Aspects of Pro-Social Behaviour
Theory and Experiments

Vasileios Kotsidis

University of Nottingham

A thesis submitted for the degree of

Doctor of Philosophy

Nottingham 2018

This thesis is dedicated to the wonderful
people who have accompanied me on this journey
and are the sole reason why i managed to traverse it in
relative sanity.

Acknowledgements

Throughout the four-year process that culminated in this thesis i have had the pleasure and privilege to interact with a number of people. This work is the result of these interactions, and their invaluable and unwavering support, that has sustained me to this point.

One of the most decisive factors in the completion of this project is, of course, the help and guidance i have received from my supervisors, Silvia Sonderegger and Daniele Nosenzo. Their wisdom, enthusiasm and patience in guiding and assisting me have been vital and truly inspiring. Working with them has been a blessing and i am immensely grateful.

I am grateful to the Centre for Decision Research and Experimental Economics (CeDEx) and to the Economic and Social Research Council (ESRC) for their financial support during the period of my research, and to CeDEx for providing the resources and facilities for running experiments.

My gratitude also extends to Robin Cubit, Daniel Seidmann, Chris Starmer, Simon Gäachter, Suzanne Robey, and the rest of the CeDEx community for their insightful suggestions at various stages of this research. I also thank participants to the conferences where i have presented my work for their helpful comments.

Finally, this work (or i, for that matter) would not exist were it not for the continual encouragement and support i have received from my friends. I am therefore deeply grateful to Lionel Roger, Martina Magli, Thomas Mutton, Joshua Davis, John Burton, Despoina Alempaki, Xueheng Li, Hanna Fromell, Marta Ronchetti, Arno Hantzsche, Till Weber, Basile Boulay, Orestis Kopsacheilis, Valeria Bourdea, Mayte Quintana, Roel Dom, Georgia Michailidou, Yuzhu Zhang, Miriam Saldana, Antonio Carillo, and Zhixian Yu. I am also deeply grateful to Giorgos Kladogenis, Ariadne Peraki, Bruno Bessi, Yannis Vaggelokostas, Dora Konstantinidou and Christos Kalogridakis, who have been a true family to me from afar. In addition, i thank the students i have had the chance to tutor and my own old professor, Christos Constantatos, who started it all. Needless to say, i am also deeply grateful to my family for their unconditional love and understanding.

Abstract

Chapter 1 introduces the work, providing an overview of the common themes underlying the research and outlining the focus and approach particular to each project.

Chapter 2 proposes a game-theoretic model that shows how moral preferences can emerge endogenously to promote material outcomes and traces their relationships with the fundamentals of the environment. The analysis indicates that the instilling of moral values can act as a commitment mechanism that counteracts the detrimental effects of behavioural biases. The greater the effect of such biases on the agents' decisions (and, thus, payoffs), the more expanded the scope for morality.

The study in chapter 3 tests the performance of a leading account of social preferences, namely the model of inequality aversion proposed by Fehr and Schmidt (1999), in tracking behaviour. It does so through an appropriately designed experiment. The aim is to evaluate if the account can consistently anticipate people's behaviour. The results suggest that the model performs well only with respect to people that exhibit either very high or very low aversion to advantageous payoff inequality.

The study in chapter 4 repeats the exercise reported in chapter 3, this time with respect to an account of social preferences that builds on the idea of social norm compliance, in particular, the one proposed by Krupka and Weber (2013). The aim is again to evaluate if the model performs well in consistently tracking people's behaviour. The results do not offer much support for the explanatory power of the model. The individuals that exhibit the least concern about adhering to social norms and are choosing the payoff-maximising options are the only ones the actions of whom match the model's predictions.

Chapter 5 summarises the findings of this thesis and concludes.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General introduction

This thesis is a collection of three chapters, which report studies that contribute to research in game theory and experimental economics. Chapter 2 is entirely self contained and can be read independently of the other two. Chapters 3 and 4 are linked by section 3.3, but otherwise they are also self-contained. All three of them, however, investigate different aspects of the same subject matter, namely pro-social behaviour, and can, as such, be viewed within a unified framework.

The focus on the overarching theme of pro-social behaviour is motivated by a large and expanding literature of experimental evidence on strategic decision-making. While traditional game-theoretic accounts rely on the assumption that players are solely concerned about their own material payoffs, the choices of people in appropriately designed laboratory experiments reveal that substantial proportions of them are willing to go against their material interests, in order to uphold some social principles, such as fairness, reciprocity, and altruism. In order to account for these behavioural patterns within the framework of rational choice, economists have proposed a number of models of social preferences (see, e.g., Camerer, 2003; Fehr and Schmidt, 2003, 2006; Gächter, 2007 for overviews of the experimental data and the accounts proposed). On a parallel development, the experimental findings have fuelled the discussion on the foundations of rational-choice theory (see, e.g., Stigler and Becker, 1977; Hollis and Sugden, 1993; Dietrich and List, 2013; and the debate between Binmore and Shaked, 2010, on the one hand, and Fehr and Schmidt, 2010, and Eckel and Gintis, 2010, on the other).

Some concerns that are commonly expressed in this discourse relate to the properties of preferences that are not exclusively expressed over one's own material payoff. The three studies reported in this thesis contribute to the dialogue in two distinct ways. The first is the examination of some conditions under which non-material preferences may arise *in addition* to purely materialist concerns and the implications of their emergence for public-policy design. The second is the evaluation of the performance of two different models of social preferences in accounting for people's behaviour.

The model proposed in Chapter 2 relates to the first of these two lines of inquiry. It demonstrates that non-materialist preferences may in fact be beneficial from a materialist point of view, if they are used to countervail a pre-existing bias. In doing so, it combines insights from different strands of the game-theoretic literature, as well as notions related to the psychology of decision-making. More specifically, it studies a process of preference indoctrination in an intertemporal-choice setting, where there is a discrepancy between the agents' discount factors. This discrepancy is caused by present-bias, a tendency to overweight present consequences relative to future ones (see e.g. Ainslie, 1975, 1992; Laibson, 1997). The concept of present bias is particularly appealing, because it can be shown to have an evolutionary rationale (using a mechanism similar to that of Samuelson and Swinkels, 2006). The character and degree of the resulting non-materialist preferences are tied to the objective conditions of the environment. Thus, the setup yields important implications for the design of public policies that aim to affect these preferences.

Chapters 3 and 4 report experiments that are designed to investigate the performance of two different accounts involving pro-social preferences in accurately tracking behaviour across a series of settings. This is a matter of preference consistency, so long as preferences have been correctly identified (which is an issue for each model itself). Consistency here requires that every preference-ordering of the various alternatives made by the decision-maker uses the same version of a parametrised model. Thus, preferences are time-invariant and independent of irrelevant alternatives. Intuitively, a model of social behaviour will provide meaningful predictions about an agent's social behaviour to the extent that the agent's social sensitivities, as defined by the model, remain stable or, at the very least, their variation is accounted for.

Models with other-regarding preferences have been shown to be capable of organising the behavioural regularities commonly observed in many laboratory experiments well (see Fehr and Schmidt, 2006 for a review). However, their ability to track individual behaviour across different settings is questionable at best (see e.g. Blanco et al., 2011; Bruhin et al., 2016). The two experimental studies reported in this thesis address precisely this question, using a design that allows for a more accurate distinction between social preferences and strategic considerations. The models that are being evaluated have been shown to be very effective in accounting for aggregate behavioural patterns in many stylised games and are, thus, good candidates for the 'stricter' test of within-subject consistency. The first is the account of inequality aversion proposed by Fehr and Schmidt (1999). It postulates that, in addition to their personal material payoffs, people prefer, to idiosyncratic extents, equitable distributions of payoffs to non-equitable ones. The second is the account of social-norm adherence championed by Krupka and Weber (2013). It posits that people care about their own material payoffs and the degree to which their actions are deemed socially appropriate.

Two crucial differences between these two models are important to notice at the outset. The Krupka-Weber model allows for a more general class of social maxims (other than payoff-equality) and for setting-specific classifications of normative behaviour (by allowing the relative influence of different norms to vary across settings). The Fehr-Schmidt model is more restrictive in both these dimensions, but, accordingly, it is more specific and imposes fewer epistemic requirements. The focus here lies on whether either (or both) of these two accounts is able to trace individual behaviour through a series of different games, in the absence of strategic considerations related to other people's choices. If a model exhibits consistently high performance in doing so, this constitutes evidence that it captures some of the principles underlying behaviour accurately.

1.2 Thesis outline

Chapter 2, titled 'Endogenous moral preferences - A simple theoretical analysis', reports a theoretical account of endogenous preference formation within a framework of Parent-Child interaction. Parents are assumed to care solely about the material welfare of themselves and that of their children. Their preferences are time-consistent. The children's preferences, on the other hand, are characterised by present bias, a tendency to overweight present events relative to future ones. Each parent can, at a personal cost, instil a direct preference for a particular type of behaviour into her child's preferences. The analysis demonstrates that in this setting even fully materialist parents may optimally endow their children with preferences for certain behaviours. The study explores the relationship between such preferences and the parameters in the environment, and enhances the analysis by introducing a stochastic component. The results have interesting implications for the design of public policy. The design can also be applied to intertemporal-choice problems of single individuals, under the interpretation of habit formation.

Chapter 3, titled 'Endogenous moral preferences - The case of aversion to advantageous inequality', reports an experimental study designed to evaluate the performance of the Fehr-Schmidt (1999) model of inequality aversion. The subjects are asked to participate in a series of games that do not involve strategic uncertainty, in the sense that they are aware of all the actions taken by the other players upon making their decisions. With this design it is possible to isolate the effect of their preferences on their behaviour, since strategic considerations are removed. The study elicits the individual-specific parameters of advantageous-inequality aversion (guilt) based on their decisions in the first game (a variant of the dictator game). It then uses the model to predict their behaviour in two other games (a trust and a lying game). The results indicate that the performance of the model in predicting people's behaviour varies considerably with the strength of their preferences. That is, it performs significantly better with respect to the people who exhibit either very high or very low aversion to advantageous payoff inequality. It appears that particularly selfish and egalitarian types behave consistently so throughout, whereas people with moderate concerns about payoff inequality appear confused with respect to their preferences.

Chapter 4, titled 'The curious case of the rational homo sociologicus - Consistency of normative preferences', examines social behaviour from a socially normative perspective. People's strategic decisions appear sensitive to changes in the environment within which they are expressed. One way to account for such dependencies is to postulate that individuals are intrinsically driven to comply with some socially determined rules, the relative prevalence of which differs across settings. This study evaluates the ability of one such account, namely that proposed by Krupka and Weber (2013), to consistently track behaviour. Their model is tested using the data from the experiment in chapter 3, along with some additional data that are particular to this investigation. The results offer little support for the predictive power of the model. Individual sensitivities towards norm compliance vary substantially across the three games. In addition, the results obtained in situations where different norms are in conflict differ markedly from those observed in situations where a single norm prevails. Contrary to the narrative of the model, it appears that some people adhere to specific ideals, which they hold on to even in situations where doing so is considered socially inappropriate. The rest, for the most part, exhibit non-stable degrees of sensitivity towards norm compliance.

Finally, Chapter 5 summarises of the main points from Chapter 2 and the results of Chapters 3 and 4. It concludes by pointing out some limitations of the analysis and suggesting avenues for future research.

1.3 References

Ainslie, G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin, 82*(4), 463.

Ainslie, G. (1992). *Picoeconomics: The strategic interaction of succes-*

sive motivational states within the person. Cambridge University Press.

