H players and 0 for L players. Demographic controls include employment shock due to Covid-19, age, gender, household size, education, region, subjective social status, and fairness of task. All regressions control for order effects.

³³See Appendix B.2, Tables B.13 and B.14 for an analysis of order effects. The results also hold when restricting the sample to only the first decision of each individual (see Appendix B.2, Table B.15).

allocate on average more to themselves than L players.³⁴

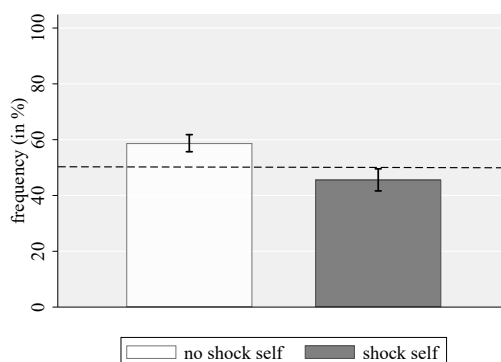
When it comes to the weight placed on own and other's shocks, Figure 2.8 already indicates that the reaction to other's shock is much more pronounced than the reaction to own shocks. This is confirmed by Wald tests ($p < 0.001$ in columns (1) and (2), $p = 0.01$ in column (3)). However, a direct comparison between own and other's shocks might be again misleading. First, the information about the other's shock is varied exogenously, while the own Covid-19 experience depends on personal characteristics. Second, participants are only asked about own experience of real world shocks in the ex-post survey, making it less salient during the experiment. Finally, the majority of participants (66%) state in the questionnaire that the other person was affected more than they were by Covid-19. Even among participants who state they suffered a significant loss in livelihood, 35% believe that the other player was affected more. The data thus suggests that if people perceive that others are affected more than themselves, they take this into account and put a higher weight on the other's shock.³⁵

5.2.3 The role of incomplete information

Finally, I explore what happens if there is uncertainty about the other's real world shock. Under incomplete information, participants are told that the other “*might or might not have suffered a significant loss in livelihood due to Covid-19*”. They do know, however, that the probability for a shock is 50% among all participants. When asked to guess the other's shock, half of the participants should thus guess that the other did suffer a shock and half that they did not.

On aggregate, the share of participants who believe that the other suffered a shock is very close to 50% and if anything slightly above (t-test, $p = 0.08$). Even though Figure 2.9 shows

Figure 2.9: Share of players believing that the other suffered a Covid shock



Note: The dashed line represents the actual share of players suffering a shock (50%). Whiskers represent 95% confidence intervals.

³⁴Since I also have information about Covid shocks for participants in the previous study, I test how own Covid shocks affect behaviour in that experiment. Table B.16, Appendix B.2 shows that also in Study 1 individuals with a Covid shock allocate more to themselves.

³⁵For an analysis of affect reactions and closeness in Study 2 see Appendix B.2.9.

that there is some heterogeneity, with individuals who did not suffer a shock themselves being more likely to acknowledge a shock for others than those who suffered a shock themselves (t-test, $p = 0.003$), overall there is no evidence for self-serving beliefs in the data.³⁶ In line with this, allocation decisions under incomplete information lie exactly between the decisions when it is known that the other did or did not suffer a shock (see Appendix B.2 B.17).

These findings are contrary to Result 4 in Study 1. As the real world shock is so large, it might be psychologically too costly to assume that the other did not suffer a shock and behave selfishly towards someone who really suffered a significant loss in livelihood.

6 Conclusion

In this paper, I study how redistributive preferences are affected by negative economic shocks and differences in relative income, allowing a better understanding of how redistributive preferences are formed and interact with socioeconomic variables. To structure my empirical analysis, I first develop a theoretical framework that is based on reference dependence and a self-serving bias in fairness perceptions. I then use an experiment that exogenously varies individuals' experience of shocks as well as their relative income to isolate the effect on redistributive decisions.

The results provide evidence for the notion that support for redistribution depends on an individual's position within the income position. Poorer individuals who benefit from redistribution tend to distribute resources in an egalitarian way, while richer individuals who benefit from the status quo distribute resources proportionate to individual inputs. Moreover, relative income moderates reactions to shocks, with only richer participants responding to shocks. They become less supportive of redistribution after the experience of own shocks, but more supportive when learning that a poorer person was hit by a shock. This shows that depending on which socioeconomic groups are affected by economic shocks, times of crises can both generate support for or resistance to redistributive policies. Lastly, I find that if there is uncertainty about the experience of others, participants downplay the possibility of a shock to avoid having to distribute more towards them. It thus matters how aware individuals are of the consequences of a crisis for different groups. This may be particularly important if groups are affected differently by economic shocks.

An interesting implication of the interaction between responses to shocks and relative income is that there seem to be limits to self-serving interpretations of fairness. While individuals behave in line with the fairness view that is most beneficial to them, they do not re-evaluate this view in the presence of shocks or self-servingly only react to own shocks. Individuals either compensate both own and others shocks or no shocks at all. This suggests interesting directions for further research on the limits of self-serving or motivated beliefs.

The paper also highlights that the context and scale of economic shocks matter. In a follow-up study, I incorporate real world shocks into the experiment, using data on whether people

³⁶The same pattern holds when examining H and L players separately. See Appendix B.2, Figure B.7.

suffered a *significant loss in livelihood* due to Covid-19. An interesting characteristic of the Covid shock is that like income shocks in the experiment they may be interpreted as random, with individuals not being accountable for their situation. The results confirm that own real world shocks lead to larger allocations to self while shocks to others lead to higher redistribution towards those who suffered a shock. However, income differences induced in the lab become less important in the presence of Covid shocks. Similarly, participants do not downplay the possibility that the other suffered a Covid shock under incomplete information. Intuitively, the follow-up study does not only add real world elements, but also increases the scale of shocks, making them harder to ignore.

Overall, this paper shows that economic shocks have a significant effect on peoples' support for redistribution. Taking changes of socio-economic contexts into account can thus provide a better understanding of the heterogeneity in redistributive preferences.

References

- Adams-Prassl, A., Boneva, T., Golin, M., & Rauh, C. (2020). Inequality in the impact of the coronavirus shock: Evidence from real time surveys. *Journal of Public Economics*, 189(2020), 104245.
- Adler, N. E., Epel, E. S., Castellazzo, G., & Ickovics, J. R. (2000). Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy, white women. *Health Psychology*, 19(6), 586.
- Alesina, A., Cozzi, G., & Mantovan, N. (2012). The evolution of ideology, fairness and redistribution. *The Economic Journal*, 122(565), 1244–1261.
- Alesina, A. & Giuliano, P. (2011). Preferences for redistribution. In *Handbook of Social Economics*, volume 1 (pp. 93–131). Elsevier.
- Alesina, A. & La Ferrara, E. (2005). Preferences for redistribution in the land of opportunities. *Journal of Public Economics*, 89(5-6), 897–931.
- Almås, I., Cappelen, A. W., & Tungodden, B. (2020). Cutthroat capitalism versus cuddly socialism: Are americans more meritocratic and efficiency-seeking than scandinavians? *Journal of Political Economy*, 128(5), 1753–1788.
- Amasino, D., Pace, D., & van der Weele, J. J. (2021). Fair shares and selective attention. *Tinbergen Institute Discussion Paper*, TI 2021-066/I.
- Andre, P. (2021). Shallow meritocracy: An experiment on fairness views. *ECONtribute Discussion Paper*, No. 15.
- Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, 63(4), 596.

- Babcock, L. & Loewenstein, G. (1997). Explaining bargaining impasse: The role of self-serving biases. *Journal of Economic Perspectives*, 11(1), 109–126.
- Barr, A., Burns, J., Miller, L., & Shaw, I. (2015). Economic status and acknowledgement of earned entitlement. *Journal of Economic Behavior & Organization*, 118(2015), 55–68.
- Barr, A., Miller, L., & Ubeda, P. (2016). Moral consequences of becoming unemployed. *Proceedings of the National Academy of Sciences*, 113(17), 4676–4681.
- Bénabou, R. (2015). The economics of motivated beliefs. *Revue d'Économie Politique*, 125(5), 665–685.
- Bénabou, R. & Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, 126(2), 805–855.
- Bénabou, R. & Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3), 141–64.
- Brekke, K. A., Kverndokk, S., & Nyborg, K. (2003). An economic model of moral motivation. *Journal of Public Economics*, 87(9-10), 1967–1983.
- Butler, P. (2021). Majority think Covid has increased UK social inequality, survey shows. Social Mobility Commission says results show need for urgent action to stop gap growing even wider. *The Guardian*, March 11, 2021. Available at: <https://www.theguardian.com/inequality/2021/mar/11/majority-think-covid-has-increased-uk-social-inequality-survey-shows>.
- Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.
- Cameron, L. A. (1999). Raising the stakes in the ultimatum game: Experimental evidence from indonesia. *Economic Inquiry*, 37(1), 47–59.
- Cappelen, A. W., Falch, R., Sørensen, E. Ø., & Tungodden, B. (2021). Solidarity and fairness in times of crisis. *Journal of Economic Behavior & Organization*, 186(2021), 1–11.
- Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., & Tungodden, B. (2007). The pluralism of fairness ideals: an experimental approach. *American Economic Review*, 97(3), 818–827.
- Cappelen, A. W., Konow, J., Sørensen, E. Ø., & Tungodden, B. (2013a). Just luck: An experimental study of risk-taking and fairness. *American Economic Review*, 103(4), 1398–1413.
- Cappelen, A. W., Moene, K. O., Sørensen, E. Ø., & Tungodden, B. (2013b). Needs versus entitlements – an international fairness experiment. *Journal of the European Economic Association*, 11(3), 574–598.
- Cassar, L. & Klein, A. H. (2019). A matter of perspective: How failure shapes distributive preferences. *Management Science*, 65(11), 5050–5064.

- Charité, J., Fisman, R., Kuziemko, I., & Zhang, K. (2022). Reference points and redistributive preferences: Experimental evidence. *Journal of Public Economics*, 216(2022), 104761.
- Cialdini, R. B., Trost, M. R., & Newsom, J. T. (1995). Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications. *Journal of Personality and Social Psychology*, 69(2), 318.
- Cruces, G., Perez-Truglia, R., & Tetaz, M. (2013). Biased perceptions of income distribution and preferences for redistribution: Evidence from a survey experiment. *Journal of Public Economics*, 98(2013), 100–112.
- Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1), 67–80.
- Deffains, B., Espinosa, R., & Thöni, C. (2016). Political self-serving bias and redistribution. *Journal of Public Economics*, 134, 67–74.
- DellaVigna, S., Lindner, A., Reizer, B., & Schmieder, J. F. (2017). Reference-dependent job search: Evidence from Hungary. *The Quarterly Journal of Economics*, 132(4), 1969–2018.
- Desmet, P., Overbeeke, K., & Tax, S. (2001). Designing products with added emotional value: Development and application of an approach for research through design. *The Design Journal*, 4(1), 32–47.
- Doherty, D., Gerber, A. S., & Green, D. P. (2006). Personal income and attitudes toward redistribution: A study of lottery winners. *Political Psychology*, 27(3), 441–458.
- Durante, R., Putterman, L., & Van der Weele, J. (2014). Preferences for redistribution and perception of fairness: An experimental study. *Journal of the European Economic Association*, 12(4), 1059–1086.
- Erkal, N., Gangadharan, L., & Nikiforakis, N. (2011). Relative earnings and giving in a real-effort experiment. *American Economic Review*, 101(7), 3330–48.
- Falk, A. & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, 326(5952), 535–538.
- Falk, A. & Zimmermann, F. (2011). A taste for consistency and survey response behavior. *CEifo Economic Studies*, 59(1), 181–193.
- Fehr, E. & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3), 817–868.
- Fischbacher, U. & Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, 11(3), 525–547.
- Fisman, R., Jakiela, P., & Kariv, S. (2015). How did distributional preferences change during the great recession? *Journal of Public Economics*, 128(2015), 84–95.

- Gagnon, N., Saulle, R., & Zaunbrecher, H. W. (2021). Decreasing incomes increase selfishness. *Working paper, available at: arXiv preprint arXiv:2107.02888*.
- Gallenstein, R. (2021). Wealth inequality and endogenous fairness views. *Working Paper, available at SSRN 3886795*.
- Gerhards, L. & Heinz, M. (2017). In good times and bad—reciprocal behavior at the workplace in times of economic crises. *Journal of Economic Behavior & Organization*, 134, 228–239.
- Giamattei, M., Yahosseini, K. S., Gächter, S., & Molleman, L. (2020). Lioness lab: a free web-based platform for conducting interactive experiments online. *Journal of the Economic Science Association*, 6(1), 95–111.
- Gneezy, U., Kajackaite, A., & Sobel, J. (2018). Lying aversion and the size of the lie. *American Economic Review*, 108(2), 419–53.
- Grossman, Z. & Van der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, 15(1), 173–217.
- Haisley, E. C. & Weber, R. A. (2010). Self-serving interpretations of ambiguity in other-regarding behavior. *Games and Economic Behavior*, 68(2), 614–625.
- Haushofer, J., Schunk, D., & Fehr, E. (2019). Negative income shocks increase discount rates. *Working Paper*.
- Hoffman, E., McCabe, K., Shachat, K., & Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, 7(3), 346–380.
- Huck, S. & Oechssler, J. (1999). The indirect evolutionary approach to explaining fair allocations. *Games and Economic Behavior*, 28(1), 13–24.
- Hvidberg, K. B., Kreiner, C., & Stantcheva, S. (2023). Social position and fairness views on inequality. *Review of Economic Studies*, forthcoming.
- Jakiela, P. (2015). How fair shares compare: Experimental evidence from two cultures. *Journal of Economic Behavior & Organization*, 118(2015), 40–54.
- Jia, T. (2022). Empathy, motivated reasoning, and redistribution. *Working Paper*.
- Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–292.
- Karadja, M., Mollerstrom, J., & Seim, D. (2017). Richer (and holier) than thou? the effect of relative income improvements on demand for redistribution. *Review of Economics and Statistics*, 99(2), 201–212.
- Konow, J. (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. *American Economic Review*, 90(4), 1072–1091.

- Konow, J. (2003). Which is the fairest one of all? a positive analysis of justice theories. *Journal of Economic Literature*, 41(4), 1188–1239.
- Kőszegi, B. & Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4), 1133–1165.
- Larney, A., Rotella, A., & Barclay, P. (2019). Stake size effects in ultimatum game and dictator game offers: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 151(2019), 61–72.
- Lotti, L. & Pethiyagoda, S. (2022). Generosity during covid-19: investigating socioeconomic shocks and game framing. *Humanities and Social Sciences Communications*, 9(1), 1–10.
- Margalit, Y. (2019). Political responses to economic shocks. *Annual Review of Political Science*, 22, 277–295.
- Martén, L. (2019). Demand for redistribution: Individuals’ responses to economic setbacks. *The Scandinavian Journal of Economics*, 121(1), 225–242.
- Mengel, F. & Weidenholzer, E. (2022). Preferences for redistribution. *Journal of Economic Surveys*, 2022, 1–18.
- Mérola, V. & Helgason, A. F. (2016). Are we in the same boat or not? the opposite effects of absolute and relative income shifts on redistributive preferences. *The Journal of Politics*, 78(4), 1107–1123.
- Naumann, E., Buss, C., & Bähr, J. (2016). How unemployment experience affects support for the welfare state: a real panel approach. *European Sociological Review*, 32(1), 81–92.
- Nettle, D. & Saxe, R. (2020). Preferences for redistribution are sensitive to perceived luck, social homogeneity, war and scarcity. *Cognition*, 198(2020), 104234.
- Neuber, T. E. (2021). Egocentric norm adoption. *ECONtribute Discussion Paper*, No. 116.
- Piketty, T. (1995). Social mobility and redistributive politics. *The Quarterly Journal of Economics*, 110(3), 551–584.
- Powdthavee, N. & Oswald, A. J. (2014). Does money make people right-wing and inegalitarian? A longitudinal study of lottery winners. *IZA Discussion Paper*, 7934.
- Preuss, M., Reyes, G., Somerville, J., & Wu, J. (2022). Inequality of opportunity and income redistribution. *Beiträge zur Jahrestagung des Vereins für Socialpolitik 2022: Big Data in Economics*, ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg.
- Raihani, N. J., Mace, R., & Lamba, S. (2013). The effect of 1, 5 and 10 stakes in an online dictator game. *PloS one*, 8(8), e73131.
- Regner, T. (2018). Reciprocity under moral wiggle room: Is it a preference or a constraint? *Experimental Economics*, 21(4), 779–792.

- Roth, C. & Wohlfart, J. (2016). Experienced inequality and preferences for redistribution. *CE-Sifo Working Paper*, No. 6251.
- Schildberg-Hörisch, H. (2010). Is the veil of ignorance only a concept about risk? An experiment. *Journal of Public Economics*, 94(11-12), 1062–1066.
- Selten, R. & Ockenfels, A. (1998). An experimental solidarity game. *Journal of Economic Behavior & Organization*, 34(4), 517–539.
- Siemplenski, J. (2022). Ukraine economic shock. *European Investment Bank*, June 14, 2022. Available at: <https://www.eib.org/en/stories/ukraine-trade-inflation>.
- Stein, W. & Weisser, R. A. (2022). Direct shock experience vs. tangential shock exposure: Indirect effects of flood shocks on well-being and preferences. *The World Bank Economic Review*, 36(4), 909–933.
- Sutter, M. & Weck-Hannemann, H. (2003). Taxation and the veil of ignorance—A real effort experiment on the laffer curve. *Public Choice*, 115(1), 217–240.
- Ten, S., Yonzan, A., Narayan, N., Cojocaru, A., Agrawal, A., Bundervoet, S., Davalos, T., Garcia, M. E., Lakner, N., Mahler, C., et al. (2022). Covid-19 and economic inequality: Short-term impacts with long-term consequences. *World Bank Policy Research Working Paper Series*, No 9902.
- The Economist (2021). Might the pandemic pave the way for a universal basic income? A true UBI seems far off. But more experimentation is likely. *The Economist*, March 4, 2021. Available at: <https://www.economist.com/finance-and-economics/2021/03/02/might-the-pandemic-pave-the-way-for-a-universal-basic-income>.
- Trump, K.-S. (2020). When and why is economic inequality seen as fair. *Current Opinion in Behavioral Sciences*, 34(2020), 46–51.
- Ubeda, P. (2014). The consistency of fairness rules: An experimental study. *Journal of Economic Psychology*, 41(2014), 88–100.
- United Nations (2022). Global impact of the war in ukraine: Billions of people face the greatest cost-of-living crisis in a generation. *UN global crisis response group on food, energy and finance*. Brief No.2, June 8, 2022. Available at: https://news.un.org/pages/wp-content/uploads/2022/06/GCRG_2nd-Brief_Jun8_2022_FINAL.pdf.
- Zhu, M., Yang, Y., & Hsee, C. K. (2018). The mere urgency effect. *Journal of Consumer Research*, 45(3), 673–690.

CHAPTER 3

Strategic Behaviour with Tight, Loose, and Polarised Norms[†]

1 Introduction

Descriptive norms — the behaviour of other individuals in one’s reference group — play a key role in shaping individual decisions (see e.g. [Cialdini & Trost, 1998](#); [Bicchieri, 2005](#); [Fisman & Miguel, 2007](#); [Perez-Truglia & Cruces, 2017](#); [Fehr & Schurtenberger, 2018](#); [Bicchieri et al., 2022b](#)). When characterising the behaviour of others, a common approach in the existing literature is to focus on *mean* or *modal* behaviour (see e.g. [Kandel & Lazear, 1992](#); [Grout et al., 2015](#); [Feldhaus et al., 2019](#)).¹ In this paper, we argue that this approach fails to account for important features of the distribution of behaviour — namely its *variance* or *shape* — that may play a vital role in how people react to a descriptive norm in the presence of strategic interactions.

The role of variance in characterising descriptive norms is emphasised by [Gelfand et al. \(2011\)](#) and [Gelfand \(2021\)](#), who distinguish between tight and loose norms, a distinction that can help to understand systematic differences across cultures. Tight cultures are characterised by well-defined behaviour, while loose cultures show patterns of greater behavioural variance. Implicit in this approach is the idea that, when faced with a loose norm, people’s reactions exhibit more variation and vice versa for tight norms, generating multiple equilibria that can be expressed in different cultural characteristics. In line with this, variations in the tightness of norms have also been linked to cross-cultural differences in norm-following and norm stability ([Aycinena et al., 2022](#); [Kimbrough et al., 2022](#)).

In this paper we develop and test a model that — inspired by the tightness-looseness framework — shows how differences in the variance of descriptive norms lead to different degrees

[†]This chapter is based on joint work with Eugen Dimant, Michele Gelfand, and Silvia Sonderegger. I thank CeDEX members and participants at several conferences for their feedback and comments. I am also very grateful for financial support from Stanford University and the German Research Foundation (DFG) under Germany’s Excellence Strategy (EXC2126/1–390838866). The study received ethical approval from the Nottingham School of Economics Research Ethics Committee on 11/10/2021 and the Stanford University Human Subjects Research Board on 23/11/2021.

¹This of course does not imply that other aspects of distribution have been entirely neglected (see e.g., [Bicchieri & Xiao \(2009\)](#); [Krupka & Weber \(2013\)](#); [Adriani & Sonderegger \(2019\)](#); [Michaeli & Spiro \(2017, 2015\)](#)). However, as we will argue, the literature currently lacks a systematic investigation of how the variance and shape of the descriptive norm affect individual behaviour in strategic environments.

of strategic uncertainty and ultimately to different behavioural responses.² In addition to variance, we investigate another important feature that characterises the distribution of behaviour, namely its shape (unimodal versus u-shaped or polarised). Understanding how people react to polarised environments has become increasingly important, especially in the face of increasing political polarisation and its detrimental societal outcomes.³

The focus of this paper is to shed light on how people react to different features of descriptive norms in a context that is arguably one of the cornerstones of human cooperation and ubiquitous in all environments involving social interactions: public goods provision. The existing literature has documented substantial heterogeneity in the distribution of contributions in public good dilemmas across cultures (Henrich et al., 2001a,b; Gächter et al., 2010), as well as differences in norm-sensitivity (Kimbrough & Vostroknutov, 2016), making it a particularly interesting case study for examining the effects of different descriptive norms. We consider a one-shot public good game where two individuals need to sacrifice some of their self-interest to further joint welfare. It is well known that, when deciding on their contributions in social dilemmas, people show strong social preferences in the form of conditional reciprocity (see e.g. Gächter et al., 2010, 2017b; Kölle & Quercia, 2021). For example, when facing a high contributor, they tend to react by choosing a high contribution themselves, while when facing a low contributor they react by contributing little. However, people typically do not know their co-player's contribution when making their choice. What they do know in our set-up is that their co-player's contribution will be drawn from a given probability distribution. This mimics the idea that individuals can observe the prevailing descriptive norm in the society around them, but face strategic uncertainty when predicting the actions of a specific individual.

To see why the shape and the variance of the distribution may matter for an individual's optimal response, consider a collective action problem where individuals contribute on average 2 out of a maximum of 4 tokens. This may reflect a situation where either everyone contributes 2, where each contribution level is selected by an equal share of the population, or where half of the population contributes nothing and half contributes everything. Although these scenarios generate the same average contribution, they clearly depict very different social environments. While a mean-focused approach suggests that contributions should be the same in all three scenarios, our premise is that different distributions entail very different degrees of strategic uncertainty. In very tight environments, where there is low variance in behaviour, strategic uncertainty is minimal, while in loose or polarised environments where there is high variance, the behaviour of other individuals is less predictable and strategic uncertainty is substantial. Using a theoretical model based on reciprocity concerns, we show that strategic uncertainty matters for how individuals respond to descriptive norms. The larger the

²We use the term descriptive norm to refer to the *distribution of behaviour*. We thus deviate slightly from definitions e.g. by Cialdini et al. (1990) or Bicchieri (2005) who define descriptive norms as *what people commonly do*.

³See Iyengar & Westwood (2015); McConnell et al. (2018); Iyengar et al. (2019); Enke (2020); Dimant et al. (2022); Dimant (2023); Ren et al. (2023); Robbett & Matthews (2021); Nunn (2022); Callander & Carbajal (2022). What we call polarised norms can be equivalently interpreted as reflecting the coexistence of two different norms in the population. Thus, while we use the term *polarised* descriptive norm to indicate a polarised distribution of behaviour, this is simply a semantic choice and should not be seen as conflicting with a dual norm interpretation.

variance of a given distribution, the larger the variance of individuals' responses. The distribution of descriptive norms is thus self-sustaining.

We confirm the importance of the shape and variance of descriptive norms experimentally through the lens of a well-powered ($N=1203$) and pre-registered study using a representative sample of the U.S. population.⁴ We do so by introducing a variant of the established public goods game (PGG) with two players as used by Gächter et al. (2017b). Players receive a number of tokens at the beginning of the game and can decide to keep them for themselves or invest them in a public good that is then multiplied by a positive factor and shared equally among both players. The experiment is divided into two parts. In Part I, we use the ABC strategy method (Fischbacher et al., 2001; Gächter et al., 2017b) to elicit the participants' underlying cooperative propensity, as well as their beliefs about their co-player's contribution. In Part II we then present the participants with the distribution from which the co-player's contribution will be drawn. In a between-subject design, we implement six different treatments that vary the mean (*high/low*) and variance/shape (*tight/loose/polarised*) of the co-player's distribution.

Our results confirm that considering the mean as a sufficient statistic when characterising descriptive norms provides an incomplete picture. We find that exposure to descriptive norms that share the same average behaviour generates very different responses depending on their precise nature (*tight, loose, polarised*). In line with our theoretical examination, the distribution of participants' contributions roughly replicates the descriptive norm they are confronted with. When the descriptive norm they face is *tight*, the participants' contributions are narrowly distributed. When the descriptive norm they face is *loose*, the participants' contributions are spread out. Similarly, when confronted with a *polarised* descriptive norm, the participants' reactions are concentrated at the extremes of the distribution: they either choose to contribute a lot or very little. Interestingly, we find that in *polarised* environments a large share of participants choose the maximum contribution, thus exposing themselves to considerable strategic risk: depending on who they are matched with, their contribution will either perfectly match that of their co-player or will be as far as possible from it. These participants could shield themselves from strategic risk by choosing a middle-of-the-road contribution, but they do not. Rather than focusing on minimising strategic risk, they appear to be primarily concerned with not letting high contributors down. This suggests that, in the social domain, people are tolerant of and may even embrace risk.

Our results also point to an interaction between strategic uncertainty and the role of personal traits: When strategic uncertainty is high, people turn to their preferences and personal values to decide how to behave (see also Elster & Gelfand, 2021). We find that *sucker aversion* (the aversion to contributing more than the co-player), *free-riding aversion* (the aversion to contributing less than the co-player), and *personal values* (what people perceive to be the "right thing to do") play a larger role in determining behaviour when individuals are confronted with a *loose/polarised* environment compared to a *tight* one. However, descriptive norms even matter after accounting for personal values and preferences. This underlines the importance of

⁴See <https://aspredicted.org/pm7fu>.

considering both personal values and descriptive norms when making behavioural predictions (see e.g. [Bicchieri, 2005](#); [Bicchieri & Dimant, 2019](#); [Capraro et al., 2019](#); [Bašić & Verrina, 2020](#)). This has clear practical implications, e.g. for policy-makers interested in assessing the role of norms and values in influencing behaviour (see e.g. [Barr et al., 2020](#), for a discussion on the importance of distinguishing norms from personal values for facilitating behavioural change).

Taken together, our findings contribute to the existing literature in various ways. First, we add to a growing body of research that explores the effect of descriptive norms on individual behaviour. Many studies have shown in different contexts that providing information on what other people did in a given situation influences individual decisions. This has been found in both nonstrategic settings such as dictator games ([Bicchieri & Xiao, 2009](#); [Gächter et al., 2017a](#)), voluntary payments ([Shang & Croson, 2009](#); [Feldhaus et al., 2019](#)) and donations to charities ([Dimant, 2019](#); [Bicchieri et al., 2022a](#)), as well as in strategic interactions such as public goods provision ([Chaudhuri et al., 2006](#); [Kerr et al., 2009](#); [Chen et al., 2010](#)). These studies typically communicate information about others using mean or modal behaviour. Consistent with this literature, our study also finds that differences in means have a significant effect on subsequent decisions. However, we extend these findings by exploring differences between whole distributions. Related to the importance of variance, recent literature has shown that uncertainty about norms matters for how individuals respond to it ([Fosgaard et al., 2020](#); [Ciranka & van den Bos, 2020](#)) and that variance plays a role in non-strategic interactions ([d’Adda et al., 2020](#)).

We also add to the literature that stresses conditional cooperation as a powerful motive for explaining contributions to a public good (see e.g. [Fischbacher & Gächter, 2010](#); [Bowles & Gintis, 2013](#); [Gächter et al., 2017b](#)). A number of studies examine how heterogeneity in contributions in one’s interacting group affects (conditional) cooperation ([Chaudhuri et al., 2006](#); [Kerr et al., 2009](#); [Croson, 2007](#); [Wolff, 2017](#)). Using the strategy method, [Cheung \(2014\)](#) and [Hartig et al. \(2015\)](#) consider a set-up where the actions of all group members are known, showing that people do not only react to averages but the whole profile of individual contributions. Our results extend these findings by studying the effects of behavioural heterogeneity in a set-up characterised by strategic uncertainty, where participants know the distribution of behaviour but not their partner’s actions in advance.

A final contribution of our work is that we develop and test a novel norm elicitation approach that allows us to measure not only beliefs about the mean as is the case in established elicitation methods such as [Bicchieri & Chavez \(2010\)](#) or [Krupka & Weber \(2013\)](#), but also regarding the entire distribution in an incentive-compatible way (for a discussion, see [Dimant, 2022](#)). We therefore provide a fine-grained measurement tool to develop a better understanding of descriptive norms and their impact on behaviour that can be used in future research.⁵

The remainder of this paper is structured as follows. In Section 2, we outline our theoretical framework and hypotheses. Section 3 describes the experimental design. Section 4 presents the empirical results, while Section 5 provides a discussion and concludes.

⁵The software eliciting norm-related beliefs using the distribution builder can be downloaded under <https://osf.io/xkc5w/>.

2 Theoretical framework

It is well known that, in strategic environments, reciprocity plays an important role in determining an individual's choice of action (Fehr & Gächter, 2000). The literature on public goods games extensively documents the presence of reciprocity motives (see e.g. Fischbacher & Gächter, 2010; Bowles & Gintis, 2013; Gächter et al., 2017b). When faced with a high contributor, participants contribute a lot, while when faced with a low contributor, they contribute little. This shows that individuals are strongly concerned with matching the behaviour of others, and substantiates our investigation on the effect of descriptive norms.

The underlying idea is that, consistent with their concerns for reciprocity, individuals incur a psychological loss whenever their contribution differs from that of their co-player. Consequently, they adapt their behaviour to the behaviour of their co-player. At the same time, when engaging in everyday interactions people cannot fully anticipate what their counterparts will do, in spite of being aware of the typical distribution of behaviour within society (the descriptive norm). Since their co-player's contribution is unobserved at the time when they choose their action, agents are exposed to strategic uncertainty. Their choice needs to trade off the risk of contributing too little and the risk of contributing too much (relative to their co-player). These competing factors determine the optimal contribution for an individual when confronted with a distribution of co-player's actions.

As remarked in the introduction, a standard approach in economics is to use the mean of the distribution of others' behaviour as a sufficient statistic to determine an individual's optimal reaction. This would suggest that people are indifferent to the variance/shape of the distribution. An alternative view is that people do react to the variance/shape of the distribution they are facing and that their reactions mimic the initial distribution, generating multiple equilibria.

In what follows, we show that these competing views can be captured by employing two canonical loss functions widely used in statistics (see e.g. DeGroot, 2005) as well as economics to model the psychological cost of a mismatch between own and co-player's contribution: (i) *quadratic*: cost is proportional to the square of the difference between contributions, and (ii) *absolute value*: cost is proportional to the absolute value of the difference between contributions. Quadratic loss functions are commonly used, for instance, in models of conformity or coordination (see e.g. Kandel & Lazear, 1992; Grout et al., 2015). The absolute value loss function is widely used following a seminal contribution by Fehr & Schmidt (1999).

Let x_i denote one's own contribution, x_j one's co-player's contribution and X the endowment. We are interested in a set-up where individuals do not observe their co-player's behaviour before selecting their action, but know that the action of their co-player is drawn from a distribution $f(x)$ on $[0, \bar{x}]$ with mean μ and variance σ^2 . Note that, for ease of exposition, in this analysis we focus on continuous approximations of the discrete distributions we use in our experiment, which are depicted in Figure 3.1 of Section 3. All proofs can be found in Appendix C.1.

2.1 Case (i): Individuals react only to the mean

Consider the following stylised model of reciprocal preferences:

$$u_i = X - x_i + \gamma(x_i + x_j) - \frac{\eta_i}{2} (x_i - x_j)^2 - \frac{\delta_i}{2} (x_i - x_i^a)^2. \quad (1)$$

where $X - x_i + \gamma(x_i + x_j)$ is material payoff (for some $\frac{1}{2} < \gamma < 1$), $\eta_i \geq 0$ parametrises i 's reciprocity concerns, $x_i^a \in [0, \bar{x}]$ is what i considers the “right thing to do” (their personal value) and $\delta_i \geq 0$ captures the importance that i ascribes to acting in accordance to their personal value.⁶ The first quadratic term in Equation (1) captures the desire to minimise the psychological loss incurred whenever the player's contribution differs from that of the co-player (“mismatch loss”), while the last quadratic term in Equation (1) models the psychological cost incurred by i when deviating from their personal value. Each individual i selects x_i to maximise their expected utility, where the expectation is taken with respect to x_j . We denote i 's optimal contribution as x_i^* .

Proposition 1: *When utility is given by (1), we have (i) $x_i^* = 0$ if $\eta_i < (1 - \gamma - \delta_i x_i^a)/\mu$, and (ii) $x_i^* = [\eta_i \mu + \delta_i x_i^a - (1 - \gamma)]/(\eta_i + \delta_i)$ otherwise.*

Intuitively, when the mismatch loss from selecting a contribution that is different from the co-player's is quadratic, individuals are averse to strategic risk. They choose their contribution to minimize the strategic risk they are exposed to. The optimal solution to this problem indexes i 's contribution to μ , the co-player's mean contribution. This ensures that the difference between x_i and x_j is never too large. Crucially, it implies that i 's choice only depends on $f(x)$ through μ , and is independent of the other features of the distribution of co-player's behaviour.

2.2 Case (ii): Individuals react to the whole distribution

Suppose now that utility is

$$u_i = X - x_i + \gamma(x_i + x_j) - \alpha_i (x_i - x_j) \mid_{x_j < x_i} - \beta_i (x_j - x_i) \mid_{x_j > x_i} - \frac{\delta_i}{2} (x_i - x_i^a)^2. \quad (2)$$

This utility function differs from Equation (1) in that the mismatch loss incurred by individuals is proportional to the absolute value of the difference between their contribution and that of their co-player. The parameter $\alpha_i \geq 0$ ($\beta_i \geq 0$) measures the marginal disutility obtained from selecting a contribution that exceeds (is lower than) the co-player's contribution. In the following we refer to α_i as *sucker* and β_i as *free-riding aversion*.

Proposition 2: *Let $\phi_i \equiv \beta_i - (1 - \gamma) + \delta_i x_i^a$. When utility is given by Equation (2), we have*

⁶In the main body we focus on the simple case where the psychological loss from deviating from one's personal norm is quadratic. In Appendix C.1 we discuss the case where this loss depends on the absolute value. Finally, it is worth noting that, although the disutility from contributing a different amount from the co-player will typically depend on the degree of intentionality in the co-player's action, this is immaterial here since in our design intentionality is the same across all treatments.

(i) $x_i^* = 0$ if $\phi_i \leq 0$, while (ii) x_i^* satisfies $\delta_i x_i^* + F(x_i^*) (\alpha_i + \beta_i) = \phi_i$ otherwise.