Binmore, K., & Shaked, A. (2010). Experimental economics: Where next?. *Journal of Economic Behavior & Organization, 73*(1), 87-100. Chicago

Binmore, K., & Shaked, A. (2010). Experimental Economics: Where Next? Rejoinder. *Journal of Economic Behavior & Organization, 73*(1), 120-121.

Blanco, M., Engelmann, D., & Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior, 72*(2), 321-338.

Bruhin, A., Fehr, E., & Schunk, D. (2016). *The Many Faces of Human Sociality: Uncovering the Distribution and Stability of Social Preferences* (No. 5744). CESifo Group Munich.

Dietrich, F., & List, C. (2013). A Reason-Based Theory of Rational Choice. *Nous, 47*(1), 104-134. Chicago

Eckel, C., & Gintis, H. (2010). Blaming the messenger: Notes on the current state of experimental economics. *Journal of Economic Behavior & Organization, 73*(1), 109-119.

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

Fehr, E., & Schmidt, K. M. (2003). Theories of fairness and reciprocity-evidence and economic applications. *Advances in Economics and Econometrics*, p.208-257.

Fehr, E., & Schmidt, K. M. (2010). On inequity aversion: A reply to Binmore and Shaked. *Journal of Economic Behavior & Organization,* *73*(1), 101-108.

Gächter, S. (2007). Conditional cooperation: Behavioral regularities from the lab and the field and their policy implications. na.

Hollis, M., & Sugden, R. (1993). Rationality in action. *Mind, 102*(405), 1-35. Chicago

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.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics, 112*(2), 443-478.

Stigler, G. J., & Becker, G. S. (1977). De gustibus non est disputandum. *The American Economic Review, 67*(2), 76-90.

Swinkels, J. M., & Samuelson, L. (2006). Information, evolution and utility. *Theoretical Economics, 1*(1), 119-142. Chicago

Chapter 2

Endogenous moral preferences - A simple theoretical analysis

2.1 Introduction

Standard economic theory postulates that preferences are given and immutable. Hobbes prompts us to think of humans as if they were mushrooms, attaining full development prior to engaging in any form of interaction with each other (Hobbes, 1949). His position has been widely adhered to by traditional economic approaches. In the view of Stigler and Becker (1977) tastes tend to be relatively stable and qualitatively similar across people. As such, they are prone to being considered as constant in the analysis of economic behaviour. This view of preferences can lead to important insights into the causal mechanisms underlying behaviour.

However, the conception of stable, universal preferences is becoming increasingly challenged in the economics literature. Bowles remarks that thinking of preferences in this way does result in the simplification of the task facing economists, but also compromises economic analysis in terms of explanatory power, relevance, and ethical consistency (Bowles, 1998). Indeed, to the extent that preferences are, even partly, affected by the environment where the individuals live and interact, the implications for economic theory and the design of public policy can be quite significant.

Today there are many game-theoretic accounts of endogenous preference formation. Examples include the evolution of homo moralis (Alger and Weibull, 2012,2013 - see also Hamilton, 1964a,1964b), history and leadership (Acemoglu and Jackson, 2011), and parenting (e.g. Bisin and Verdier, 2001a; Cosconati, 2009). Although often markedly different in their founding principles and structure, they all propose ways in which certain preferences emerge through the interplay among the individuals.

A major contribution to our understanding of preference formation was made by Samuelson and Swinkels (2006). They deploy a setting where Nature acts as a benevolent parent to maximise the utility of the agents (humans). They show that if the agents' prior understanding of the causal and statistical structure of the world is imperfect, Nature will optimally endow them with preferences for certain actions, so as to correct for marginal errors that may ensue due to incorrect information processing. Building on the same logic of preference indoctrination, Adriani and Sonderegger (2009) propose a similar situation, where parents may endow their children with pro-social preferences. Here the choice of each parent to instil such preferences is dependent on the choices of the rest. Again, the fact that certain pieces of information about the environment are available to the parents but not the children implies that instilling values that are seemingly in conflict with material welfare may actually be promoting it. Based on these arguments, we ask how such values vary in response to changes in the environment where they arise.

We address this question in a framework of rationality, through a sequential game. Following Adriani and Sonderegger (2009), we construct a model in the spirit of Tabellini (2008), who applies the imperfect-empathy setup of Bisin and Verdier (2001a) to the transmission of pro-social values across generations. This is a model of Parent-Child interaction. The assumptions that they make are that a) parents can affect the deep preferences of their children and b) parents try to maximise a notion of utility of their children that departs from pure material welfare. This general framework of Parent-Child interaction (with alternatives to imperfect empathy) is becoming increasingly popular as a means of explaining social dynamics and cultural change (see e.g. Doepke and Zilibotti, 2007,2012).

A powerful feature of such models is that they facilitate preference heterogeneity in the strategic interplay between the different agents and institutions over time. For example, Lizzeri and Siniscalchi (2006) focus on the issue of asymmetric information between the parents and their children. In their context parents can intervene to affect the payoffs of their children, so as to protect them from harmful choices. The tradeoff is that this limits the children's ability to learn from experience. Adriani and Sonderegger (2009) also assume that parents are better informed than their children, but they assume that the former can manipulate the deep preferences of the latter, in order to promote their welfare.

In our model the children exhibit present-bias, which results in discounting future consequences unreasonably heavily in favour of present ones. Simply put, they assign a very high weight on present outcomes, to the detriment of their future welfare. Present bias is an increasingly popular notion in the economics literature.[1] In sub-section 2.2.2 we discuss this feature of our model in greater detail. Parents do not suffer from

[1] See e.g. Meier and Sprenger (2010); Benhabib et al. (2010) for experimental studies of the phenomenon and Laibson (1997); O'Donoghue and Rabin (1999); Gul and Pesendorfer (2001); Bénabou and Tirole (2002) for formal accounts.

present bias, but exhibit semi-altruistic preferences: they care about the joint maximisation of their own and their children's material welfare. We show that in this setup even materialist parents will opt for instilling moral values into their children's deep preferences. We then argue that measures of public policy that affect the parameters of our setup may crowd out the parents' private incentives, thus working against their stated goals. Our conclusions are akin to those reported by Bohnet et al. (2001), who analyse the non-monotonic effect of variations in contract enforcement on (endogenously determined) trustworthiness.

We view our paper as closest to that by Adriani and Sonderegger (2009), in that they focus on a different aspect (informational asymmetry) and use the same mechanism to account for the problem. Another setup that can be deemed as complementary to ours is the one proposed by Lindbeck and Nyberg (2006), where altruist parents decide how much to invest in their children's upbringing, in order to influence their future effort choices and, thus, the likelihood that they will need financial support.[2] We instead express the problem in terms of a bias that affects time-discounting and allow for a more general interpretation of preferences attached on actions. Our model is also conceptually close to that of Bhatt and Ogaki (2012), who propose an account of tough love. In their model children are assumed to be more impatient the more they consume. We depart from their setup in that we do not impose any assumptions that link the agents' preferences with their consumption and rely solely on present bias to support our conclusions.

Abstracting from the literature on cultural transmission, our paper also relates to time-inconsistent decision making (Laibson, 1997). Specifically, it can be applied to situations where people choose to exert self-

[2] On the deployment of strategic bequests by altruistic parents, see also Bernheim et al. (1985), Lindbeck and Weibull (1986), and Wilhelm (1996) among others.

control. We introduce a direct preference for an action as a commitment mechanism. We show that the tradeoff between the relative costs and benefits of the *'desirable, yet potentially harmful'* action has important implications for the individual's incentives and, thus, for the design of public policy.

The remainder of the paper proceeds as follows. Section 2.2 contains the setup of our model, a discussion about some of its core features, and the analysis of equilibrium. In section 2.3 we discuss policy implications and consider a number of extensions and alternative readings of the model analysed in section 2.2. Section 2.4 concludes.

2.2 Model

2.2.1 Parent - Youngster setup

Consider a two-player sequential game, $\mathcal{G}$, spanning across three periods, denoted by $t \in 0, 1, 2$. The first player, the parent (P), is the first to move, at $t = 0$. The second player, the youngster (Y), observes the parent's move and subsequently makes his own, at $t = 1$. The youngster must select an action, $\alpha \in \{B, F\}$ (smoke/do not smoke, be extravagant/be thrifty, break/follow the law, etc.). Each of these two actions yields a consumption payoff. The consumption payoff of action F is normalised to zero.[3] Selecting action B generates an immediate consumption benefit, $b_1 \in \mathbb{R}^{++}$, as well as a delayed cost, $b_2 \in \mathbb{R}^{++}$.[4]

[3] This is without loss of generality. Given any $\pi^Y_{\bar{t}}(F)$ and $\pi^Y_{\bar{t}}(B)$ in some $\bar{t} \in \{1, 2\}$, where $\pi^Y_{\bar{t}}(.)$ is the material-payoff function of agent Y in period $\bar{t}$, subtracting $\pi^Y_{\bar{t}}(F)$ from both will not alter Y's decision.

[4] The same relationship could have been achieved by restricting both b_1 and b_2 to be negative. Indeed, the important element is that they are of the same sign. In section 2.3 we examine this alternative case, where a present loss is weighted against a future benefit. We show that this scenario is a reflection of ours. Owing to the symmetric

The youngster decides with the aim to maximise his utility, which is given by the present discounted value of his consumption payoff over periods 1 and 2, as well as a hedonic component, which is manipulated by the parent (more on this later). There is no hedonic component associated with action B. By choosing F, on the other hand, the youngster experiences a (net) degree of intrinsic gratification, denoted by $n \in \mathbb{R}^{+}$. We will refer to n as the level of 'morality' player Y is endowed with.

Definition 2.2.1. *Morality* The degree of moral preference, n, for action α is the level of intrinsic (non-material) utility player Y receives upon choosing α. This is additional to the material payoff resulting from action α, but relevant only to the 'moral agent', i.e. player Y.

For the ease of exposition, we will use a working example. Let action F be labelled as 'being frugal' and action B as 'being extravagant' with respect to one's monetary expenditure. Then, his problem becomes clear. By being frugal he can save some money in period 1, so as to be able to spend them in period 2, augmented by the interest rate on savings. By being extravagant, on the other hand, he increases his period-1 utility (by consuming more) at the expense of the additional augmented period-2 income that would have resulted from his savings. We will use this interpretation of actions B and F throughout our analysis. Note, however, that this is only an example, designed to facilitate a more immediate understanding of the problem. The domain of application of our theory is much more general and includes all instances where one-shot decisions can have consequences at multiple points in time.

An important difference between the parent and the youngster lies in their degrees of patience. In particular, the youngster's preferences are presently biased, while those of the parent are not. Let $\delta^{Y} = \beta\delta$ represent

structure of the analysis, our results are invariant across the two.

the youngster's discount factor, where $0 < \beta < 1$ and $0 < \delta \leq 1$.[5] Then, his utility function can be written as:

$$U^Y = \begin{cases} b_1 - \beta\delta b_2 & \text{if } \alpha = B \\ n & \text{if } \alpha = F \end{cases} \tag{2.2.1.1}$$

It is worth noting that present bias is not a necessary assumption within our framework. What needs to be the case is that the youngster discounts the future more heavily than the parent does. We invoke the assumption of present bias to reinforce the connection between this parent-youngster framework and that of the intertemporal self, who has to anticipate her/his future choices when making decisions in the present. Simply assuming that the two agents have different discount factors might be plausible in the case of the parent-youngster framework, but it does not appear quite so plausible in the case of the intertemporal self. By invoking present bias, we are able to readily adapt our analysis in both frameworks. In addition, present bias is theoretically appealing as a potentially robust feature of preferences on evolutionary grounds (this can be seen in the context of the framework proposed by Samuelson and Swinkels, 2006). We discuss present bias and its implications in greater detail in sub-section 2.2.2.

As stated before, the parent moves first, at $t = 0$. Her objective is to maximise the joint welfare of herself and the youngster. She does so by determining the value of n, at a cost. This is captured by $C : \mathbb{R}^+ \to \mathbb{R}^+$, which associates each action available to the parent with a material loss she has to incur to take that action. We postulate that no such loss occurs by default, i.e. $C(0) = 0$. We also assume that this loss is increasing linearly in the degree of the parent's interference, i.e. that $C'(n) = \frac{dC(n)}{dn} = c > 0$. The

[5] Here, δ is the standard discount factor, while β is an additional weight that the youngster attaches on *all* future consequences. We say that the youngster exhibits quasi-hyperbolic, time-inconsistent preferences.

linearity assumption here is only imposed for simplicity. Our results would be no different in a qualitative sense under an exponentially increasing cost function.[6] Let $\delta^P = \delta$ represent the parent's discount factor.[7] Then, the parent's utility evaluated at $t = 0$ can be described as follows:

$$U^P = \begin{cases} b_1 - \delta b_2 - \frac{C(n)}{\delta} & \text{if } \alpha = B \\ -\frac{C(n)}{\delta} & \text{if } \alpha = F \end{cases} \tag{2.2.1.2}$$

Notice that U^P has been divided by δ, in order to maintain uniformity and simplicity in the representation. This is necessary, because the parent is deciding at t_0 and, thus, she discounts the youngster's future decision by δ, whereas she has to incur $C(n)$ immediately.

Importantly, the difference between the discount factor of the parent and that of the youngster can create a conflict of interest. Intuitively, our specification captures the notion that the youngster is more impatient than the parent. Furthermore, the parent does not internalise fully the youngster's preferences, but instead applies imperfect empathy. That is, she evaluates the youngster's material payoff through the lens of her own preferences (this is quite standard in the literature, see Bisin and Verdier, 2001). Hence, the conflict of interests arises: *the parent would like the youngster to be more patient than he actually is.* To correct for this, given her inability to address the youngster's present bias directly, she can opt instead to imbue him with some intrinsic (moral) preference for one of the actions.

[6] Indeed, $C(n)$ is assumed weakly convex for our proofs in the Appendix.

[7] We say that the parent exhibits time-consistent preferences by discounting the future exponentially. Notice that her standard discount factor is the same with that of the youngster. It is worth repeating that this does not need to be the case. So long as the two players exhibit different degrees of patience, our analysis applies. In our framework the youngster is not simply impatient (i.e. exhibits a lower discount factor). Instead, he attaches a pronounced significance on *present* consumption.

Equations 2.2.1.1 and 2.2.1.2 highlight this potential for discrepancy between the choice favoured by the youngster and the one the parent would prefer. To see this, consider the following example, where $n = 0$. Here, P would prefer Y to choose B iff:

$$U^P(0, B) \geq U^P(0, F) \Rightarrow b_1 - \delta b_2 \geq 0 \Rightarrow b_2 \leq \frac{b_1}{\delta}$$

On the other hand, Y will opt for B iff:

$$U^Y(0, B) \geq U^Y(0, F) \Rightarrow b_1 - \beta\delta b_2 \geq 0 \Rightarrow b_2 \leq \frac{b_1}{\beta\delta}$$

Thus, the youngster would switch from B to F at a higher threshold value for b_2. From the point of view of the parent that would be sub-optimal. In the context of our working example, the parent would prefer the youngster to behave frugally (choose action F), provided that the return to his savings (b_2) is at least equal to $\frac{b_1}{\delta}$. In simple terms, she would like him to be frugal, so long as the period-1 value of the return to his savings surpasses the period-1 value of the amount he has to save. On the other hand, the youngster would demand a return equal to at least $\frac{b_1}{\beta\delta}$ in order to give up part of his period-1 expenditure. That is, he would be too 'lavish' (and short-sighted) in the parent's opinion: due to his presently biased preferences, he would assign an unreasonably high weight on his period-1 utility. This situation, where the parent does not interfere with the youngster's preferences at all ($n = 0$), is illustrated in Figure 2.1.

Suppose now that the parent chooses instead to instil a direct preference for action F at $t = 0$. Let $n > 0$. That will induce the youngster to lower his threshold for switching from B to F. Consider, again, our working example. The parent is trying to instil a moral code in the young-

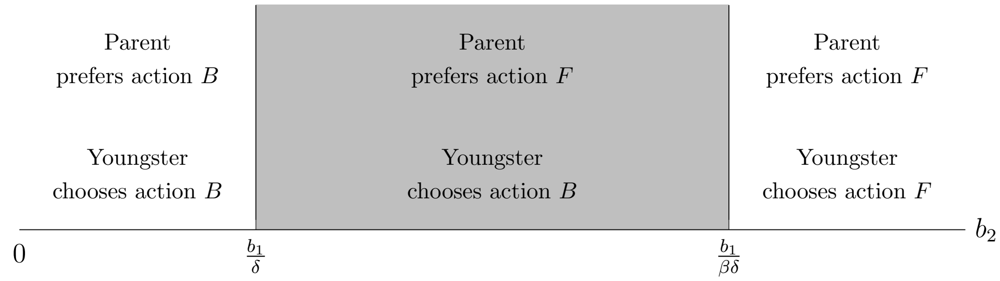


Figure 2.1: $n = 0$: no preference for a particular action

ster: to instruct him that he should behave frugally not because it yields large material benefits, but because it is *the right thing to do, in and of itself.* That is, she chooses to imbue action F with a moral content that is additional to its material consequences.[8] This does not affect the material consequences implied by the choices available to the youngster or his present bias, but it does affect his utility. In this way, it counteracts the effect of his impatience and brings his preferences closer to those of the parent. In other words, the youngster behaves more frugally not because he has grown more patient, but because he is morally incentivised to do so. The resulting situation looks like the one depicted in Figure 2.2.

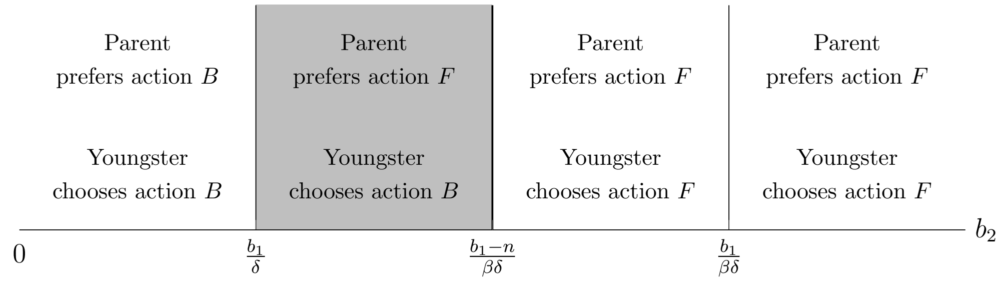


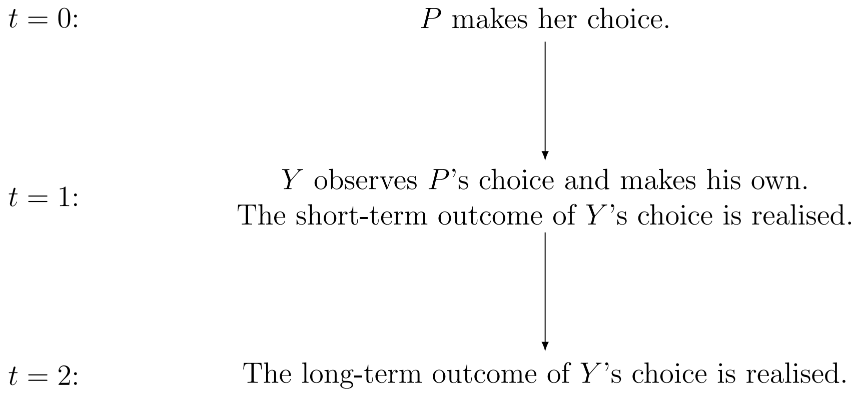
Figure 2.2: $n > 0$ assigned on action F

Notice that so far the magnitudes of b_1 and b_2 are both deterministic,

[8] Notice that in our characterisation the morality assigned to an action is *dependent* on its material consequences. The level of n is chosen by the parent in order to account for the youngster's present bias and not because she actually believes that morality is meaningful in any way. One way to think about this instrumentalist approach would be to consider that virtually any action can be imbued with a moral content, so long as the parent prefers it more than the youngster does. Note, however, that for the latter morality *is* meaningful, in the sense that his utility increases by n whenever he chooses the morally superior option. The appeal of such an extreme scenario is precisely that even if people did think an act like this, there would still be scope for moral values to arise.

that is, there is no uncertainty associated with any of them. We start from this case, in sub-section 2.2.3, because it is useful as a basis for comparison. In 2.2.4 we consider a more realistic scenario, by allowing for uncertainty over b_2.

Finally, it is useful to summarise the timing of this game.



$t = 0$: P makes her choice.

$t = 1$: Y observes P's choice and makes his own.
The short-term outcome of Y's choice is realised.

$t = 2$: The long-term outcome of Y's choice is realised.

Figure 2.3: Timeline of events

In period $t = 0$ the parent selects n so as to maximise the joint utility of herself and the youngster, evaluated according to her preferences at that time. The youngster observes the parent's move and subsequently makes his own, at $t = 1$. The youngster's choice yields both a short- and a long-term outcome. The short-term outcome is realised immediately upon his choice, i.e. at $t = 1$. The long-term outcome is realised in the following period, i.e. at $t = 2$. A timeline of the events is provided in Figure 2.3.

2.2.2 Discussion of the model

Before we continue with our analysis, we deem it meaningful to discuss three features of our design in greater detail. The first is present bias. Rational-choice theory models intertemporal decision making using exponential discounting for future periods. In this way, the decisions made by the individual are time-consistent. However, when choosing among alternative options, people typically manifest a strong preference for present outcomes, which leads to time-inconsistency. Following the seminal contributions of Ainslie in the domain of temptation and self-control (see Ainslie, 1975, 1992), many experimental studies have documented the phenomenon in economics (Meier and Sprenger, 2010; Benhabib et al., 2010 are two recent examples). This led to a growing literature of formal accounts that have established the phenomenon as a feature of people's preferences (see e.g. Laibson, 1997; Bénabou and Tirole, 2002).

In our parent-child context we incorporate present bias as a feature of the preferences of the child, but not the parent. This distinction is maintained for its plausibility and to reinforce the connection with the relevant literature, which highlights the discrepancy between the preferences of the parents and those of their children. However, this particular preference configuration is not essential for our results. Notice that the choices of the parents correspond to future consequences, which are discounted altogether. Thus, endowing the parents with present bias as well would not have a qualitative impact on our results. Notice also that we could instead have started from an impatient parent and a patient child and our conclusions would be the same. Our choice of set-up demonstrates an intuitively simple idea. That the anticipation of impulsive behaviour by the child may affect the incentives of a parent who only has material-welfare concerns and induce her to intervene.

Present bias also has a theoretical rationale as a feature of humans' preferences in an evolutionary sense. If the information reception and processing mechanisms of humans are imperfect (as in the context of Samuelson and Swinkels, 2006), then their uncertainty about the environment may induce them to place a lot of weight on present consumption. Finally, present bias allows our model to also be read from the viewpoint of the intertemporal self exercising self control, as we discuss in section 2.3.