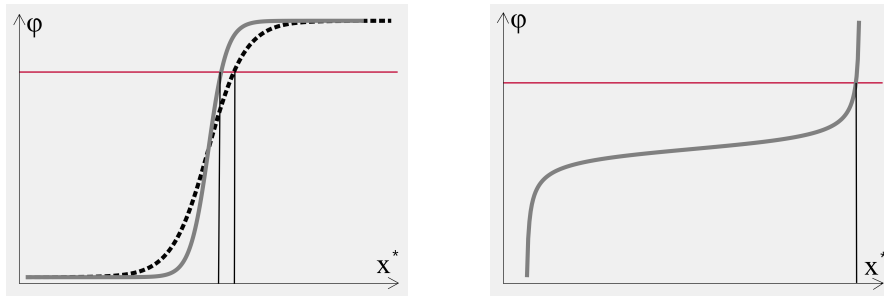
To fix ideas, consider the simple case where $\delta_i = 0$ so that, when interior, x_i^* satisfies $F(x_i^*) = \phi_i$, where ϕ_i is defined as $\phi_i \equiv \frac{\Phi_i}{\alpha_i + \beta_i}$. Figure 3.1 represents the function $F(x)$ for the case of single-peaked distributions (LHS) and polarised, i.e. u-shaped distributions (RHS). In the left graph, the solid line represents a distribution with a smaller variance compared to the dashed line. The horizontal straight lines represent ϕ_i .

As can be seen from Figure 3.1, the point where $F(x)$ and ϕ_i cross depends on the nature of the distribution of co-player contributions. For instance, when $f(\cdot)$ is polarised (RHS), $F(x)$ is steep at the extremes and flat in the middle. This implies that, typically, $F(x)$ and ϕ_i will cross when x takes extreme values – either very low or very high (Figure 3.1 illustrates the latter possibility). When facing a polarised distribution, individuals thus exhibit strategic risk-taking behaviour: they prefer to take a gamble and risk ending up in a completely mismatched position vis-à-vis their co-player rather than opting for a “middle of the road” contribution level which would minimize risk.

In contrast, when the distribution of co-player contributions is single-peaked, $F(x)$ is flat at the extremes and steep in the middle. Consequently, $F(x)$ and ϕ_i will tend to cross when x takes intermediate values. As the variance of $f(\cdot)$ increases, though, x_i^* will tend to become progressively more extreme, as can be seen by comparing the solid and the dashed lines in the left graph of Figure 3.1.

The following result formalises the notion that, as the variance of co-player contribution increases, individuals tend to select more extreme contributions. Consider two distributions f_0 and f_1 with the same mean μ and suppose that f_0 is single-crossing stochastic dominant over f_1 (Machina & Pratt, 1997) so that, for some $\hat{x} \in (0, \bar{x})$, the following holds: $F_1(x) > F_0(x)$ for $x < \hat{x}$ and $F_1(x) < F_0(x)$ for $x > \hat{x}$. In our experiment, this is satisfied for all pairwise comparisons of descriptive norms sharing the same mean (see Appendix C.2, Figure C.1), with $\hat{x} = 2$ in all cases. Note that this condition implies that f_1 is a mean-preserving spread of f_0 . Denoting the optimal contribution under f_k as x_{ik}^* , the following holds.

Figure 3.1: $F(x)$ for single peaked (LHS) and polarised (RHS) distributions



Corollary 1: (i) For all individuals i for whom $x_{i0}^* \in (0, \hat{x}) : x_{i0}^* > x_{i1}^*$. (ii) For all individuals i for whom $x_{i0}^* \in (\hat{x}, \bar{x}) : x_{i0}^* < x_{i1}^*$.

In other words, when individuals are confronted with a distribution that is more spread out, their responses are also more spread out. People who choose a low contribution when facing f_0 choose an even lower contribution when confronted with the more spread out distribution f_1 . Vice versa, those who choose a high contribution when facing f_0 choose an even higher contribution when confronted with f_1 . We now look at the role of individual traits, α_i , β_i and x_i^a , in determining individual contributions. From Equation 2, we can see that contributions increase with the free-riding aversion β_i and personal values x_i^a , but decrease with the sucker aversion α_i .

Corollary 2 When interior, optimal contributions are decreasing in α_i , increasing in β_i and increasing in x_i^a .

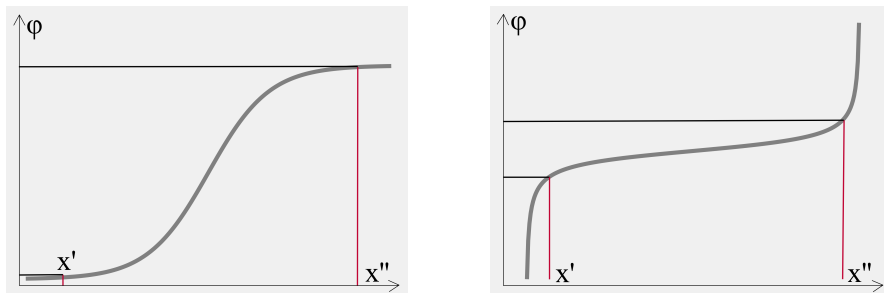
Next, we compare the effect of a change in individual traits on the optimal contribution when the individual is confronted with f_0 vs the more spread out distribution f_1 . The underlying question is whether the nature of the descriptive norm would affect the extent to which personal traits influence contributions.

Corollary 3 Consider $x'' > \hat{x} > x'$. The parameter shift (either in α_i , β_i or x_i^a) needed to generate a change in optimal contribution from x' to x'' is larger under f_0 than under f_1 .

Figure 3.2 illustrates Corollary 3 for a tight versus a polarised distribution, in the easy-to-depict case where $\delta_i = 0$. The change in φ_i needed to generate a shift in optimal contribution from x' to x'' is much larger when the individual faces a tight distribution (LHS) compared to the case of a polarised distribution (RHS).

Note that Corollary 3 does not state that parameter changes always have a larger effect on contributions when individuals face more dispersed empirical norms. If $x' > \hat{x}$ or $x'' < \hat{x}$, for instance, it is possible that this might not be the case. However, the result highlighted in Corollary 3 provides a rationale for why parameter shifts may have a larger effect when individuals are confronted with more dispersed descriptive norms. This is intuitive: When individuals face a dispersed descriptive norm, optimal contributions are dispersed. This implies that there is more scope for differences in parameter values to generate large swings in contributions. This intuition is corroborated by [Elster & Gelfand \(2021\)](#)'s cross-cultural analysis

Figure 3.2: Change in optimal contributions for single-peaked (LHS) and polarised (RHS) distributions



of the World Value Survey, which finds that, in loose cultures, personal values play a greater role in determining civic and pro-environmental involvement compared to tight cultures.⁷

Our final corollary moves away from individual traits and instead compares the effect of a change in the mean of the descriptive norm on contributions.

Corollary 4: *Suppose that f_2 first-order stochastically dominates f_3 . Then, $x_{i2}^* \geq x_{i3}^*$ with strict inequality whenever $F_2(x_{i3}^*) < F_3(x_{i3}^*)$.*

As shown in Appendix C.2, Figure C.1, first-order stochastic dominance applies to all pairwise comparisons of descriptive norms with the same variance but different means in our experiment. Accordingly, Corollary 4 argues that the optimal contribution of an individual confronted with a norm exhibiting a higher mean will be higher.

2.3 Hypotheses

We can now lay out the hypotheses that follow from our theoretical framework. As shown above, there are two alternative hypotheses regarding the impact of the distribution's shape and variance, depending on the assumptions about the underlying loss function.

Hypothesis 1a. *Individuals only react to the mean of a descriptive norm, independently of its variance and shape.*

Hypothesis 1b. *Keeping everything else equal, contributions exhibit larger variance when individuals face a descriptive norm with larger variance, and tend to be polarised when individuals face a polarised descriptive norm.*

Hypothesis 1a follows directly from Proposition 1, assuming a quadratic loss function, while Hypothesis 1b follows from Proposition 2 and Corollary 1, assuming a loss function based on absolute values.

If Hypothesis 1b holds, we can derive further hypotheses for how personal traits affect contributions. As we will describe in greater detail below (in Section 3.3), we elicit both a measure of sucker aversion (α_i) and free-riding aversion (β_i) in the experiment. x_i^a , an individual's personal values, are also measured directly within the experiment. From Corollary 2, we expect that contributions increase with x_i^a and α_i , but decrease with β_i .

Hypothesis 2 *Suppose that Hypothesis 1b holds. Then, contributions are (i) decreasing in sucker aversion, (ii) increasing in free-riding aversion, and (iii) increasing in personal values.*

As stated in Corollary 3 and in line with the findings of Elster & Gelfand (2021), we expect the effect of personal traits to differ across different descriptive norms. It should be larger, the higher the variance.

Hypothesis 3. *Suppose that Hypothesis 1b holds. Keeping everything else equal, the effect of sucker aversion, free-riding aversion, and personal values on individual contributions is larger when individuals face descriptive norms with larger variance.*

⁷See also d'Adda et al. (2020) who find a similar result, albeit within a context which lacks strategic uncertainty.

Finally, in line with Corollary 4 and previous experimental evidence, we expect individuals to react to the mean of a given distribution, contributing more the higher the mean.

Hypothesis 4. *Keeping everything else equal, contributions are larger when individuals face descriptive norms with a higher mean.*

3 Experimental design

3.1 Basic set-up and treatment conditions

To empirically test the hypotheses derived from our theoretical framework, our experiment exogenously varies the *mean* and *variance/shape* of the co-player’s behaviour in a two-person PGG. Table 3.1 gives an overview of the different treatments, including their mean and variance. The corresponding distributions are visualised in Figure 3.3. Full instructions can be found in Appendix C.4. The experiment consists of two parts, and participants learn the details of the second part only upon completion of the first.

- In **Part I**, which is identical across all treatments, we use the ABC strategy method (Fischbacher & Gächter, 2010; Gächter et al., 2017b) to elicit the underlying cooperative attitudes, when individuals interact with a randomly chosen co-player without knowing anything about the underlying distribution of behaviour. These attitudes can then be used as controls for analysing behaviour in Part II.
- At the beginning of **Part II**, participants are randomly assigned to one of six treatment conditions (between participants). In each treatment, participants see a different distribution of behaviour and are informed that the contribution of their co-player for Part II will be randomly drawn from this distribution. The distributions vary with respect to both their mean (high and low), as well as their variance/shape (low variance, high variance, u-shaped), resulting in six treatment conditions.

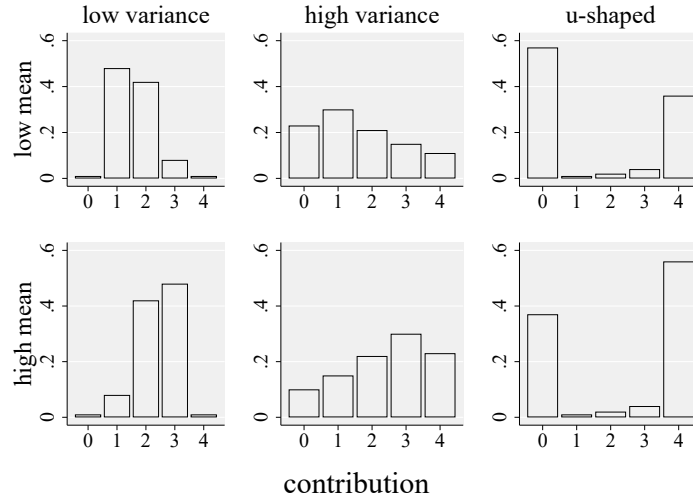
The distributions of co-player behaviour in Part II are constructed through non-random sampling from a previous session (see e.g. Frey & Meier, 2004; Bicchieri & Xiao, 2009; Krupka & Weber, 2009; Bursztyn et al., 2020, for the use of similar approaches).⁸ Participants are aware that the distributions do not represent overall behaviour in a PGG, but only the behaviour of

Table 3.1: Experimental conditions

	<i>Single-peaked</i>						<i>Double-peaked</i>		
	Low variance			High variance			u-shaped		
	mean	var	mode	mean	var	mode	mean	var	mode
Low mean	1.6	0.5	1	1.6	1.6	1	1.6	3.6	0
High mean	2.4	0.5	3	2.4	1.6	3	2.4	3.6	4

⁸See also Charness et al. (2022) and Bardsley (2000) on the use of non-random samples in experiments.

Figure 3.3: Experimental conditions: distribution of co-player’s contribution



a selected subgroup which we constructed using real choices of participants from a previous session. Participants understood that their behaviour is incentive-compatible in that it would affect the payoffs of those previous participants, with one of which they would be paired at random.

3.2 Two-player PGG and beliefs elicitation

In both Part I and Part II, we use a two-player variant of the PGG in which each participant can contribute up to 4 tokens (Ledyard, 1995; Fischbacher et al., 2001). Participants are paid for both their decision in Parts I and II, but receive no information about their payoffs between parts to reduce potential hedging. Tokens invested in the public good are multiplied by 1.4 and shared between both participants. The game embodies the classic tension between private and collective interest: while fully contributing to the public good maximises joint payoffs, each player’s self-interest is maximised by contributing nothing.

To assess underlying cooperativeness, we apply the “ABC of cooperation” (Attitudes-Beliefs-Contribution) in Part I, a method developed by Fischbacher et al. (2001) that aims to disentangle the underlying motives to contribute in a PGG. It embodies three distinct elicitation:

- contribution choices conditional on each possible level of co-player’s contribution to measure cooperative attitude (**A**);
- belief-elicitation task to measure expectations about co-player’s contribution (**B**);
- contribution choice without being informed about co-player contribution (**C**).

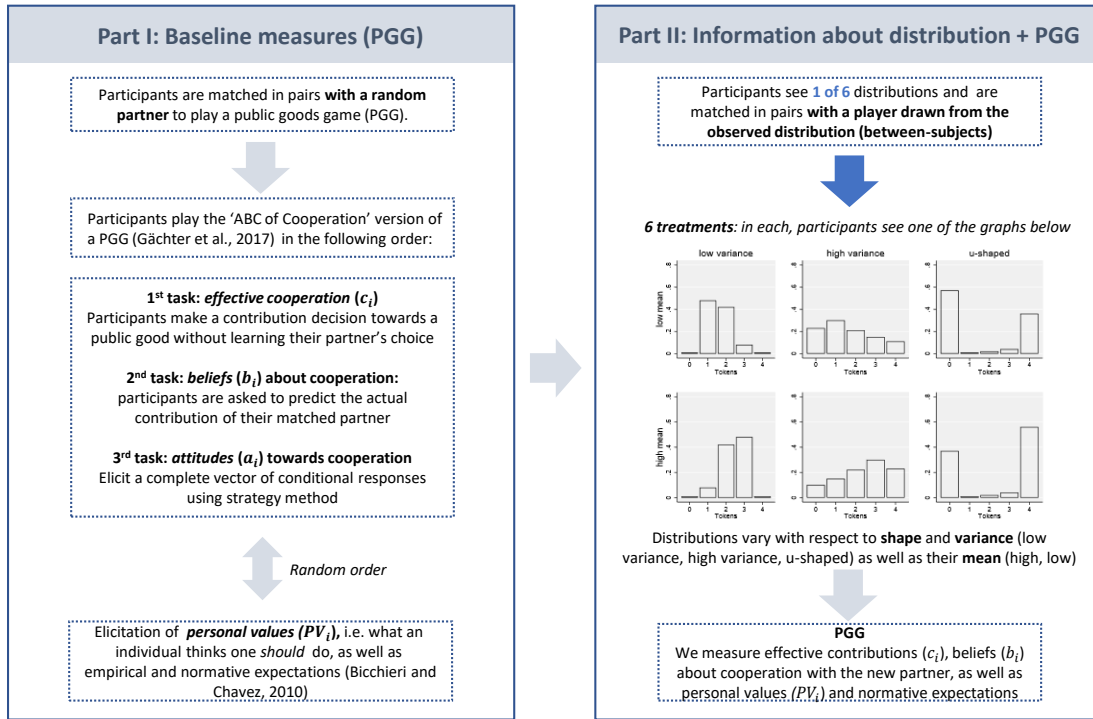
We always elicit unconditional contributions first, followed by beliefs about co-player behaviour and conditional contribution attitudes. The elicited attitudes give us a conditional

contribution vector that we use to classify participants into different cooperation “types”.⁹ In addition, participants are asked to express their personal values (what an individual thinks is the “right thing to do”).¹⁰

In line with our research question, we measure participants’ beliefs as a whole distribution rather than just eliciting their beliefs about the expected co-player’s contribution. To do so, we follow the approach discussed in Dimant (2022) and ask participants to allocate points across all possible co-player contributions (see Appendix C.2, Figure C.2).¹¹ The more likely participants think a contribution is, the more points they should allocate to it. To incentivise decisions, we use a quadratic scoring rule adapted from Artinger et al. (2010) and coupled with an intuitive visual interface developed by Quentin (2016). See Appendix C.3 for more details.

In Part II of the experiment, we ask participants to play a one-shot PGG with a randomly chosen co-player whose contribution is drawn from the shown distribution and again elicit the participants’ personal values and beliefs about their co-player’s contribution. Clearly enough,

Figure 3.4: Overview of the experimental design



⁹Following Fischbacher et al. (2001) we distinguish between conditional cooperators, unconditional cooperators, free-riders, triangle cooperators, and others.

¹⁰We also measure empirical and normative expectations about what most people do and what most people think one should do, in a randomised order. As shown in Appendix C.2, Figure C.6, these are highly correlated with personal values and beliefs about the co-player’s behaviour.

¹¹This method is inspired by existing work on eliciting distributions of subjective beliefs (e.g., Lau et al., 1998; Goldstein & Rothschild, 2014; Harrison et al., 2017) as well as an oral presentation given by Don Ross at the 2019 “Norms and behaviour Change” (NoBeC) workshop. For the purposes of our investigation, we apply those insights to our context of tight, loose, and polarised environments.

if the treatment is successful, then the participants' beliefs about their co-player's contribution should reflect the distribution they have been shown. The decisions participants take in the PGG in Part II have real consequences. They determine the size of the bonus for participants from the previous sessions used to construct the distributions, as well as their own bonus. Figure 3.4 gives an overview of the design.

3.3 Sample and data collection

We programmed the experiment using [Qualtrics \(2005\)](#) and recruited participants online via Prolific in December 2021. The experiment and our hypotheses were pre-registered in November 2021. In total, we recruited a sample of about 1200 US participants that are representative in terms of age, gender, and ethnicity, resulting in about 200 observations per treatment. The chosen sample size was determined using data from a pilot and allows us to detect an effect size of $\eta^2 = 0.01$ at a 5% significance level with 90% power. On average, participants needed 17 minutes to complete the study and earned \$11.29 per hour. To construct the six distributions, we collected data from 685 MTurkers in September 2021. They initially received a show-up fee and then earned an additional bonus depending on the decisions of the participants in the main experiment.

After the main experiment, participants completed an ex post survey that provided additional demographic controls. In addition, we asked participants about the perceived average and variance of the observed distributions as well as their difficulty in interpreting them. As we told participants that the distribution from which their co-player's contribution is drawn was taken from one of the six subgroups we constructed from previous sessions, we also asked how common they think this behaviour is.¹² Finally, we introduced two questions to proxy the participants' aversion towards contributing more and less than their co-player (sucker and free-riding aversion).

4 Results

4.1 Behaviour in Part I

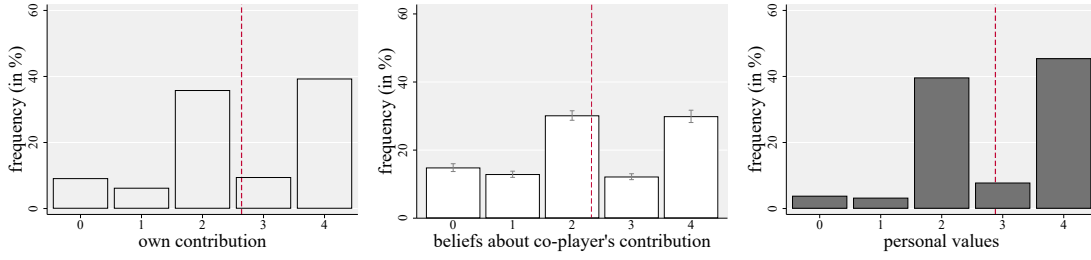
First, we provide an overview of our Part I results. Consistent with the existing literature, the contribution schedule in our data reveals a strong pattern of conditional cooperation among participants. Following the type definitions developed by [Fischbacher et al. \(2001\)](#) and [Thöni & Volk \(2018\)](#), we classify 84% of all participants as conditional cooperators.¹³ The distribution of types is independent of our treatments (χ^2 -test, $p = 0.35$).

Figure 3.5 shows contributions, aggregate beliefs about co-player contributions, and personal values in Part I. The most frequent contribution levels are 2 and 4 tokens, revealing

¹²Figure C.3 in Appendix C.2 shows that participants have a correct interpretation of the observed distributions and state that they were relatively easy to understand. Moreover, the perception of how difficult to interpret and how common the observed behaviour is in a wider population is similar across conditions.

¹³The rest consists of 5% unconditional cooperators, 3% free-riders, 5% triangle cooperators and 4% others.

Figure 3.5: Contributions, aggregate beliefs about the co-player’s contribution, and personal values in Part I



Note: The dashed lines represent averages. Whiskers show 95% confidence intervals.

relatively high levels of cooperation. The same is true for beliefs about the other player’s contribution and personal values about what is the “right thing to do”.¹⁴

The data from Part I already allows us to gain some insight into our research question. Although we do not have exogenous variation in co-player behaviour, we can look at the relationship between participants’ contributions and the variance of their beliefs about their co-player’s contribution. We find that for participants who show a greater variation in beliefs, contributions also vary more (F-test, $p = 0.001$). While this analysis cannot provide causal evidence, it gives initial anecdotal support for the notion that variance of one’s co-player’s behaviour matters and that looser environments may generate more dispersed responses.¹⁵ In the next section, we turn to a more rigorous test of our hypotheses that builds on exogenous variation in co-player’s behaviour.

4.2 Effect of variance and shape of the descriptive norm

This section addresses how exogenous differences in the variance and shape of the descriptive norm presented to participants affect individual responses. We start by looking at the participants’ beliefs about their co-player’s contribution. Figure 3.6 shows that these beliefs closely mirror the distribution of co-player behaviour that was presented to participants.¹⁶

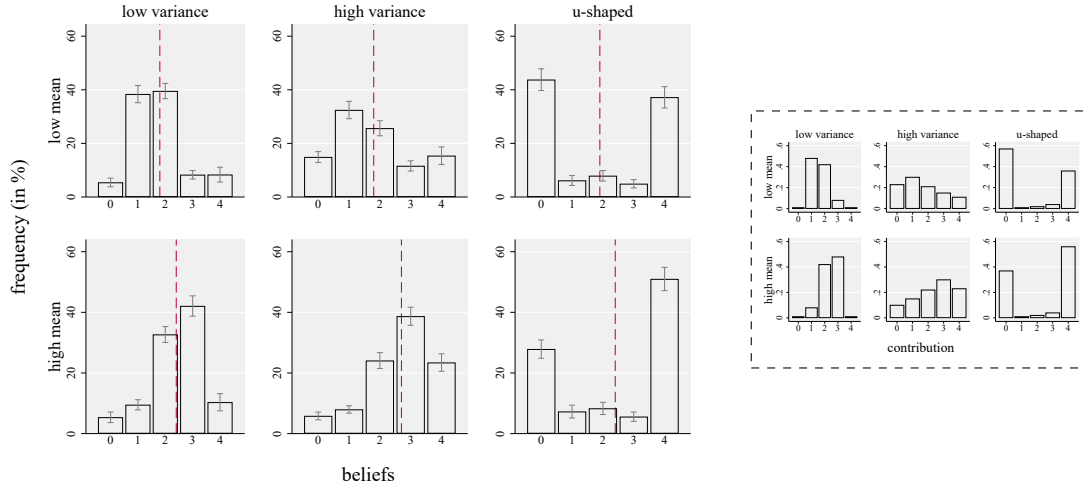
Next, we turn to our main research question of whether differences in the variance and shape of norms affect individual contribution behaviour. Figure 3.7 shows the distribution of contributions in Part II for each experimental condition, confirming that there is indeed a stark difference between treatments. In particular, we see that in tight environments (low variance), participants choose contribution levels that are tightly centered around the mean of the shown distribution ($\sigma^2 = 1.26$). In loose environments (high variance), by contrast, we see a much larger variation in behaviour ($\sigma^2 = 1.65$). In other words, loose behaviour generates loose

¹⁴In Appendix C.2 (Figures C.4 and C.5) we provide a more detailed characterization of the participants’ beliefs about their co-player’s contribution in Part I, showing that the aggregate distribution in Figure 3.5 hides substantial heterogeneity.

¹⁵Similarly, we find a significant positive correlation between average beliefs and contributions ($r=0.65$, $p < 0.001$).

¹⁶Figure C.10 in Appendix C.2 visualises how participants change their beliefs between parts as a reaction to the provided descriptive norm.

Figure 3.6: Aggregate beliefs about co-player's contribution in Part II by treatment

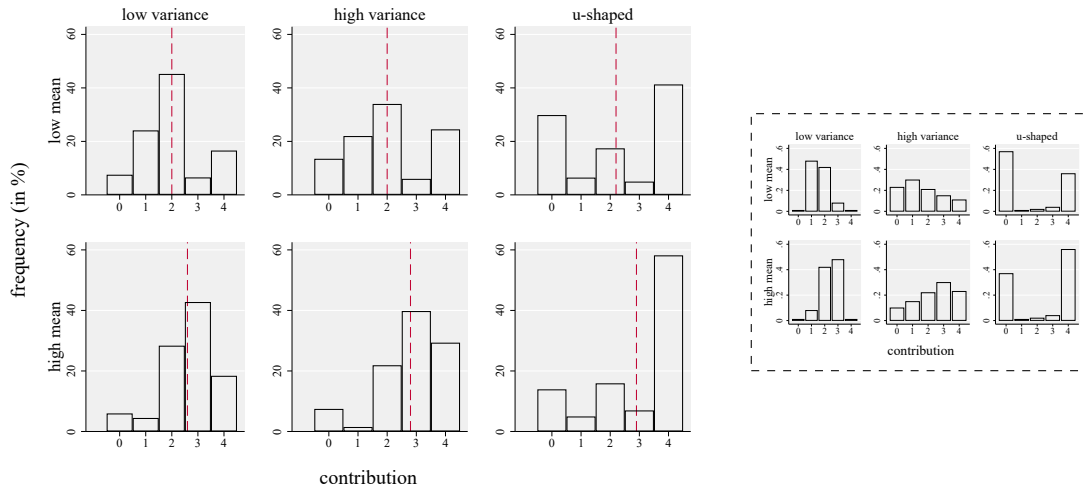


Note: The dashed lines represent average beliefs about co-player's contribution. Whiskers show 95% confidence intervals. The box on the right depicts the distributions shown in each treatment.

responses while tight behaviour generates tight responses. Finally, Figure 3.7 shows that the polarised norm induces polarisation in subsequent contribution behaviour ($\sigma^2 = 2.68$). In Section 4.3 we provide a discussion of the personal traits that drive this heterogeneity.

To test the first visual impression of treatment effects, we perform pairwise F-tests for the equality of standard deviations between treatments. Overall, we find that the variance in contributions is significantly higher in polarised than in tight or loose environments (for both

Figure 3.7: Distribution of participants' contributions in Part II by treatment



Note: The dashed lines represent average contributions. The righthand box depicts the distributions shown in each treatment.

$F < 0.001$).¹⁷ Loose conditions in turn generate a significantly higher variance in participants' contributions than tight conditions ($F = 0.007$). Pairwise χ^2 tests confirm that also the distribution of contributions is significantly different between treatments ($p < 0.001$, $p < 0.001$, and $p = 0.002$ respectively). Moreover, the distribution of contributions in Part II is significantly different from the distribution in Part I ($p < 0.001$). As a further test of treatment differences, in Appendix C.2 we split the sample into two groups, high contributions ($x_i > 2$) and low contributions ($x_i \leq 2$), and show that a higher variance in co-player behaviour increases high contributions, while it decreases low ones (see Table C.1). This confirms that the variance of behaviour increases when participants are faced with more spread-out descriptive norms.

Taken together, our results confirm the importance of both the variance and shape of the observed behaviour for individual decisions. Therefore, we reject Hypothesis 1a and accept Hypothesis 1b. Different environments generate very different responses. Loose, tight, and polarised behaviours reproduce themselves.

Result 1. *Looser descriptive norms lead to a larger variance in contributions. Polarised descriptive norms generate extreme contributions and lead to further polarisation.*

4.3 Personal traits

We now look at the effect of personal values, free-riding, and sucker aversion on contributions. In Table 3.2 we regress contributions on these personal traits, treatment indicators for the shape/variance of the observed distribution and their interaction. To account for the censored nature of our data (Tobin, 1958) and in line with previous PGG studies (see e.g. Fehr & Gächter, 2000; Fischbacher & Gächter, 2010; Chaudhuri et al., 2017), we use Tobit regressions to analyse contributions.¹⁸ Consistent with Hypothesis 2, Table 3.2 confirms that personal values and free-riding aversion increase contributions, while sucker aversion decreases them.

Result 2. *Sucker-aversion lowers contributions while free-riding aversion and personal values increase contributions.*

Consider now Hypothesis 3: Personal traits should matter more in loose or polarised environments compared to tight ones. Intuitively, when strategic uncertainty is high, individuals face a trade-off. If they increase their contribution, they face a higher probability of looking like a sucker, by contributing more than their co-player. On the other hand, by decreasing their contribution they face a higher chance of looking like a free-rider, by contributing less than their co-player. Their degree of sucker aversion and free-riding aversion, as well as their personal values, determine the outcome of this trade-off.

¹⁷Figure C.7 in Appendix C.2 shows the distribution of contributions across different variance conditions, independent of the mean. Figure C.8 shows that the effect of variance/shape holds when restricting the sample to conditional cooperators.

¹⁸Alternative estimation methods such as OLS or ordered probits yield qualitatively similar results (see Appendix C.2, Table C.2). Table C.3 shows that the results also hold when focusing on the sub-sample of conditional cooperators.

Table 3.2: Effect of personal traits on contributions in Part II

	(1)	(2)
High mean	0.96*** (0.12)	0.97*** (0.12)
<i>Variance (baseline = low)</i>		
High variance	-0.51 (0.52)	-0.59 (0.51)
U-shaped	-0.41 (0.55)	-0.51 (0.54)
<i>Sucker aversion</i>	-0.16*** (0.06)	-0.16*** (0.06)
Sucker aversion x high variance	-0.04 (0.08)	-0.00 (0.08)
Sucker aversion x u-shaped	-0.26*** (0.08)	-0.24*** (0.08)
<i>Free-riding aversion</i>	0.17*** (0.05)	0.17*** (0.05)
Free-rider aversion x high variance	0.03 (0.07)	0.01 (0.07)
Free-rider aversion x u-shaped	0.27*** (0.07)	0.27*** (0.07)
<i>Personal values in part I</i>	0.52*** (0.09)	0.50*** (0.09)
High variance x PVs in part I	0.29** (0.12)	0.29** (0.12)
PVs in part I x u-shaped	0.37*** (0.13)	0.38*** (0.13)
Constant	0.65* (0.38)	-1.45** (0.70)
Demographic Controls	No	Yes
N observations	1203	1188
Pseudo R ²	0.10	0.12

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Note: Results of two tobit regressions (data is censored at 0 and 4). The dependent variable is *contributions in Part II*. *Personal values* are measured in Part I and can take values between 0 and 4. *High mean* is a binary variable with 0 = low and 1 = high mean. *Variance* is a categorical variable with 0 = low variance, 1 = high variance and 2 = u-shaped. Sucker and free-riding aversion are measured on a Likert scale from 1 to 7. Demographic controls include age, gender, education, acceptance of risk, trust, and measures for negative and positive reciprocity. All regressions control for order effects.

As can be seen from Table 3.2, personal traits guide individual contributions more in environments characterised by high strategic uncertainty. The interactions between the u-shaped environment and sucker aversion, as well as between u-shaped environment and free-riding aversion are highly significant in the expected direction. The interaction with personal values is both significant for the high variance and the u-shaped environment, confirming that personal values have a stronger effect on decisions in these conditions (see Hypothesis 3).¹⁹

Result 3. *Personal traits (sucker-version, free-riding aversion, and personal values) matter more when participants are confronted with loose or polarised descriptive norms than in the presence of tight descriptive norms.*

¹⁹We use personal values in Part I as these are collected “in a vacuum” and are therefore not affected by the treatments. Our results hold when controlling for the change in personal values and their interaction with treatment indicators (see Appendix C.2, Table C.4).

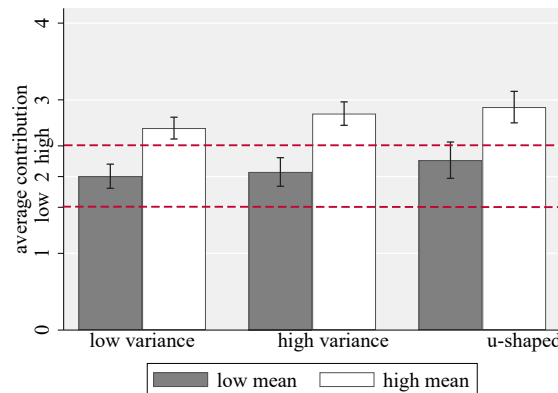
The finding that personal traits moderate participants' reactions to environments characterised by high strategic uncertainty explains the greater heterogeneity observed in those environments relative to tight environments, where strategic uncertainty is very low. Applied to the polarised scenarios, this means that whether a participant's contribution appears to respond primarily to (potential) co-player contributions at the lower or higher end of the distribution seems to be at least partly driven by the participant's personal traits.²⁰

4.4 Effect of high and low means

Finally, we look at the difference between descriptive norms with high and low means (see Hypothesis 4). In line with previous literature, contributions are significantly higher in high mean conditions (Wilcoxon-Mann-Whitney test, $p < 0.001$). This is true independent of the shape and variance of the observed distribution (see Figure 3.8).

Table 3.3 tests Hypothesis 4 formally, by regressing contributions on treatment indicators for the mean and variance/shape of the observed distribution. Columns (4) – (6) moreover include an interaction term between mean and variance/shape. As expected, contributions are significantly higher in the high mean conditions. This finding also holds when controlling for baseline behaviour and including demographic controls. Moreover, columns (1) – (3) show that, overall, contributions are higher in the high-variance and polarised conditions than in the low-variance conditions. This is an interesting finding worth emphasizing: In loose and polarised environments, participants could in principle use the non-negligible probability of facing a free-rider as an “alibi” to justify selfish behaviour. However, our data suggests that this is not the case. Greater strategic uncertainty promotes on average higher contributions. Following the findings discussed in Result 2, this appears to be driven by many participants

Figure 3.8: Effect of a high or low mean on contributions



Note: The dashed lines represent the mean of the observed distributions (high = 2.4, low = 1.6). Whiskers show 95% confidence intervals.

²⁰In Appendix C.2, Table C.5, we look at the correlation between personal traits and demographic characteristics. Women score higher than men both in terms of free-riding and sucker aversion, while older participants show lower levels of both types of aversion.

Table 3.3: Effect of high and low mean conditions on contributions in Part II

	No interaction			Interaction		
	(1)	(2)	(3)	(4)	(5)	(6)
High mean	1.00*** (0.14)	0.99*** (0.11)	0.93*** (0.10)	0.70*** (0.23)	0.64*** (0.18)	0.58*** (0.17)
<i>Variance (baseline = low)</i>						
High variance	0.23 (0.16)	0.26** (0.13)	0.26** (0.12)	0.08 (0.23)	0.09 (0.18)	0.09 (0.17)
U-shaped	0.61*** (0.17)	0.66*** (0.13)	0.61*** (0.13)	0.29 (0.23)	0.28 (0.18)	0.24 (0.18)
<i>Interactions</i>						
High mean x high variance				0.29 (0.33)	0.33 (0.26)	0.35 (0.25)
High mean x u-shaped				0.65* (0.34)	0.80*** (0.27)	0.78*** (0.26)
Constant	2.05*** (0.16)	-1.10*** (0.21)	-1.77*** (0.58)	2.21*** (0.19)	-0.93*** (0.22)	-1.56*** (0.58)
Baseline controls	No	Yes	Yes	No	Yes	Yes
Demographic controls	No	No	Yes	No	No	Yes
N observations	1203	1203	1188	1203	1203	1188
Pseudo R ²	0.02	0.14	0.17	0.02	0.14	0.17

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Note: Results of tobit regressions (data is censored at 0 and 4). The dependent variable is *contributions in Part II*.