It is critical for our account that the parent cannot address the youngster's present bias directly. At first glance, this might seem arbitrary. Why should the parent not simply invest in eliminating this feature from the youngster's preferences? One argument is that our model would still apply in a situation where the parent could indeed influence β, but only to some extent or at too high a cost. A stronger argument can be made about the nature of each source of motivation. In our model we have described present bias as an innate characteristic, an impulse similar to the drive for profit. As we have argued in the previous paragraph, such an impulse may emerge as an evolutionarily optimal feature of preferences under certain conditions. By contrast, we have described the parent's intervention as cultural indoctrination. That is, the parent is still able to interfere with the youngster's preferences to some extent, but by *instilling an element of culture*, rather than *embedding an impulse*. Even if she wanted to influence the youngster's discounting directly, she would have to *teach* the youngster the virtues of patience, not *eradicate* his innate impatience. Thus, our model would still apply. As a final point, such constraints are common in this literature (see e.g. Samuelson and Swinkels, 2006 on the constraints in information processing).

The second feature of our model is our definition of morality. A remark on our choice of terminology is important. A generic preference to act in a particular way can be accommodated within various frameworks, that are not necessarily compatible with each other. For example, what may be construed as a moral motive may also be conceivable as a desire for social conformity. Our aim here is not to provide a clear-cut distinction on how to separate different sources of motivation. Rather, we are moving in the opposite direction: Given the innate disagreements among these different sources of motivation, we are mapping a way in which they can be thought to affect people's behaviour. To do so, we focus on their effects on preferences, by postulating that any non-material motive implies a direct preference for a particular action.[9] Consequently, the label *morality* in definition 2.2.1 is merely illustrative of the type of motivation we refer to and should not be taken as exhaustive. In principle, variable n refers to any non-material increment that is added on the youngster's utility, irrespectively of its definition (so long as it is chosen strategically by the parent).

Finally, a word of caution. In our framework we adopt the assumption that parents can manipulate their children's preferences at will. This claim is quite contentious. There is a long-standing debate on the effectiveness of parenting in shaping children's preferences, which is part of the greater debate between *nature* and *nurture*.[10] Addressing this debate lies far outside the scope of this paper. In support of our approach, we advance two arguments. The first is that this debate is still ongoing and the results from the different studies cannot typically account for the whole spectrum of environmental influence (Pinker (2003), p.325). To the extent that parents can have *any* effect on their children's preferences (irrespective of parenting style, which we do not specify), our model can be applied. The second is that by 'manipulation of deep preferences' we do not refer to a radical change in the behavioural traits towards an extreme. In technical terms,

[9] However, the moral imperative should not be viewed as an isolated prescription. Instead, it should have a wider interpretation, in terms of a typology of behaviour. For instance, a preference for fair allocations should be present not only when an individual is on the receiving end, but also when (s)he is called to allocate. These are not merely different idle positions. They involve different actions, which have to be taken strategically, and yet the same *type* of behaviour must emerge. More generally, such a preference should be active in all cases where allocations are to be made, irrespectively of their specifics.

[10] See e.g. Pinker (2003), pp.13-14 for an overview on parenting, pp.324-326 for a refutation of environmental effects on behavioural traits - but notice potential causes of bias in p.25. For conclusions in support of the opposite view see Heckman et al. (2006); Algan et al. (2011).

the deep preference for an action does not constitute an omnipotent argument in the child's utility function. In fact, that would be sub-optimal given our framework. Instead, it is instilled as a measure of choice, capturing the extent to which the parent herself wants the child to adhere to the relevant action. As such, it remains in conflict with the objective magnitudes that define the payoffs (which one can readily generalise to reflect genetic pre-dispositions). The unconvinced reader may still want to consider the alternative readings of our model outlined above.

We shall now proceed to characterise the value for n that constitutes the solution to the parent's problem.

2.2.3 Baseline

Some important remarks are in order. To start with, notice that the parent would have no incentive to set $n > (1-\beta)b_1$, as that would not only be more costly for her, but also counter-productive. Indeed, such a value for n would induce the youngster to choose action F even in instances where the parent would want him to opt for B. In addition, the parent would have no incentive to instil a preference for action B instead.[11] Doing so would also be counter-productive, as it would increase the discrepancy between the two players' preferences.

Lastly, it can be easily shown that the sequences of actions in tables 2.1 and 2.2 would be reversed if it was the case that $b_1, b_2 < 0$. That is,

[11] In this paper we focus on positive values for n in an effort to determine the action that will be *chosen*, as opposed to that which will be *avoided*. The two are equivalent n our framework, where the youngster faces a binary-choice problem. However, in a situation with three or more available actions assigning a negative n to an action (aversion towards a certain type of behaviour) does not generally ensure that the desired action will be chosen. A comparison between the cost of discouraging certain types of behaviour and that of encouraging others is an interesting research project itself. We leave this for the future and focus instead on positive education (encouragement of a particular behaviour).

if action B led to a present cost and a future benefit, then both players would favour F for $|b_2| \leq |\frac{b_1}{\delta}|$ and both would choose B for $|b_2| \geq |\frac{b_1}{\beta\delta}|$. For $|b_2| \in (|\frac{b_1}{\delta}|, |\frac{b_1}{\beta\delta}|)$ they would disagree, with the parent favouring B and the youngster choosing F. Then, the former would find it optimal to assign $n > 0$ to action B. Taking these observations into account, we can form the following proposition.

Proposition 2.2.1. *In any equilibrium of game $\mathcal{G}$, $n \in [0, (1-\beta)b_1)$*

Proof. Formally, this can be proved by contradiction. Consider first the case where $b_1, b_2 > 0$ and, thus, P assigns n to action F.

i. Suppose $n < 0$: Then, $\forall b_2 \in [\frac{b_1}{\beta\delta}, \frac{b_1 - n}{\beta\delta})$ it would be true that $\frac{b_1-n}{\beta\delta} - b_2 > 0$. Thus, Y would choose action B and P would have been better off setting $n = 0$.

ii. Suppose $n > (1-\beta)b_1$: Then, $\forall b_2 \in (\frac{b_1-n}{\beta\delta}, \frac{b_1}{\delta}]$ it would be true that $\frac{b_1-n}{\beta\delta} - b_2 < 0$. Thus, Y would choose action F, even though P would prefer action B. Therefore, P would have been better off setting $n = (1-\beta)b_1$.

iii. Suppose $n = (1-\beta)b_1$: For $b_2 \in [\frac{b_1}{\delta}, \frac{b_1}{\beta\delta})$ Y would choose action F, in line with P's preferences. If $b_2 = \frac{b_1}{\delta}$, P would be indifferent between actions F and B, as they would both result in $U^Y = 0$. Setting $n = (1-\beta)b_1$ would render Y indifferent between the two actions at a positive cost to P. Thus, P would be better off setting n slightly below $(1-\beta)b_1$, so as to avoid the unnecessary expenditure in the case where $b_2 = \frac{b_1}{\delta}$.

An equivalent argument holds in the case where $b_1, b_2 < 0$ and P attaches n on action B. □

Proposition 2.2.1 describes the upper and lower bound for n. In simple terms, it determines the values of n which it makes sense for the parent to consider.

Consider, now, the situation outlined in sub-section 2.2.1 from the parent's perspective at $t = 0$. The parent knows that in period 1 the youngster will choose based on:

$$n \gtreqless b_1 - \beta\delta b_2 \Rightarrow b_2 \gtreqless \frac{b_1 - n}{\beta\delta}$$

If the future cost from action B is such, that the preferences of the youngster are at odds with those of the parent, then the latter may find it optimal to engage in some moral instilling. In other words, if $\frac{b_1}{\delta} < b_2 < \frac{b_1}{\beta\delta}$, then P may optimally assign $n > 0$ on action F, so as to induce Y to choose it at $t = 1$. This depends on the cost of inspiring that moral code. To simplify the analysis, suppose that when the youngster's preferences render him indifferent between the two options, he always chooses action F. Then, the various different cases are summarised in the following proposition.

Proposition 2.2.2. *Given game $\mathcal{G}$ with $b_1, b_2 > 0$, P assigns n^* to action F such, that:*

i. if $b_2 > \frac{b_1}{\beta\delta}$, then $n^* = 0$ and Y will choose action F.

ii. if $b_2 < \frac{b_1}{\delta}$, then $n^* = 0$ and Y will choose action B.

iii. $$\text{if } \frac{b_1}{\delta} < b_2 < \frac{b_1}{\beta\delta}, \text{ then } n^* = \begin{cases} b_1 - \beta\delta b_2 & \text{if } \frac{C(b_1 - \beta\delta b_2)}{\delta} < \delta b_2 - b_1 \\ & \text{and } Y \text{ will choose action } F. \\ 0 & \text{if } \frac{C(b_1 - \beta\delta b_2)}{\delta} > \delta b_2 - b_1 \\ & \text{and } Y \text{ will choose action } B. \end{cases}$$

Proof. The proof of this proposition is straightforward. Trying to maximise their joint welfare, the parent compares the material gain that results from n^* with the cost of instilling it into the youngster. When they both agree on which action the latter should take, there is no need for a value system ($n^* = 0$). When they do not, if $n^* > 0$, then it is precisely such that it makes the youngster indifferent between F and B (given the assumption stated above, that in such cases the youngster opts for F). Any higher or lower value would incur an additional cost to the parent with no added benefit. Thus. the parent has to compare what she gets by setting $n^* = b_1 - \beta\delta b_2$ with the cost, $C(b_1 - \beta\delta b_2)$, of doing so. If the benefit surpasses the cost, then n^* is set equal to $b_1 - \beta\delta b_2$, otherwise it is set equal to 0. □

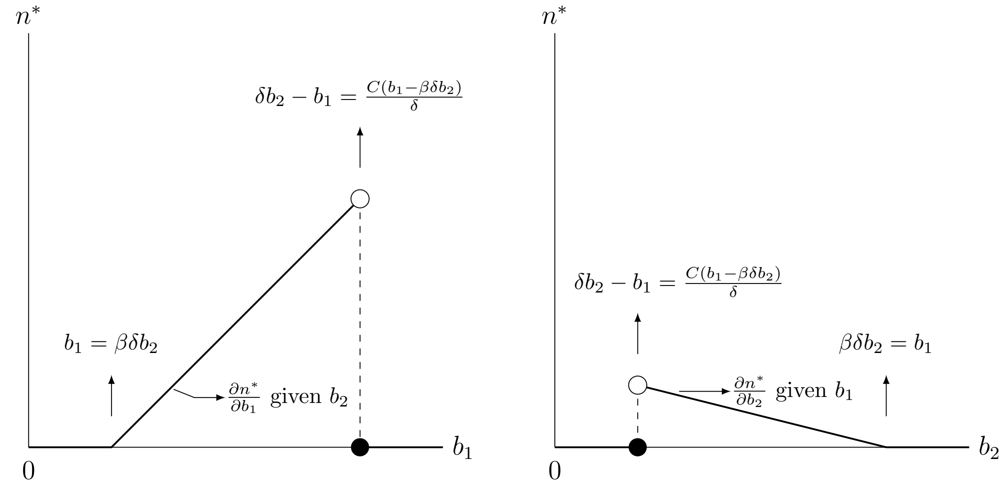
The content of proposition 2.2.2 may be best described by application to our working example. Recall that this is a situation where the parent knows the exact value of the material benefit the youngster can obtain in period 2 by being frugal in period 1. If this material benefit is so low that P herself would prefer Y to not be frugal, then she would not assign any moral underpinning to parsimony. Equally, if the return to savings is so large that Y will save some of his wealth anyway, then there is no use, and, thus, no scope for a value function. Indeed, a moral connotation is relevant only when the parent considers the investment worthwhile, whereas the youngster's impatience favours an extravagant behaviour. In that case, provided that the cost is sufficiently low, the parent will engage in moral indoctrination. Furthermore, she will set the utility from being prudent so as to make the youngster precisely indifferent between acting frugally and acting extravagantly. A higher or lower level of 'moral' utility will be costly for the parent without adding anything to the youngster's welfare.

The instrumental view of morality championed in our paper gives rise to a rich structure of variations. Recall that the level of moral preference the parent optimally attaches onto an action is dependent on the material consequences implied by that action relative to those implied by the other actions available. In our simple scenario, the degree of moral inclination towards behaving frugally varies with the net benefit/cost of being extravagant. The latter is expressed as a comparison between b_1 and b_2. The following corollaries summarize how changes in these two parameters affect n^*.

Corollary 2.2.3. *Consider game $\mathcal{G}$ with $b_1, b_2 > 0$ and n^* assigned on action F. The relationship between n^* and b_1 is non-monotonic. That is, $\exists \ \bar{b}_1 : n^*_{\hat{b}_1} = 0 \ \forall \ \hat{b}_1 \geq \bar{b}_1, \ \ n^*_{\tilde{b}_1} < n^*_{\breve{b}_1} \ \ \forall \ \tilde{b}_1 < \breve{b}_1 < \bar{b}_1$. In particular, an increase in b_1 will encourage the parent to increase the level of n^* at a one-to-one rate, so long as b_1 remains lower than $\delta b_2 - \frac{C(b_1 - \beta\delta b_2)}{\delta}$. If b_1 becomes equal to or greater than $\delta b_2 - \frac{C(b_1 - \beta\delta b_2)}{\delta}$, the value of n^* will drop to zero.*

Corollary 2.2.4. *Consider game $\mathcal{G}$ with $b_1, b_2 > 0$ and n^* assigned on action F. The relationship between n^* and b_2 is non-monotonic. That is, $\exists \ \bar{b}_2 : n^*_{\hat{b}_2} = 0 \ \forall \ \hat{b}_2 \leq \bar{b}_2, \ \ n^*_{\tilde{b}_2} > n^*_{\breve{b}_2} \ \ \forall \ \bar{b}_2 < \tilde{b}_2 < \breve{b}_2$. In particular, an increase in b_2 past $\frac{1}{\delta}\big(b_1 + \frac{C(b_1 - \beta\delta b_2)}{\delta}\big)$ will encourage the parent to decrease n^* at a rate lower than one-to-one (equal to $\beta\delta$), unless n^* is already equal to zero. For b_2 values lower than or equal to $\frac{1}{\delta}\big(b_1 + \frac{C(b_1 - \beta\delta b_2)}{\delta}\big)$, n^* will be equal to zero.*

An increase in b_1 implies that the temptation to behave extravagantly is now higher for the youngster. Therefore, if the parent still thinks that such behaviour is non-optimal, she will need to invest in a higher level of moral indoctrination to prevent it. As b_1 increases, there comes a point where such an investment is sub-optimal from the parent's point of view: What the youngster gains by behaving frugally is not enough to justify the cost of the moral education necessary to induce him to do so. From



(a) Relationship between n^* and b_1 given b_2 and $C(n)$: so long as there is conflict of preferences between P and Y and the cost of indoctrination is sufficiently low, morality gets stronger as temptation increases.

(b) Relationship between n^* and b_2 given b_1 and $C(n)$: given that there is conflict of preferences between P and Y and the cost of indoctrination is sufficiently low, morality gets weaker as the cost of temptation increases.

that point onward, the only sensible option for the parent is to not invest in instilling a moral value at all. Similarly, a diminishing b_2 implies that the future cost of impulsive behaviour gets lower. Therefore, the parent needs to increase her moral investment to ensure that the youngster will remain prudent. As b_2 keeps dwindling, however, there comes a point where the material benefit of prudence does not cover the cost of her investment. From that point onward, further reductions in b_2 will be accompanied by an equilibrium level of morality equal to zero. Figures 2.4a and 2.4b illustrate these two cases.

We can describe the variations in n^*, the optimum level of morality, as responding to variations in the parent's total utility. Recall that her utility depends on hers and the youngster's joint material payoff. This, in turn, is determined by her decision on n and the youngster's choice between actions F and B. Based on our previous analysis, the optimal value for n will be either equal to zero or such that will render the youngster exactly indifferent between F and B. This is true for any pair of values, b_1 and b_2, preference parameters, δ and β, and linear cost function, $C(n)$. We can, thus, describe the equilibrium level of morality, n^*, as a function of the difference in P's utility between the following two combinations of choices:

$$dU^P \equiv U^P(\bar{n}, F) - U^P(0, B) = \delta b_2 - b_1 - \frac{C(\bar{n})}{\delta}, \qquad \bar{n} > 0 \qquad (2.2.3.1)$$

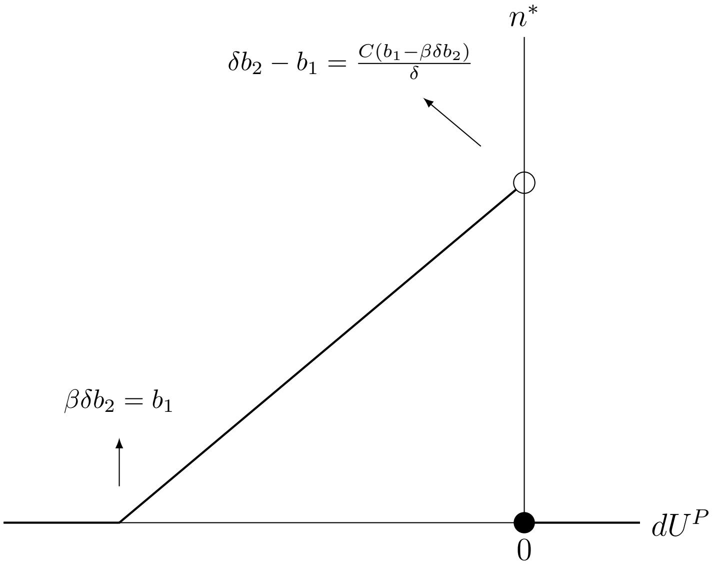


Figure 2.5: Relationship between n^* and dU^P: morality is at its highest when financial prudence (minus the cost of instilling it) is only marginally more beneficial than improvidence.

Figure 2.5 illustrates how changes in dU^P affect the optimal level of morality, n^*. It is worth noting that moral indoctrination attains its highest levels in our framework for dU^P values close to zero. This is true when the total cost from action B from the parent's point of view is only marginally higher than the cost of the moral education necessary to avert it. In other words, a relatively high degree of morality is needed when action B is sub-optimal, but only just so.

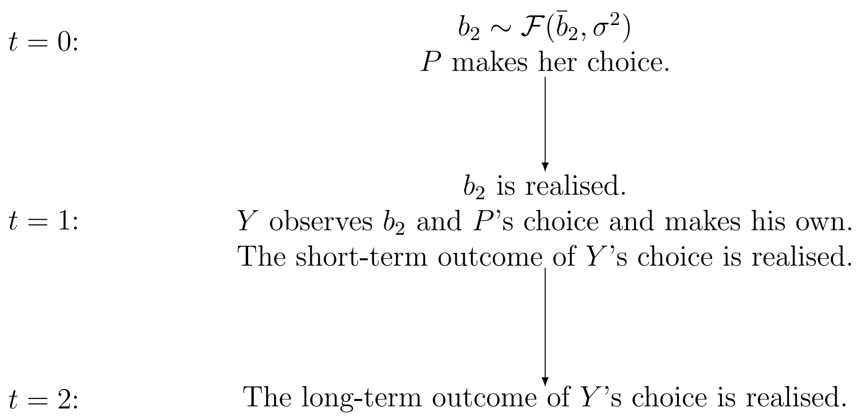
To clarify this point, consider again our working example. Our framework implies that, given the cost of moral education, for the parent to be willing to invest a lot in it, the return to frugality should be only slightly higher than the return to extravagance. It is in this case that temptation to overspend and, thus, the need for strict moral discipline is at its highest. Intuitively, given the youngster's degree of impatience, when the difference in returns is sizeable, little self-control is needed to refrain from overspending. As this difference shrinks, the youngster has to exercise progressively more self-discipline to ignore his impulse. This requires a stronger commitment to his moral position.

We now turn to examine the case where the parent does not know b_2 ex ante, only that it follows a certain distribution, $\mathcal{F}(b_2)$.

2.2.4 Probabilistic future cost

In this sub-section we allow for some information asymmetry to arise over the value of b_2, the future consequence of action B. Specifically, the parent is now unaware of the actual value of b_2 when she makes her decision. She only knows that it follows a specific distribution, with a positive mean and a certain variance. The youngster, on the other hand, knows its exact value when he makes his choice. Suppose that b_2 is normally distributed in $\mathbb{R}^+$ and let $\mathcal{F}(\bar{b}_2, \sigma^2)$ be the cumulative distribution function, with the corresponding probability-density function represented by $f(b_2)$. Then, the timeline of the events is akin to that in Figure 2.6.

This new structure enhances the generality of our results. To see this, note that our framework accommodates cases where b_2 is ex ante definite as instances where $\sigma^2 = 0$. In addition, we view it as intuitively plausible. Indeed, the parent can be fairly certain about the degree of gratification the youngster can expect instantaneously upon making a decision. However, future consequences related to that decision are inherently compromised by environmental volatility - changes in exogenous factors the parent may



$t = 0$: $b_2 \sim \mathcal{F}(\bar{b}_2, \sigma^2)$ P makes her choice.

$t = 1$: b_2 is realised. Y observes b_2 and P's choice and makes his own. The short-term outcome of Y's choice is realised.

$t = 2$: The long-term outcome of Y's choice is realised.

Figure 2.6: Timeline of events - b_2 uncertain at $t = 0$

not even be aware of, let alone able to influence. In this sense, the youngster has an informational advantage simply by being closer to these future consequences. In the context of our working example, the parent may well be aware in period 0 of the amount of wealth the youngster will have at his disposal in period 1. However, she is unlikely to be aware of the interest rate that may accrue on the youngster's savings. Thus, the material payoff of the youngster will feature in her utility in expected terms.

$$U^P = \begin{cases} \int_0^\infty (b_1 - \delta b_2) f(b_2) db_2 - \frac{C(n)}{\delta} & \text{if } \alpha^Y = B \\ 0 - \frac{C(n)}{\delta} & \text{if } \alpha^Y = F \end{cases} \tag{2.2.4.1}$$

The youngster, on the other hand, will be offered a specific interest rate before he makes his decision. Therefore, the parent's information problem is irrelevant to him. That is, his utility is still represented by equation 2.2.1.1. Taking equations 2.2.4.1 and 2.2.1.1 into account, the parent's problem can be stated as follows.

$$\max_n U^P = \pi^Y - \frac{C(n)}{\delta} = \int_0^{\frac{b_1 - n}{\beta\delta}} (b_1 - \delta b_2) f(b_2) db_2 - \frac{C(n)}{\delta} \tag{2.2.4.2}$$

Here, $\pi^Y = \pi_1^Y + \pi_2^Y$ is the youngster's total material payoff across periods 1 and 2. The particular functional form of the distribution of b_2 may imply more than one local maxima for 2.2.4.2. To maintain simplicity, we impose two technical assumptions, which jointly ensure that the maximum is unique.

Assumption 2.2.5. *Given game $\mathcal{G}$, let $f(.)$ denote the density function according to which b_2 is distributed. Then, $f(.)$ is quasi-concave in b_2.*

Assumption 2.2.6. *In any game $\mathcal{G}$, $\beta^2 \delta C'(0) < [(1-\beta)b_1]f(\frac{b_1}{\beta\delta})$.*

Assumption 2.2.5 implies that the marginal gain from n will not increase again once it has started decreasing. Given that $C(.)$ is increasing in n, a unique maximum point is implied. Assumption 2.2.6 precludes the possibility of a minimum. This would be possible if, for example, for n sufficiently small, the cost of increasing it surpassed its additional benefit. Assumptions 2.2.5 and 2.2.6 together ensure that P's problem attains a unique optimum solution, which confers the maximum return to n.