High mean is a binary variable with 0 = low and 1 = high mean. *Variance* is a categorical variable with 0 = low variance, 1 = high variance and 2 = u-shaped. Baseline controls include contributions, beliefs, and personal values in Part I and can take values between 0 and 4. Demographic controls include age, gender, education, acceptance of risk, trust, sucker aversion, free-riding aversion, and measures for negative and positive reciprocity. All regressions control for order effects.

experiencing high levels of free-riding aversion, as well as high personal values for cooperation. It would be thus interesting to investigate this relationship in other contexts where personal values may diverge from our sample. Finally, columns (4) – (6) show that the effect of the mean seems to be even larger in the polarised condition, as indicated by the significant interaction.

Result 4. *Participants contribute significantly more when the descriptive norm has a higher mean.*

5 Discussion and conclusion

In this study, we investigate how different descriptive norms of cooperative behaviour affect an individual's own willingness to cooperate. Our contribution lies in the causal evaluation of the tightness-looseness framework of norms that has its origin in cross-cultural psychology (Gelfand et al., 2011) through the lens of an economic model that we test with a behavioural experiment. We first develop a theoretical framework that is based on the notion that individuals are motivated by reciprocity concerns and that differences in the variance/shape of descriptive norms generate different degrees of strategic uncertainty, which in turn affect individual behaviour. We then test our framework empirically in the context of a public goods game where we vary both the *mean* (high/low) as well as the *variance* and *shape* (tight/loose/polarised) of the distribution from which the co-player's contribution is drawn.

Our results confirm previous research showing that information about average behaviour has an important effect on subsequent decisions. Individuals contribute significantly more in high mean conditions than in low mean conditions. Most importantly, though, we show that the mean is not the only relevant feature of the distribution. In line with our theoretical framework, we find that loose norms generate a larger variance in individual responses compared to tight norms, and that polarised environments generate polarised behaviour. In other words, “tight breeds tight”, “loose breeds loose”, and “polarised breeds polarised”.

Interestingly, we find that a large share of participants makes choices that expose them to considerable strategic risk. Consider, for instance, a situation where a share $p < 1/2$ of the population contributes 4 and a share $1 - p$ contributes 0. This is an approximation of the descriptive norm that we utilised in our polarised treatment with a low mean. In this context, choosing a contribution of 2 allows to eliminate all strategic risk, since it generates a sure mismatch of 2 between one’s own contribution and the co-player’s. Anyone who is risk-averse in terms of mismatches will therefore not choose to contribute more than 2, as this exposes them to avoidable strategic risk, while also lowering material payoff.²¹ Yet, we find that a sizeable share of our participants (46%, of which 41% contribute 4) does precisely that. Although (as indicated in Part I of the experiment) the overwhelming majority of participants are strongly motivated by reciprocity concerns and thus want to match their co-player’s contribution, they are not mismatch-risk-averse. This suggests that, when it comes to the social domain, individuals may exhibit different attitudes to risk compared to what we are accustomed to seeing in the monetary domain, which is typically characterised by risk-averse behaviour.²² In future research, one could investigate this hypothesis further to fully appreciate the potential discrepancy between attitudes toward strategic risk in the social and attitudes toward risk in the monetary domain.

Another key finding of our analysis is that an individual’s reaction to high strategic uncertainty is moderated by their personal traits. Relative to environments where strategic uncertainty is low, decisions under high uncertainty are more strongly influenced by personal traits such as personal values, i.e. what an individual thinks is the “right thing to do” in a given situation. This has practical implications for policy-makers and behavioural interventions. Current interventions are often directed at both personal values and beliefs, as well as norms to achieve change (Dimant & Shalvi, 2022). Our results suggest that depending on the relative tightness or looseness of the norm, different approaches might be more fruitful. For example, when intervening in contexts with loose or polarised norms, a focus on personal values might be more successful, whereas when intervening in contexts with tight norms, it may be better to focus on the behaviour of others.

As is universally true for experimental research, the existence and role of experimenter demand effects should be examined (Zizzo, 2010). We consider the potential presence of ex-

²¹For instance, choosing to contribute 4 exposes a participant to a lottery where with probability $1 - p$ the mismatch between their own and the co-player’s contribution is 4, and with probability p it is 0. It is easy to see that $v(2) < (1 - p)v(4) + pv(0)$ for any increasing convex loss function $v(\cdot)$. This follows since, by Jensen’s inequality, $(1 - p)v(4) + pv(0) > v(4(1 - p)) > v(2)$ where the last inequality is due to $p < 1/2$.

²²See e.g. Chetty (2006) and more generally the vast literature on monetary risk aversion.

perimeter demand effects to not be a concern in our set-up for two reasons. First, the finding that loose norms lead to loose responses and polarised norms to polarised ones does not rely on the behaviour of one individual but results from the behaviour of all participants. Thus, different people respond very differently to the observed distribution. This makes demand effects with respect to variance improbable. Second, demand effects would not be able to explain the interaction between variance and personal values that we observe in our results.

Our analysis provides many other avenues for future research. For example, existing literature suggests that the enforcement of norms through punishment is an important part of sustaining existing norms (Balafoutas & Nikiforakis, 2012; Balafoutas et al., 2014). To test our research question, the inclusion of a punishment opportunity was not essential. Our results show that tight, loose, and polarised descriptive norms generate different responses even in the absence of a norm enforcement mechanism. The study thus represents a conservative test of our theory. Future research could investigate the role of these behavioural patterns in a set-up that incorporates punishment. Just as we have shown that individuals are sensitive to distributions of contributions, they are also likely to be responsive to distributions of punishment and adjust their behaviour accordingly. Another potential extension of our work would be to move beyond WEIRD (Western, Educated, Industrialised, Rich, Democratic, see Henrich et al., 2010) samples to test the generalisability of our results.

Although the analysis is static, we believe that our results can speak to the long-run sustainability of different descriptive norms (tight/loose/polarised). A necessary condition for a norm to be self-sustaining in the sense of the standard definition of a stationary distribution (see e.g. Ross et al., 1996) is that, when individuals are confronted with the norm, their reactions should reproduce the norm itself. This is very much in line with our findings. Our participants strongly react to the shape of the descriptive norm they are presented with and the resulting distributions look remarkably similar to the original norms. In this sense, our results can be seen as providing suggestive evidence that – all else equal – there may be multiple equilibria, involving tight, loose, or polarised distributions of behaviour. This ties our findings to the (primarily theoretical) literature on norms as equilibrium selection devices (see e.g. Binmore & Samuelson, 1994; Basu, 1998; Young, 2015) and also opens the door for future research on exploring tightness and looseness in a dynamic and fully endogenous setting.

Overall, we show in this paper that considering the whole distribution instead of focusing only on average behaviour provides substantial analytical richness. This can form the basis for a better appreciation of different behavioural patterns observed across societies. We hope that our work will pave the way to a wider understanding of the interplay between norms and behaviour that encompasses less-studied aspects such as variance and shape, generating a fertile agenda for future research.

References

- Adriani, F. & Sonderegger, S. (2019). A theory of esteem based peer pressure. *Games and Economic Behavior*, 115(2019), 314–335.
- Artinger, F., Exadaktylos, F., Koppel, H., & Sääksvuori, L. (2010). Applying quadratic scoring rule transparently in multiple choice settings: A note. *Jena Economic Research Papers*, (2010,021).
- Aycinena, D., Rentschler, L., Beranek, B., & Schulz, J. F. (2022). Social norms and dishonesty across societies. *Proceedings of the National Academy of Sciences*, 119(31).
- Balafoutas, L. & Nikiforakis, N. (2012). Norm enforcement in the city: A natural field experiment. *European Economic Review*, 56(8), 1773–1785.
- Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*, 111(45), 15924–15927.
- Bardsley, N. (2000). Control without deception: Individual behaviour in free-riding experiments revisited. *Experimental Economics*, 3(3), 215–240.
- Barr, A., Dekker, M., Mwansa, F., & Zuze, T. L. (2020). Financial decision-making, gender and social norms in zambia: Preliminary report on the quantitative data generation, analysis and results. *CeDEx Discussion Paper Series*, (2020-06).
- Bašić, Z. & Verrina, E. (2020). Personal norms—and not only social norms—shape economic behavior. *MPI Collective Goods Discussion Paper*, (2020/25).
- Basu, K. (1998). Social norms and the law. In P. Newman (Ed.), *The New Palgrave Dictionary of Economics and the Law*. Palgrave Macmillan London.
- Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.
- Bicchieri, C. & Chavez, A. (2010). Behaving as expected: Public information and fairness norms. *Journal of Behavioral Decision Making*, 23(2), 161–178.
- Bicchieri, C. & Dimant, E. (2019). Nudging with care: The risks and benefits of social information. *Public Choice*, 191(2022), 1–22.
- Bicchieri, C., Dimant, E., Gächter, S., & Nosenzo, D. (2022a). Social proximity and the erosion of norm compliance. *Games and Economic Behavior*, 132(2022), 59–72.
- Bicchieri, C., Dimant, E., Gelfand, M., & Sonderegger, S. (2022b). Social norms and behavior change: The interdisciplinary research frontier. *Journal of Economic Behavior & Organization*, 205(2022), A4–A7.
- Bicchieri, C. & Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, 22(2), 191–208.

- Binmore, K. & Samuelson, L. (1994). An economist's perspective on the evolution of norms. *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, (pp. 45–63).
- Bowles, S. & Gintis, H. (2013). *A Cooperative Species: Human reciprocity and its evolution*. Princeton University Press.
- Bursztyn, L., Egorov, G., & Fiorin, S. (2020). From extreme to mainstream: The erosion of social norms. *American Economic Review*, 110(11), 3522–48.
- Callander, S. & Carbajal, J. C. (2022). Cause and effect in political polarization: A dynamic analysis. *Journal of Political Economy*, 130(4), 825–880.
- Capraro, V., Jagfeld, G., Klein, R., Mul, M., & de Pol, I. v. (2019). Increasing altruistic and cooperative behaviour with simple moral nudges. *Scientific Reports*, 9(1), 1–11.
- Charness, G., Samek, A., & van de Ven, J. (2022). What is considered deception in experimental economics? *Experimental Economics*, 25(2), 385–412.
- Chaudhuri, A., Paichayontvijit, T., et al. (2006). Conditional cooperation and voluntary contributions to a public good. *Economics Bulletin*, 3(8), 1–14.
- Chaudhuri, A., Paichayontvijit, T., & Smith, A. (2017). Belief heterogeneity and contributions decay among conditional cooperators in public goods games. *Journal of Economic Psychology*, 58(2017), 15–30.
- Chen, Y., Harper, F. M., Konstan, J., & Li, S. X. (2010). Social comparisons and contributions to online communities: A field experiment on movielens. *American Economic Review*, 100(4), 1358–98.
- Chetty, R. (2006). A new method of estimating risk aversion. *American Economic Review*, 96(5), 1821–1834.
- Cheung, S. L. (2014). New insights into conditional cooperation and punishment from a strategy method experiment. *Experimental Economics*, 17(1), 129–153.
- Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6), 1015–1026.
- Cialdini, R. B. & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology* (pp. 151–192). New York: McGraw-Hill.
- Ciranka, S. & van den Bos, W. (2020). A bayesian model of social influence under risk and uncertainty. *PsyArXiv*, 26 May 2020. Web.
- Croson, R. T. (2007). Theories of commitment, altruism and reciprocity: Evidence from linear public goods games. *Economic Inquiry*, 45(2), 199–216.

- d'Adda, G., Dufwenberg, M., Passarelli, F., & Tabellini, G. (2020). Social norms with private values: Theory and experiments. *Games and Economic Behavior*, 124(2020), 288–304.
- DeGroot, M. H. (2005). *Optimal statistical decisions*. John Wiley & Sons.
- Dimant, E. (2019). Contagion of pro-and anti-social behavior among peers and the role of social proximity. *Journal of Economic Psychology*, 73, 66–88.
- Dimant, E. (2022). Distributions matter: Measuring the tightness and looseness of social norms. *Working Paper, available at SSRN 4107802*.
- Dimant, E. (2023). Hate trumps love: The impact of political polarization on social preferences. *Management Science*, forthcoming.
- Dimant, E., Clemente, E. G., Pieper, D., Dreber, A., & Gelfand, M. (2022). Politicizing mask-wearing: predicting the success of behavioral interventions among republicans and democrats in the us. *Scientific Reports*, 12(1), 1–12.
- Dimant, E. & Shalvi, S. (2022). Meta-nudging honesty: Past, present, and future of the research frontier. *Current Opinion in Psychology*, 47(2022), 101426.
- Elster, A. & Gelfand, M. J. (2021). When guiding principles do not guide: The moderating effects of cultural tightness on value-behavior links. *Journal of Personality*, 89(2), 325–337.
- Enke, B. (2020). Moral values and voting. *Journal of Political Economy*, 128(10), 3679–3729.
- Fehr, E. & Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 14(3), 159–181.
- Fehr, E. & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3), 817–868.
- Fehr, E. & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, 2(7), 458–468.
- Feldhaus, C., Sobotta, T., & Werner, P. (2019). Norm uncertainty and voluntary payments in the field. *Management Science*, 65(4), 1855–1866.
- Fischbacher, U. & Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, 100(1), 541–56.
- Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, 71(3), 397–404.
- Fisman, R. & Miguel, E. (2007). Corruption, norms, and legal enforcement: Evidence from diplomatic parking tickets. *Journal of Political Economy*, 115(6), 1020–1048.
- Fosgaard, T. R., Gårn Hansen, L., & Wengström, E. (2020). Norm compliance in an uncertain world. *IFRO Working Paper*.

- Frey, B. S. & Meier, S. (2004). Social comparisons and pro-social behavior: Testing "conditional cooperation" in a field experiment. *American Economic Review*, 94(5), 1717–1722.
- Gächter, S., Gerhards, L., & Nosenzo, D. (2017a). The importance of peers for compliance with norms of fair sharing. *European Economic Review*, 97(2017), 72–86.
- Gächter, S., Herrmann, B., & Thöni, C. (2010). Culture and cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1553), 2651–2661.
- Gächter, S., Kölle, F., & Quercia, S. (2017b). Reciprocity and the tragedies of maintaining and providing the commons. *Nature Human Behaviour*, 1(9), 650–656.
- Gelfand, M. J. (2021). Cultural evolutionary mismatches in response to collective threat. *Current Directions in Psychological Science*, 30(5), 401–409.
- Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliah, A., Ang, S., Arnadottir, J., et al. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, 332(6033), 1100–1104.
- Goldstein, D. G. & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment & Decision Making*, 9(1), 1–14.
- Grout, P. A., Mitraillie, S., & Sonderegger, S. (2015). The costs and benefits of coordinating with a different group. *Journal of Economic Theory*, 160(2015), 517–535.
- Harrison, G. W., Martínez-Correa, J., Swarthout, J. T., & Ulm, E. R. (2017). Scoring rules for subjective probability distributions. *Journal of Economic Behavior & Organization*, 134(2017), 430–448.
- Hartig, B., Irlenbusch, B., & Kölle, F. (2015). Conditioning on what? heterogeneous contributions and conditional cooperation. *Journal of Behavioral and Experimental Economics*, 55(2015), 48–64.
- Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001a). In search of homo economicus: behavioral experiments in 15 small-scale societies. *American Economic Review*, 91(2), 73–78.
- Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., et al. (2001b). Cooperation, reciprocity and punishment in fifteen small-scale societies. *American Economic Review*, 91(2), 73–78.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3), 61–83.
- Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, 22, 129–146.

- Iyengar, S. & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, 59(3), 690–707.
- Kandel, E. & Lazear, E. P. (1992). Peer pressure and partnerships. *Journal of Political Economy*, 100(4), 801–817.
- Kerr, N. L., Rumble, A. C., Park, E. S., Ouwerkerk, J. W., Parks, C. D., Gallucci, M., & Van Lange, P. A. (2009). “how many bad apples does it take to spoil the whole barrel?”: Social exclusion and toleration for bad apples. *Journal of Experimental Social Psychology*, 45(4), 603–613.
- Kimbrough, E. O., Krupka, E. L., Kumar, R., Murray, J., Ramalingam, A., Sánchez-Franco, S., Sarmiento, O. L., Kee, F., & Hunter, R. (2022). On the stability of norms and norm-following propensity: A cross-cultural panel study with adolescents. *Working paper, available at SSRN 4025407*.
- Kimbrough, E. O. & Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, 14(3), 608–638.
- Kölle, F. & Quercia, S. (2021). The influence of empirical and normative expectations on cooperation. *Journal of Economic Behavior & Organization*, 190(2021), 691–703.
- Krupka, E. & Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology*, 30(3), 307–320.
- Krupka, E. L. & 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.
- Lau, H.-S., Lau, A. H.-L., & Ho, C.-J. (1998). Improved moment-estimation formulas using more than three subjective fractiles. *Management Science*, 44(3), 346–351.
- Ledyard, J. O. (1995). Public goods: A survey of experimental research. In A. Roth & J. Kagel (Eds.), *Handbook of Experimental Economics*. Princeton University Press.
- Machina, M. & Pratt, J. (1997). Increasing risk: some direct constructions. *Journal of Risk and Uncertainty*, 14(2), 103–127.
- McConnell, C., Margalit, Y., Malhotra, N., & Levendusky, M. (2018). The economic consequences of partisanship in a polarized era. *American Journal of Political Science*, 62(1), 5–18.
- Michaeli, M. & Spiro, D. (2015). Norm conformity across societies. *Journal of Public Economics*, 132(2015), 51–65.
- Michaeli, M. & Spiro, D. (2017). From peer pressure to biased norms. *American Economic Journal: Microeconomics*, 9(1), 152–216.
- Murphy, A. H. & Winkler, R. L. (1970). Scoring rules in probability assessment and evaluation. *Acta Psychologica*, 34, 273–286.

- Nunn, N. (2022). On the dynamics of human behavior: The past, present, and future of culture, conflict, and cooperation. *AEA Papers and Proceedings*, 112(2022), 15–37.
- Perez-Truglia, R. & Cruces, G. (2017). Partisan interactions: Evidence from a field experiment in the united states. *Journal of Political Economy*, 125(4), 1208–1243.
- Qualtrics (2005). Qualtrics software, Version [January 2021] of Qualtrics. Copyright ©[2021]. Qualtrics, Provo, Utha, USA. Available at: <https://www.qualtrics.com>.
- Quentin, A. (2016). distBuilder v 1.1.0 (1.1.0). Zenodo. <https://doi.org/10.5281/zenodo.166736>.
- Ren, Z. B., Dimant, E., & Schweitzer, M. (2023). Beyond belief: How social engagement motives influence the spread of conspiracy theories. *Journal of Experimental Social Psychology*, 104(2023), 104421.
- Robbett, A. & Matthews, P. H. (2021). Polarization and group cooperation. *Working paper, available at SSRN 3770834*.
- Ross, S. M., Kelly, J. J., Sullivan, R. J., Perry, W. J., Mercer, D., Davis, R. M., Washburn, T. D., Sager, E. V., Boyce, J. B., & Bristow, V. L. (1996). *Stochastic Processes*, volume 2. Wiley New York.
- Schlag, K. H. & van der Weele, J. J. (2015). A method to elicit beliefs as most likely intervals. *Judgment and Decision Making*, 10(5), 456–468.
- Shang, J. & Croson, R. (2009). A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods. *The Economic Journal*, 119(540), 1422–1439.
- Thöni, C. & Volk, S. (2018). Conditional cooperation: Review and refinement. *Economics Letters*, 171(2018), 37–40.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society*, 26(1), 24–36.
- Wolff, I. (2017). What are the equilibria in public-good experiments? *Economics Letters*, 150(2017), 83–85.
- Young, H. P. (2015). The evolution of social norms. *Annual Review of Economics*, 7(1), 359–387.
- Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, 13(1), 75–98.

APPENDICES

A Appendix for Chapter 1

A.1 Additional analysis

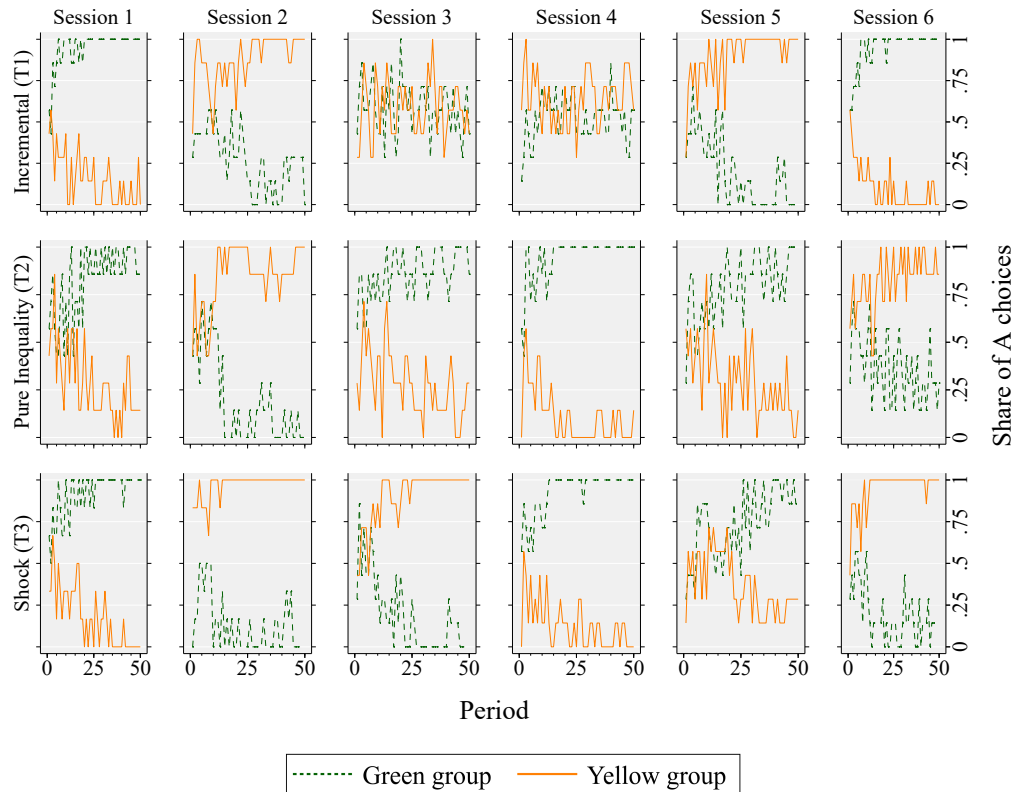
All subsequent analyses are referenced in the document. They are structured by section within the chapter. Brief descriptions are provided for each analysis.

A.1.1 Emergence of an unequal convention

Emergence per session

Over the first 50 periods, the majority of sessions in treatments with low inequality (T1–T3) coordinate on an unequal convention. This is reflected in a significant and positive increase of SSI over time, as shown in the main text. By looking at the individual sessions, we can confirm that stability is not achieved by taking turns between actions but that we observe the emergence of the same repeated pure NE. Figure A.1 shows the share of A choices over time for the yellow and the green group, confirming that in most sessions groups specialise on different actions. Only in two sessions play concurs with a mixed NE, which predicts for

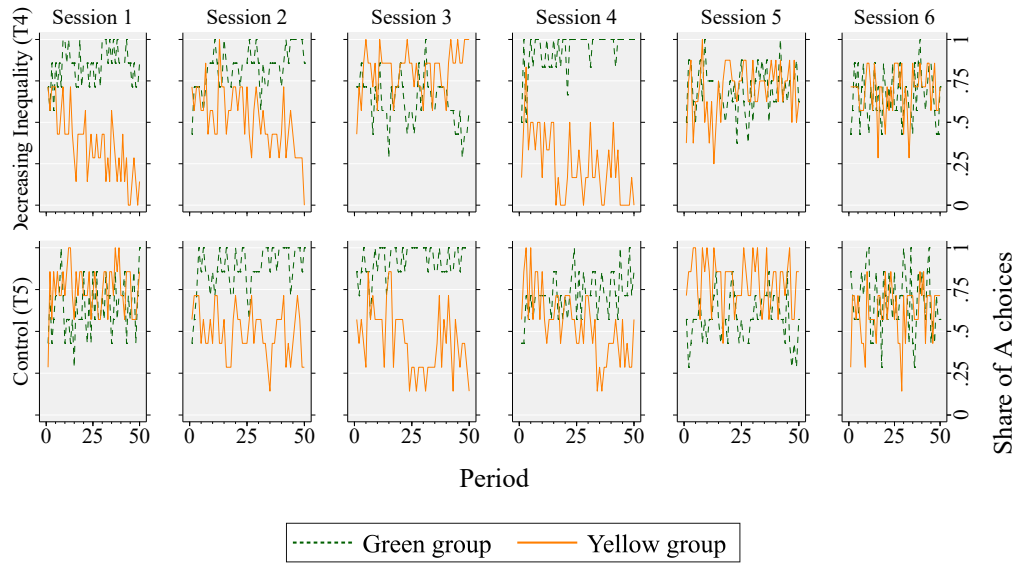
Figure A.1: Emergence under low inequality per session



$\Delta = 25$ that participants choose A in 60% of periods.

Under high inequality (T4, T5), by contrast, the emergence of a pure NE is less common. Figure A.2 shows that in T4 and T5 there is much more chaos with groups not achieving to coordinate on different actions in several sessions. Moreover, even if there is a tendency for groups to specialise on different actions, it is much less pronounced than under low inequality.

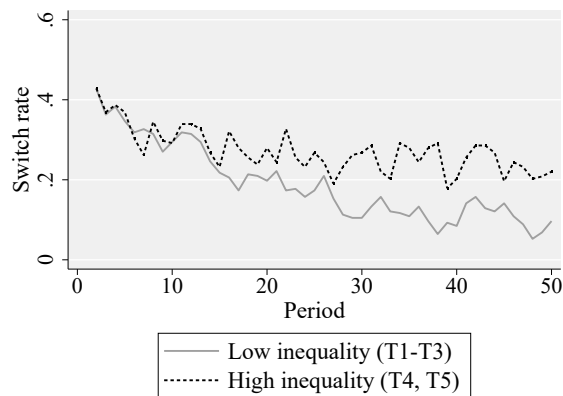
Figure A.2: Emergence under high inequality per session



Individual choices — reduction in switches over time

The emergence of an unequal convention is also reflected in a reduction of switches between actions A and B over the first 50 periods. Figure A.3 shows that the number of individuals who switch decreases during the first 50 periods of the experiment. Again, this pattern is more pronounced under low inequality (T1-T3) than under high inequality (T4, T5).

Figure A.3: Decrease in switches between A and B over the first 50 periods



A.1.2 Does inequality threaten stability?

Stability before and after $t=50$ — robustness checks

In the main text, we exclude the first 20 periods from the analysis of how stability changes after $t = 50$. The reason for this is that we want to test how an established convention is affected by increases in inequality. However, in the first 20 periods of the game, stability is still relatively low (average $SSI = 0.37$), indicating that a convention has not yet emerged. Using all periods before $t = 50$ can thus confound results, as it includes initial periods of miscoordination. Table A.1 shows that if we only exclude the first 10 periods as an initial phase of emergence, where the average $SSI = 0.29$, all coefficients keep the same sign and all but the coefficient for the *Pure Inequality* (T2) treatment remain significant. If we do not exclude any period, again all signs remain the same, but now the *Shock* (T3) treatment also loses its significance (see Table A.2).

Table A.1: Changes in stability after $t = 50$ — excluding first 10 periods

	(T1) Incremental	(T2) Shock	(T3) Pure inequality	(T4) Decreasing inequality	(T5) Control
After $t=50$	-0.18*** (0.06)	-0.10 (0.09)	-0.18** (0.09)	0.23*** (0.07)	0.12** (0.05)
Constant	0.61*** (0.13)	0.69*** (0.06)	0.81*** (0.07)	0.38*** (0.09)	0.31*** (0.05)
N observations	2700				
N clusters	30				
N periods	90				
R^2	0.19				

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the session level.

Note: Results of a single random effects panel regression. The dependent variable is the SSI in a given session and period and varies between 0 and 1. *After $t = 50$* is a dummy variable that takes the value 0 for the periods before $t = 50$ and 1 for periods 51–100. The regression controls for treatment indicators and their interaction with *After $t = 50$* . Each column shows the results for a different treatment as the base category. Treatment differences are omitted from the table for conciseness. The first 10 periods are excluded, as it takes time for the convention to emerge and the first periods still exhibit a large degree of switching behaviour (average $SSI = 0.29$).

Table A.2: Changes in stability after $t = 50$ — full sample

	(T1) Incremental	(T2) Shock	(T3) Pure inequality	(T4) Decreasing inequality	(T5) Control
After $t=50$	-0.13*** (0.04)	-0.02 (0.09)	-0.09 (0.09)	0.26*** (0.07)	0.13** (0.06)
Constant	0.56*** (0.11)	0.61*** (0.05)	0.72*** (0.06)	0.35*** (0.08)	0.30*** (0.04)
N observations	3000				
N clusters	30				
N periods	100				
R^2	0.16				

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the session level.

Note: Results of a single random effects panel regression. The dependent variable is the SSI in a given session and period and varies between 0 and 1. *After $t = 50$* is a dummy variable that takes the value 0 for the periods before $t = 50$ and 1 for periods 51–100. The regression controls for treatment indicators and their interaction with *After $t = 50$* . Each column shows the results for a different treatment as the base category. Treatment differences are omitted from the table for conciseness.

Probability of choosing A before and after $t=50$

The change in behaviour between the first and the second part of the experiment also becomes clear when looking at the individual choice data. As the SSI is based on the difference in A choices between the yellow and green group, an increase in the probability of choosing A for each individual is equivalent to a decrease in stability. Table A.3 shows the results of a single random effects panel model, taking a binary variable that takes the value 1 if an individual chooses A and 0 otherwise as the dependent variable and the *After t = 50* dummy as the explanatory variable of interest. As in the analysis of the SSI, the regression includes treatment indicators and their interaction with *After t = 50* as controls. Each column in Table A.3 shows the results for taking a different treatment as the base category to present the effect for each treatment. As can be seen from Table A.3, the individual probability of choosing A differs significantly between the first and the second 50 periods for all treatments. In all three treatments where inequality increases (T1–T3), we see a significant increase in the probability of choosing A after $t = 50$. This shows again that stability is negatively affected by increasing inequality. In treatments where inequality is not increasing (T4, T5), by contrast, the probability of choosing A decreases after $t = 50$, indicating an increase in stability.

Table A.3: Changes in the probability of choosing A after $t = 50$

	(T1) Incremental	(T2) Shock	(T3) Pure inequality	(T4) Decreasing inequality	(T5) Control
After $t=50$	0.12*** (0.01)	0.07*** (0.02)	0.07*** (0.02)	-0.11*** (0.01)	-0.04** (0.02)
Constant	0.38*** (0.10)	0.37*** (0.10)	0.38*** (0.09)	0.48*** (0.11)	0.50*** (0.10)
N observations	41600				
N subjects	416				
N clusters	30				
N periods	100				
R ²	0.01				

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the session level.

Note: Results of a single random effects panel regression. The dependent variable is a binary variable that takes the value of 0 if an individual chooses action B and 1 if they choose action A. *After t = 50* is a dummy variable that takes the value 0 for the first 50 periods of the experiment and 1 for periods 51–100. The regression controls for treatment indicators and their interaction with *After t = 50*. Each column shows the results for a different treatment as the base category. Treatment differences are omitted from the table for conciseness. The regression controls for group size.

Deviations from an established convention

Table A.4 explores whether certain individual characteristics predict the probability to initialise deviations from an established convention. As deviations happen mainly under increasing inequality, the analysis is restricted to T1–T3. Initialising a deviation is thereby defined as choosing an action that goes against the established convention, while the majority of the own group is still following it. An individual is thus classified as an initiator if the majority of their group still chooses B (A), but they decide to choose A (B). As the first 50 periods are treated as the emergence stage of a convention after which one group specialises in action A and the other in action B, the analysis is restricted to periods larger than 50.

Table A.4 shows the results of random effects panel regressions with a binary variable that takes the value 1 if an individual deviates and 0 otherwise as the dependent variable and different individual characteristics as the explanatory variables. Column (1) reports results for a pooled regression across T1–T3, while columns (2) – (4) show results by treatment. The results show that the probability of deviating is lower for higher levels of stability (higher SSI). This is intuitive, as in these cases the risk of receiving zero is particularly high. Moreover, individuals who are more conscientious and report a higher negative affect are more likely to deviate. There is also some evidence that individuals with a lower level of risk aversion are more likely to deviate (only significant in T1).

Table A.4: Probability of deviating from a convention ($t > 50$)

	(T1-T3) Pooled	(T1) Incremental	(T2) Pure inequality	(T3) Shock
SSI	-0.411*** (0.014)	-0.407*** (0.019)	-0.440*** (0.024)	-0.405*** (0.019)
Female	0.011 (0.021)	-0.023 (0.028)	0.016 (0.066)	-0.011 (0.027)
Age	-0.000 (0.002)	0.002 (0.003)	0.001 (0.008)	-0.006* (0.003)
Prosocial	-0.010 (0.017)	-0.028 (0.033)	0.013 (0.042)	0.018 (0.031)
Risk aversion	-0.006 (0.004)	-0.019** (0.007)	-0.001 (0.010)	-0.008 (0.009)
<i>Big 5</i>				
Openness	0.009 (0.009)	0.031 (0.021)	-0.003 (0.017)	-0.011 (0.018)
Conscientiousness	0.016** (0.008)	-0.005 (0.015)	0.020 (0.015)	0.036*** (0.008)
Extroversion	0.008 (0.007)	0.006 (0.017)	0.029** (0.015)	-0.002 (0.015)
Agreeableness	-0.009 (0.010)	-0.002 (0.008)	0.016 (0.024)	-0.037* (0.022)
Neuroticism	-0.004 (0.006)	-0.004 (0.011)	-0.005 (0.011)	0.015 (0.010)
Negative affect	0.002*** (0.000)	0.003*** (0.001)	0.003*** (0.001)	0.000 (0.001)
Constant	0.193* (0.111)	0.266 (0.182)	0.085 (0.296)	0.479* (0.265)
N observations	20800	20800	20800	20800
N subjects	416	416	416	416
N clusters	30	30	30	30
N periods	50	50	50	50
R ²	0.12	0.13	0.13	0.13

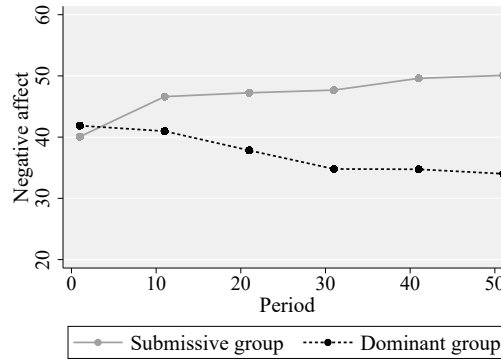
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the session level.

Note: Results of random effects panel regressions for $t > 50$. The dependent variable is a binary variable that takes the value of 1 if an individual chooses action A despite the majority of their group choosing B or vice versa and 0 otherwise. Analysis is restricted to periods larger than 50, as the first 50 periods are treated as the emergence stage of a convention. Regressions control for group size.

Affect responses

Figure A.4 shows how negative affect develops over time across dominant and submissive groups. With growing stability but also growing inequality in earnings between groups, we observe that a gap in affect emerges. Dominant groups show more positive emotional affect over time, while submissive groups show more negative affect, leading to a significant difference over the first 50 periods (t-test, $p < 0.001$).