Assumptions 2.2.5 and 2.2.6 are rather restrictive, but their purpose is to maintain the analysis simple. Note that the set of values for b_2 that are relevant to P's problem is bounded: $(\frac{b_1}{\delta}, \frac{b_1}{\beta\delta})$. Thus, a solution would be attainable even with a different functional form for $f(.)$. The additional complication would be a comparison across all local maxima to determine the global one(s). Moreover, the same would be true even in the presence of local minima. We simply chose to sidestep these additional complexities, in order to refrain from further obscuring our analysis.

Bearing the above in mind, we can now proceed to characterise the solution to P's problem in the face of uncertainty. Proposition 2.2.7 presents this result.

Proposition 2.2.7. *Consider game $\mathcal{G}$ with $f(b_2)$ and $C(n)$ in line with* *assumptions 2.2.5 and 2.2.6. Then, the optimal n satisfies:*

$$n^* = (1-\beta)b_1 - \frac{\beta^2}{f(\frac{b_1-n^*}{\beta\delta})}C^{'}(n^*) \tag{2.2.4.3}$$

The proof can be found in section A.1 of the appendix. The result is, by construction, consistent with the analytical perspective of methodological individualism: n will be assigned a positive value only if it is instrumental to the achievement of P's goal, and only to the extent that it has a higher rate of return compared to its cost. We, thus, see that the instrumental character of morality does not change when uncertainty is introduced. The solution to P's problem is qualitatively similar to the one in our baseline version.

What about the youngster's decision? In our baseline scenario the value of n^* would be such, that he would always be exactly indifferent between actions B and F, and would eventually choose F in line with the parent's preference.[12] In this new scenario, however, it is possible that the youngster's choice will not reflect the parent's preference, even given her investment in n.The reason is that the actual realisation of b_2 may be so low, that he may find it profitable to choose action B even after he has considered his moral attachment to action F. Figure 2.7 illustrates such a scenario.

To motivate this situation, we turn again to our working example. When the parent invests in moral instilling the *future* return to savings (the opportunity cost of lavish behaviour) is not necessarily known. Indeed,

[12] The same would be true in expected terms, if the cost of action B was uncertain for both players. So long as P and Y had the same distribution of b_2 in mind, Y's choice would be anticipated by P: They would both form the same expectation about b_2. Thus, even if the *actual* value of b_2 eventually proved to be different than what they had expected, their choices would coincide.

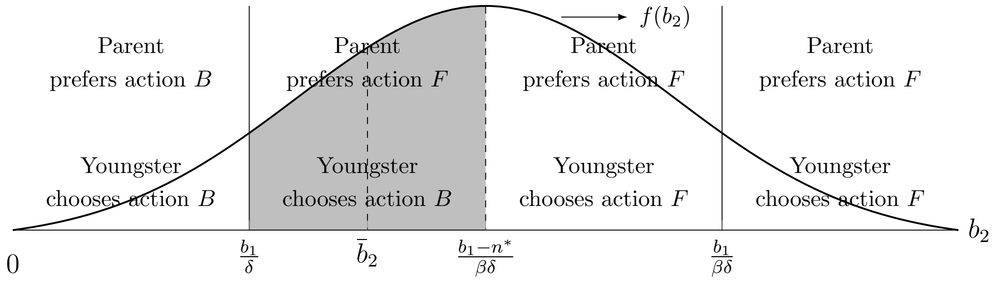

$f(b_2)$
Parent prefers action B
Parent prefers action F
Parent prefers action F
Parent prefers action F
Youngster chooses action B
Youngster chooses action B
Youngster chooses action F
Youngster chooses action F
b_2
0
$\frac{b_1}{\delta}$
$\bar{b}_2$
$\frac{b_1-n^*}{\beta\delta}$
$\frac{b_1}{\beta\delta}$

Figure 2.7: *Misalignment of preferences* Player P has optimally assigned n^* on action F knowing that b_2 is drawn from $\mathcal{F}(b_2)$, but the realised value, $\bar{b}_2$, induces Y to opt for action B. The shaded area is the cumulative probability of all such b_2 values.

in forming a prediction on what the interest rate on savings will be when the time comes for the youngster to make his choice, the parent may only be able to observe past interest rates. In the next period, however, when the youngster is called to decide, he will be given a definite one-period interest rate. As a result, he will know precisely what the opportunity cost of overspending is. That interest rate may indeed be drawn from the distribution that the parent had in mind. However, this does not preclude the possibility that its value will be too low to induce the youngster to be frugal, even given his moral commitment.

Given that the possibility is now open for the youngster's choice to be different than what the parent would want, we can also assess how the probability of this scenario varies with b_1 and the distribution of b_2. To do so, we need to formally distinguish between cases where the choice of Y agrees with P's preference and cases where the two differ.

Definition 2.2.2 (*Compliance*)**.** The degree of conformity following P's choice of $\hat{n}^*$ is the cumulative probability that Y's choice will agree with P's preference given $\hat{n}^*$.

Using definitions 2.2.1 and 2.2.2, we now turn to examine how morality and compliance are affected by changes in b_1 and $\mathcal{F}(b_2)$.

Corollary 2.2.8. *Consider game $\mathcal{G}$ satisfying assumptions 2.2.5 and 2.2.6.*

An increase in the value of b_1, from $\bar{b}_1$ to $\hat{b}_1$ may lead to a higher n^*, so long as assumption 2.2.6 remains satisfied. However, compliance may be lower as a result of the increase in b_1.

Proof. See section A.2 in the Appendices. □

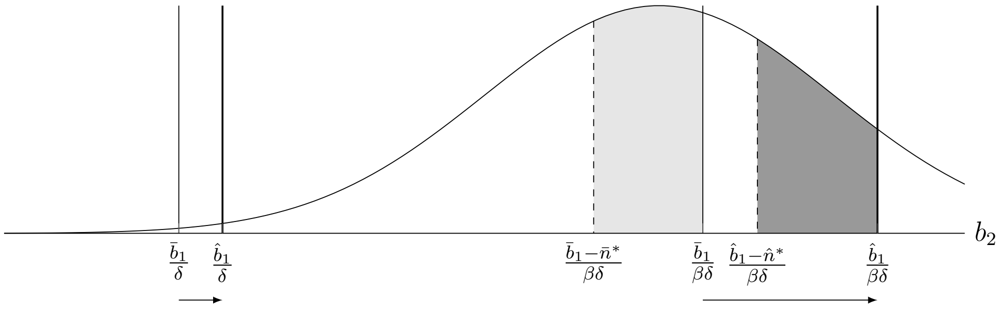

$\frac{\bar{b}_1}{\delta}$
$\frac{\hat{b}_1}{\delta}$
$\frac{\bar{b}_1-\bar{n}^*}{\beta\delta}$
$\frac{\bar{b}_1}{\beta\delta}$
$\frac{\hat{b}_1-\hat{n}^*}{\beta\delta}$
$\frac{\hat{b}_1}{\beta\delta}$
b_2

Figure 2.8: $\hat{b}_1 > \bar{b}_1$: The immediate consequence from option B is relatively larger and so is the level of n^*. If the mass of additional b_2 values that fall to the left of the first cut-off point as a result of the change is sufficiently small, then the total proportion of b_2 values for which Y's choice will conform with P's preference will be lower.

Corollary 2.2.8 points out that there is potential for moral reinforcement in the face of increased temptation. Suppose that b_1 increases. This implies that both players will be more inclined to opt for action B than before. However, the discrepancy between their preferences increases. To see this, notice that the youngster's switching threshold changes by a greater margin than the parent's one does. Therefore, the range of b_2 values for which their preferences are conflicting is now larger. As a result, if the parent still prefers action F, then the previous level of n^* is no longer optimal. In particular, the increase in b_1 induces her to increase n, in order to account for the additional appeal of action B relative to action F.

It is important to bear in mind that in adjusting n^* to account for the change, the parent is interested in its *marginal* benefit, not what she gets out of it *on average*. It may well be the case that on average the youngster will choose action B, contrary to the parent's preference. However, it may still make sense for her to invest in instilling some degree of morality, so long as what she gets from doing so (in expected terms) is more than what she spends.

Figure 2.8 illustrates this situation, given a linear cost function and a normal distribution for b_2. In this scenario, an increase in b_1 results in a higher n^*, although compliance is lower under the new level of morality. In the context of our example, a relatively higher benefit from lavish behaviour[13] may result in stricter indoctrination about the moral value of frugality, despite the fact that the youngster is more likely to make the 'morally wrong' choice.

Additionally, the positive relation between b_1 and n^* implies that a decrease in the youngster's temptation will likely be followed by a reduction in the level of morality. Intuitively, the decrease in b_1 makes option B less appealing and, therefore, encourages the parent to reduce the level of moral education, so as to lower its cost. We, thus, observe a trade-off between the exogenous incentive to opt for the option that the parent favours and the endogenous deep preference she instils herself.

Corollary 2.2.9. *Consider game $\mathcal{G}$ satisfying assumptions 2.2.5 and 2.2.6. A parallel rightward shift of $\mathcal{F}(b_2)$, which increases $E[b_2]$ from $\bar{b}_2$ to $\hat{b}_2$, where $\frac{b_1}{\delta} < \bar{b}_2 < \hat{b}_2$, may induce player P to invest less in morality. However, such a shift will always result in greater compliance.*

Proof. See section A.2 in the Appendices. □

An increase in the magnitude of the expected future consequence can lead to a lower level of moral preference. The intuition behind this result is straightforward. As the increase in $E[b_2]$ renders option B less attractive,

[13] This can occur, for example, through a drop in the level of prices in period 1.

the parent will eventually be discouraged from investing in n. The reason is that the instrumentality of the moral preference dwindles. As the youngster becomes more likely to avoid B anyway, investing in n and assuming the cost of doing so gets progressively counter-productive.

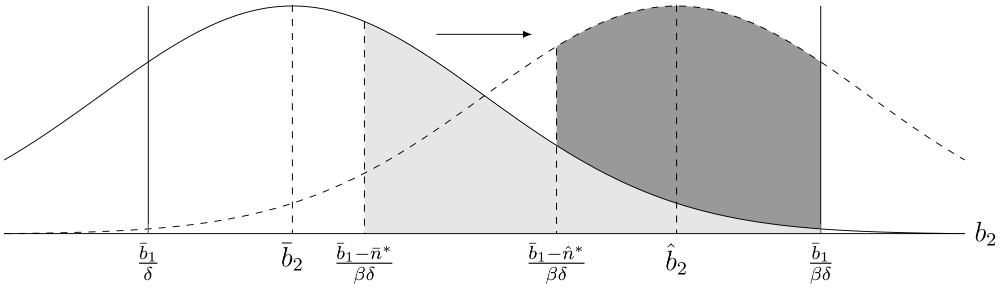

$\frac{\bar{b}_1}{\delta}$
$\bar{b}_2$
$\frac{\bar{b}_1-\bar{n}^*}{\beta\delta}$
$\frac{\bar{b}_1-\hat{n}^*}{\beta\delta}$
$\hat{b}_2$
$\frac{\bar{b}_1}{\beta\delta}$
b_2

Figure 2.9: $\hat{b}_2 > \bar{b}_2$: The expected future consequence is larger, the level of n^* is lower, and the probability of compliance is higher.

Thus, the increase in $E[b_2]$ may be partially crowded out by the decrease in the incentive to instil a given level of n. The same trade-off ensues between the youngster's extrinsic and intrinsic incentives to act in a particular way. In the face of higher exogenous motivation, his esoteric desire to uphold certain values dwindles, because it is no longer relevant.

It is worth noting that this is also true when the magnitude of the expected future consequence goes towards the opposite direction. The reasoning is the same as before. A reduction in $E[b_2]$ may induce the parent to compensate by increasing n. However, successive reductions will eventually discourage her from increasing n, as the preference discrepancy becomes progressively less relevant.

In line with the previous arguments, the youngster's degree of compliance with the parent's preference depends on the initial distribution of b_2. If $E[b_2] > \frac{\bar{b}_1}{\delta}$ in the first place, then any subsequent increase will lead to higher compliance. Figure 2.9 presents a situation where a higher $E[b_2]$ results in both a lower n^* and a higher degree of compliance.

Notice that the crowding out of the moral value by the material benefit is always accompanied by enhanced compliance. To see why, consider a situation where the expected cost of lavish behaviour is such, that the parent should optimally assign $n^* > 0$ to action F. If $E[b_2]$ increases, then the parent will only settle for a lower level of morality if it confers a greater return that the previous one. Investing in moral education is not more expensive than it was before. If anything, she could still invest in it to the extent she did before. If she chooses to undercut her investment, it is because this is the optimal response: she gets a higher return even with a lower degree of morality.

Notice that Corollary 2.2.9 describes a variance-preserving switch. That is, it refers to a shift in the distribution of b_2 to a higher expected value, but with the same *degree of uncertainty.* This is important for our analysis, as our conclusion that the increase in $E[b_2]$ always results in an increased degree of compliance does not necessarily hold if we allow for simultaneous changes in its variance. To see this, consider a situation where an exogenous shift affects both $\bar{b}_2$ and σ^2. Since n^* is affected by both, the effects of this change may actually counteract each other. We explore this possibility in the following Proposition.

Proposition 2.2.10. *Consider game $\mathcal{G}$ satisfying assumptions 2.2.5 and 2.2.6. Suppose that an exogenous shock changes the distribution of b_2 to one that has a higher mean and a higher variance. In other words, it increases both the expected value of b_2 and its degree of dispersion. Suck a shock may induce player P to invest less in morality and may also lead to lower compliance.*

Proof. See section A.3 in the Appendices for a proof by example. □

Proposition 2.2.10 highlights the potential conflict between two effects that result from the distributional change. One of these effects comes as a

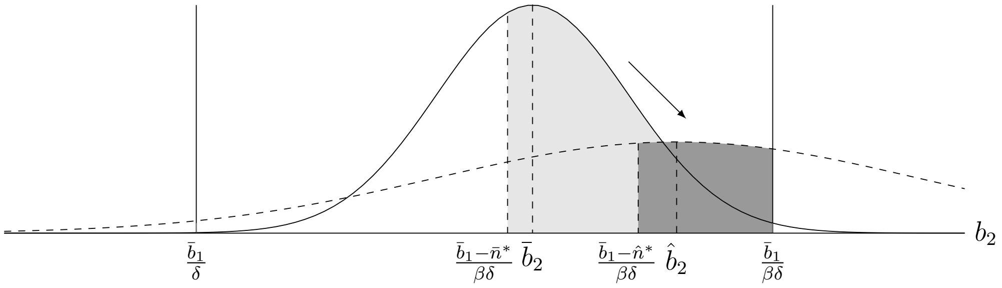


Figure 2.10: $\hat{b}_2 > \bar{b}_2, \hat{\sigma}^2 > \bar{\sigma}^2$: The expected future consequence is larger and more uncertain. The level of n^* and the degree of compliance are both lower.

result of the higher expected future consequence. The other follows from the increased uncertainty about that consequence. The net effect on n^* and the degree of compliance can be surprising.

As it has already been argued (see corollary 2.2.9), an increase in $E[b_2]$ may reduce the parent's incentive to invest in n, thus resulting in a lower level of morality in equilibrium. However, the parent's incentive is crowded out due to the fact that given the new $E[b_2]$ even a lower n makes the youngster more likely to comply. Thus, the increase in $E[b_2]$ (given b_1) leads to a higher degree of compliance.

The increase in σ^2, on the other hand, may induce n to fall even further. The reason is that as the future consequence becomes more volatile, the marginal return that the parent receives by increasing n gets progressively lower. Intuitively, the increased uncertainty implies that the most likely b_2 values are now less probable. As a result, it becomes difficult for the parent to pinpoint a level for n that is highly likely to be optimal.[14] Given that the cost of providing n has not changed, the parent may find it better to reduce n in the face of the increased uncertainty.

The final outcome may, thus, resemble the one illustrated in Figure 2.10. This is a case where a shift towards a higher but more volatile $E[b_2]$ re-

[14] Recall that the optimal level of n would render the youngster exactly indifferent between options B and F.

sults in a reduced probability of the youngster choosing in accordance with the parent's preference. Consequently, apart from crowding out 'morality', as captured by n, the change also renders the youngster more susceptible to present bias. This result is all the more striking when considered in light of the intuition that a higher $E[b_2]$ on its own would have the exact opposite effect.

We have thus far examined the changes in n^* and the degree of compliance induced by changes in b_1 and the distribution of b_2. As a final remark, we note that n^* varies monotonically with $C^{'}(.)$, to which it is inversely related. That is, other things being equal, an increase in the marginal cost of instilling morality always leads to a lower level of moral indoctrination and vice versa. In particular, there is no crowding-out related to the parent's incentives: a reduction in $C^{'}(n)$ will render her unambiguously more willing to provide a higher n^*. The same is true with respect to compliance.

2.3 Discussion

2.3.1 Policy implications

We now turn to examine some consequences of our analysis for the design of public policy. Our aim is to demonstrate that, owing to the strategic interplay analysed in section 2.2, the results of policy measures may be very different from those originally expected. To do so, we use examples of policies that may prove inefficient, given the policymaker's stated goals.

Consider, thus, a policy aimed at encouraging more people to save some of their income, e.g. an increase in the interest rate, taking effect at $t = 1$. Such a policy will have an effect on the amount of period-2 consumption one has to forfeit in order spend more money in period 1. In the context of our model, it amounts to an exogenous increase in $E[b_2]$. Should we expect that this policy will be successful, and to what extent? One factor that may limit the policy's effectiveness is the change in the culture of parsimony that its announcement initiates. As corollary 2.2.9 points out, a greater $E[b_2]$ may induce parents to invest less in instilling an intrinsic value for behaving frugally. Thus, even in the absence of additional effects stemming from the announcement of the policy, the resulting increase in the proportion of savers may not be as high as initially expected.

Suppose, now, that the government aims to discourage tax avoidance while in the midst of an austerity programme. To do so, it imposes stronger sanctions to perpetrators. However, owing to the need for austerity, it is also required to cut back on audits. What does the resulting situation look like? The announcement of stricter penalties (higher $E[b_2]$) is set to increase compliance, although it is also expected to discourage a culture of duty to pay one's taxes (lower n). The reduction in oversight, however, results in these penalties being more unlikely than before. As a result, it mitigates both moral education and compliance. Proposition 2.2.10 suggests that the resulting effect on taxes may well be negative.

Lastly, consider a policy that aims to reduce carbon emissions. One way of doing so would be to collect research on the adverse consequences for the environment and, thus, the society's future prospects. Then, this research would be disseminated, perhaps in the form of short advertisements, in a bid to increase environmental awareness among the population. Our analysis shows that there may be a caveat in this reasoning. Specifically, if the research appears inconclusive, so that many possible future scenarios seem likely but none is deemed particularly probable, the policy may backfire. Furthermore, as proposition 2.2.10 points out, this can be true even if the additional information results in the situation appearing more dire on average. Thus, our framework suggests that caution must be exercised in the release of information as part of a policy measure.

The three examples outlined above highlight the tradeoff between people's (exogenous) material incentives and their (endogenous) intrinsic motivation. By providing extrinsic incentives, public policies may end up crowding out private moral indoctrination. In doing so, they are compromising, at least partly, their own effects. Our analysis indicates that caution needs to be exercised when assessing the potential effects of a proposed policy measure.

2.3.2 Extensions

In this sub-section we explore some elements of our framework that give rise to additional features of interest. To start with, notice that although we focused on situations with $b_1 > 0$ throughout section 2.2, our results hold more generally. In particular, even with $b_1 < 0$ the same incentive structure emerges, with the sole difference that $n \in [0, |(1-\beta)b_1)|$ now needs to be assigned to action B (instead of F). To see this, recall that the payoff from action F in each period $t = 1, 2$ is normalised to zero. Consider, then, the following proposition.

Proposition 2.3.1. *Consider game $\mathcal{G}$ with $\bar{b}_1 > 0$, $\bar{b}_2 \sim \mathcal{F}(\bar{b}_2, \bar{\sigma}^2)$, and $\bar{n}^*$ assigned on action F. Let $f(.)$ denote the probability density function of distribution $\mathcal{F}(.)$. Suppose that $\bar{b}_1$ is replaced with $-\bar{b}_1$ and $\bar{b}_2$ with $-\bar{b}_2$. Then, P assigns $\bar{n}$ on action B in equilibrium and the degree of compliance is the same as before.*

Proof. Action B is compared to action F, the payoffs of which have been normalised to zero. Notice that $-\bar{b}_1$ is symmetric to $\bar{b}_1$ about zero. Thus, the absolute magnitude of the difference in payoffs between action B and action F is preserved. Owing to the change in the signs of b_1 and $E[b_2]$, the preferences of the players are reversed. Now, P would prefer Y to choose B for a larger set of realisations of b_2 than what Y is willing to accept. Formally, P favours action B for every $|b_2|$ value higher than $|\frac{b_1}{\delta}|$, while Y only chooses B if $|b_2| \geq |\frac{b_1}{\beta\delta}|$. Note that $f(.)$ has not changed. To account, then, for this discrepancy, P optimally assigns $\bar{n}$ to action B, so as to induce Y to chose B for $|b_2| \geq |\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}|$. Given $f(.)$, the share of b_2 values for which Y's choice complies with P's preference is the same as before. □

Intuitively, the change resulted in action B yielding an immediate cost and a future benefit. Owing to his presently biased preferences, the youngster discounts the future benefit more than the parent does. For this reason, there are some values of b_2 for which the youngster will choose F, while the parent would prefer him to choose B. This is why it now makes sense for P to attach n on action B. Since the distribution of b_2 and the cost function of n have not changed, the change in the signs of b_1 and b_2 brings about a symmetrically opposite situation. Therefore, in equilibrium the parent will attach $n = \bar{n}$ on action B and the youngster will adhere to the parent's preference with the same probability as before the change.

So far we have analysed preferences for actions within a parent-youngster framework. We can also evaluate the scope for such preferences under a different perspective, namely that of the intertemporal self. To that end, consider a game $\mathcal{G}$ that is being played among the various instances of the same person, acting at different points in time. Then, our analysis focuses on the action of her self at $t = 0$ and the choice of her self at $t = 1$. Suppose that this person is initially characterised solely by preferences over outcomes and that she also exhibits present bias. Suppose, further, in line with our previous set-up, that while she knows about her bias, she cannot eliminate it per se. Then, in trying to maximise her intertemporal utility, she would optimally set $n \in [0, |(1-\beta)b_1|)$.

How can such a result be interpreted? From the point of view of the self at $t = 0$ it is (weakly) optimal to commit to preferring an action over another. She knows that if she is equipped only with materialistic preferences, then it is probable that in the face of temptation she will make an ill-preferred choice. To reduce this probability, she may want to commit to a particular code of conduct, so as to enhance the appeal of the other option.