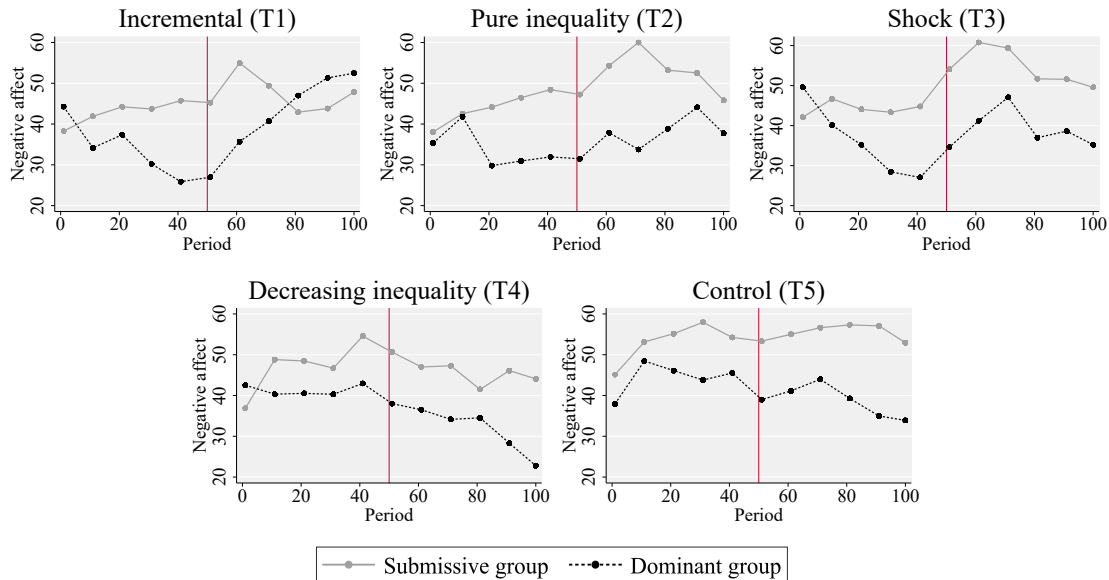
Figure A.4: Affect responses for $t \leq 50$



Note: Negative affect is measured on a scale from 0 to 100, with higher values indicating more negative emotions. Negative affect is measured every tenth round.

Figure A.5 shows how negative affect develops for dominant and submissive groups over the whole experiment for the different treatments. We can see that overall the gap between dominant and submissive groups persists throughout the experiment. Moreover, affect depends on the level of inequality. An increase in inequality (T1, T2, T3) initially leads to an

Figure A.5: Affect responses across treatments



increase in negative affect, while a decrease in inequality (T4) results in more positive affect responses.²³

A.1.3 Treatment differences

Coordination under high inequality given different histories

In both the *Shock* (T3) and the *Control* (T5) treatment, participants face high inequality ($\Delta = 75$) for $t > 50$. The difference is however that in T3, participants experienced high stability in the first 50 periods while participants in the *Control* (T5) treatment did not. In Table A.5, we regress the SSI as the dependent variable on a time trend, treatment indicators, and their interaction. We thereby focus on the periods 51-100, where inequality is identical in both T3 and T5. While the *Control* (T5) treatment does not show a statistically different SSI at $t = 51$ from the *Shock* (T3) treatment, we can see that the increase in stability is significantly lower. This shows that past experiences of instability can undermine coordination even in environments with low inequality. The regression results are robust to the inclusion of controls in column (2) and a fixed effects specification in column (3).

Table A.5: The effect of different histories on coordination under high inequality ($\Delta = 75$)

	(1)	(2)	(3)
<i>Baseline = Shock (T2)</i>			
Control (T5)	0.385 (0.236)	0.248 (0.296)	
Time trend	0.010*** (0.003)	0.010*** (0.003)	0.010*** (0.003)
Control (T5) x time trend	-0.008** (0.004)	-0.008** (0.004)	-0.008** (0.004)
Constant	-0.102 (0.213)	-0.287 (0.270)	0.364*** (0.088)
Demographic controls	No	Yes	FE
N observations	1500	1500	1500
N clusters	30	30	30
N periods	50	50	50
R ²	0.13	0.33	0.05

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the session level.

Note: Results of three panel regressions for $t > 50$. The dependent variable is the SSI in a given session and period and varies between 0 and 1. Demographic controls include the difference in the share of female participants as well as in average measures of Big 5, risk aversion, prosociality and age between the groups in a session. All regressions include treatment indicators for T1, T2, and T4. As the relevant comparison is between T3 and T5, they are omitted from the table. Columns (1) and (2) are random effects models, while column (3) reports results of a fixed effects estimation.

²³At later periods, we observe more positive responses in T2 and T3 and even a closure of the gap between submissive and dominant groups in T1. This has to do with the specific dynamics of each session and the existence of reversals (see Section 5.3.4).

Coordination under low inequality given different histories

Table A.6 regresses the SSI in a given session and period on treatment indicators, a time trend, and their interaction. Our comparison of interest is between the *Decreasing Inequality* (T4) treatment for $t > 50$ and the treatments with low initial inequality (T1–T3) for $t \leq 50$. For these intervals, the treatments share low levels of inequality ($\Delta = 25$), but T4 has a history of instability that is absent in the other treatments. Column (1) in Table A.6 compares the SSI by treatment, taking T4 as the baseline. While for all treatments stability increases over time, we see that the time trend is stronger in T1–T3 (even though not statistically significant for T1). In columns (2) – (4), we pool all three treatments with initially low inequality. Again the latter show a significantly more positive time trend than T4. This result is robust to the inclusion of controls in column (3) and a fixed effects in column (4).

Table A.6: The effect of different histories on coordination under low inequality ($\Delta = 25$)

	By treatment	T1–T3 pooled		
	(1)	(2)	(3)	(4)
<i>Baseline = Decreasing inequality (T4)</i>				
Incremental (T1)	-0.124 (0.147)			
Shock (T3)	-0.091 (0.158)			
Pure inequality (T2)	-0.182 (0.146)			
<i>Time trend</i>	0.005*** (0.002)			
Incremental (T1) x time trend	0.003 (0.003)			
Shock (T3) x time trend	0.008*** (0.002)			
Pure inequality (T2) x time trend	0.007*** (0.002)			
Low inequality (T1–T3)		-0.132 (0.140)	-0.111 (0.142)	
<i>Time trend</i>		0.005*** (0.002)	0.005*** (0.002)	0.005*** (0.002)
Low inequality (T1–T3) x time trend		0.006*** (0.002)	0.006*** (0.002)	0.006*** (0.002)
Constant	0.488*** (0.135)	0.488*** (0.135)	0.388* (0.216)	0.367*** (0.033)
Demographic controls	No	No	Yes	FE
N observations	1500	1500	1500	1500
N clusters	30	30	30	30
N periods	50	50	50	50
R ²	0.23	0.20	0.24	0.16

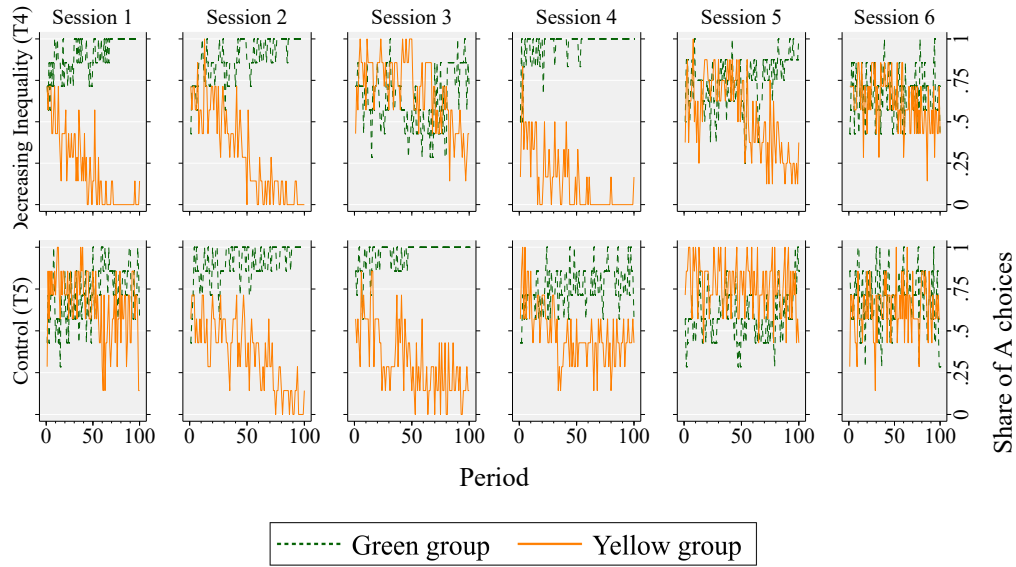
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the session level.

Note: Results of four panel regressions. The dependent variable is the SSI in a given session and period and varies between 0 and 1. All regressions include a treatment indicator for T5. As the relevant comparison is between T4 and T1–T3, it is omitted from the table. Demographic controls include the difference in the share of female participants as well as in average measures of Big 5, risk aversion, prosociality and age between the groups interacting with each other in a session. Column (1), (2) and (3) are random effects models, while column (4) reports results of a fixed effects estimation.

Behaviour in treatments with non-increasing inequality

Figure A.6 shows the share of A choices for yellow and green groups by session for treatments with non-increasing inequality (T4, T5). In particular, in the *Decreasing Inequality* (T4) treatment — but also to a lesser extent in the *Control* (T5) treatment — average stability is larger for $t > 50$ than in the first 50 periods, as indicated by the increasing coordination on different actions between groups. As inequality is either decreasing or stable, it is not surprising that we do not see any attempts at reversals in these treatments.

Figure A.6: Behaviour in treatments with non-increasing inequality (by session)



A.2 Theory revisited — simulation results

Simulations for a model where θ_i decreases with H and L

Figure A.7: Simulation of the *Incremental* (T1) treatment if θ_i decreases with H and L

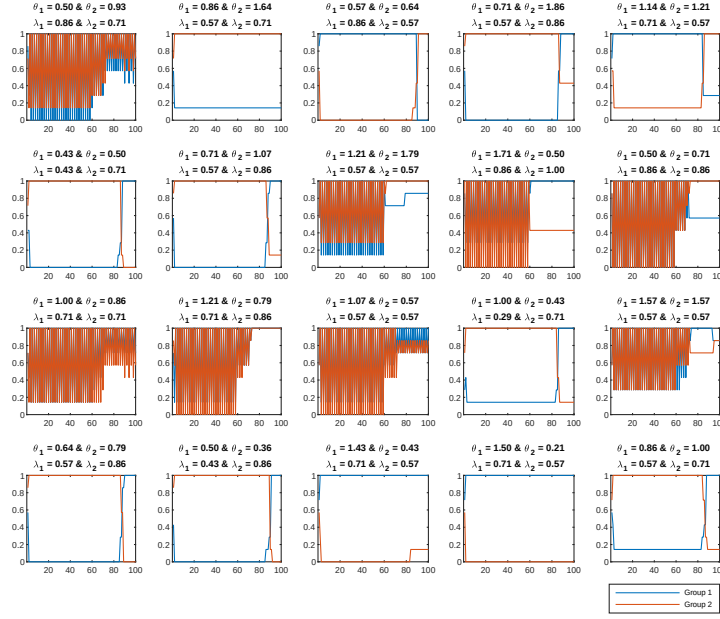


Figure A.8: Simulation of the *Pure Inequality* (T2) treatment if θ_i decreases with H and L

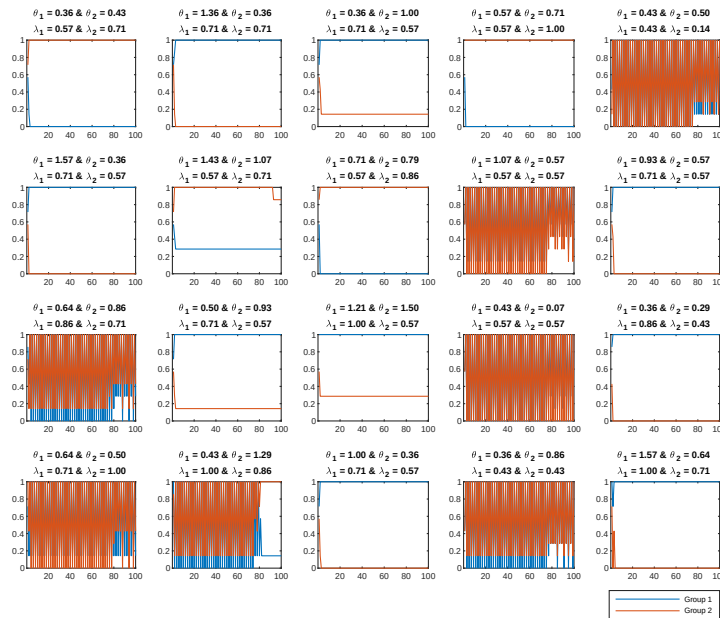
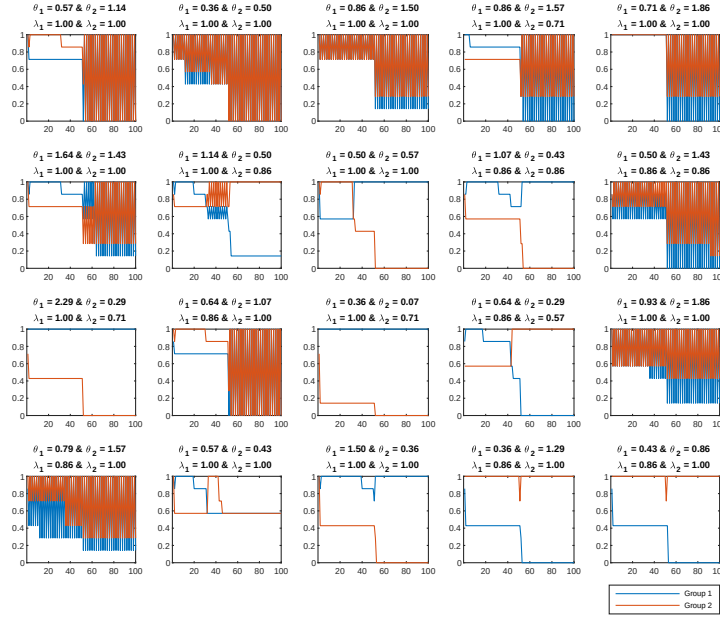
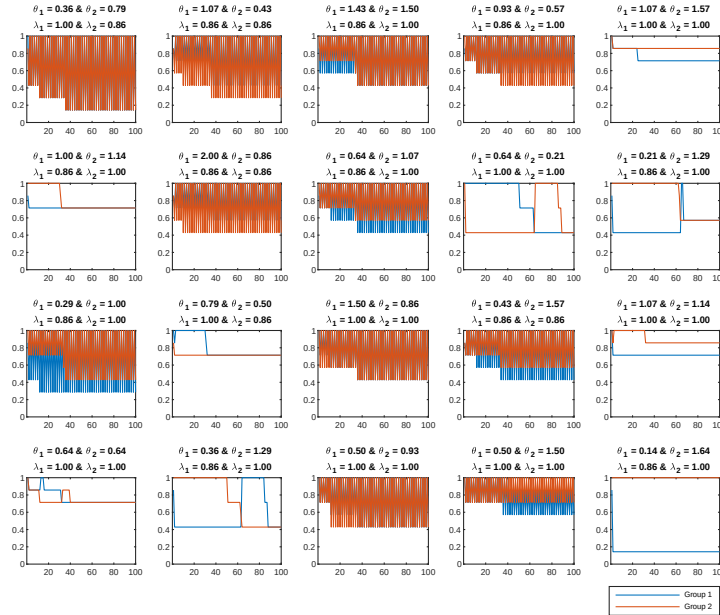


Figure A.9: Simulation of the *Decreasing Inequality* (T4) treatment if θ_i decreases with H and LFigure A.10: Simulation of the *Control* (T5) treatment if θ_i decreases with H and L

Simulations for a model where θ_i decreases with Z

Figure A.11: Simulation of the *Incremental* (T1) treatment if θ_i decreases with Z

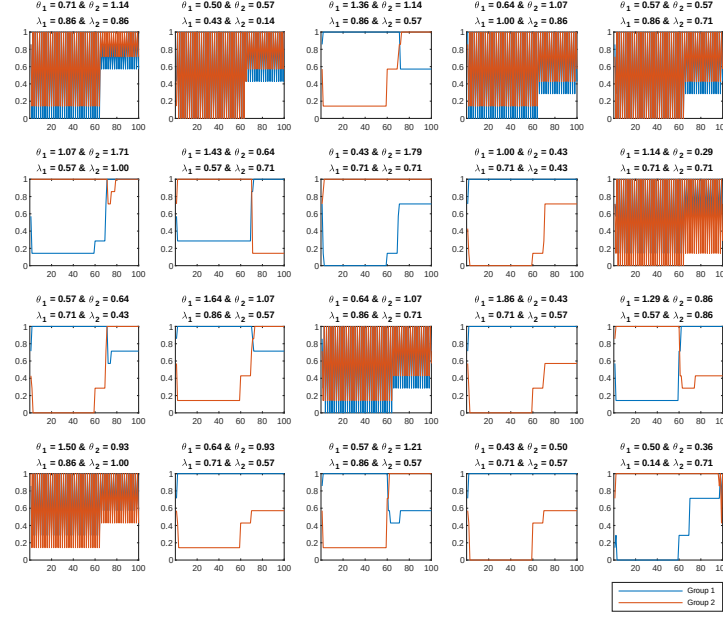


Figure A.12: Simulation of the *Pure Inequality* (T2) treatment if θ_i decreases with Z

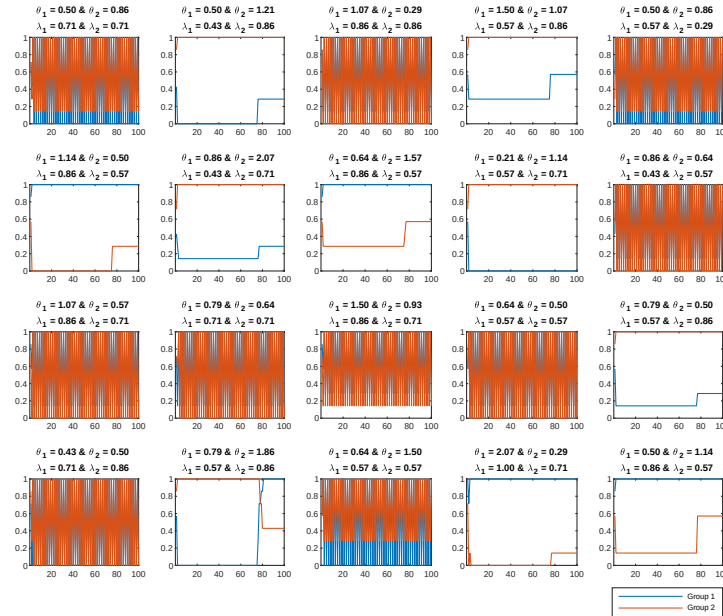
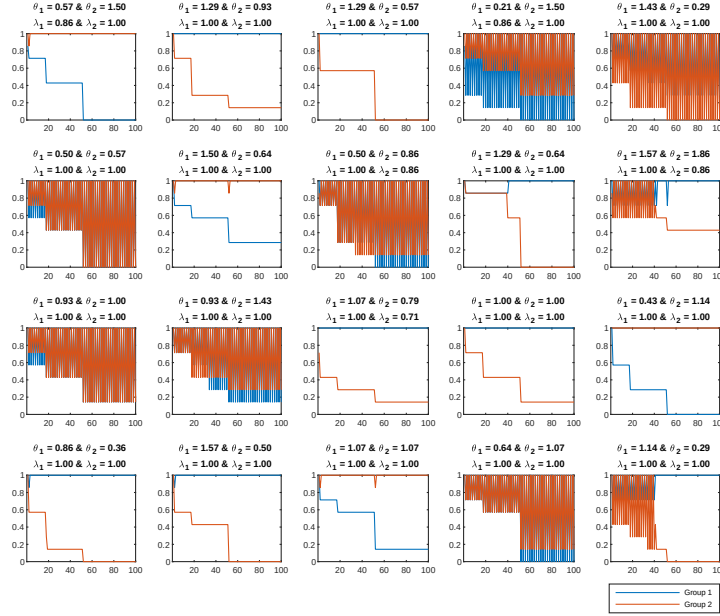
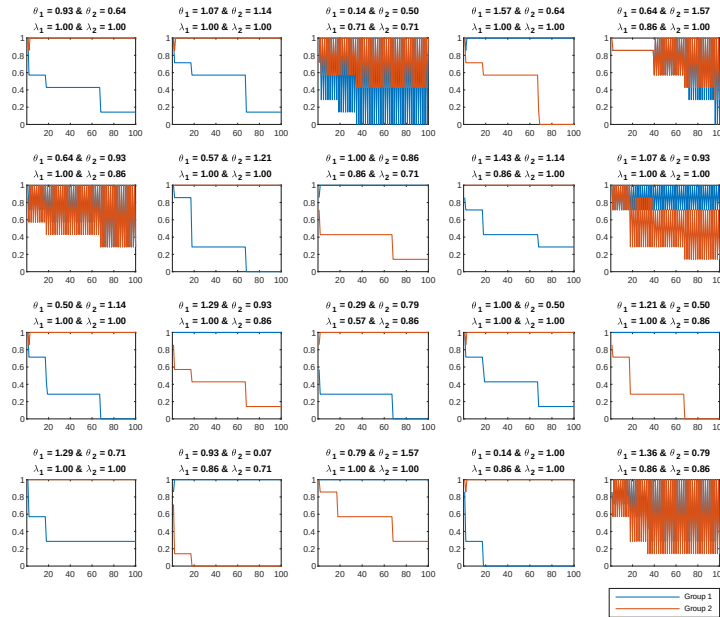


Figure A.13: Simulation of the *Decreasing Inequality* (T4) treatment if θ_i decreases with ZFigure A.14: Simulation of the *Control* (T5) treatment if θ_i decreases with Z

A.3 Instructions

Main experiment

All participants received a hard-copy of the instructions for the main experiment at the beginning of the session. The following text shows the instructions for *Incremental*, *Pure Inequality* and *Shock* treatment (T1 – T3). Instructions for *Control* and *Decreasing Inequality* treatment (T4, T5) use different payoffs and are presented in brackets.

General instructions for participants

We warmly welcome you to this experimental study.

Please read the following instructions carefully. During this experiment, depending on your decisions and those of other participants, you can earn some money over and above your show-up fee of £3. It is very important that you read all the instructions carefully, so that you understand the potential consequences of your decisions. If you have any questions, please raise your hand and an experimenter will come to you. During the experiment, please do not try to communicate with any of the other participants and please do not use mobile phones. If you do not follow these rules, you will be excluded from the study and will not be paid. Below, we describe the experiment you are going to participate in during today's session. The anonymity of all the decisions you make during the experiment is guaranteed.

Detailed information about the study

Earnings

During the experiment, you will earn points. Then, at the end of the experiment, your total points will be converted into pounds using the following conversion rate: **100 points = £0.15** The resulting amount plus your show-up fee of £3 will be given to you in cash at the end of today's session.

Allocation into groups

At the beginning of the experiment, **you and each of the other participants in the session will be randomly assigned to a group** by the draw of a coloured ball (with probability 0.5). Half of the participants will be assigned to a **green** group, the other half to a **yellow** group. **Each participant will stay in the same group for the whole experiment.**

The decision situation

In this experiment you will play a game 100 times. We will refer to each time you play as a round. The game is played in pairs. **In each round, you will play the game with someone**

randomly selected from a differently coloured group. New random selections will be made for each round. So, if you are in a **green** group, you will be randomly and newly paired with a player from a **yellow** group in each round. And, if you are in a **yellow** group, you will be randomly and newly paired with a player from a **green** group in each round. **The group you are paired with stays the same** for all 100 rounds. You will never know the identity of the people you play each game with and they will never know yours.

In a round, you and your playing partner for that round each have to choose between **Action A** and **Action B**. The consequence for you of the action that you choose depends also on the action chosen by your partner. To understand exactly what this means, take a look at the **Decision table** below.

Decision table

		Choice of yellow player	
		Action A	Action B
Choice of green player	Action A	0 points	a points
	Action B	b points	0 points

The Decision table can be read as follows:

If the **green** player chooses A and the **yellow** player chooses A, then both players receive 0 points. If the **green** player chooses B and the **yellow** player chooses B, then both players receive 0 points.

If the **green** player chooses A and the **yellow** player chooses B, then the **green** player receives **a** points and the **yellow** player receives **b** points. If the **green** player chooses B and the **yellow** player chooses A, then the **green** player receives **b** points and the **yellow** player receives **a** points. During the experiment the amounts for **a** and **b** will vary, so keep an eye on the decision table when making your choices. At regular intervals we are going to highlight the decision table to help you remember to check it. However, the amounts of **a** and **b** could change at any time. So, you need to quickly check the decision table, before making your choice in each round.

As you can see, the lowest paying situation for the players is if both make the **same choice**,

that is, if both players choose A or both players choose B.

Both players make their decision whether to choose A or B simultaneously. **So, in each round, you will not know your partner's decision before you make your decision.**

Below, is a screenshot of what you will see on your computer in the **first round** of the experiment.²⁴ The screen is set up assuming that you are a member of a **green** group. At the top of the screen is a reminder of your group and the group to which your playing partner for that round belongs. On the left-hand side is the Decision table (same as you saw above). Note that for example in this round **a=75 and b=50** [**a=100 and b=25**]. You indicate your choice for the round in the box in the bottom right hand corner of the screen. In each round you must choose either **Action A** or **Action B**.

The table on the top left reminds you of the possible outcomes of the task. In the box at the bottom right, please enter your choice of action for this round.

Remember that you are a member of a **green** group and your partner is a member of a **yellow** group.

Decision table

Choice of **yellow** player

	Action A	Action B	
Choice of green player (you)	Action A	0 points	50 points
	Action B	75 points	0 points

Make your choice

What action do you want to take in this round?

☐ Action A
☐ Action B

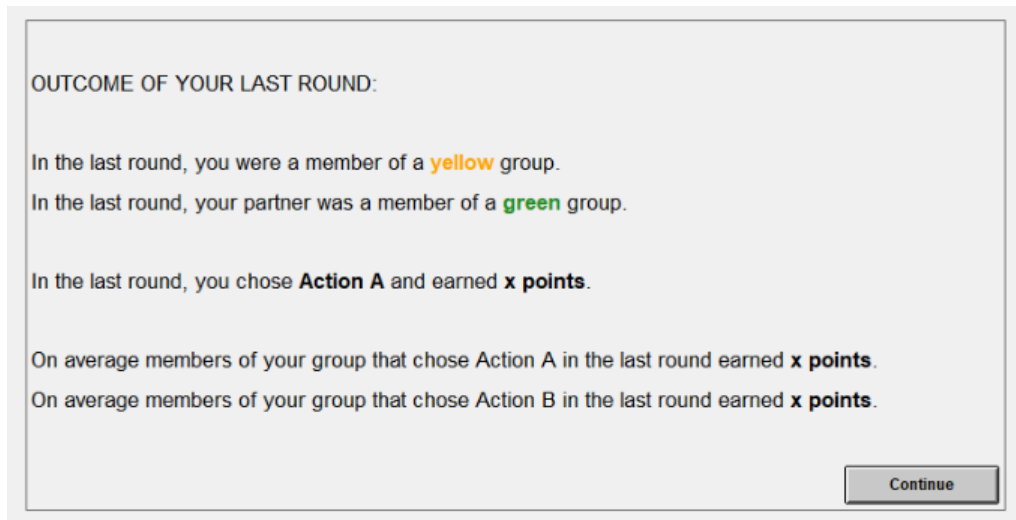
Submit

After each round, before you start playing the next round, you will be given the following information, summarising the last round:

- a reminder of your group affiliation and the group affiliation of the person you played with in the previous round
- the choice you made in the previous round and how many points you earned in the previous round
- how many points players from YOUR group who chose A in the previous round earned on average in that round
- how many points players from YOUR group who chose B in the previous round earned on average in that round.

Below is an **example** for the screen summarizing the last round, assuming that you are a member of a **yellow** group:

²⁴The screenshot for the *Decreasing Inequality* (T4) and the *Control* (T5) treatment shows a=75 and b=50.



Note that where you see “**x points**” in the screenshot above, positive numbers will appear, as appropriate, during the experiment. After round 100 you will receive a summary of the whole experiment. Then, we will ask you to complete a **questionnaire**, which consists of three parts. You will receive more detailed information about the questionnaire after you have completed the experimental task. Finally, you will be paid.

Summary

Participants are randomly assigned to a **yellow** or a **green** group. Participants’ group affiliation remain the same for all 100 rounds.

- In each round you will be newly, randomly matched with a participant from a differently coloured group. You will always be paired with someone from the **same other** group.
- You and your partner each have to choose either A or B and you do this simultaneously.
- If you and your partner both choose A or both choose B, each of you will earn zero. If one of you chooses A and the other B, the one choosing A earns **a** points and the one choosing B earns **b** points.
- The amounts for **a** and **b** can change during the experiment. It is thus important that you quickly check the decision table, before making your choice in each round.

If you have completely understood the instructions, please answer the control questions on screen. If you have any questions please raise your hand and an experimenter will come to you.

Risk elicitation task

The instructions for this part of the experiment were provided on screen (see screenshot below). This is an example of the instructions for a green player. The instructions for a yellow player were adjusted accordingly. The text on top of the screen reads as follows:

Hypothetical games (1 of 6)

Now you are not playing against another participant, but against the computer. Further, as well as making its own choice between A and B, the computer makes your choice as well. The only choice you have is which game to play. You are going to be asked to make six choices. Each time, you will be choosing one out of two games. You are still the **green** player, the computer has taken over the role of the **yellow** player. Look at **Game 1** first. In **Game 1** the likelihood of the computer playing A is **100%** and the likelihood of it playing B is **0%** and the computer has chosen B for you. Then look at **Game 2**. In **Game 2** the likelihood of the computer playing A is **50%** and the likelihood of it playing B is **50%** and the computer has chosen A for you. Please tell us which of the games you would prefer to be played. Once you have chosen your six games, one will be played out for real money.

Hypothetical games (1 of 6)

Now you are not playing against another participant, but against the computer. You are still the **green** player, the computer is taking over the role of the **yellow** player. Look at **Game 1** first. You can see that the computer is playing A with **100%** and B with **0%**. Knowing how the computer decides, please choose which action you want to take in Game 1. Then look at **Game 2** and make your decision taking the probabilities of Game 2 into account. Finally, tell us which of the games you would prefer to play. You will face 6 decisions between two games. One of the games you chose will be played for real money. You told us your action, the computer's action will be determined using the respective probabilities.

Game 1

Choice of computer

	Action A (100%)	Action B (0%)
Choice of green player (you)		
Action A	0 points	75 points
Action B	50 points	0 points

What action do you want to take in **Game 1**?

☐ Action A
☐ Action B

Game 2

Choice of computer

	Action A (50%)	Action B (50%)
Choice of green player (you)		
Action A	0 points	75 points
Action B	50 points	0 points

What action do you want to take in **Game 2**?

☐ Action A
☐ Action B

Make your choice

Which game do you prefer to play?

☐ Game 1
☐ Game 2

Submit

Questionnaire

1) Below you see a number of statements. For each statement, please indicate how much you agree with this. I SEE MYSELF AS SOMEONE WHO...

[Answer options were on a 7 point Likert scale from strongly disagree to strongly agree]

- ...does a thorough job
- ...is communicative, talkative
- ...is sometimes somewhat rude to others
- ...is original, comes up with new ideas
- ...worries a lot
- ...has a forgiving nature
- ...tends to be lazy
- ...is outgoing, sociable
- ...values artistic experiences
- ...gets nervous easily
- ...does things effectively and efficiently
- ...is reserved
- ...is considerate and kind to others
- ...has an active imagination
- ...is relaxed, handles stress well

2) What is your gender? (male/ female/ other)

3) What is your age?

4) Which year of university are you in? (Undergraduate first year/second year/ third year/ fourth year or further/ Master/ PhD/ other)

5) Which faculty do you belong to? (Arts/ Engineering/ Medicine and Health Sciences/ Sciences/ Economics or Business School/ Social Sciences/ None of the above)

6) How many participants of this experiment have you known beforehand?

7) At which round did the payoffs change?

8) How often did you take the decision table into account before making your decision?

9) Did you hear any details about this experiment from other students before participating?
(Yes/ No)

10) Do you have any other comments on this experiment? If yes, you can give us feedback here. If not, just type no into the box.

Thank you very much for participating in this experiment!

Your total payoff from this experiment is £x.

You will receive your payment in a moment. All payments will be rounded up to the next decimal. Please wait until you are called up.

B Appendix for Chapter 2

B.1 Theoretical appendix

B.1.1 Proofs

Proof of Proposition 1 Substituting $m_{it} = \alpha_i x_{it} + (1 - \alpha_i) \frac{X_t}{2}$ into Equation (2) and maximising the expression with respect to α_i yields

$$\max y_{it}^* = \underbrace{\alpha_i x_{it} + (1 - \alpha_i) \frac{X_t}{2}}_{m_{it}} + \frac{X_t}{\beta_i} \text{ w.r.t } \alpha_i. \quad (\text{B.1})$$

Solving Equation (B.1) results in $\frac{\partial y_{it}^*}{\partial \alpha_i} = x_{it} - \frac{X_t}{2}$, implying the following corner solutions:

$$\alpha_i^* = \begin{cases} 1, & \text{if } x_{it} > \frac{X_t}{2}. \\ 0, & \text{if } x_{it} < \frac{X_t}{2}. \end{cases} \quad (\text{B.2})$$

Given differences in relative income, where H refers to a relatively higher income and L to a relatively lower income, $\frac{X_t}{2}$ can be written as $\frac{X_t}{2} = \frac{x_{Ht} + x_{Lt}}{2}$. By definition $x_{Ht} > x_{Lt}$. Consequently, for richer individuals we have $x_{it} > \frac{X_t}{2}$ resulting in $\alpha_i^* = 1$. For poorer individuals we have $x_{it} < \frac{X_t}{2}$ resulting in $\alpha_i^* = 0$. ■

Proof of Proposition 2 Using the result from Proposition 1 that $m_{kt} = x_{kt}$ for richer individuals and substituting it into Equation (5a) results in

$$m_{i2}^* = (1 - \rho_i) x_{i2} + \rho_i \underbrace{[\gamma_i x_{i1} + (1 - \gamma_i)(X_2 - x_{j1})]}_{\text{feasible } m_{i1} \text{ satisfying } X_2 = x_{i1} + x_{j1}}. \quad (\text{B.3})$$

m_{j2}^* can be derived analogously and is therefore omitted from the rest of the proof. Note that X_2 can be written as $x_{i2} + x_{j2}$. Equation (B.3) can be simplified to

$$m_{i2}^* = x_{i2} + \rho_i \gamma_i (x_{i1} - x_{i2}) - \rho_i (1 - \gamma_i) (x_{j1} - x_{j2}). \quad (\text{B.4})$$

Reactions to own shocks ($x_{i1} - x_{i2}$) and the shocks of others ($x_{j1} - x_{j2}$) can be analysed by evaluating the following FOC:

$$\frac{\partial m_{i2}^*}{\partial (x_{i1} - x_{i2})} = \rho_i \gamma_i \geq 0 \text{ and } \frac{\partial m_{i2}^*}{\partial (x_{j1} - x_{j2})} = -\rho_i (1 - \gamma_i) \leq 0 \quad (\text{B.5})$$

This is true as $\rho_i, \gamma_i \in [0, 1]$. If $\rho_i > 0$ and $0 < \gamma_i < 1$ these inequalities become strict. ■

Proof of Proposition 3 Substituting $m_{kt} = \frac{X_t}{2}$ into Equation (5a) yields

$$m_{i2}^* = (1 - \rho_i) \frac{X_2}{2} + \rho_i \underbrace{\left[\gamma_i \frac{X_1}{2} + (1 - \gamma_i) \left(X_2 - \frac{X_1}{2} \right) \right]}_{\text{feasible } m_{i1} \text{ satisfying } X_2 = 2 \frac{X_1}{2}}. \quad (\text{B.6})$$

The fair allocation for the other person m_{j2}^* is again omitted as it is derived analogously. Simplifying Equation (B.6) leads to

$$m_{i2}^* = \frac{X_2}{2} + \rho_i \gamma_i \frac{(X_1 - X_2)}{2} - \rho_i (1 - \gamma_i) \frac{(X_1 - X_2)}{2}. \quad (\text{B.7})$$

Both a shock to self and a shock to other ultimately lead to a reduction in $\frac{(X_t)}{2}$. The FOC is then $\frac{\partial m_{i2}^*}{\partial (X_1/2 - X_2/2)} = \rho_i \gamma_i - \rho_i (1 - \gamma_i)$. m_{i2}^* is thus independent of whether there is an own shock, a shock to other or a shock to both. If $\gamma_i = 0.5$, FOC=0. ■

B.1.2 Model variation: Shares as reference points

In the main text, the fair allocation is expressed in absolute terms. An alternative is to use shares of the total endowment. In this section, I show that the key predictions of the framework are robust to this modification.

In line with Barr et al. (2015), the utility function used in Cappelen et al. (2007) can be slightly adjusted such that it does not depend on absolute income but the share of income that individuals allocate to themselves ($\tilde{y}_{it}$):

$$\tilde{U}_{it} = \tilde{y}_{it} - \beta_i \frac{(\tilde{y}_{it} - m_{it})^2}{2}, \beta_i \geq 0 \quad (\text{B.8})$$

The optimal share allocated to self is then: $\tilde{y}_{it}^* = m_{it} + \frac{1}{\beta_i}$. Similarly, it is possible to express the allocation that is perceived as fair in terms of shares: $m_{it} = \alpha_i \tilde{x}_{it} + (1 - \alpha_i) 1/2$, where $\tilde{x}_{it}$ is the share that individual i contributed to total earnings.

Proposition 1a *Fairness views depend on relative income. For richer individuals $\alpha_i^* = 1$, implying that they hold a contribution-based fairness view ($m_{kt} = \tilde{x}_{kt}$). Poorer individuals, with $\alpha_i^* = 0$ hold an egalitarian fairness view ($m_{kt} = \frac{1}{2}$).*

Proof of Proposition 1a The proof is analogously to the one in B.1.1. Maximising $\tilde{y}_{it}^*$ with respect to α_i

$$\max \tilde{y}_{it}^* = \underbrace{\alpha_i \tilde{x}_{it} + (1 - \alpha_i) 1/2}_{m_{it}} + \frac{1}{\beta_i} \text{ w.r.t } \alpha_i \quad (\text{B.9})$$

results in the following corner solutions

$$\alpha_i^* = \begin{cases} 1, & \text{if } \tilde{x}_{it} > 1/2. \\ 0, & \text{if } \tilde{x}_{it} < 1/2. \end{cases} \quad (\text{B.10})$$

If there exist differences in earnings with $\tilde{x}_{Ht} > \tilde{x}_{Lt}$, then for individuals with higher earnings (H) by definition $\tilde{x}_{it} > 1/2$ and vice versa for individuals with lower earnings (L). Richer individuals are thus predicted to hold contribution-based views ($\alpha_i^* = 1$) and poorer individuals to hold egalitarian views ($\alpha_i^* = 0$). ■

Proposition 2a: *Richer individuals react to negative income shocks. For own shocks $\frac{\partial m_{i2}^*}{\partial (\tilde{x}_{i1} - \tilde{x}_{i2})} = \rho_i \geq 0$, while for shocks to others $\frac{\partial m_{i2}^*}{\partial (\tilde{x}_{j1} - \tilde{x}_{j2})} = -\rho_i \leq 0$.*

Proof of Proposition 2a Analogously to Equations (5a) and (5b) in Section 2 the allocation perceived as fair after a shock can be modelled as a weighted average between the pre-shock fairness judgement that serves as a reference point and the hypothetical judgement of the post-shock situation:

$$m_{i2}^* = (1 - \rho_i)m_{i2} + \rho_i m_{i1} \text{ and } m_{j2}^* = 1 - m_{i2}^* \quad (\text{B.11})$$

Note that when focusing on shares, a shock that reduces the own share of earnings automatically increase the other's share and vice versa. This has two implications. First, it is theoretically not possible to isolate a negative own shock from a positive shock to the other. Second, m_{i1} will always be feasible, as shares can be calculated independently of a reduction in total resources. It is thus not necessary to introduce weights for own and other's reference points (γ_i). As the fair allocation to the other individual is automatically given by $m_{j2}^* = 1 - m_{i2}^*$ it is omitted in the rest of this section.

For richer individuals, where $m_{it} = \tilde{x}_{it}$, Equation (B.11) can be simplified to

$$m_{i2}^* = \tilde{x}_{i2} + \rho_i(\tilde{x}_{i1} - \tilde{x}_{i2}). \quad (\text{B.12})$$

Given that individuals care about pre-shock reference points ($\rho_i > 0$), the response to an own negative income shock, $\frac{\partial m_{i2}^*}{\partial(\tilde{x}_{i1} - \tilde{x}_{i2})} = \rho_i$ will be positive, implying an increase in the share individuals view as fair for themselves. By contrast, if there is a negative shock to the other person, $\frac{\partial m_{i2}^*}{\partial(\tilde{x}_{j1} - \tilde{x}_{j2})} = -\rho_i$ will be negative, and the share an individual views as fair for themselves decreases. Proposition 2 thus also holds when using shares as reference points. ■

Proposition 3a: *Poorer individuals take the same allocation decision independent of who suffers a shock. The effect of a negative shock is equal to zero.*

Proof of Proposition 3a Using Equation (B.11) and the result from Proposition 1a that for poorer individuals $m_{it} = 1/2$, the allocation perceived as fair can be written as

$$m_{i2}^* = (1 - \rho_i)1/2 + \rho_i 1/2 = 1/2. \quad (\text{B.13})$$

Consequently, $\frac{\partial m_{i2}^*}{\partial(\tilde{x}_{k1} - \tilde{x}_{k2})} = 0$. For poorer individuals following an egalitarian fairness view, the allocation that is perceived as fair is independent of shocks. ■

B.1.3 Application to experimental setting

In the following, I apply the theoretical framework to the experimental design. An individual is predicted to hold contribution-based fairness views ($\alpha_i = 1$) if $x_{it} > \frac{X_t}{2}$ and egalitarian views ($\alpha_i = 0$) otherwise. As $\frac{X_t}{2} = x_{Ht} + x_{Lt}$ and by definition $x_{Ht} > x_{Lt}$ this implies that for H players $\alpha_i^* = 1$, while for L players $\alpha_i^* = 0$.

For H players, the final allocation considered as fair is then given by $m_{i2}^* = x_{i2} + \rho_i \gamma_i (x_{i1} - x_{i2}) - \rho_i (1 - \gamma_i) (x_{j1} - x_{j2})$ (see Equation (B.4)). For L players it is given by $m_{i2}^* = \frac{X_2}{2} + \rho_i \gamma_i \frac{(X_1 - X_2)}{2} - \rho_i (1 - \gamma_i) \frac{(X_1 - X_2)}{2}$ (see Equation (B.7)). Depending on the matching between participants, we have situations in which there is a) no shock, b) only a

shock to self, c) only a shock to the other, d) a shock to both. In the main experiment, absolute shocks are identical for H and L players with $x_{H1} - x_{H2} = x_{L1} - x_{L2} = 100$. For the variation, where shocks are identical in relative terms we have $x_{H1} - x_{H2} = 300$ and $x_{L1} - x_{L2} = 100$. Table B.1 presents the predicted fair allocations to self (m_{i2}^*) for these values. The changes resulting from the variation with $x_{H1} - x_{H2} = 300$ are provided in brackets. Fair allocations to the other player (m_{j2}^*) can simply be calculated by using $X_2 - m_{i2}^*$ and are thus omitted from the table.

Table B.1: Fair allocations to self (m_{i2}^*)

	H players	L players
No shock	300	200
Shock to H	$300 + \rho_i \gamma_i 100(300)$	$200 + \rho_i \gamma_i 50(150) - \rho_i (1 - \gamma_i) 50(150)$
Shock to L	$300 - \rho_i (1 - \gamma_i) 100$	$200 + \rho_i \gamma_i 50 - \rho_i (1 - \gamma_i) 50$
Shock to both	$300 + \rho_i \gamma_i 100(300) - \rho_i (1 - \gamma_i) 100$	$200 + \rho_i \gamma_i 100(200) - \rho_i (1 - \gamma_i) 100(200)$

Note: Values in brackets refer to fair allocations under constant relative shocks.

Table B.1 illustrates that for H players conditional on $\gamma_i, \rho_i > 0$ a shock to self increases the allocation to self individuals view as fair, while a shock to the other player reduces this amount. As the effects are additive, if both players are hit by a shock the final evaluation depends on γ_i . Moreover, in the constant relative shocks variation (in brackets) the absolute shock is larger for H players. If participants think in absolute terms it is thus likely that they show a larger reaction to own shocks.

By contrast, independent of the shock, L players, perceive similar amounts as fair (200).²⁵ As final allocation decisions depend on fairness judgements, the framework thus predicts that H players, but not L players, will react to negative income shocks.

As Section B.1.2 shows, these predictions also hold when expressing the utility function in terms of shares. Applied to the experimental design, this means that L players will always perceive an allocation of 1/2 as fair. For H players, an own shock increases the allocation perceived as fair ($\frac{\partial m_{i2}^*}{\partial (\bar{x}_{i1} - \bar{x}_{i2})} = \rho_i \geq 0$), while a shock to the other player decreases it ($\frac{\partial m_{i2}^*}{\partial (\bar{x}_{j1} - \bar{x}_{j2})} = -\rho_i \leq 0$).

B.1.4 Extension: self-serving belief formation

So far I have assumed that individuals have perfect knowledge about shocks. While this is a reasonable assumption for own shocks, there might be uncertainty about the shocks of others. In this section, I thus outline an extension of the theoretical framework that incorporates incomplete information about the other's shock.

As before, individuals learn about their own relative income position at $t = 0$ and form α_i^* . Before deciding on the allocation they perceive as fair at $t = 1$, under incomplete infor-

²⁵As ρ_i and γ_i are exogenous, any deviation from 200 will be identical in the shock to H, shock to L, and shock to both condition.

mation, they first need to form expectations about the other's shock (at $t = 0.5$). I assume that while individuals do not know whether the other suffered a shock, they do know the share of individuals that have been hit by a shock within the population (σ). When forming beliefs about the likelihood that another individual that is drawn from this population suffers a shock (p_s), a rational and unbiased expectation is thus to guess $p_s = \sigma$.

However, individuals might benefit from distorting their beliefs. In doing so, they trade off the costs of deviating from an honest assessment against the benefits of deviating from σ . This is similar to standard models of lying aversion (see e.g. Fischbacher & Föllmi-Heusi, 2013; Gneezy et al., 2018). The benefit of deviating is to avoid lower allocations to self as a fair response to the other's shock (b_i). The cost is a psychological cost of not being honest (c_i). An individual would thus deviate from guessing $p_s = \sigma$ iff

$$(\sigma - p_s)b_i > c_i, \text{ with } c_i \geq 0. \quad (\text{B.14})$$

Corollary *It is never optimal for individuals to deviate from $p_s = \sigma$, unless $b_i > c_i$.*

Intuitively, individuals would never guess $p_s > \sigma$, as this involves costs of not being honest but no benefits. The question is thus, when would they downplay their beliefs that the other suffered a shock? To inform the subsequent predictions, I use the experimental results from Section 4.2, showing that richer individuals allocate less to themselves after they suffer a shock (Result 2ii), while poorer individuals do not respond to shocks (Result 3). This is in line with the theoretical framework for $\rho_i > 0$ and $\gamma_i = 0.5$.

Formally, after a shock to the other, for richer individuals the fair amount for themselves decreases by $\rho_i(1 - \gamma_i)(x_{1j} - x_{2j})$. This corresponds to the amount that can be gained if they believe the other did not suffer a shock (b_i). As long as i is not indifferent about a shock to the other ($\gamma_i < 1, \rho_i > 0$) and $b_i > c_i$, i should choose $p_s < \sigma$. From $U(p_s) = (\sigma - p_s)b_i - c_i$, it can be seen that an individual's utility is maximised at $p_s^* = 0$.

For poorer individuals, the experiment shows that $b_i = 0$, as they do not change their allocations after shocks, implying $b_i \leq c_i$. Equation (B.14) can thus never hold for poorer individuals and it is optimal for them to guess $p_s = \sigma$.

Proposition: *Richer individuals can benefit from deviating from σ with $p_s^* = 0$ if $b_i > c_i$. For poorer individuals $b_i \leq c_i$, implying that they will never find it optimal to deviate from $p_s = \sigma$.*

B.2 Additional analysis

All subsequent analyses are referenced in the document. They are structured by section within the chapter. Brief descriptions are provided for each analysis.

B.2.1 Sample and procedures

Balance checks

Table B.2: Sample characteristics and balance test across shocks

	Total	No shock to self	Shock to self	Difference
Age	31.82	31.62	32.02	-0.40
Gender (Female=1)	0.44	0.43	0.45	-0.02
<i>Education level</i>				
No formal education	0.01	0.01	0.01	0.00
Secondary school/GCSE	0.11	0.13	0.08	0.05*
College/A levels	0.17	0.17	0.17	0.00
Undergraduate degree	0.34	0.34	0.34	0.00
Graduate degree	0.34	0.30	0.39	-0.09**
PhD	0.03	0.05	0.02	0.03*
<i>Employment status</i>				
Full-time	0.50	0.50	0.50	0.00
Unemployed	0.19	0.19	0.19	0.00
Part-time	0.31	0.31	0.31	0.00
Income bin (1-8)	3.62	3.79	3.45	0.34**
Household size	2.79	2.80	2.79	0.01
Social ladder (1-10)	5.45	5.50	5.40	0.11
Loss in livelihood (1-7)	3.55	3.52	3.59	-0.07
Financial struggles (1-7)	3.59	3.59	3.60	-0.01
Fairness task (1-7)	4.08	4.07	4.09	-0.02
<i>Region</i>				
Europe	0.80	0.78	0.81	-0.03
North America	0.09	0.09	0.09	0.00
Other	0.11	0.12	0.10	0.02
N	536	268	268	536

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B.2 shows the demographic composition of the sample. As can be seen, participants who suffer or do not suffer a shock within the experiment are balanced across most variables. The only small differences are with respect to education, where participants without a shock are slightly more likely to have a secondary school degree or a PhD and less likely to have a graduate degree. Moreover, participants who do not suffer a shock are in a slightly higher income bin. All differences are however very small.

Similarly, table B.3 shows that the sample is balanced across all observable demographics with respect to H and L players. The only significant difference lies in the fairness perception of the task. This, however, is an intuitive result as L players are assigned lower initial earnings and might thus feel disadvantaged compared to H players.

Table B.3: Sample characteristics and balance test across player types

	Total	H players	L players	Difference
Age	31.82	31.94	31.70	0.24
Gender (Female=1)	0.44	0.47	0.41	0.07
<i>Education level</i>				
No formal education	0.01	0.01	0.00	0.01
Secondary school/GCSE	0.11	0.12	0.09	0.04
College/A levels	0.17	0.17	0.16	0.01
Undergraduate degree	0.34	0.32	0.36	-0.04
Graduate degree	0.34	0.34	0.35	-0.01
PhD	0.03	0.03	0.04	-0.02
<i>Employment status</i>				
Full-time	0.50	0.50	0.50	0.00
Unemployed	0.19	0.19	0.19	0.00
Part-time	0.31	0.31	0.31	0.00
Income bin (1-8)	3.62	3.61	3.63	-0.02
Household size	2.79	2.74	2.85	-0.11
Social ladder (1-10)	5.45	5.54	5.35	0.19
Loss in livelihood (1-7)	3.55	3.55	3.56	0.00
Financial struggles (1-7)	3.59	3.58	3.60	-0.02
Fairness task (1-7)	4.08	4.27	3.88	0.39***
<i>Region</i>				
Europe	0.80	0.80	0.79	0.01
North America	0.09	0.09	0.09	0.00
Other	0.11	0.11	0.12	-0.01
N	536	268	268	536

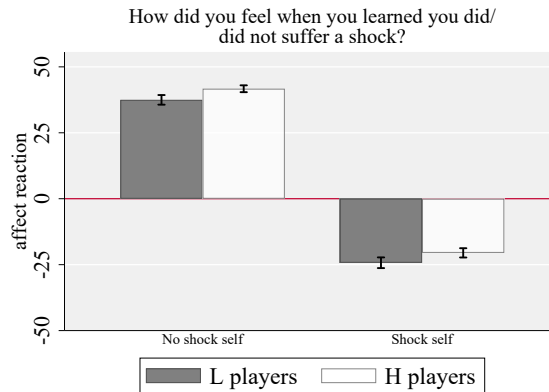
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B.2.2 Effect of negative income shocks

Affect reaction to own shock

Figure B.1 shows how people feel after learning they did/did not experience a negative shock within the experiment. Affect is measured on a scale from -50 to $+50$, where higher values are associated with more positive reactions. For both player types receiving a negative shock leads to significantly more negative affect reactions ($p < 0.001$). This provides evidence that the shock matters to participants and is perceived as a negative event.

Figure B.1: Affect reactions to finding out about a shock/no shock to self



Note: Whiskers represent 95% confidence intervals.

Testing for order effects

For the regression analysis in Section 4.2, order effects may be a concern as participants take repeated allocation decisions in different scenarios. Table B.4 shows the results regressing allocations to self on indicators for own and other's shock, an order dummy, and their interaction. Individually, the order effect is statistically insignificant. I then perform a test for the joint significance of the order dummy and the interaction with treatment indicators. Again, there are no significant results.

Table B.4: The effect of shocks on allocation to self (y_i). Testing for order effects.

	H players		L players	
	(1)	(2)	(3)	(4)
Shock self	13.10* (7.87)	5.13 (10.76)	4.00 (11.64)	-9.34 (16.81)
Shock other	-13.99*** (4.70)	-14.56** (5.82)	2.79 (5.37)	6.00 (7.47)
<i>Decision order</i>				
Shock other first	-5.69 (8.83)	-11.68 (12.64)	17.26 (12.29)	0.50 (17.61)
Shock other x Shock other first	4.68 (6.40)	-1.43 (8.15)	-10.20 (7.31)	-7.56 (10.43)
Shock self x Shock other		1.16 (9.51)		-6.03 (10.72)
Shock self x Shock other x Shock other first		23.64 (17.77)		27.37 (23.79)
Constant	278.58*** (6.93)	282.43*** (7.74)	208.69*** (10.33)	215.79*** (12.12)
N observations	536	536	534	534
N clusters	268	268	267	267
R ²	0.02	0.02	0.00	0.01

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the individual level.

Note: Results of OLS regressions. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Shock self* is a binary variable that takes the value of 1 if an individual i suffered a shock themselves and 0 otherwise. *Shock other* is a binary variable that takes the value of 1 if the player i has been matched with suffered a shock and 0 otherwise. *Shock other first* is a binary variable that takes the value of 1 if an individual was first faced with the decision situation where the other player had suffered a shock and 0 otherwise.

Looking at first decisions only

Despite the insignificant order effects, an additional robustness check that can be done is to restrict the sample to the first decision of each individual. This gets rid of the within-subject variation. Table B.5 shows the results of regressing allocations to self on indicators for own and other's shock for this restricted sample. In line with the full analysis, for H players the effect of an own shock is positive, while it is negative if the other suffers a shock. However, the coefficient on shock to other only becomes significant when including an interaction term between own and other's shock in column (2). As the coefficient on *shock self* in column (1) is similar in size to the effect when including all three periods ($b = -12.06$, see Table 2.2), the insignificance is likely due to the larger standard errors in the smaller sample. The significant interaction term for H players indicates that individuals only become more generous after others suffer a shock if they do not suffer a shock themselves. For L players, by contrast, reaction

to shocks are insignificant in all specifications. Overall, the analysis of the first decision thus confirms the results of the analysis in the full sample.

Table B.5: The effect of shocks on allocation to self (y_i). Sub-sample of first decisions.

	H players		L players	
	(1)	(2)	(3)	(4)
Shock self	23.44** (10.27)	8.68 (12.49)	13.51 (14.17)	-8.89 (19.59)
Shock other	-11.69 (10.84)	-31.65** (14.87)	18.59 (14.11)	-2.54 (19.76)
Shock self x Shock other		38.68* (21.47)		41.62 (28.11)
Constant	274.54*** (8.20)	281.86*** (8.99)	204.26*** (12.00)	215.98*** (13.53)
N observations	183	183	186	186
R ²	0.03	0.05	0.01	0.03

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Note: Results of OLS regressions. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Shock self* is a binary variable that takes the value of 1 if an individual i suffered a shock themselves and 0 otherwise. *Shock other* is a binary variable that takes the value of 1 if the player i has been matched with suffered a shock and 0 otherwise. All regressions are restricted to the first decision of each participant.

Differences between player types in their reaction to shocks

Table B.6 shows the results of a single regression of allocations to self, indicators for own and other's shock, Player types and their interaction. Column (1) shows the results for using H players as the base category. Column (2) does the same for L players. The difference in coefficients between H and L players is reported in column (3). As Table B.6 shows, there is a significant difference in both the intercept as well as the slope for *shock other* between H and L players. This is also confirmed by a Chow test ($p < 0.01$ respectively). Although the difference in the slope of *shock to self* is of similar size than the one of *shock to other* it is not statistically significant due to larger standard errors. Nevertheless, overall the results confirm

Table B.6: Difference between player types in allocations to self (y_i)

	(1) H players	(2) L players	(3) Difference
Shock self	13.10* (7.86)	3.10 (11.62)	-10.00 (14.03)
Shock other	-12.06*** (3.29)	-2.29 (3.66)	9.76** (4.92)
Constant	277.61*** (6.73)	211.72*** (10.02)	-65.89*** (12.07)
N observations		1072	
N clusters		536	
R ²		0.11	

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the individual level.

Note: Results of a single OLS regression. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Shock self* is a binary variable that takes the value of 1 if an individual i suffered a shock themselves and 0 otherwise. *Shock other* is a binary variable that takes the value of 1 if the player i has been matched with suffered a shock and 0 otherwise. The regression controls for player type and its interaction with shock indicators. Columns (1) and (2) show the results for a different player type as the base category. The regression controls for order effects.

that H and L players show significantly different reactions to shocks in the experiment.

Reaction to shocks depending on fairness views

As Figure 2.3 shows, there is some heterogeneity in fairness views within types. Even though most H players behave in line with a contribution-based view, some seem to follow an egalitarian view and vice versa for L players. As a robustness check, I thus test how individuals who follow egalitarian or contribution-based fairness views (independent of whether they are H or L players) respond to shocks. I thereby classify individuals as holding a contribution-based view if — in the absence of a shock — they choose an allocation of 300 for the H and 100 for the L player. They are classified as egalitarians if they choose an allocation of 200-200.²⁶ Table B.7 regresses allocations to self on indicators for own and other's shock for individuals with contribution-based or egalitarian fairness views. In line with theory, the results show that under contribution-based fairness views individuals respond to shocks, becoming more selfish after a shock to self and more generous after a shock to other. Egalitarians, by contrast, do not respond to any type of shock.

Table B.7: The effect of shocks on individuals with different fairness views

	Contribution-based		Egalitarian	
	(1)	(2)	(3)	(4)
Shock self	25.12* (14.00)	28.65* (15.12)	0.89 (3.44)	0.06 (0.35)
Shock other	-10.34** (4.93)	-7.01 (7.95)	5.51 (3.40)	4.71 (3.59)
Shock self x Shock other		-7.06 (9.66)		1.67 (6.91)
Constant	243.80*** (11.65)	242.13*** (12.16)	201.14*** (1.99)	201.55*** (1.64)
N observations	280	280	356	356
N clusters	140	140	178	178
R ²	0.07	0.07	0.01	0.01

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (in parentheses) are clustered at the individual level.

Note: Results of OLS regressions. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Shock self* is a binary variable that takes the value of 1 if an individual i suffered a shock themselves and 0 otherwise. *Shock other* is a binary variable that takes the value of 1 if the player i has been matched with suffered a shock and 0 otherwise. All regressions control for order effects.

Controlling for the perception of shocks

Table B.8 shows that the effect of shocks on allocation decisions is robust to controlling for perceptions of shocks. In the post-experimental survey, participants are asked to judge their own shock relative to the shock to the other person. They can either say that they were more, less, or equally affected by the shock compared to their co-player. This analysis can only be done for the sub-sample of participants who suffered a shock themselves, which is why *shock self* is not included in the regression. As can be seen, H players show a negative reaction to the

²⁶To ensure a sufficient sample size and account for some error, I allow allocation decisions to deviate by 10 tokens from the ideal contribution-based and egalitarian split in either direction.

shock to others independent of their relative perception of shocks. All interaction terms are insignificant. Moreover, the relative perception of shocks does not matter for allocation decisions. The same holds true for L players who unlike H players show no reaction to shocks.²⁷

Table B.8: The effect of shocks on allocation to self (y_i) after controlling for relative perception of shocks

	H players		L players	
	(1)	(2)	(3)	(4)
Shock other	-15.91** (6.70)	-14.09** (6.86)	1.57 (6.38)	-2.47 (5.61)
<i>Perception of shock (ref.: same)</i>				
Self more affected	5.64 (27.12)	10.04 (27.08)	10.82 (17.99)	12.36 (19.35)
Other more affected	-0.89 (12.50)	-5.68 (13.03)	-30.18* (16.73)	25.53 (44.24)
<i>Interactions</i>				
Shock other x self more affected	37.34 (29.66)	35.52 (30.91)	-14.02 (9.70)	-5.56 (8.94)
Shock other x other more affected	8.33 (9.78)	6.07 (10.21)	-5.57 (6.38)	-1.53 (5.61)
Constant	285.37*** (9.10)	321.37*** (46.83)	203.17*** (16.23)	263.29*** (58.64)
Demographic controls	No	Yes	No	Yes
N observations	268	258	268	250
N clusters	134	129	134	125
R ²	0.01	0.16	0.02	0.17

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the individual level.

Note: Results of OLS regressions. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Shock other* is a binary variable that takes the value of 1 if the player i has been matched with suffered a shock and 0 otherwise. *Perception of shock* is a categorical variable that takes the value of 0 if shock to self and shock to other are perceived to be identical, 1 if shock to self is perceived as larger and 2 if shock to other is perceived as larger. Demographic controls include age, gender, income, household size, education, region, subjective social status, and fairness of task. All regressions control for order effects.

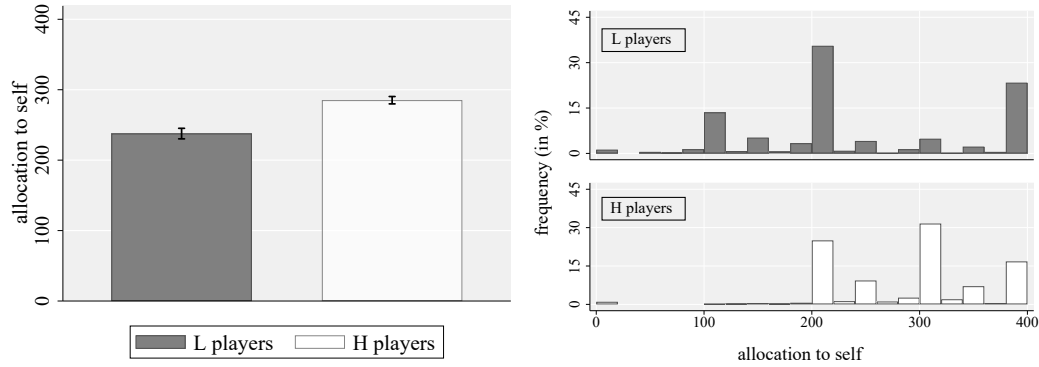
B.2.3 Constant relative shocks

Relative income and fairness views

Figure B.2 shows that the relationship between allocations and relative income is replicated when using constant relative shocks. The most common allocation for L players is 200, in line with the egalitarian fairness view. H players, by contrast, are most likely to choose allocations in line with individual contributions, resulting in 300 for themselves. The distribution of allocations differs significantly between types (Kolmogorov-Smirnov test, $p < 0.001$). This difference also translates into significantly lower allocations to self for L players (t-test, $p < 0.001$).

²⁷ While Table B.8 shows that L players allocate significantly less to themselves if they feel the other has been affected more by the shock, there are only 3 participants who expressed this feeling.

Figure B.2: Average allocation to self (LHS) and distribution (RHS) under constant relative shocks

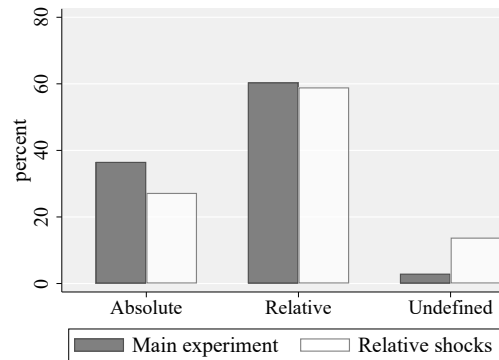


Note: Whiskers represent 95% confidence intervals.

Relative vs absolute perception of shocks

Figure B.3 visualises how participants perceive shocks in the post-experimental survey. Participants could state in the post-experimental survey whether they were more, less or identically affected by shocks compared to their co-player. This allows to assess whether participants view shocks in relative or absolute terms. For instance, in the main experiment, where absolute shocks are held constant participants are classified as viewing shocks in absolute terms if they say both players are affected in the same way. If they state that L players are affected more, they are classified as viewing shocks in relative terms. An analogous classification can be done for the variation where shocks are identical in relative terms.

Figure B.3: Perception of shocks within the experiment



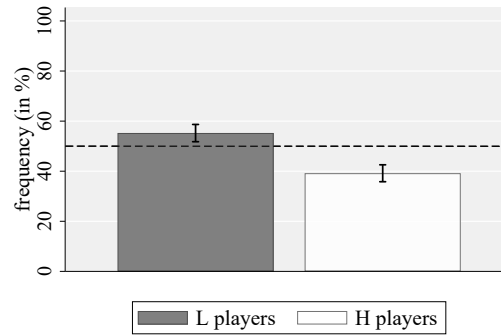
As Figure B.3 shows, both absolute and relative perceptions of shocks are present in the study. However, the majority seems to perceive shocks in relative terms. Interestingly, participants are significantly more likely to perceive shocks in absolute terms in the main study (t-test, $p < 0.001$). This suggests that it matters how information about shocks is presented.

B.2.4 The role of incomplete information

Replication of biased beliefs with constant relative shocks

Figure B.4 shows that the bias in beliefs for H players can be replicated in the treatment with constant relative shocks. Significantly less than 50% of H players believe that the other suffered a shock (t-test, $p < 0.001$). L players do not show this bias. If anything slightly more than 50% state that the other suffered a shock (t-test, $p < 0.09$).

Figure B.4: Share of players believing that the other suffered a shock under constant relative shocks



Note: Whiskers represent 95% confidence intervals.

Effect of incomplete information on allocation decisions

Table B.9 shows the results of regressing allocations to self on shock to self and the available information about the other player. The table shows that under no information, H players behave as if the other player did not suffer a shock. Similarly, a Wald test shows that allocations to self are significantly higher under incomplete information than if it is known that the other suffered a shock ($p < 0.001$). For L players, by contrast, allocation decisions are indistinguishable across all three scenarios.

Note that under constant relative shocks, this result is less robust. H players allocate significantly more to the other after incomplete information compared to knowing that the other did not suffer a shock. The difference becomes insignificant after controlling for an interaction with perceived fairness in initial income differences.

Table B.9: The effect of shocks and information on allocation to self (y_i)

	H players		L players	
	(1)	(2)	(3)	(4)
Shock self	11.74 (7.81)	13.14* (7.60)	5.62 (11.80)	4.90 (11.61)
<i>Information (ref.: Info = no shock other)</i>				
Info = Shock other	-12.06*** (3.30)	-11.71*** (3.37)	-2.67 (3.66)	-2.20 (3.60)
Info = No information	-1.53 (3.26)	-0.52 (3.30)	1.79 (3.56)	1.06 (3.71)
Constant	272.74*** (11.98)	301.58*** (35.57)	216.60*** (15.38)	260.46*** (41.79)
Demographic controls	No	Yes	No	Yes
N observations	804	777	801	762
N clusters	268	259	267	254
R ²	0.02	0.12	0.02	0.14

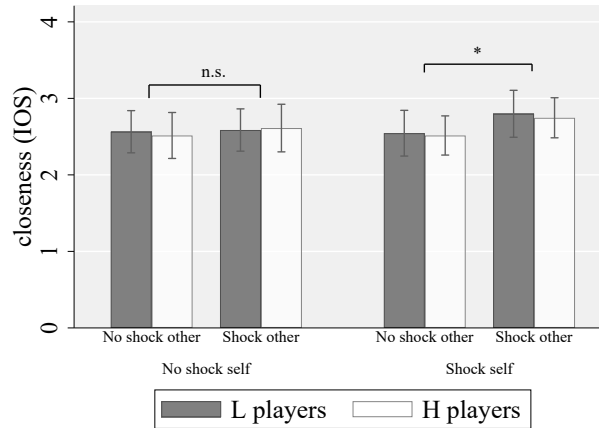
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the individual level.

Note: Results of OLS regressions. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Shock self* is a binary variable that takes the value of 1 if an individual i suffered a shock themselves and 0 otherwise. *Information* is a categorical variable that takes the value of 0 if player i is informed that the other did not suffer a shock, 1 if player i is informed that the other did suffer a shock and 2 if the other's shock is unknown. Demographic controls include employment shock due to Covid-19, age, gender, income, household size, education, region, subjective social status, and fairness of task. All regressions control for order effects.

B.2.5 Relationship closeness and affect

While the main focus of the experiment is understanding the effect of negative shocks and relative income on redistributive preferences, I also explore whether shocks change perceptions of closeness between participants and positive/negative affect responses.²⁸ I elicit closeness between participants using the IOS scale (Aron et al., 1992) after each allocation decision.²⁹

Figure B.5: Reported closeness between players conditional on shocks



Note: Whiskers represent 95% confidence intervals.

²⁸The analysis of closeness and affect under constant relative shocks yields qualitatively similar results.

²⁹The IOS scale measures closeness on a scale from 1 to 7, represented by pairs of circles with different degrees of overlap. A larger value thereby represents a higher degree of closeness.

For the affect questions, I ask participants how they felt when they learned that they/the other player did or did not suffer a shock. Affect was measured on a scale from -50 to 50 , corresponding to very negative or very positive affect reactions. I thereby use a variation of the pictorial assessment scale developed by [Desmet et al. \(2001\)](#).

When it comes to relationship closeness, I find that overall the experience of shocks does not have large effects (see Figure B.5). While a shock to the other player slightly increases closeness, this effect is only marginally significant if players suffer a shock themselves (t-test, $p = 0.09$). In addition, closeness is very similar for both player types. While participants' reported closeness to each other did not vary systematically across experience of shocks, I do find that there is a significant correlation between allocation decisions and closeness. Both H and L players who report a higher level of closeness to their co-player allocate significantly less to themselves. As Table B.10 shows the effect of shocks on allocation decisions is robust to the inclusion of IOS scores as an additional control variable. H players allocate more to themselves after experiencing a shock and less to themselves after the other does so. L players, by contrast, still show no reaction to any shocks.

Table B.10: The effect of shocks on allocation to self (y_i) after controlling for perceived closeness

	H players		L players	
	(1)	(2)	(3)	(4)
Shock self	14.15* (7.22)	15.38** (7.28)	4.64 (11.49)	3.10 (11.77)
Shock other	-9.25*** (2.99)	-8.87*** (3.09)	-1.38 (3.65)	-0.94 (3.66)
Closeness (IOS)	-17.09*** (1.98)	-16.36*** (2.03)	-9.34*** (3.50)	-9.35** (3.72)
Constant	318.95*** (8.56)	324.61*** (33.41)	237.49*** (15.62)	309.03*** (42.05)
Demographic controls	No	Yes	No	Yes
N observations	536	518	534	508
N clusters	268	259	267	254
R ²	0.18	0.26	0.03	0.16

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the individual level.