An appealing feature of this account of preference formation is its general applicability. Note that the aforementioned code of conduct can be grounded on various premises, such as moral principles, social norms, reputation, and habitual or conventional decision-making. All such concerns can be seen to be instrumental from a purely materialist viewpoint. Thus, such preferences can also emerge through an evolutionary process, assuming that present bias is also at play (through a reasoning similar to Samuelson and Swinkels, 2006).

Finally, notice that present bias is essential for the intertemporal-self interpretation of our model. In the parent-youngster set-up it is not necessary that the latter suffers from present bias, only that his discount factor is different than that of the former. In the intertemporal-self version the self is the same across the different periods and has, thus, a single discount factor. Present bias allows us to create the internal conflict that corresponds to two agents exhibiting different preferences.

2.4 Concluding Remarks

We propose a game-theoretic model of moral preferences, where parental indoctrination is optimally counterbalancing presently biased preferences. We build on the idea that preferences are, to a certain extent, malleable. We then investigate the relationship between material incentives and intrinsic motivation. Our analysis indicates that the relationship between the two is non-monotonic. Our results are especially relevant in the domain of policy analysis.

The theory presented here describes how the instilling of an intrinsic value can be optimal from a materially rational perspective. We depict the dialectics between parameter variations and individual incentives and show how the effects of the former can sometimes crowd out the latter. These effects are important, both with respect to cultural transmission and the exercise of self-control. The effectiveness of a policy is demonstrated to depend, at least to some extent, on it providing the right incentives to the agents.

The paper does not consider the intergenerational dynamics that ensue in such a context. This is a fascinating research question in its own right. Here, instead, we propose two main arguments: that preferences for actions can be rationally instilled and that preference formation should be taken into account when considering the effects of changes in the underlying economic environment.

2.5 References

Acemoglu, D. and Jackson, M. O. (2011). History, expectations, and leadership in the evolution of social norms. Technical report, National Bureau of Economic Research.

Adriani, F. and Sonderegger, S. (2009). Why do parents socialize their children to behave pro-socially? an information-based theory. *Journal of Public Economics*, 93:1119–1124.

Ainslie, G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin*, 82:463–496.

Ainslie, G. (1992). *Picoeconomics: The strategic interaction of successive motivational states within the person.* Cambridge University Press.

Algan, Y., Cahuc, P., and Shleifer, A. (2011). Teaching practices and social capital. Technical report, National Bureau of Economic Research.

Alger, I. and Weibull, J. W. (2012). A generalisation of hamilton's rule: Love others how much? *Journal of Theoretical Biology*, 299:42–54.

Alger, I. and Weilbull, J. W. (2013). Homo moralis - preference evolution under incomplete information and assortative matching. *Econometrica*, 81:2269–2302.

Andreoni, J. and Bernheim, B. D. (2009). Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5):1607–1636.

Arrow, K. J. (1994). Methodological individualism and social knowledge. *The American Economic Review*, 84:1–9.

Becker, G. S. (1976). Altruism, egoism, and genetic fitness: Economics and sociobiology. *Journal of Economic Literature*, 14(3):817–826.

Becker, L. C. (1999). Crimes against autonomy: Gerald dworkin on the enforcement of morality. *William and Mary Law Review*, 40:959–973.

Bénabou, R. and Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, 117(3):871–915.

Bénabou, R. and Tirole, J. (2003). Intrinsic and extrinsic motivation. *The Review of Economic Studies*, 70(3):489–520.

Bendor, J. and Swistak, P. (2001). The evolution of norms. *American Journal of Sociology*, 106(6):1493–1545.

Benhabib, J., Bisin, A., and Schotter, A. (2010). Present-bias, quasi-hyperbolic discounting, and fixed costs. *Games and Economic Behavior*, 69:205–223.

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

Bernheim, B. D., Shleifer, A., and Summers, L. H. (1985). The strategic bequest motive. *Journal of Political Economy*, 93(6):1045–1076.

Bhatt, V. and Ogaki, M. (2012). Tough love and intergenerational altruism. *International Economic Review*, 53(3):791–814.

Bicchieri, C. (2006). *The Grammar of Society: the Nature and Dynamics of Social Norms*. Cambridge University Press.

Bicchieri, C. (2010). Norms, preferences, and conditional behavior. *Politics Philosophy Economics*, 9:297–313.

Bicchieri, C. and Xiao, E. (2009). Do the right thing: But only if others do so. *Journal of Behavioral Decision Making*, 22:191–208.

Binmore, K. (1987). Modeling rational players: Part i. *Economics and Philosophy*, 3:179 – 214.

Binmore, K. (1988). Modeling rational players: Part ii. *Economics and Philosophy*, 4:9–55.

Binmore, K. (2010). Social norms or social preferences? *Mind*, 9:139–157.

Bisin, A. and Verdier, T. (2001a). The economics of cultural transmission and the dynamics of preferences. *Journal of Economic Theory*, 97:298–319.

Bisin, A. and Verdier, T. (2001b). The economics of cultural transmission and the dynamics of preferences. *Journal of Economic Theory*, 97:298–319.

Bohnet, I., Frey, B. S., and Huck, S. (2001). More order with less law: On contract enforcement, trust, and crowding. *American Political Science Review*, 95(1):131–144.

Bowles, S. (1998). Endogenous preferences: The cultural consequences of markets and other economic institutions. *Journal of Economic Literature*, 36(1):75–111.

Cosconati, M. (2009). Parenting style and the development of human capital in children. *Job Market Paper, University of Pennsylvania*, -:–.

Deci, E. L. and Koestner, R. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6):627–668.

Doepke, M. and Zilibotti, F. (2007). Occupational choice and the spirit of capitalism. Technical report, National Bureau of Economic Research.

Doepke, M. and Zilibotti, F. (2012). Parenting with style: Altruism and paternalism in intergenerational preference transmission. Technical report, Forschungsinstitut zur Zukunft der Arbeit.

Dufwenberg, M. and Güth, W. (1999). Indirect evolution vs. strategic delegation: a comparison of two approaches to explaining economic institutions. *European Journal of Political Economy*, 15:281–295.

Ellickson, R. C. (1998). Law and economics discovers social norms. *The Journal of Legal Studies*, 27:537–552.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669.

Elster, J. (1989). Social norms and economic theory. *The Journal of Economic Perspectives*, 3(4):99–117.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54:293–315.

Fehr, E. and Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, 8:185–190.

Fudenberg, D. and Levine, D. K. (2006). A dual-self model of impulse control. *The American Economic Review*, Vol. 96(5):1449–1476.

Gächter, S. and Falk, A. (2002). Reputation and reciprocity: Consequences for the labour relation. *Scandinavian Journal of Economics*, 104:1–26.

Gershoff, E. T. (2002). Corporal punishment by parents and associated child behaviors and experiences: A meta-analytic and theoretical review. *Psychological Bulletin*, 128(4):539–579.

Gintis, H. (2009). *The bounds of reason: Game theory and the unification of the behavioral sciences.* Princeton University Press.

Gul, F. and Pesendorfer, W. (2001). Temptation and self-control. *Econometrica*, 69(6):1403–1435.

H. Wold, G. L. S. S. and Savage, L. J. (1952). Ordinal preferences or cardinal utility? *Econometrica*, 20(4):661–664.

Hamilton, W. D. (1964a). The genetical evolution of social behaviour. i. *Journal of Theoretical Biology*, 7:1–16.

Hamilton, W. D. (1964b). The genetical evolution of social behaviour. ii. *Journal of Theoretical Biology*, 7:17–52.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. Technical report, National Bureau of Economic Research.

Herold, F. (2012). Carrot or stick? the evolution of reciprocal preferences in a haystack model. *The American Economic Review*, 102(2):914–40.

Hobbes, T. (1949). De cive (the citizen)[1651]. *New York: Appleton-Century-Crofts.*

Hoffman, M. L. (1975). Moral internalization, parental power, and the nature of parent-child interaction. *Developmental Psychology*, 11(2):228–239.

Hollis, M. and Sugden, R. (1993). Rationality in action. *Mind*, 102(405):1–35.

Hudson, J. L. and Rapee, R. M. (2001). Parent-child interactions and anxiety disorders: an observational study. *Behaviour Research and Therapy*, 39:1411–1427.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–477.

Lepper, M. R. and Cordova, D. I. (1992). A desire to be taught: Instructional consequences of intrinsic motivation. *Motivation and Emotion*, 16(3):187–208.

Lewis, D. K. (1969). *Convention: A Philosophical Study*. Blackwell Publishers, 2002 edition.

Lindbeck, A. and Nyberg, S. (2006). Raising children to work hard: Altruism, work norms, and social insurance. *The Quarterly Journal of Economics*, 121(4):1473–1503.

Lindbeck, A. and Weibull, J. W. (1986). Intergenerational aspects of public transfers, borrowing and debt. *The Scandinavian Journal of Economics*, pages 239–267.

Lizzeri, A. and Siniscalchi, M. (2006). Parental guidance and supervised learning. Technical report, Discussion paper//Center for Mathematical Studies in Economics and Management Science.

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

McKelvey, R. D. and Parlfrey, T. R. (1992). An experimental study of the centipede game. *Econometrica*, 60(4):803–836.

Meier, S. and Sprenger, C. (2010). Present-biased preferences and credit card borrowing. *American Economic Journal: Applied Economics*, 2:193–210.

Nowak, M. and Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, 364:56–58.

O'Donoghue, T. and Rabin, M. (1999). Doing it now or later. *The American Economic Review*, 89(1):103–124.

Paternotte, C. and Grose, J. (2012). Social norms and game theory: Harmony or discord? *The British Journal for the Philosophy of Science*, 0:1–37.

Pinker, S. (2003). *The blank slate: The modern denial of human nature.* Penguin.

Plant, E. A. and Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, 75(3):811–832.

Raz, J. (1986). *The Morality of Freedom.* Oxford University Press.

Ryan, R. M. and Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25:54–67.

Samuelson, L. and Swinkels, J. (2006). Information, evolution and utility. *Theoretical Economics*, 1:119–142.

Savage, L. J. (1954). *The foundations of statistics.* Courier Dover Publications, 1972 edition.

Shafer-Landau, R., editor (2013). *Ethical Theory: An Anthology.* John Wiley & Sons, Inc. & Blackwell Publishers Ltd, second edition.

Simon, H. A. (1976). From substantive to procedural rationality. In *25 Years of Economic Theory*, pages 65–86. Springer US.

Singer, P. (1972). Famine, affluence, and morality. *Philosophy & Public Affairs*, 1(3):229–243.

Siong, C., Brass, S. M., Heinze, H.-J., and Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11(5):543–545.

Sobel, J. (2002). Putting altruism in context. *Behavioral and Brain Sciences*, 25:275–276.

Stigler, G. J. and Becker, G. S. (1977). De gustibus non est disputandum. *The American Economic Review*, 67(2):76–90.

Sugden, R. (1989). Spontaneous order. *The Journal of Economic Perspectives*, 3(4):85–97.

Sugden, R. (1993). Thinking as a team: Towards an explanation of nonselfish behavior. *Social Philosophy & Policy*, 10(1):69–89.

Tabellini, G. (2008). The scope of cooperation: Values and incentives. *The Quarterly Journal of Economics*, 123:905–950.

Wilhelm, M. O. (1996). Bequest behavior and the effect of heirs' earnings: Testing the altruistic model of bequests. *The American Economic Review*, pages 874–892.

Zafirovski, M. (2003). Human rational behavior and economic rationality. *Electronic Journal of Sociology*, 7:1–40.

Chapter 3

Consistency of pro-social preferences - The case of aversion to advantageous inequality

3.1 Introduction

The