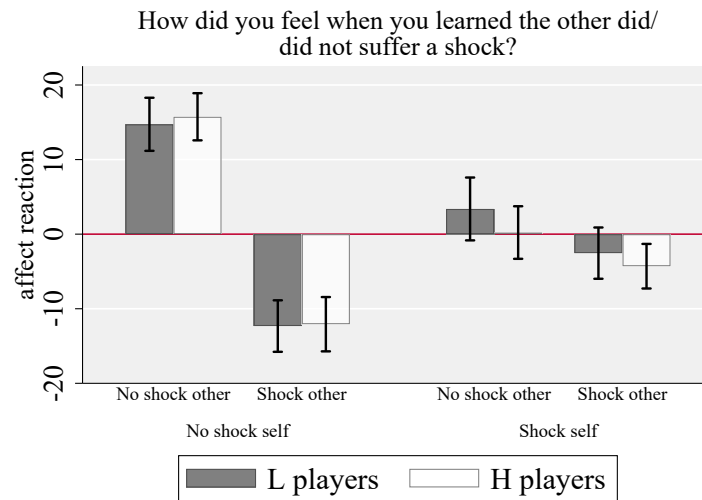
Note: Results of OLS regressions. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Shock self* is a binary variable that takes the value of 1 if an individual i suffered a shock themselves and 0 otherwise. *Shock other* is a binary variable that takes the value of 1 if the player i has been matched with suffered a shock and 0 otherwise. *Closeness* is measured via the IOS scale, where a value of 1 indicates the lowest and a level of 7 the highest degree of perceived closeness between players. Demographic controls include employment shock due to Covid-19, age, gender, income, household size, education, region, subjective social status, and fairness of task. All regressions control for order effects.

In addition to oneness, there are also other significant correlates of allocation decisions. Attitudes towards solidarity and behaviour during a crisis affect allocations in the experiment. People who state that compassion is a crucial virtue ($p = 0.01$), that in a crisis people become more compassionate ($p = 0.07$) and that one should give priority to society's instead of individual problems allocate significantly less to themselves ($p = 0.06$). In terms of demographics, women and older participants allocate significantly less to themselves ($p = 0.08$ and $p = 0.03$ respectively).

Finally, Figure B.6 shows affect reactions to the information that the other person did/did

not suffer a shock. A shock to the other leads to a significantly more negative reaction than learning that the other did not suffer a shock (t-test, $p < 0.001$). However, the difference is much larger if players themselves did not suffer a shock. They are significantly less happy (sad) for the other player avoiding (suffering) the shock if they were hit by a shock themselves (t-tests, $p < 0.001$ respectively).

Figure B.6: Affect responses to other's shock



Note: Whiskers represent 95% confidence intervals.

B.2.6 Study 2: Sample and data collection

Balance check

Table B.11 shows the sample characteristics for Study 2. As expected, suffering a significant loss in livelihood is not orthogonal to demographic characteristics. In particular, participants who state they suffered a significant loss in livelihood are more likely to be female, to have lost their employment, to be from a lower income category, to have a lower perceived social status, to have recently faced financial struggles and to be from the US. Moreover, they are less likely to have a graduate degree, be full-time employed or to come from a European country.

Table B.11: Sample characteristics and balance test across the experience of Covid shocks

	Total	No Covid shock	Covid shock	Difference
Age	31.21	31.08	31.42	-0.34
Gender (Female=1)	0.43	0.39	0.49	-0.10**
<i>Education level</i>				
No formal education	0.01	0.00	0.01	-0.01
Secondary school/GCSE	0.15	0.13	0.18	-0.05
College/A levels	0.16	0.16	0.17	-0.01
Undergraduate degree	0.34	0.33	0.35	-0.02
Graduate degree	0.31	0.34	0.26	0.08*
PhD	0.04	0.04	0.03	-0.01
<i>Employment status (prolific)</i>				
Full-time	0.50	0.67	0.22	0.45***
Unemployed	0.50	0.33	0.78	-0.45***
<i>Employment status (survey)</i>				
Full-time	0.42	0.57	0.19	0.38***
Unemployed	0.25	0.11	0.46	-0.35***
Other	0.33	0.32	0.35	-0.04
Income bin (1-8)	3.47	3.65	3.18	0.47***
Household size	2.87	2.83	2.94	-0.11
Social ladder (1-10)	5.26	5.58	4.75	0.83***
Financial struggles (1-7)	3.94	2.96	5.55	2.59***
Fairness task (1-7)	4.09	4.13	4.03	-0.10
<i>Region</i>				
Europe	0.77	0.81	0.71	0.11***
North America	0.10	0.05	0.19	-0.14***
Other	0.12	0.14	0.10	0.03
N	536	332	204	536

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B.2.7 Study 2: Effect of negative income shocks

Can player types be pooled?

To test whether H and L players can be pooled, I run a single regression of allocations to self on indicators for own and other's shock, player types and their interaction (see Table B.12). Columns (1) and (2) respectively present results for taking H and L players as the base category, while column (3) shows the difference in coefficients between player types. As the regression shows, the only significant difference between H and L players is with respect to levels. Their reactions to shocks, by contrast, are statistically indistinguishable. Are Chow-test confirms that the slopes for *shock self* and *shock other* are not significantly different between player types ($p = 0.34$ and $p = 0.84$ respectively). I thus pool player types for the rest of the analysis, while controlling for level differences in allocations to self.

Table B.12: The difference between player types in allocations to self (y_i) after Covid shocks

	(1) H players	(2) L players	(3) Difference
Shock self	9.10 (9.38)	23.77* (12.19)	-14.68 (15.37)
Shock other	-57.56*** (4.26)	-56.21*** (4.91)	-1.34 (6.50)
Constant	276.78*** (6.71)	209.14*** (8.60)	67.64*** (9.54)
N observations		1072	
N clusters		536	
R ²		0.17	

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the individual level.

Note: Results of a single OLS regression. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Shock self* is a binary variable that takes the value of 1 if an individual i suffered a shock themselves and 0 otherwise. *Shock other* is a binary variable that takes the value of 1 if the player i has been matched with suffered a shock and 0 otherwise. The regression controls for player type and its interaction with shock indicators. Columns (1) and (2) show the results for a different player type as the base category. The regression controls for order effects.

Order effects

As participants are taking several allocation decisions for different participants I control for order effects. Table B.13 shows the results of regressing allocations to self on indicators for own and other's shock and their interaction with decision order. For the effect of a shock to

Table B.13: The effect of Covid shocks on allocation to self (y_i). Testing for order effects.

	(1)	(2)
Shock self	16.47** (7.71)	28.41** (11.39)
Shock other	-67.59*** (4.65)	-66.09*** (6.29)
Shock self x Shock other		-3.71 (9.28)
<i>Decision order</i>		
Shock other first	1.31 (7.81)	11.55 (9.63)
Shock other x Shock other first	21.26*** (6.44)	16.41* (8.48)
Shock self x Shock other x Shock other first		-14.00 (17.13)
H Player	61.38*** (7.46)	61.28*** (7.46)
Constant	217.60*** (7.85)	212.80*** (8.33)
N observations	1072	1072
Demographic controls	No	No
N clusters	536	536
R ²	0.18	0.18

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the individual level.

Note: Results of OLS regressions. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Shock self* is a binary variable that takes the value of 1 if an individual i suffered a shock themselves and 0 otherwise. *Shock other* is a binary variable that takes the value of 1 if the player i has been matched with suffered a shock and 0 otherwise. *Shock other first* is a binary variable controlling for order effects that takes the value 1 if a participant was first matched with another player who did suffer a shock and 0 if they were first matched with another player who did not suffer a shock. *H Player* is a binary variable that takes the value of 1 for H players and 0 for L players.

other, decision order plays a significant role. Moreover, a Wald test confirms that order effects are jointly significant ($p = 0.05$ in column (1), $p = 0.002$ in column (2)). I therefore test whether the effect of shocks on allocation decision holds independent of the decision order. As Table B.14 shows the effect of a shock to other is smaller when participants first face the situation where the other suffered a shock and then the situation where the other did not suffer a shock. Nevertheless, independent of the decision order, a shock to the other player always leads to a significant reduction in allocations to self.

Table B.14: The effect of Covid shocks on allocation to self (y_i). Interaction with order effects.

	(1)	(2)	(3)
Shock self	16.47** (7.71)	16.85* (9.01)	28.52** (12.54)
<i>Shock other</i>			
Decision order: no shock other first	-67.59*** (4.65)	-67.58*** (4.82)	-66.35*** (6.52)
Decision order: shock other first	-46.33*** (4.46)	-46.38*** (4.49)	-51.16* (5.68)
<i>Shock self x Shock other</i>			
Decision order: no shock other first			-3.05 (9.62)
Decision order: shock other first			13.75 (9.21)
Decision order	1.31 (7.81)	0.46 (8.15)	11.44 (10.09)
H Player	61.38*** (7.46)	64.00*** (7.77)	63.98*** (7.78)
Constant	217.60*** (7.85)	229.09*** (31.03)	221.80*** (31.21)
Demographic controls	No	Yes	Yes
N observations	1072	1026	1026
N clusters	536	513	513
R-squared	0.18	0.21	0.21

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the individual level.

Note: Results of OLS regressions. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Shock self* is a binary variable that takes the value of 1 if an individual i suffered a shock themselves and 0 otherwise. *Shock other* is a binary variable that takes the value of 1 if the player i has been matched with suffered a shock and 0 otherwise. *H Player* is a binary variable that takes the value of 1 for H players and 0 for L players. *Decision order* is a binary variable controlling for order effects that takes the value 1 if a participant was first matched with another player who did suffer a shock and 0 if they were first matched with another player who did not suffer a shock. Demographic controls include employment shock due to Covid-19, age, gender, household size, education, region, subjective social status, and fairness of task.

Looking at first decisions only

As an additional robustness check, I restrict the sample to the first decision of each individual when regressing allocation decisions on indicators for own and other's shock. Table B.15 shows that in line with the full analysis, the effect of an own shock leads to a positive and significant effect on allocations to self, while the effect of a shock to other leads to a negative, significant effect. The analysis of the first decision thus confirms the results of the analysis in the full sample.

Table B.15: The effect of Covid shocks on allocations to self (y_i). Sub-sample of first decisions.

	(1)	(2)
Shock self	21.36** (10.34)	22.02 (14.19)
Shock other	-40.81*** (9.94)	-40.33*** (12.53)
Shock self x Shock other		-1.27 (20.60)
Constant	224.79*** (10.27)	224.54*** (10.86)
N observations	355	355
R ²	0.10	0.10

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Note: Results of OLS regressions. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Shock self* is a binary variable that takes the value of 1 if an individual i suffered a shock themselves and 0 otherwise. *Shock other* is a binary variable that takes the value of 1 if the player i has been matched with suffered a shock and 0 otherwise. Regressions are restricted to the first decision of each participant.

The effect of real world shocks in Study 1

Table B.16 regresses allocation decisions in Study 1 on indicators for own and other's shock in the experiment as well as own real world shocks. A real world shock is defined as a *significant loss in livelihood due to Covid-19*. The results show that H players who experienced a real world shock allocate more to themselves. This effect becomes significant after including controls in column (2) and is of a similar, statistically indistinguishable, size as the effect of shocks that happen within the experiment.

Table B.16: The effect of Covid shocks on allocation to self (y_i) in Study 1

	H players		L players	
	(1)	(2)	(3)	(4)
Real world shock	12.07 (8.00)	15.26* (8.33)	-0.72 (12.02)	-6.06 (13.18)
Shock self	12.54 (7.82)	14.61* (7.78)	4.04 (11.63)	3.80 (11.80)
Shock other	-12.06*** (3.29)	-11.71*** (3.39)	-2.67 (3.66)	-2.20 (3.61)
Constant	272.95*** (7.71)	317.88*** (34.94)	211.67*** (10.91)	291.08*** (42.74)
Demographic controls	No	Yes	No	Yes
N observations	536	518	534	508
N clusters	268	259	267	254
R ²	0.023	0.105	0.004	0.132

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the individual level.

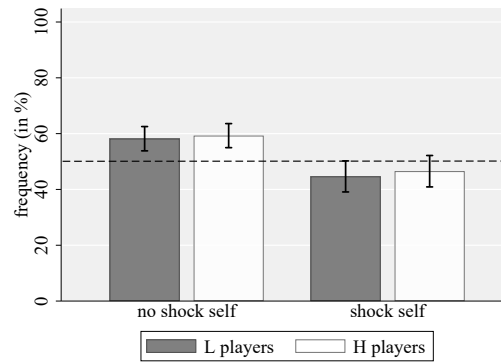
Note: Results of OLS regressions. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Real world shock* is a binary variable that takes the value 1 if an individual states they suffered a significant loss in livelihood due to Covid-19 and 0 otherwise. *Shock self* is a binary variable that takes the value of 1 if an individual i suffered a shock themselves and 0 otherwise. *Shock other* is a binary variable that takes the value of 1 if the player i has been matched with suffered a shock and 0 otherwise. Demographic controls include age, gender, income, household size, education, region, subjective social status, and fairness of task. All regressions control for order effects.

B.2.8 Study 2: The role of incomplete information

Beliefs by player type

The belief pattern about the other player holds when looking separately at H and L players. If players suffer a shock themselves, their aggregate beliefs of whether the other did suffer a real world shock are not statistically different from 50% (see Figure B.7). If they did not suffer a shock themselves players beliefs lie even significantly above 50% (t-tests, $p = 0.02$ for H and $p = 0.04$ for L players).

Figure B.7: Share of players believing that the other suffered a Covid shock by player type



Note: Whiskers represent 95% confidence intervals.

Effect of incomplete information on allocation decisions

In addition to beliefs, I test the effect of incomplete information on distributive preferences by regressing allocations to self on shock to self and the available information about the other player. As Table B.17 shows, if there is no information, participants allocate significantly less to themselves than if they know the other did not suffer a significant loss in livelihood. However, the reaction is not as strong as if they know for sure the other suffered a shock. A Wald test confirms that the coefficients are statistically different ($p < 0.001$). Allocation decisions under incomplete information thus lie between the decisions when it is known that the other did or did not suffer a shock. This stresses again that there is no self-serving bias among participants.

Table B.17: The effect of Covid shocks and information on allocation to self (y_i)

	(1)	(2)
Shock self	10.90 (7.56)	12.37 (8.70)
<i>Type of information (ref.: Info = No shock other)</i>		
Info = Shock other	-56.88*** (3.25)	-57.00*** (3.32)
Info = No information	-21.38*** (2.27)	-21.78*** (2.34)
H Player	59.20*** (7.38)	61.51*** (7.63)
Constant	228.94*** (11.92)	241.72*** (31.30)
Demographic controls	No	Yes
N observations	1608	1539
N clusters	536	513
R ²	0.15	0.20

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the individual level.

Note: Results of OLS regressions. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Shock self* is a binary variable that takes the value of 1 if an individual i suffered a shock themselves and 0 otherwise. *Information* is a categorical variable that takes the value of 0 if player i is informed that the other did not suffer a shock, 1 if player i is informed that the other did suffer a shock and 2 if the other's shock is unknown. *H Player* is a binary variable that takes the value of 1 for H players and 0 for L players. Demographic controls include employment shock due to Covid-19, age, gender, income, household size, education, region, subjective social status, and fairness of task. All regressions control for order effects.

B.2.9 Study 2: Relationship closeness and affect

I find that the experience of shocks does not only affect redistributive preferences, but also closeness between players and affect reactions. Figure B.8 shows how reported closeness as measured by the IOS scale (1-7) changes with the experience of real world shocks. While a shock to the other player has no effect on perceived closeness if players did not suffered a shock themselves, it increases closeness if players suffer a shock themselves ($p < 0.001$). As in the previous study, participants who report a higher level of closeness to their co-player allocate significantly less to themselves. The effect of shocks on allocation decisions is robust to the inclusion of IOS scores as an additional control (see Table B.18).

Figure B.8: Reported closeness between players conditional on Covid shocks

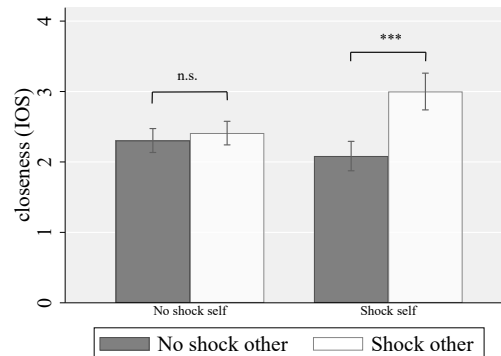


Table B.18: The effect of Covid shocks on allocations to self (y_i) after controlling for perceived closeness

	(1)	(2)
Shock self	19.13*** (7.36)	18.39** (8.60)
Shock other	-50.44*** (3.45)	-50.60*** (3.53)
Closeness (IOS)	-15.55*** (2.04)	-15.77*** (2.15)
H Player	59.86*** (7.19)	63.29*** (7.47)
Constant	248.35*** (9.71)	256.36*** (29.95)
Demographic controls	No	Yes
N observations	1072	1026
N clusters	536	513
R ²	0.23	0.26

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the individual level.

Note: Results of OLS regressions. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Shock self* is a binary variable that takes the value of 1 if an individual i suffered a shock themselves and 0 otherwise. *Shock other* is a binary variable that takes the value of 1 if the player i has been matched with suffered a shock and 0 otherwise. *Closeness* is measured via the IOS scale, where a value of 1 indicates the lowest and a level of 7 the highest degree of perceived closeness between players. *H Player* is a binary variable that takes the value of 1 for H players and 0 for L players. Demographic controls include employment shock due to Covid-19, age, gender, income, household size, education, region, subjective social status, and fairness of task. All regressions control for order effects.

Figure B.9 shows participants' reactions to the information that the other person did/did not suffer a shock. A shock to other always leads to a significantly more negative reaction than learning that the other did not suffer a shock (t-tests, $p < 0.001$).

Figure B.9: Affect responses to other's Covid shock



B.3 Introducing uncertainty about entitlements

B.3.1 Motivation and Design

In the main study, participants know how much each player contributes to the total earnings, allowing for differences in entitlement between player types. In the following, I describe a treatment variation that introduces uncertainty about entitlements by only telling participants who did the easier or more difficult task but not how much they earned individually (Study 3). Instead, participants only learn how much they and their matched partner earned together and in a next step whether their joint earnings are hit by a shock or not. As before, post-shock earnings are held constant across conditions to identify the effect of a negative shock.

In particular, while pairs in the shock condition start with 600 tokens and lose 200 due to a negative shock, pairs in the no shock condition directly start with 400 tokens. As I do not provide information about individual pre-shock earnings, I cannot explore the effect of a shock to self or other, but only compare a situation with no shock to one with common shocks.³⁰

Not knowing individual contributions introduces uncertainty about entitlements, which could weaken differences between player types. Especially L players can now justify higher allocations to themselves, as the labels *easier* and *more difficult* task leave plenty of room for interpretation and it is in their best interest to play down any differences. In the main study, by contrast, this is harder as there is a concrete number attached to these differences.

Hypothesis A: *If there is uncertainty about entitlements differences between player types will decrease. In particular, I expect L players to allocate more to themselves in this set-up.*

Another motivation for this treatment is to control for stake size effects. Whenever one of the two players suffers a negative shock, inevitably the total amount that can be distributed between players is shrinking. Previous literature has discussed the importance of stake size effects in dictator and ultimatum games (Cameron, 1999; Raihani et al., 2013; Larney et al., 2019), suggesting that under higher stakes receivers tend to accept lower offers in ultimatum games and proposers make smaller offers in dictator games. However, it is not clear how a *shock* to the stake size affects allocation decisions. The additional treatment is designed to measure this effect. In particular, if there are no stake size effects, there should be no difference in allocations to self between the shock and no shock condition.

Hypothesis B: *If there is no stake size effect, I expect no difference in allocations to self between shock and no shock conditions.*

B.3.2 Data collection and results

In line with power calculations for the main study, I recruited 134 participants for the shock condition and 134 for the no shock condition via Prolific in July 2021. The experiment was shorter than the main study, as participants are only matched with one other player. The

³⁰Note that if the pair suffers a shock they lose 200 tokens. This is identical to the size of a shock to both players in the main study.

median completion time was 12.1 minutes and participants earned on average £9.34 per hour. I again stratify with respect to the experience of real world employment shocks during Covid to get a balanced sample with respect to this dimension. Moreover, I use the same ex-post questionnaire as in the main study and elicit affect measures as well as closeness between participants.

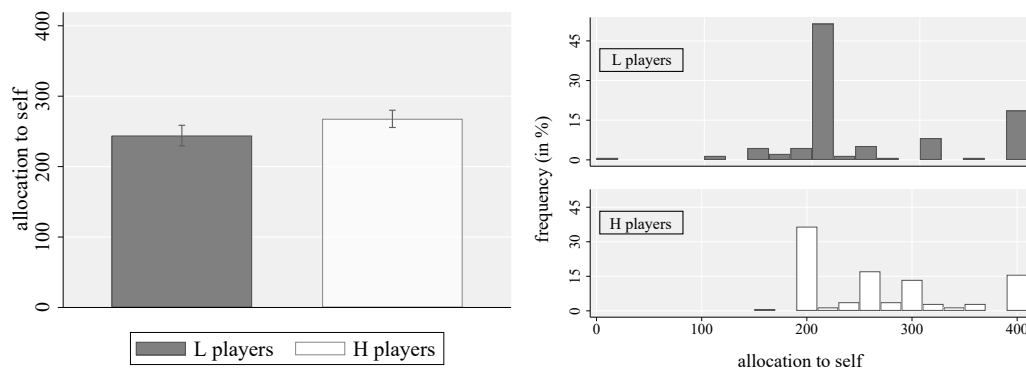
The importance of relative income differences

I first examine how uncertainty about initial endowments affects differences between H and L players. As the left graph in Figure B.10 shows, H players still allocate significantly more to themselves than L players ($p < 0.05$). This confirms that the different difficulty levels justify different allocations and already generate separate feelings of entitlement between player types. However, this difference has been significantly reduced compared to Study 1. In Study 1, H players allocate on average 58 tokens more to themselves than L players, while they only allocate 24 tokens more to themselves once uncertainty is introduced. This difference is mainly driven by L players. While H players hardly and insignificantly adjust their allocations to self downwards if there is no information about individual pre-shock earnings (-10 tokens), L players show a very strong reaction, increasing their allocations to self significantly under uncertainty (+24 tokens, $p < 0.01$). The results thus confirm Hypothesis A that uncertainty decreases differences between player types and that particularly L players use this uncertainty to justify larger allocations to themselves.

Result A: *If there is uncertainty about entitlements, differences between player types diminish. This is mainly driven by L players allocating more to themselves.*

The right graph in Figure B.10 shows the distribution of allocations to self for both player types. Interestingly, now the equal split seems to be the most salient allocation for both H and L players. Without any information about individual earnings, this seems to be in fact the most intuitive sharing rule and confirms in line with previous studies the importance of the egalitarian fairness view.

Figure B.10: Average allocation to self (LHS) and distribution (RHS) by player type (Study 3)



Effect of negative income shocks

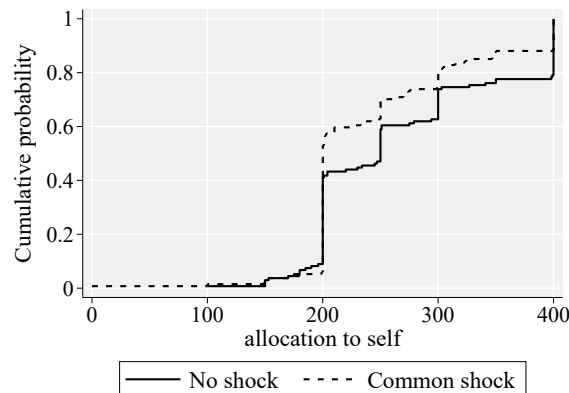
A Chow-test confirms that both H and L players react in the same way to a negative shock. I thus pool the data for H and L players for all subsequent analysis. This also gives more power to detect differences between the shock and no shock condition as participants only take one decision, reducing the total number of observations. Figure B.11 shows the cumulative distribution of allocations to self after no shock and a common shock to both players. Contrary to Hypothesis B, there is a clear shift of the distribution to the left under a common shock, meaning participants allocate less to themselves in this case. In particular, there is a strong increase in allocating 200 to self. The proportion of participants choosing this allocation increases from 32% without a shock to 47% after a common shock (proportion test, $p = 0.01$).

This finding also holds when regressing the number of tokens individuals allocate to themselves on the experience of a common shock.³¹ Table B.19 confirms that a negative common shock significantly decreases allocations to self. This finding is robust to the inclusion of controls in column (2). It thus appears that individuals do not only react to own and others' shocks but also to a general reduction of the pie.

Result B: *Negative common shocks have an effect on allocation decisions, with both H and L players allocating significantly less to themselves. In particular, there is a significant increase in the proportion of participants choosing an egalitarian split.*

As mentioned above, the uncertainty treatment aims to test the importance of stake size effects. However, this seems to be only part of what is going on in this treatment. In addition to the uncertainty about initial endowments, shocks introduce a further layer of uncertainty, as it is not specified whether one person was affected more by a shock but only inform participants about the reduction in joint earnings. Given this double-layer of uncertainty, it might be more difficult to decide what is the appropriate allocation, driving participants towards the salient and simple egalitarian split. Another difference between Study 1 and this treatment is how

Figure B.11: Reaction to a common shock



³¹As each individual only takes one decision, I neither cluster standard errors at the individual level nor need to control for order effects.

Table B.19: The effect of common shocks on allocations to self (y_i)

	(1)	(2)
Common shock	-23.70** (9.59)	-23.17** (9.93)
H Player	23.73** (9.59)	24.18** (9.98)
Constant	255.91*** (9.16)	296.76*** (33.93)
Demographic controls	No	Yes
N observations	268	256
R ²	0.04	0.14

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Note: Results of OLS regressions. The dependent variable is the number of tokens individuals allocate to themselves (y_i). *Common shock* is a binary variable that takes the value of 1 if there is a shock to joint earnings and 0 otherwise. *H Player* is a binary variable that takes the value of 1 for H players and 0 for L players. Demographic controls include employment shock due to Covid-19, age, gender, income, household size, education, region, subjective social status, and fairness of task.

close participants feel to their matched partner. I find that perceived closeness is significantly higher in the uncertainty treatment than in Study 1 (t-test, $p = 0.08$). While in Study 1 participants learn about their individual earnings and shocks, in the uncertainty treatment they are informed about joint earnings and common shock. This seems to generate a feeling of *we*, which in turn favours an egalitarian split. Moreover, after participants suffer a common shock they report higher (although not significantly higher) levels of closeness than without a shock, explaining the higher occurrence of equal splits.

Taken together I conclude that the uncertainty treatment does not allow to isolate the effect of a change in stake sizes, but mainly allows to explore what happens when introducing uncertainty about entitlements. The results show that this reduces differences between player types and leads to an increased focus on the egalitarian split, in particular after participants experience a common shock.

B.4 Supplementary materials

B.4.1 Instructions

In the following I provide instructions and screenshots for Study 1. As instructions only differ slightly for H and L players, I provide one set of instructions from the perspective of H players and add the corresponding text for L players in brackets when required.

Instructions

Please read the instructions carefully.

This study consists of three stages: an earning stage, a decision stage, and a short questionnaire. You will be given details about the different stages below. Throughout the study you can earn Experimental Tokens, which are later converted into a bonus (£) at a pre-defined exchange rate (100 tokens = £0.40). Your final bonus will depend partly on your own decision, partly on the decision of another Prolific participant and partly on external factors. How your bonus is determined will be outlined below in greater detail.

Earning stage: In this stage, you have to work in order to earn tokens that are then used in the second stage of the study. As a task, you will be given randomly generated strings of letters that you need to type in reverse order. As an example: If you see the string “rlgowsahc”, the correct answer will be “chaswoglr”. In order to earn your tokens, you need to correctly solve 10 strings. All participants are randomly assigned to different difficulty levels and earn a different number of tokens. For a higher difficulty level you will earn more tokens. This will help us to correctly calibrate different difficulty levels for another study. **You have been assigned to a more difficult (easier) level for which you will earn 400 (200) tokens.** After finishing the string reversal task, you will be given more information about external factors that could affect the number of tokens you take into the second stage.

Decision stage: The second stage of the study consists of **three separate decision rounds**. In each round, you will be matched with a different Prolific participant who is participating in the same study but has been assigned to an **easier (more difficult) level**. In each round, you have to decide how you want to divide the total number of tokens that you and the other participant took into the second stage between **yourself** and the **other participant**. The other participant will do the same. The screenshot below shows what the decision screen will look like:

Your choice

The total number of tokens you and the other participant earned to take into the decision stage is **X**. How many of those X tokens do you want to allocate to yourself? How many to the other participant? **Please make your choice by moving the sliders below.** Then click on continue to confirm your decision.



At the end of the experiment, there will be a lottery that selects one of the three decision rounds to count towards your bonus. In other words, each of the three decision rounds can be relevant for your final payment. Once a round has been selected, another lottery will decide whether **your choice** or the **other participant's choice** will determine your final bonuses. With a 50% chance, your choice will determine your and the other participant's final bonuses, with a 50% chance, the other participant's choice will determine your and their final bonuses. All interactions in this study are completely anonymous. You will never know the identity of the other participants, and they will never know yours.

B.4.2 Screenshots experiment

Real effort task

Earning stage (1/2)

In order to earn **your tokens** for the decision stage, you need to correctly reverse **10 strings**. As an example: If you see "rigowsahc", the correct answer will be "chaswogl". **Once you have completed the task, you can move on to the next part of the study.**

To submit a reversed string, either press the "Submit" button or your "Enter" key.

Total strings attempted:

Number of correct strings:

Text sequence:

nfcslyew

Your answer:

Submit

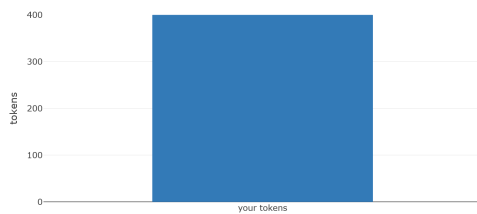
Learning about shocks

a) H players with a shock

Earning stage (2/2)

Thank you! You have completed the task and earned **400 tokens**. The graph below shows the total number of tokens that you possess at the moment.

As mentioned in the instructions, the number of tokens you take into the decision stage depends partly on the task and partly on external factors. Each participant could suffer a **negative shock and lose 100 tokens**. **Half of the participants in this study are hit by a shock and half of the participants are not hit by a shock.**

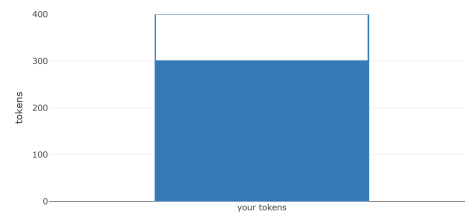


Please click on the button below to find out whether you have been hit by a shock:

Learn about shock

Oh no, you have been hit by a shock!

You initially earned **400 tokens**. You **lost 100 tokens** due to a shock and have thus left **300 tokens** to take into the decision stage. The graph below shows your tokens before and after the shock.



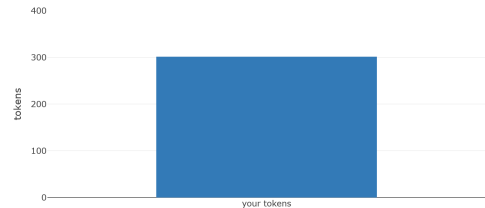
Continue

b) H players without a shock

Earning stage (2/2)

Thank you! You have completed the task and earned 300 tokens. The graph below shows the total number of tokens that you possess at the moment.

As mentioned in the instructions, the number of tokens you take into the decision stage depends partly on the task and partly on external factors. Each participant could suffer a negative shock and lose 100 tokens. Half of the participants in this study are hit by a shock and half of the participants are not hit by a shock.

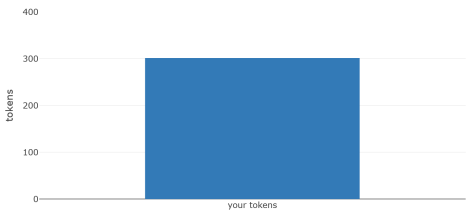


Please click on the button below to find out whether you have been hit by a shock:

Learn about shock

Phew, you have not been hit by a shock!

You initially earned 300 tokens. You did not suffer a shock and have thus left all 300 tokens to take into the decision stage. The graph below shows your tokens before and after the shock.



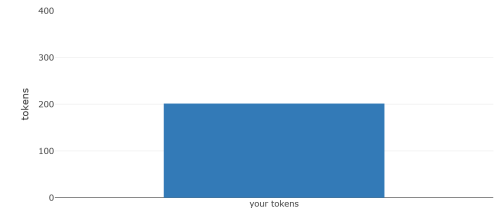
Continue

c) L players with a shock

Earning stage (2/2)

Thank you! You have completed the task and earned 200 tokens. The graph below shows the total number of tokens that you possess at the moment.

As mentioned in the instructions, the number of tokens you take into the decision stage depends partly on the task and partly on external factors. Each participant could suffer a negative shock and lose 100 tokens. Half of the participants in this study are hit by a shock and half of the participants are not hit by a shock.

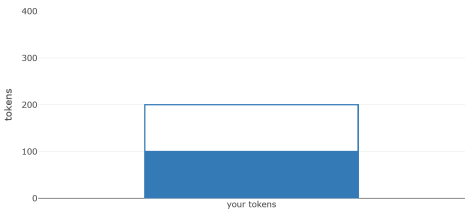


Please click on the button below to find out whether you have been hit by a shock:

Learn about shock

Oh no, you have been hit by a shock!

You initially earned 200 tokens. You lost 100 tokens due to a shock and have thus left 100 tokens to take into the decision stage. The graph below shows your tokens before and after the shock.



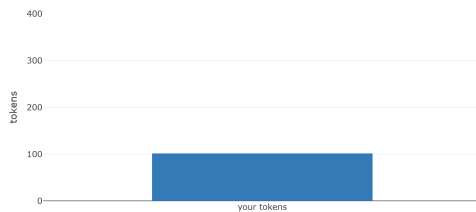
Continue

d) L players without a shock

Earning stage (2/2)

Thank you! You have completed the task and earned **100 tokens**. The graph below shows the total number of tokens that you possess at the moment.

As mentioned in the instructions, the number of tokens you take into the decision stage depends partly on the task and partly on external factors. Each participant could suffer a **negative shock and lose 100 tokens**. **Half of the participants in this study are hit by a shock and half of the participants are not hit by a shock.**

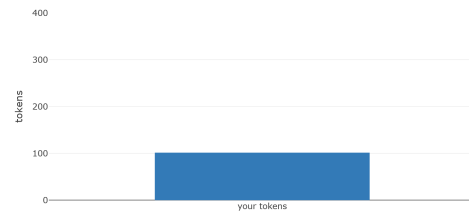


Please click on the button below to find out whether you have been hit by a shock:

Learn about shock

Phew, you have not been hit by a shock!

You initially earned **100 tokens**. You did not suffer a shock and have thus left all **100 tokens** to take into the decision stage. The graph below shows your tokens before and after the shock.



Continue

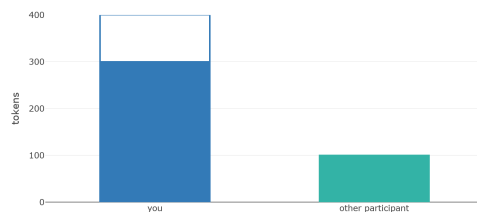
Decision screen

The decision screen is shown from the perspective of an H player who did suffer a shock. The screen is adjusted accordingly for different scenarios (order and colour of scenarios is randomised). In each of the 3 rounds, participants receive information about the participant they are being matched with.

Information about Participant 1

The **first player** you have been matched with initially earned **100 tokens**. They **did not suffer a shock** and, so, took all **100 tokens** into the second stage.

In comparison, **you** initially earned **400 tokens**. You **lost 100 tokens** due to a shock and, so, had **300 tokens** left to take into the second stage. The graph below shows both your and the other participant's tokens (before and after the shock).



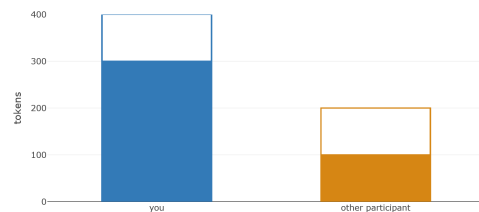
Click continue to proceed to the decision page.

Continue

Information about Participant 2

The **second participant** you have been matched with initially earned **200 tokens**. They **lost 100 tokens** due to a shock and, so, had **100 tokens** left to take into the second stage.

In comparison, **you** initially earned **400 tokens**. You **lost 100 tokens** due to a shock and, so, had **300 tokens** left to take into the second stage. The graph below shows both your and the other participant's tokens (before and after the shock).



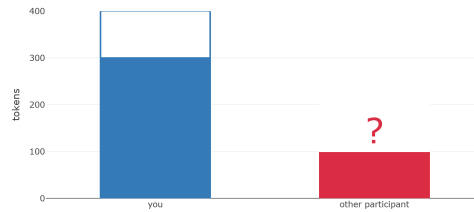
Click continue to proceed to the decision page.

Continue

Information about Participant 3

The **third participant** you have been matched with initially earned **either 200 or 100 tokens**. You don't know whether they suffered a shock. The only thing you know for sure is that they took **100 tokens** into the second stage.

In comparison, **you** initially earned **400 tokens**. You lost **100 tokens** due to a shock and so, had **300 tokens** left to take into the second stage. The graph below shows both your and the other participant's tokens (before and after the shock).

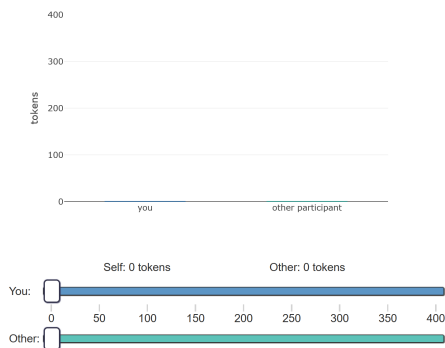


Continue

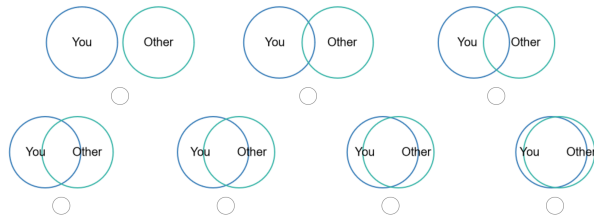
Next, they take a decision about how they want to allocate the total earnings between themselves and the other player (left graph below) and are then asked about how close they feel to the other player.

Your choice

The total amount of tokens you and the other participant earned to take into the decision stage is **400**. How many of those 400 tokens do you want to allocate to yourself? How many to the other participant? **Please make your choice by moving the sliders below**. Then click on continue to confirm your decision.



Please consider which of the pairs of circles best represents your connection with the **other participant**. By selecting a pair of circles, please indicate to what extent **you** and the other participant are connected.



Continue

Under incomplete information (here participant 3), participants answer either first the allocation decision or the question about their beliefs about the shock of the other player.

As mentioned before, **you don't know whether the other participant suffered a shock**. However, we would like you to guess what happened. If your guess is correct an additional **50 tokens** will be added to your bonus at the end of the experiment. Please make your guess by clicking on one of the options below:

The other participant suffered a shock and initially earned 200 tokens.

The other participant did not suffer a shock and initially earned 100 tokens.

Continue

Questionnaire

Again some of the questions vary with roles and experience of shocks. Here I provide the example of an H player who suffered a shock. Questions are adjusted accordingly for other players.

Questionnaire (1/3)

1) In the first stage of the study, participants were randomly assigned to different difficulty levels and **earned different amounts of tokens**. You were assigned to a more difficult level, for which **you initially earned 400 tokens**. The other participants you have been matched with **initially earned 200 or 100 tokens for an easier level**. What do you think about these differences?

very unfair



very fair

2) At the end of the earning stage, you learned that **you were hit by a shock and lost 100 tokens**. Please rate how this news made you feel by moving the slider below.



3) In one round of the decision stage you learned that **the participant you have been matched with was hit by a shock**. Please rate how this news made you feel by moving the slider below.



4) In another round of the decision stage you learned that **the participant you have been matched with was not hit by a shock**. Please rate how this news made you feel by moving the slider below.



5) You initially earned **400 tokens and lost 100 tokens** due to a shock. The other participant who suffered a shock, initially earned **200 tokens and lost 100 tokens** due to a shock. How do you judge both shocks?

We were affected in the same way.

The other participant was affected more.

I was affected more.

Continue

Questionnaire (2/3)

- 1) What is your age?
- 2) What is your gender? (male/female/prefer not to say)
- 3) What is the highest degree or level of schooling you have completed? (No formal qualification/ Secondary school or GCSE/ College or A levels/ Undergraduate degree/ Graduate degree/ PhD/ Prefer not to say)
- 4) Some people describe political affiliation on a left to right spectrum. Please indicate where you believe your political ideology lies on this spectrum.



- 5) What was your household pre-tax income last year?
- 6) Including yourself, how many people currently live in your household?
- 7) Think of the ladder below as representing where people stand in your country.

At the top of the ladder are the people who are the best off — those who have the most income, the most education, and the most respected jobs. At the bottom are the people who are worst off — who have the least money, the least education, and the least respected jobs or no jobs. The higher up you are on this ladder, the closer you are to the people at the very top; the lower you are, the closer you are to the people at the very bottom.

Where would you place yourself on this ladder?³²



³²This question was developed by Adler et al. (2000) as a measure for perceived social status.

Questionnaire (3/3)

1) How has your employment status been affected by the COVID-19 (coronavirus) pandemic?

1.1) Before the pandemic I was:

Working full-time
Working part-time
Unemployed or out of work
Self-employed
A student and not employed
Prefer not to say

1.2) My current situation is:

Working full-time
Working part-time
Working fewer part-time hours
My job has been suspended (unpaid leave or furloughed)
Unemployed or out of work
Self-employed
A student and not employed
Prefer not to say

2) To what extent do you agree with the following statements:

I suffered a significant loss in livelihood due to the Covid-19 crisis.

Strongly disagree ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

The pandemic made me struggle financially.

Strongly disagree ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

The government should aim to reduce economic differences.

Strongly disagree ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

Compassion for those who are suffering is the most crucial virtue.

Strongly disagree ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

I think during a crisis people tend to become more

Selfish ☐ ☐ ☐ ☐ ☐ ☐ ☐ Compassionate

Should you give priority to solving your own problems or should you give priority to solving your society's problems?

Own problems ☐ ☐ ☐ ☐ ☐ ☐ ☐ Society's problems

3) Any feedback regarding this study is highly appreciated. Please leave comments below:

B.4.3 Study 2: Instructions

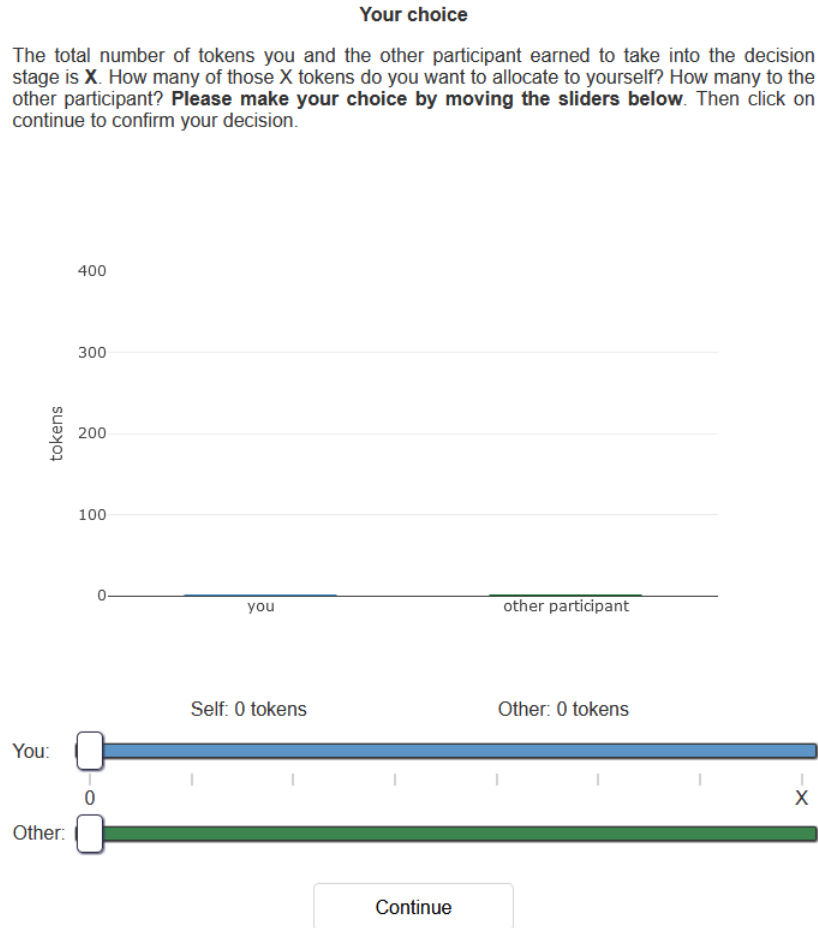
Instructions

Please read the instructions carefully.

This study consists of three stages: an earning stage, a decision stage, and a short questionnaire. You will be given details about the different stages below. Throughout the study you can earn Experimental Tokens, which are later converted into a bonus (£) at a pre-defined exchange rate (100 tokens = £0.40). Your final bonus will depend partly on your own decision, partly on the decision of another Prolific participant and partly on external factors. How your bonus is determined will be outlined in greater detail below.

Earning stage: In this stage, you have to work in order to earn tokens that are then used in the second stage of the study. As a task, you will be given randomly generated strings of letters that you need to type in reverse order. As an example: If you see the string “rlgowsahc”, the correct answer will be “chaswoglr”. In order to earn your tokens, you need to correctly solve 10 strings. All participants are randomly assigned to different difficulty levels and earn a different number of tokens. For a higher difficulty level you will earn more tokens. This will help us to correctly calibrate different difficulty levels for another study. **You have been assigned to a more difficult (easier) level for which you will earn 300 (100) tokens.**

Decision stage: The second stage of the study consists of **three separate decision rounds**. In each round, you will be matched with a different Prolific participant who is participating in the same study but has been assigned to **an easier (more difficult) level**. In each round, you have to decide how you want to divide the total number of tokens that you and the other participant earned between **yourself** and **the other participant**. The other participant will do the same. The screenshot below shows what the decision screen will look like:



In each round, before making your decision, we may show you some **background information about the other participant**. Concretely, we may tell you whether they suffered a **significant loss in livelihood** due to the Covid-19 crisis. This information comes from self-reports on Prolific. Similarly, we may inform your matched participants whether you reported a significant loss in livelihood on Prolific before they make their decisions. All interactions in this study are **completely anonymous**. You will never know the identity of the other participants, and they will never know yours.

At the end of the experiment, there will be a lottery that selects one of the three decision rounds to count towards your bonus. In other words, each of the three decision rounds can be relevant for your final payment. Once a round has been selected, another lottery will decide whether **your choice** or the **other participant's choice** will determine your final bonuses. With a 50% chance, your choice will determine your and the other participant's final bonuses, with a 50% chance, the other participant's choice will determine your and their final bonuses.

B.4.4 Study 2: Screenshots experiment

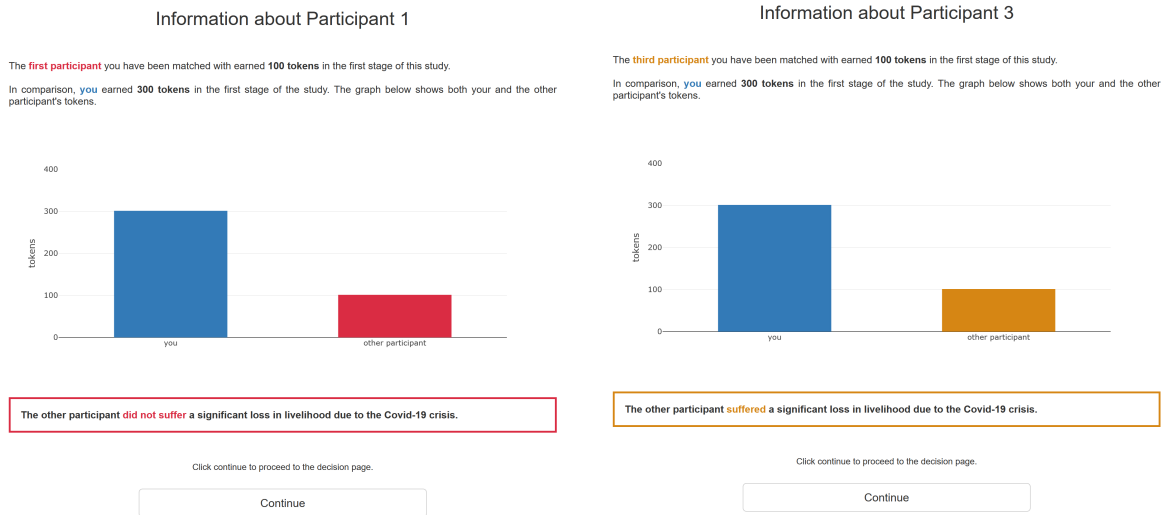
As Study 2 shares the same structure as Study 1, quite a few screens are identical across studies. For reasons of conciseness, I only show screens that differ and refer to Study 1 otherwise.

Real effort task

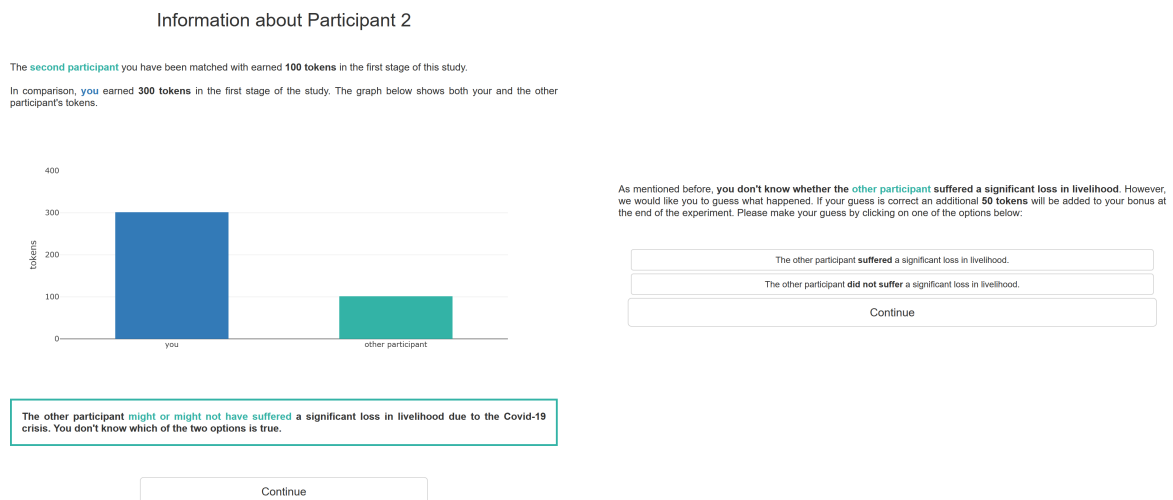
See Study 1.

Information about the other player and beliefs

In each of the 3 rounds, participants receive information about the participant they are being matched with. The order and colour of scenarios is randomised. The decision screen where participants can allocate earnings between themselves and the other player is identical to Study 1.



Under incomplete information (here participant 2), participants answer either first the allocation decision or the question about their beliefs about the real world shock of the other player (see Figure RHS below). Screenshots are made from the perspective of an H player. For L players the text and graphs are adjusted accordingly.



Questionnaire

Again the questions are phrased from the perspective of an H player. For L players, the text is adjusted accordingly. Questionnaire 2/3 and 3/3 are identical to Study 1.

Questionnaire (1/3)

1) In the first stage of the study, participants were randomly assigned to different difficulty levels and **earned different amounts of tokens**. You were assigned to a **more difficult level**, for which **you earned 300 tokens**. The other participants you have been matched with **earned 100 tokens for an easier level**. What do you think about these differences?

very unfair



very fair

2) In one round of the decision stage you learned that **the participant you have been matched with suffered a significant loss in livelihood** due to the Covid-19 crisis. Please rate how this news made you feel by moving the slider below.



3) In another round of the decision stage you learned that **the participant you have been matched with did not suffer a significant loss in livelihood** due to the Covid-19 crisis. Please rate how this news made you feel by moving the slider below.



4) **One of the participants** you have been matched with suffered a **significant loss in livelihood due to the Covid-19 crisis**. How do you judge both you and that other participant have been affected by Covid?

☐ The other participant was affected more.

☐ We were affected in the same way.

☐ I was affected more.

5) As mentioned in the instructions, the information about whether other participants suffered from a significant loss in livelihood due to the Covid-19 crisis comes from **self-reports on Prolific**. How truthfully do you believe the other participants reported their situation?

very untruthfully



very truthfully

C Appendix for Chapter 3

C.1 Theoretical appendix

C.1.1 Proofs

Proof of Proposition 1 Expected payoff is:

$$X - x_i + \gamma (x_i + x_j) - \frac{\eta_i}{2} (x_i - x_j)^2 - \frac{\delta_i}{2} (x_i - x_i^a)^2$$

which can be rewritten as

$$X - x_i + \gamma (x_i + x_j) - \frac{\eta_i}{2} \int_0^{\bar{x}} (x_i - x_j)^2 f(x_j) dx_j - \frac{\delta_i}{2} (x_i - x_i^a)^2. \quad (C.1)$$

The derivative of Equation (C.1) with respect to x_i gives

$$-(1 - \gamma) - \eta_i (x_i - \mu) - \delta_i (x_i - x_i^a). \quad (C.2)$$

Evaluated at $x_i = 0$, Equation (C.2) becomes $-(1 - \gamma) + \eta_i \mu + \delta_i x_i^a$. When $\eta_i < \frac{1 - \gamma - \delta_i x_i^a}{\mu}$ we therefore have $x_i^* = 0$. Evaluated at $x_i = \bar{x}$, Equation (C.2) becomes $-(1 - \gamma) - \eta_i (\bar{x} - \mu_x) - \delta_i (\bar{x} - x_i^a) < 0$. Finally, as highlighted in the proposition, in any interior solution Equation (C.2) is equal to 0, so that $x_i^* = \frac{\eta_i \mu + \delta_i x_i^a - (1 - \gamma)}{\eta_i + \delta_i}$. ■

Proof of Proposition 2 Expected payoff is:

$$\begin{aligned} X - x_i + \gamma (x_i + x_j) - \alpha_i F(x_i) [x_i - E(x_j | x_j < x_i)] \\ - \beta_i (1 - F(x_i)) [E(x_j | x_j > x_i) - x_i] - \frac{\delta_i}{2} (x_i - x_i^a)^2 \end{aligned}$$

which can be rewritten as

$$\begin{aligned} X - x_i + \gamma (x_i + x_j) - \alpha_i \int_0^{x_i} (x_i - x_j) f(x_j) dx_j \\ - \beta_i \int_{x_i}^{\bar{x}} (x_j - x_i) f(x_j) dx_j - \frac{\delta_i}{2} (x_i - x_i^a)^2. \end{aligned} \quad (C.3)$$

The derivative of Equation (C.3) with respect to x_i gives

$$-(1 - \gamma) - \alpha_i F(x_i) + \beta_i (1 - F(x_i)) - \delta_i (x_i - x_i^a). \quad (C.4)$$

Evaluated at $x_i = 0$, Equation (C.4) becomes $\beta_i - (1 - \gamma) + \delta_i x_i^a$. When $\beta_i < 1 - \gamma - \delta_i x_i^a$ we therefore have $x_i^* = 0$. Evaluated at $x_i = \bar{x}$, Equation (C.4) becomes $-\alpha_i - (1 - \gamma) - \delta_i (\bar{x} - x_i^a) < 0$. Finally, as highlighted in the proposition, in any interior solution Equation (C.4) is equal to 0, and hence

$$\delta_i x_i^* + F(x_i^*) (\alpha_i + \beta_i) = \beta_i - (1 - \gamma) + \delta_i x_i^a. \quad \blacksquare \quad (C.5)$$

Proof of Corollary 1 For the proof of part (i) suppose that, for some $\widehat{x} \in (0, \bar{x})$, the following holds: $F_1(x) > F_0(x)$ for $x \in (0, \widehat{x})$ and $F_1(x) < F_0(x)$ for $x \in (\widehat{x}, \bar{x})$. Consider $x_{i0}^* \in (0, \widehat{x})$. We have

$$\begin{aligned} -(1-\gamma) - \alpha_i F_1(x_{i0}^*) + \beta_i(1 - F_1(x_{i0}^*)) - \delta_i(x_i - x_i^a) &< \\ -(1-\gamma) - \alpha_i F_0(x_{i0}^*) + \beta_i(1 - F_0(x_{i0}^*)) - \delta_i(x_i - x_i^a) &= 0 \end{aligned}$$

Since $-(1-\gamma) - \alpha_i F(x) + \beta_i(1 - F(x)) - \delta_i(x - x_i^a)$ is decreasing in x , this implies that $x_{i1}^* < x_{i0}^*$. The proof of part (ii) is analogous and is therefore omitted. ■

Proof of Corollary 2. From Equation (C.5) we see that, restricting attention to interior solutions, we have

$$\begin{aligned} \frac{dx_i^*}{d\alpha_i} &= -\frac{F(x_i^*)}{\delta_i + (\alpha_i + \beta_i)f(x_i^*)} < 0 \\ \frac{dx_i^*}{d\beta_i} &= \frac{1 - F(x_i^*)}{\delta_i + (\alpha_i + \beta_i)f(x_i^*)} > 0 \\ \frac{dx_i^*}{dx_i^a} &= \frac{\delta_i}{\delta_i + (\alpha_i + \beta_i)f(x_i^*)} > 0 \quad \blacksquare \end{aligned}$$

Proof of Corollary 3 Consider first the difference in α necessary to induce a change in x_i^* from x' to x'' . Fix the other parameters to be equal to β , δ and x^a and denote as α'_k the value of α required for x' to be optimal under f_k . We have

$$F_k(x') = \frac{\varkappa}{\alpha'_k + \beta}$$

where $\varkappa \equiv \beta - (1-\gamma) + \delta x^a - \delta x'$, and, analogously,

$$F_k(x'') = \frac{\varkappa}{\alpha''_k + \beta}.$$

Recall that $F_1(x') > F_0(x')$ while $F_1(x'') < F_0(x'')$. As a result,

$$\begin{aligned} \frac{\varkappa}{\alpha'_0 + \beta} &< \frac{\varkappa}{\alpha'_1 + \beta} \text{ i.e. } \alpha'_1 < \alpha'_0 \\ \text{and similarly } \frac{\varkappa}{\alpha''_0 + \beta} &> \frac{\varkappa}{\alpha''_1 + \beta} \text{ i.e. } \alpha''_1 > \alpha''_0 \end{aligned}$$

which implies that $\alpha'_0 - \alpha''_0 > \alpha'_1 - \alpha''_1$. Consider now the difference in β necessary to induce a change in x_i^* from x' to x'' . We have

$$F_k(x') = \frac{\beta'_k + \rho(x')}{\alpha + \beta'_k}$$

where $\rho(x) \equiv -(1-\gamma) + \delta x^a - \delta x$, and, analogously,

$$F_k(x'') = \frac{\beta''_k + \rho}{\alpha + \beta''_k}.$$

Recalling that $F_1(x') > F_0(x')$ while $F_1(x'') < F_0(x'')$,

$$\frac{\beta'_0 + \rho(x')}{\alpha + \beta'_0} < \frac{\beta'_1 + \rho(x')}{\alpha + \beta'_1} \text{ i.e. } \beta'_0(\alpha - \rho(x')) < \beta'_1(\alpha - \rho(x'))$$

and similarly $\frac{\beta''_0 + \rho(x'')}{\alpha + \beta''_0} > \frac{\beta''_1 + \rho(x'')}{\alpha + \beta''_1} \text{ i.e. } \beta''_0(\alpha - \rho(x'')) > \beta''_1(\alpha - \rho(x''))$,

where $\alpha - \rho > 0$. To see this note that, from Equation (C.5), we can see that

$$\alpha_i > \beta_i(1 - F(x_i^*)) - (1 - \gamma) + \delta x_i^a - \delta x_i^* > \rho(x_i^*).$$

This implies that $\beta''_0 - \beta'_0 > \beta''_1 - \beta'_1$. Finally, the result with respect to x_i^a is derived analogously to α and β and is therefore omitted. ■

C.1.2 The disutility of deviating from x_i^a takes the absolute value form

Suppose that the disutility from selecting a contribution that differs from x_i^a is given by

$$\begin{aligned} & -\vartheta_i(x_i - x_i^a) \text{ if } x_i > x_i^a \\ & -\rho_i(x_i^a - x_i) \text{ if } x_i < x_i^a \end{aligned}$$

for some $\vartheta_i \geq 0$ and $\rho_i \geq 0$. If $x_i > x_i^a$, individual expected utility is

$$X - x_i + \gamma(x_i + x_j) - \alpha_i \int_0^{x_i} (x_i - x_j) f(x_j) dx_j - \beta_i \int_{x_i}^{\bar{x}} (x_j - x_i) f(x_j) dx_j - \vartheta_i(x_i - x_i^a)$$

while if $x_i < x_i^a$ it is

$$X - x_i + \gamma(x_i + x_j) - \alpha_i \int_0^{x_i} (x_i - x_j) f(x_j) dx_j - \beta_i \int_{x_i}^{\bar{x}} (x_j - x_i) f(x_j) dx_j - \rho_i(x_i^a - x_i)$$

The optimal contribution x_i^* satisfies

$$\begin{aligned} F(x_i^*) &= \frac{\beta_i - (1 - \gamma) - \vartheta_i}{\alpha_i + \beta_i} & \text{if } \frac{\beta_i - (1 - \gamma) - \vartheta_i}{\alpha_i + \beta_i} > F(x_i^a) \\ F(x_i^*) &= \frac{\beta_i - (1 - \gamma) + \rho_i}{\alpha_i + \beta_i} & \text{if } \frac{\beta_i - (1 - \gamma) + \rho_i}{\alpha_i + \beta_i} < F(x_i^a) \\ x_i^* &= x_i^a & \text{if } \frac{\beta_i - (1 - \gamma) + \rho_i}{\alpha_i + \beta_i} > F(x_i^a) > \frac{\beta_i - (1 - \gamma) - \vartheta_i}{\alpha_i + \beta_i} \end{aligned}$$

Consider two distributions f_0 and f_1 with the same mean and suppose that f_0 is single-crossing stochastic dominant over f_1 so that, for some $\hat{x} \in (0, \bar{x})$ the following holds: $F_1(x) > F_0(x)$ for $x < \hat{x}$ and $F_1(x) < F_0(x)$ for $x > \hat{x}$.

Clearly enough, the results outlined in Corollaries 1 and 2 continue to hold. Consider now Corollary 3. It is straightforward to check that $\alpha'_0 - \alpha''_0 > \alpha'_1 - \alpha''_1$ and $\beta'_0 - \beta''_0 > \beta'_1 - \beta''_1$ as in the proof of corollary 3. This implies that the parameter shift in α_i or in β_i needed to generate a change in optimal contribution from x' to x'' is larger under f_0 compared to f_1 . However, the prediction with respect to x_i^a is less clear-cut. Suppose for instance that $\frac{\beta_i - (1 - \gamma) - \vartheta_i}{\alpha_i + \beta_i} = x'$ and $\frac{\beta_i - (1 - \gamma) + \rho_i}{\alpha_i + \beta_i} = x''$. A change in x_i^a from $\tilde{x}' = F^{-1}(x')$ to $\tilde{x}'' = F^{-1}(x'')$ would result in the optimal contribution moving from x' to x'' . Since $F_1^{-1}(x') <$

$F_0^{-1}(x')$ and $F_1^{-1}(x'') > F_0^{-1}(x'')$, $\tilde{x}'' - \tilde{x}'$ is larger under f_1 compared to f_0 . We conclude that, if the disutility of deviating from x_i^a takes the absolute value form, the result of Corollary 3 does not hold for x_i^a (but it does for α_i and β_i).

C.2 Additional analysis

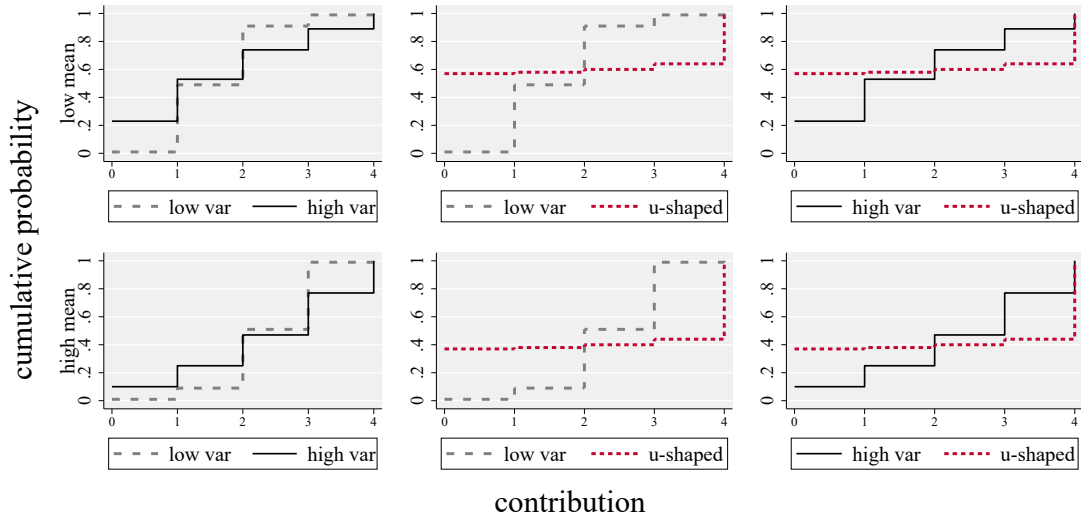
All subsequent analyses are referenced in the document. They are structured by section within the chapter. Brief descriptions are provided for each analysis.

C.2.1 Theoretical framework

Single crossing property of shown distributions

Figure C.1 plots the cumulative distributions of all six treatments. As can be seen below, the single-crossing condition discussed in Section 2 holds for all pairwise comparisons of descriptive norms sharing the same mean.

Figure C.1: Cumulative distributions of descriptive norm treatments



C.2.2 Experimental Design

Belief elicitation screen

Figure C.2 shows a screenshot of how individuals could indicate their beliefs about all available options within the experiment.

Figure C.2: Belief elicitation screen

The screenshot displays a belief elicitation interface with five columns, each representing a different number of tokens: 0 token, 1 token, 2 token, 3 token, and 4 token. Each column contains a vertical stack of small circles, with the number of circles increasing from 0 to 4 across the columns. Below each stack of circles is a row of two buttons: a '+' button and a '-' button. The entire interface is set against a blue background.

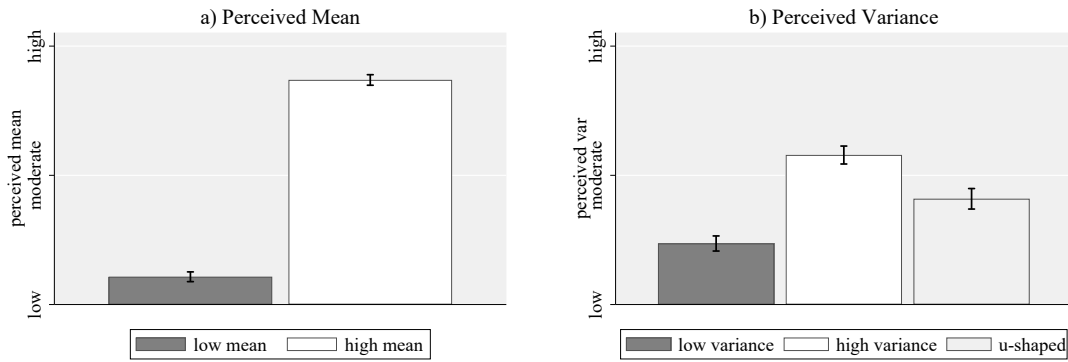
Note: Participants have to allocate a total of 10 points across all available options. The more likely they think an option is the more points they should allocate to it.

C.2.3 Sample and data collection

Manipulation checks

In the ex-post survey we ask participants about their perceptions of the shown distribution. As Figure C.3a shows, participants in high mean conditions also have a significantly higher perception of the mean (Wilcoxon-Mann-Whitney test, $p < 0.001$). In terms of the variance (Figure C.3b), participants perceive the low variance condition as significantly less varying than the high variance and the u-shaped conditions, and the u-shaped as less varying than the high variance condition (Wilcoxon-Mann-Whitney tests, for all $p < 0.001$). In addition, we ask participants about their difficulty in interpreting the distribution and how common they think this distribution is in the general population. On a scale from 1 (very easy) to 7 (very difficult) the average difficulty rating is 2.1, indicating that participants do not seem to have trouble interpreting the information. Moreover, there is no difference in the difficulty rating between high and low variance conditions or u-shaped and low variance conditions. Only the high variance condition is described as significantly harder to understand than both the u-shaped (Wilcoxon-Mann-Whitney tests, $p = 0.03$) or the low mean condition (Wilcoxon-Mann-Whitney tests, $p = 0.08$). However, the difference is very small (0.2 on a scale from 1 to 7). The average rating of how common the shown distributions are perceived to be is 5.0 on a scale from 1 (very rare) to 7 (very common). Again, there are no differences between high and low mean conditions or u-shaped and high/low variance conditions. Only the high variance condition is perceived as slightly less common than the low variance condition (Wilcoxon-Mann-Whitney tests, $p = 0.009$). Again, the difference is extremely small (0.2).

Figure C.3: Manipulation checks



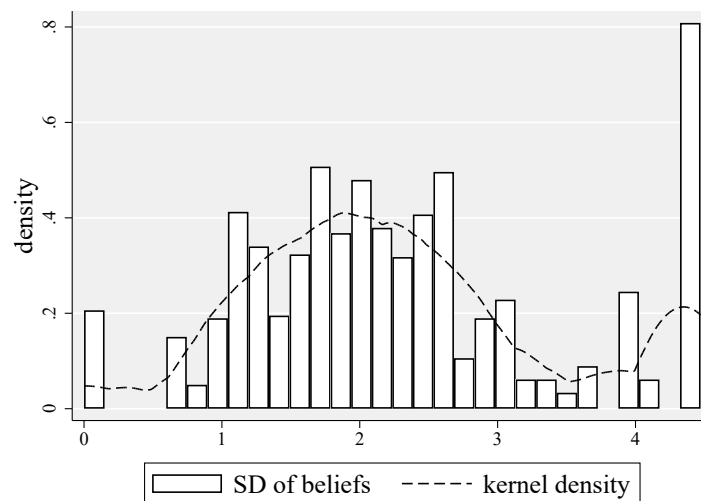
Note: Whiskers show 95% confidence intervals.

C.2.4 Behaviour in Part I

Variation of beliefs

Figure 3.5 in the main text shows the aggregate distribution of beliefs about the co-player's contribution in Part I, with most people believing that the other will either contribute 2 or 4 tokens. This aggregate distribution hides a substantial heterogeneity between participants in the distribution of beliefs. Figure C.4 shows that the standard deviation of initial beliefs varies substantially between participants (min=0, max=4.5).

Figure C.4: Variation in standard deviations of beliefs in Part I



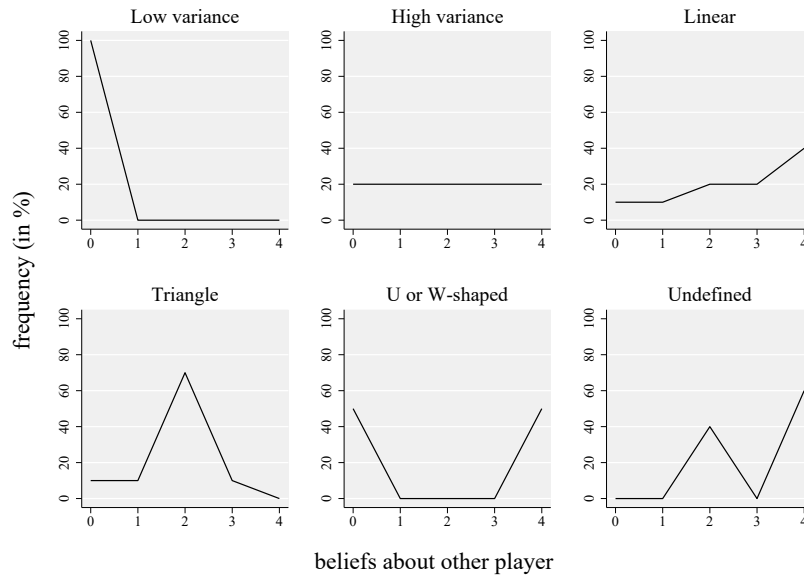
The individual distributions of beliefs can roughly be categorised into 6 types:

1. Low variance: participants who put more than 80% of points on one single option
2. High variance: participants who believe outcomes are equally likely (10–30% per option)
3. Linear: participants who have either non-decreasing or non-increasing beliefs

4. Triangle: participants with a single modal belief below 80% that is 1,2, or 3 token
5. U- or w-shaped: participants with either two modes (at 0,4) or three modes (at 0,2,4)
6. Others: not defined by the previous categories

Figure C.5 provides examples for each type. Using these rules we can classify 92% of participants. The most common types are triangles (41%), followed by low variance (21%), u-/w-shaped (15%), linear (10%), and high variance types (5%).

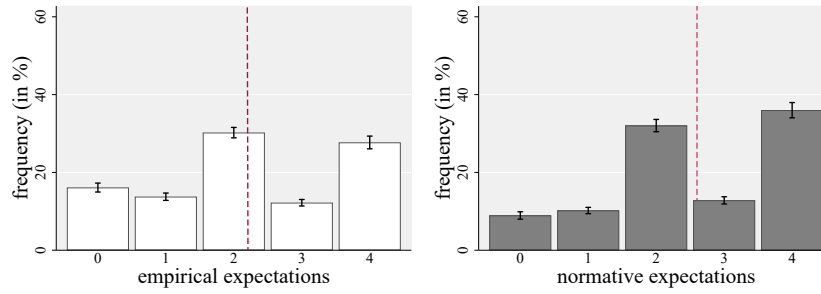
Figure C.5: Examples of individual distributions of beliefs in Part I



Empirical and normative expectations

Empirical (EE) and normative (NE) expectations in Part I — i.e. expectations about what most people do and what most people think one should do — are very similar to contributions, beliefs and PVs (see Figure C.6). Not surprisingly, when asked about what most other people actually contribute (EEs) participants answer in the same way as when asked about their co-player's likely contribution. NEs, are shifted slightly to the right, indicating that people think others contribute less than they say one should. In addition, to eliciting the distribution of expectations and beliefs, we measure the participants' confidence in their replies. Interestingly, participants also appear to be more certain about what others say one should contribute than actual contributions. Although the difference is small, our confidence measure is significantly higher for NEs than EEs about the other player (Wilcoxon signed-rank test, $p = 0.01$).

Figure C.6: EEs and NEs in Part I



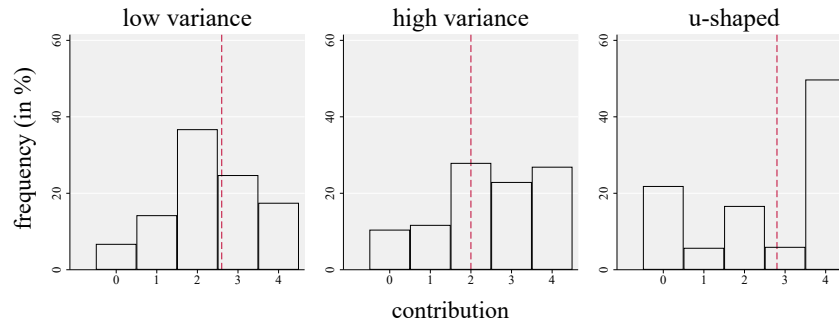
Note: The dashed lines represent averages. Whiskers show 95% confidence intervals.

C.2.5 Effect of variance and shape of the descriptive norm

Distribution of contributions by variance condition

Figure C.7 shows the distribution of contributions in Part II by variance/shape, but pooled over low and high mean conditions. As described in the main text, we can see that the variance of contributions is largest for the u-shaped conditions, followed by the loose/high variance conditions, while there is less variance in the tight/low variance conditions.

Figure C.7: Distribution of participants' contribution in Part II by variance

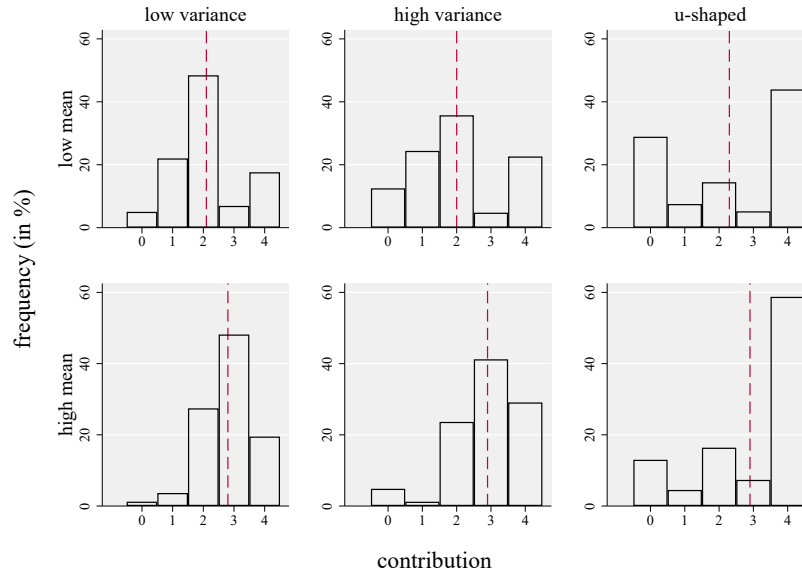


Distribution of contributions for conditional cooperators

As our theoretical framework assigns a particular role to reciprocity concerns and conditionality, analysing the effect of variance is particularly relevant for conditional contributors. As 84% of our sample are conditional cooperators, we do not analyse types separately in the main paper and report results for all participants. Figure C.8 shows that when only focusing on conditional cooperators, our results hold: tight, loose, and polarised environments induce tight, loose, and polarised responses. We also confirm that the variance of contributions in Part II among conditional cooperators in the u-shaped condition is significantly larger than in the high or low variance condition (F-tests, $p < 0.001$ respectively). The variance of contributions in the high variance conditions in turn is significantly larger than in the low variance conditions (F-test, $p < 0.001$). Also the distributions are significantly different from each

other (χ^2 -tests, $p \leq 0.002$ for all comparisons).

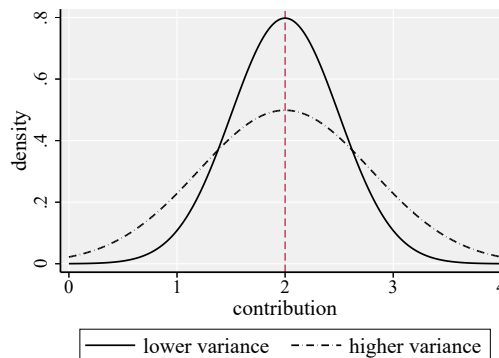
Figure C.8: Distribution of participants' contribution in Part II (conditional cooperators only)



Effect of variance on spread of contributions

A more indirect test of whether the variation in contribution increases with the observed variance is by splitting our sample in Part II in a high and a low contribution group (contributions above/below the median) and comparing their reactions to different environments. If a higher observed variation increases the overall variance in contributions, we would expect this to translate into lower contributions for the *low contribution sub-sample* and higher contributions for the *high contribution sub-sample*, resulting in an overall wider spread of contributions (see Figure C.9).

Figure C.9: Different variances and their effect on the high/low end of the distribution



Note: As the figure shows a higher variance implies more extreme contributions in both tails of the distribution

Table C.1 shows the results of two tobit models, where we regress contributions in Part II

on treatment conditions for each sub-sample. We can see that in fact in the high contribution sub-sample, both the high variance and the u-shaped conditions lead to a significant increase in contributions relative to the low variance condition. For the low contribution sub-sample, by contrast, we see the exact opposite pattern. In this case, a higher variance is associated with lower contributions. We thus confirm that both the u-shaped and the high variance condition increase the overall variation of contributions.

Table C.1: Effect of variance on contributions in Part II for different sub-samples

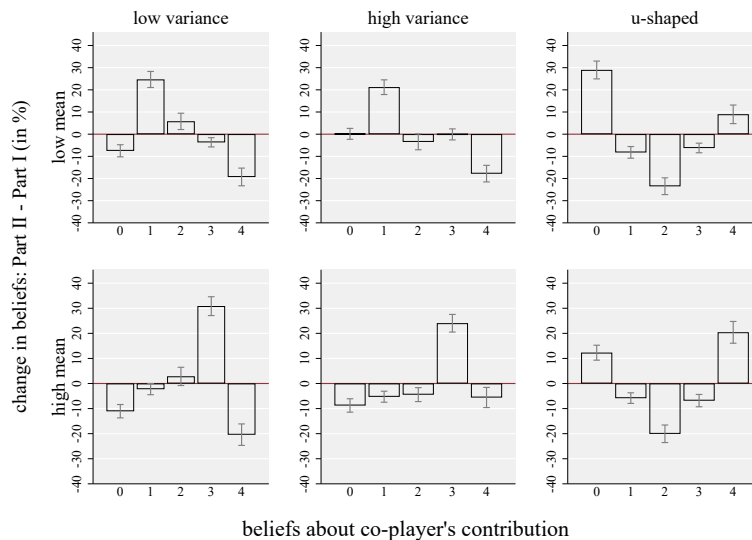
	(1) High contributions (>2)	(2) Low contributions (<=2)
<i>Variance (baseline = low)</i>		
High variance	0.33*** (0.09)	-0.14 (0.09)
U-shaped	1.20*** (0.10)	-0.77*** (0.10)
Constant	0.96** (0.44)	0.27 (0.42)
Baseline controls	Yes	Yes
Demographic controls	Yes	Yes
N observations	906	602
Pseudo R ²	0.19	0.13

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Note: Results of two tobit regressions (data is censored at 0 and 4). The dependent variable is *contributions in Part II*. *Variance* is a categorical variable with 0 = low variance, 1 = high variance and 2 = u-shaped. Baseline controls include contributions, average beliefs, and PVs in Part I and can take values between 0 and 4. Demographic controls include age, gender, education, acceptance of risk, trust sucker aversion, free-riding aversion and measures for negative and positive reciprocity. High and low contribution sub-samples are generated by dividing participants in a group with contributions in Part II above and below the median (=2).

Changes in beliefs

Figure C.10: Changes in beliefs about the co-player's distribution between parts



Note: Whiskers show 95% confidence intervals.

Figure C.10 shows changes in beliefs about their co-player's contribution between Part I and Part II. As we can see, participants change their beliefs in line with the shown distributions. For instance, for the u-shaped conditions, the beliefs that the other player contributed 0 or 4 tokens increase substantially between Part I and II of the experiment, while intermediate values (1-3) decrease. The opposite is true for the low and high variance conditions, where in line with the observed distribution, the beliefs that the other contributes 0 or 4 tokens decrease in favor of intermediate values. Intuitively, the additional information in Part II also explains that participants state a significantly higher confidence in their beliefs as compared to Part I (Wilcoxon signed-rank test, $p < 0.001$).

C.2.6 Personal traits

Robustness checks — OLS and ordered probits

Table C.2: Effect of personal traits on contributions in Part II — robustness checks

	OLS		Ordered probits	
	(1)	(2)	(3)	(4)
High mean	0.67*** (0.07)	0.67*** (0.07)	0.60*** (0.07)	0.62*** (0.07)
<i>Variance (baseline = low)</i>				
High variance	-0.22 (0.29)	-0.24 (0.29)	-0.26 (0.26)	-0.31 (0.27)
U-shaped	-0.11 (0.34)	-0.17 (0.33)	-0.21 (0.33)	-0.28 (0.34)
<i>Sucker aversion</i>	-0.10*** (0.03)	-0.10*** (0.03)	-0.10*** (0.03)	-0.10*** (0.03)
Sucker aversion x high variance	-0.01 (0.04)	0.01 (0.04)	-0.01 (0.04)	0.01 (0.04)
Sucker aversion x u-shaped	-0.09** (0.05)	-0.08* (0.05)	-0.13*** (0.05)	-0.12** (0.05)
<i>Free-riding aversion</i>	0.11*** (0.02)	0.11*** (0.02)	0.10*** (0.02)	0.10*** (0.02)
Free-rider aversion x high variance	0.01 (0.04)	0.00 (0.04)	0.01 (0.03)	0.00 (0.03)
Free-rider aversion x u-shaped	0.11*** (0.04)	0.11*** (0.04)	0.14*** (0.04)	0.14*** (0.04)
<i>Personal values in part I</i>	0.34*** (0.05)	0.33*** (0.05)	0.30*** (0.04)	0.30*** (0.04)
High variance x PVs in part I	0.13* (0.07)	0.13* (0.07)	0.14** (0.07)	0.15** (0.07)
PVs in part I x u-shaped	0.11 (0.08)	0.11 (0.08)	0.19** (0.08)	0.20** (0.09)
Demographic controls	No	Yes	No	Yes
N observations	1203	1188	1203	1188
(Pseudo) R ²	0.30	0.32	0.12	0.13

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Note: The dependent variable is *contributions in Part II* and takes values between 0 and 4. *High mean* is a binary variable with 0 = low and 1 = high mean. *Variance* is a categorical variable with 0 = low variance, 1 = high variance and 2 = u-shaped. *Personal values* can take values between 0 and 4. Sucker and free-riding aversion are measured on a Likert scale from 1 to 7. Demographic controls include age, gender, education, acceptance of risk, trust, and measures for negative and positive reciprocity. All regressions control for order effects. Columns (1) and (2) present results from OLS regressions, columns (3) and (4) from ordered probits.

Table C.2 shows that the result that personal traits are more predictive of contributions in environments with a higher variance is also confirmed when using alternative models to a tobit regression such as OLS or ordered probit models. The influence of sucker and free-riding aversion on contributions is significantly higher for the u-shaped than the low variance environment in all models. Personal values have a larger impact in both high variance and u-shaped environments (even though the latter is only significant for the ordered probit models).

Role of personal traits for conditional cooperators

Table C.3 shows that the analysis of personal traits from section 4.3 holds when restricting the sample to participants classified as conditional cooperators. We still see that sucker-aversion lowers contributions, while free-riding aversion and personal values have a positive effect on contributions. Most importantly, we confirm that the effect of these personal traits is larger

Table C.3: Effect of personal traits on contributions in Part II — conditional cooperators only

	(1)	(2)
High mean	1.00*** (0.13)	1.01*** (0.12)
<i>Variance (baseline = low)</i>		
High variance	-1.15* (0.60)	-1.20** (0.59)
U-shaped	-0.85 (0.61)	-0.86 (0.60)
Sucker aversion	-0.20*** (0.06)	-0.19*** (0.06)
Sucker aversion x high variance	0.03 (0.09)	0.06 (0.09)
Sucker aversion x u-shaped	-0.22** (0.09)	-0.21** (0.09)
Free-riding aversion	0.12** (0.05)	0.12** (0.05)
Free-riding aversion x high variance	0.08 (0.08)	0.07 (0.08)
Free-riding aversion x u-shaped	0.28*** (0.08)	0.29*** (0.08)
<i>Personal values (PVs)</i>	0.43*** (0.10)	0.43*** (0.10)
PVs x high variance	0.25* (0.14)	0.25* (0.14)
PVs x u-shaped	0.39*** (0.14)	0.40*** (0.14)
Constant	1.35*** (0.44)	-0.37 (0.76)
Demographic controls	No	Yes
N observations	1006	994
Pseudo R ²	0.09	0.11

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Note: Results of two tobit regressions (data is censored at 0 and 4). The dependent variable is *contributions in Part II*. *High mean* is a binary variable with 0 = low and 1 = high mean. *Variance* is a categorical variable with 0 = low variance, 1 = high variance and 2 = u-shaped. *Personal values* can take values between 0 and 4. Sucker and free-riding aversion are measured on a Likert scale from 1 to 7. Demographic controls include age, gender, education, acceptance of risk, trust, and measures for negative and positive reciprocity. All regressions control for order effects.

in conditions with higher variance (loose/polarised).

Controlling for changes in personal values

We measure personal values both in Part I before participants receive any information about co-player behaviour and in Part II after they see the distributions. In the main text, we use personal values in Part I as a regressor as they cannot be affected by the treatments and thus represent a cleaner representation of what participants think is the “right thing to do”. However, Table C.4 shows that our results hold after controlling for the change in personal values

Table C.4: Effect of personal traits on contributions in Part II – controlling for changes in PVs

	(1)	(2)
High mean	0.63*** (0.11)	0.66*** (0.11)
<i>Variance (baseline = low)</i>		
High variance	-0.41 (0.50)	-0.53 (0.50)
U-shaped	-0.86 (0.54)	-0.94* (0.53)
Sucker aversion	-0.13** (0.05)	-0.12** (0.05)
Sucker aversion x high variance	-0.06 (0.08)	-0.02 (0.07)
Sucker aversion x u-shaped	-0.24*** (0.08)	-0.22*** (0.08)
Free-riding aversion	0.13*** (0.05)	0.14*** (0.05)
Free-rider aversion x high variance	0.05 (0.07)	0.04 (0.07)
Free-rider aversion x u-shaped	0.28*** (0.07)	0.28*** (0.07)
<i>Personal values in part I</i>	0.75*** (0.10)	0.71*** (0.10)
High variance x PVs in part I	0.23* (0.13)	0.26** (0.13)
PVs in part I x u-shaped	0.41*** (0.14)	0.44*** (0.14)
<i>Change in personal values</i>	0.56*** (0.11)	0.51*** (0.11)
Change in PVs x high variance	0.22 (0.17)	0.24 (0.16)
Change in PVs x u-shaped	0.27* (0.14)	0.31** (0.14)
Constant	0.23 (0.37)	-1.55** (0.66)
Demographic controls	No	Yes
N observations	1203	1188
Pseudo R ²	0.14	0.15

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Note: Results of two tobit regressions (data is censored at 0 and 4). The dependent variable is *contributions in Part II*. *High mean* is a binary variable with 0 = low and 1 = high mean. *Variance* is a categorical variable with 0 = low variance, 1 = high variance and 2 = u-shaped. *Personal values* are measured in Part I and can take values between 0 and 4. Sucker and free-riding aversion are measured on a Likert scale from 1 to 7. *Change in personal values* is the difference between PVs in Part II and I and takes values between -4 and 4. Demographic controls include age, gender, education, acceptance of risk, trust, and measures for negative and positive reciprocity. All regressions control for order effects.

between parts and their interaction with treatment indicators. Note that changes in personal values are very minor. On average participants change their personal values by -0.1 (with personal values taking values between 0 and 4). Although this is statistically different from zero (Wilcoxon signed-rank test, $p < 0.001$), the median change is 0, with 69% of participants not changing their reported personal values between parts.

Personal traits and demographics

Finally, we find that some demographic variables correlate with personal traits. In Table C.5 we regress personal values in column (1), sucker aversion in column (2), and free-riding aversion in column (3) on demographic characteristics. We find that women and younger participants have both a higher sucker and a higher free-riding aversion. Personal values on the other hand are slightly lower for female participants.

Table C.5: Correlates of personal values, sucker and free-riding aversion

	Personal values (1)	Sucker aversion (2)	Free-riding aversion (3)
Gender (female = 1)	-0.144** (0.066)	0.594*** (0.152)	1.152*** (0.193)
Age	-0.000 (0.002)	-0.034*** (0.005)	-0.016*** (0.006)
<i>Education (baseline=no formal degree)</i>			
Secondary school	-0.076 (0.230)	-0.634 (0.542)	-0.621 (0.670)
University/ college	-0.044 (0.226)	-0.712 (0.533)	-1.129* (0.657)
Prefer not to say	-0.676 (0.411)	-0.451 (0.933)	-1.675 (1.179)
Constant	3.017*** (0.238)	6.520*** (0.561)	5.133*** (0.692)
N observations	1188	1188	1188
Pseudo R ²	0.00	0.01	0.01

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Note: Results of three tobit regressions. The dependent variable in column (1) is *personal values* in Part I and is censored at 0 and 4. The dependent variables in column (2) and (3) are *sucker* and *free-riding aversion* and are measured on a Likert scale from 1 to 7.

C.3 Quadratic scoring rule

The quadratic scoring rule (QSR) is used to elicit participants' beliefs about uncertain events in an incentive compatible way. As a consequence participants should reveal their beliefs truthfully.³³ The key idea of the QSR is that participants are rewarded for the correct assessment, but penalised for every wrong assessment by the squared distance between their guess and the true event, using the following equation (Murphy & Winkler, 1970):

$$Q_j(p) = \alpha + 2\beta p_j - \beta \sum_{i=1}^n (p_i)^2. \quad (\text{C.6})$$

where p_i is the probability an individual assigns to event i and p_j is the probability an individual assigns to the true event. Artinger et al. (2010) develop and test a salient and intuitive way to communicate information about payoffs to participants using a decomposed version of the equation above:

$$Q_j(p) = (\alpha + \beta) - \beta(1 - p_j)^2 - \beta \sum_{i \neq j} (p_i)^2, \quad (\text{C.7})$$

The first two terms in Equation (C.7) represent the earnings from the correct option. Participants are penalised for not assigning a probability of 1 to the correct event ($-\beta(1 - p_j)^2$). In the experiment we set $\alpha = \beta = \$0.5$. This implies that if an individual assigns full probability to the correct event they earn \$1, if they assign a zero probability to the correct event they earn \$0.5 from the first term in Equation (C.7). The last term in the equation above represents the penalty for assigning a positive probability to events that do not occur. If individuals assign zero probability to untrue events, this penalty is zero. If they assign all probabilities to wrong events, the penalty is \$0.50. This means that the maximum amount individuals can earn in the elicitation task is \$1, while the minimum is \$0.

Artinger et al. (2010) show that if payoffs are split up in this way, this significantly facilitates participants' understanding of the QSR and the implications for payoffs. Figure C.11 in the instructions shows how the QSR is presented to participants.

³³Note that the QSR may not be valid for all degrees of risk aversion (Schlag & van der Weele, 2015). However, for most empirically plausible levels of risk aversion, it seems to provide a reliable tool to elicit preferences (Harrison et al., 2017).

C.4 Instructions

Welcome

Thank you very much for participating in this study! This study consists of two parts and a questionnaire. Upon completion you will receive \$1.70 for your participation plus an additional bonus of up to \$2.36 that depends both on your decisions and the decisions of other participants.

In both parts you will face a situation in which you will be matched with one other, real participant. On the next page we will describe this situation to you in more detail.

Instructions (1/2)

In this study, you will be anonymously paired with another participant. You will each start with **4 token** in your personal **private accounts**. In addition to the private accounts, there is a **group account**. You have to decide how many of your token you want to invest in the group account (either 0, 1, 2, 3 or 4 token). The amount leftover will remain in your private account. The other player has to make the same decision.

Your income from the private account

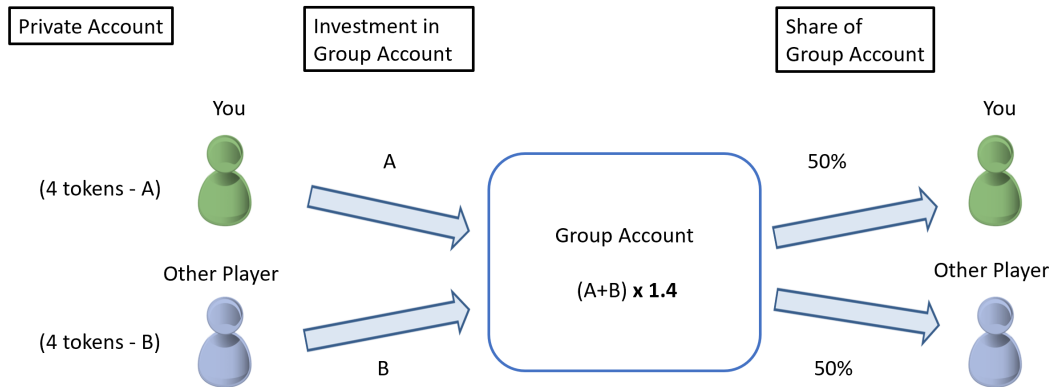
The amount in the private account is yours to keep. The other player doesn't earn anything from the token you keep in your private account. For example, if you keep 2 token in your private account, this will be your income from this account.

Your income from the group account

The amount invested in the group account will be multiplied by 1.4. That is, each token invested in the group account will yield 1.4 token for the group. The total amount in the group account will be split equally between you and your partner regardless of your individual investments. That is, each player receives half (50%) of the total amount in the group account.

Figure C.11: Representation of QSR

Points put on the correct option	Earnings from the correct option	Points put on a wrong option	Costs from a wrong option
10	100¢	10	50¢
9	99.5¢	9	40.5¢
8	98¢	8	32¢
7	95.5¢	7	24.5¢
6	92¢	6	18¢
5	87.5¢	5	12.5¢
4	82¢	4	8¢
3	75.5¢	3	4.5¢
2	68¢	2	2¢
1	59.5¢	1	0.5¢
0	50¢	0	0¢



If, for example, the sum of all investments in the group account by you and the other player is 6 token ($A+B$), then the group account yields $6 \times 1.4 = 8.4$ token. Both you and the other player would then receive $0.5 \times 8.4 = 4.2$ token from this account.

Your total income

Please note that for logistic reasons you are not interacting with other participants in real time. Once we collected all responses, we will match you with another person to calculate your and the other's total income.³⁴ The latter consists of **all the token you kept in your private account plus half of the token that you and that other participant invested in the group account.**

Your total income will then determine your bonus payment, with each token being worth \$0.10. Whatever income you earned in token will be converted at this rate into actual money at the end of the experiment and paid out as a bonus.

Instructions (2/2)

In addition to making an investment decision in this situation, we will ask you to state **your beliefs about other participants**. You will be paid for these tasks according to how accurate your beliefs are.

A brief explanation follows: let us assume, we ask you to make a guess about how many token the participant you have been matched with invested in the group account. In this case, you would have to indicate **how likely you think it is that the other participant invested 0, 1, 2, 3 or 4 token**. To make your choices you will see a screen like the one below.

³⁴Note that this is a common set-up for studies on Prolific and thus familiar to participants. The Prolific guidelines allow up to 21 days to pay participants.

0 token	1 token	2 token	3 token	4 token
				
+	+	+	+	+
-	-	-	-	-

To make your decision you have to allocate a total of 10 points across options by clicking on the plus and minus buttons. The points you allocate need to add up to 10 and **the more likely you think one option is, the more points you would allocate to it**. The points you allocate to each option will naturally reflect your beliefs about the other participant's behaviour.

The amount of money you can earn depends on how you allocated your points and what is actually true. If you put all points on the correct option, you will earn \$1 if you put all points on a wrong option you will earn \$0. In general, the more points you allocate to a correct option, the higher your earnings and the more points you allocate to a wrong option the lower your earnings. **The way your earnings are determined ensures that your best strategy is to carefully and honestly answer these questions.** If you want to have a closer look at how your earnings will be calculated click [here](#).

Let's for example assume that you think it is equally likely that the other participant invested 2 or 4 token and you put 5 points on each option. If the other participant really invested either 2 or 4 token, you would in each case earn \$0.75. If they invested 0, 1 or 3 token you would earn \$0.

What if you had instead put all your eggs in one basket and allocated 10 points on the other participant investing 2 token? If the other participant indeed invested 2 token, you earn the maximum bonus of \$1. But if any of the other options is the correct one, you would earn nothing in this task. It is thus up to you to balance the strength of your personal beliefs with the risk of them being wrong.

In total, we will ask you to state your belief on **five** different questions throughout this study. In the end, a lottery will decide **one of them to be chosen for payment**. The amount you earned in the chosen question will then be added to your bonus payment.

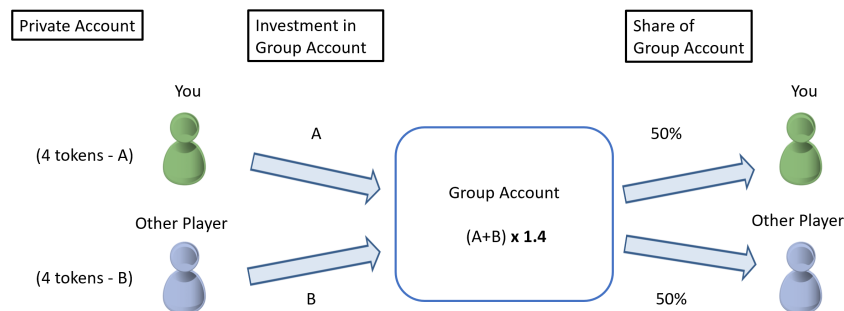
If participants click to get more information about payoffs, they see the following pop-up:

Your earnings are calculated on the basis of the table below. The more points you put on the correct option the higher your earnings. For each wrong option to which you allocate points your earnings will be reduced. The reduction is larger the more points you allocated to that option.

Points put on the correct option	Earnings from the correct option	Points put on a wrong option	Costs from a wrong option
10	100¢	10	50¢
9	99.5¢	9	40.5¢
8	98¢	8	32¢
7	95.5¢	7	24.5¢
6	92¢	6	18¢
5	87.5¢	5	12.5¢
4	82¢	4	8¢
3	75.5¢	3	4.5¢
2	68¢	2	2¢
1	59.5¢	1	0.5¢
0	50¢	0	0¢

Part 1

We are now going to ask you a number of questions that relate to the situation that you previously read (see image below). It is important that you answer these questions truthfully and as accurately as possible.³⁵



1) We asked other participants what they believe is the **most appropriate amount to invest** in the **group account**. What do you believe was the most common answer?

³⁵Either normative questions or ABC questions are asked first. The three normative questions appear in randomised order.

Appropriate here means what you personally consider to be “correct” or “moral”. You can allocate up to 10 points to an option by clicking on the plus and minus buttons below to indicate your guess. The more likely you think one option is, the more points you would allocate to it. Your guesses need to add up to 10.

Most people believe it is appropriate to invest...

0 token	1 token	2 token	3 token	4 token
				
				
				

How confident are you in your response above? (0 not very confident, 100 very confident)

2) We asked other participants to make an investment decision in this situation. How many token do you believe most people **actually invested** in the group account?

You can allocate up to 10 points to an option by clicking on the plus and minus buttons below to indicate your guess.

The more likely you think one option is, the more points you would allocate to it. Your guesses need to add up to 10.

Most people actually invested...

0 token	1 token	2 token	3 token	4 token
				
				
				

How confident are you in your response above? (0 not very confident, 100 very confident)

3) According to your own opinion and independent of the opinion of others, what is the **most appropriate amount to invest** in the group account?

Appropriate here means what you personally consider to be “correct” or “moral”.

- 0 token
- 1 token
- 2 token
- 3 token
- 4 token

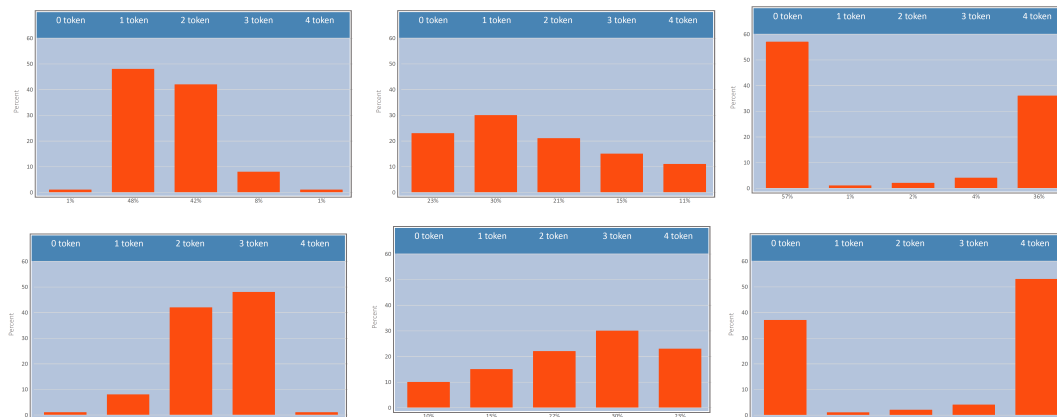
- The other player invests 0 token: _____ token
- The other player invests 1 token: _____ token
- The other player invests 2 token: _____ token
- The other player invests 3 token: _____ token
- The other player invests 4 token : _____ token

Part 2

In a previous study we asked over 600 participants to **make an investment decision** in the same situation. The possible choices were to invest 0, 1, 2, 3 or 4 token in the **group account**. From their answers we constructed different sub-groups. The graph below shows the percentage of people choosing each option in one randomly selected sub-group.

What previous participants invested in the group account:

(Participants are randomly shown one of the following six pictures.)



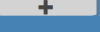
For Part 2 of the experiment you are matched with **one of the participants from the sub-group above**. You will only interact once with this person and you will never learn each other's identity.

Your and the other participant's **bonus payment for Part 2** will depend on your decisions and the decisions of this participant.³⁶

1) We asked other participants from the previous study what they believe is the **most appropriate amount to invest** in the **group account**. What do you believe was the most common answer?

Appropriate here means what you personally consider to be "correct" or "moral". You can allocate up to 10 points to an

³⁶We randomise whether participants are first asked about contributions and beliefs or about personal values and normative expectations.

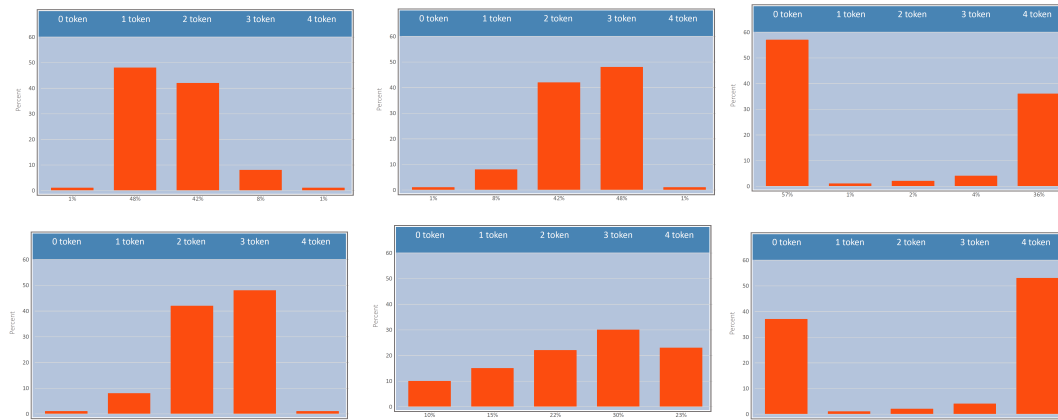
0 token	1 token	2 token	3 token	4 token
				
				
				

How confident are you in your response above? (0 not very confident, 100 very confident)

Questionnaire (1/2)

In this survey we showed you how many token a sub-group of other participants **invested in the group account**. Their answers are represented by the graph below.

What previous participants invested in the group account: *(Participants are randomly shown one of the following six pictures.)*



We will now ask you a few questions about the graph.

1) What are your thoughts on the behaviour shown above?

2) Would you say the graph shows that overall

- people invest most of their token in the group account
- people invest half of their token in the group account
- people keep most of their token in their private account

3) Would you say the graph shows that overall

- there is a strong tendency for people to invest similar amounts in the group account
- there is a moderate tendency for people to invest similar amounts in the group account
- investments in the group account are very mixed

4) How common do you think the distribution of behaviour shown above would be in other groups? (1 very rare, 7 very common)

5) How difficult was it for you to interpret the graph in Part 2, which is also shown above? (1 very easy, 7 very difficult)

6) How upset would you be if you invested everything in the group account and discovered that the participant you have been matched with invested nothing? (1 not at all upset, 7 very upset)

7) How ashamed would you be if you invested nothing in the group account and discovered that the participant you have been matched with invested everything? (1 not at all ashamed, 7 very ashamed)

Questionnaire (2/2)

1) Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?

- Need to be very careful
- Don't know
- Most people can be trusted

2) In how far do you agree with the following statement: "When someone does me a favour, I will return it." (1 don't agree at all, 7 completely agree)

3) In how far do you agree with the following statement: "If I am treated very unjustly, I will take revenge, even if there is a cost to do so." (1 don't agree at all, 7 completely agree)

4) Please tell me, in general, how willing or unwilling you are to take risks. (1 very unwilling to take risks, 11 very willing to take risks)

5) What is your age?

6) Which gender do you identify with?

- Female
- Male
- Non-binary
- Other
- Prefer not to say

7) What is the highest level of schooling you completed?

- No formal qualifications
- Secondary school
- University/ college degree
- Prefer not to say

Thanks a lot for participating in this survey! If you have any feedback for us you can write it here:
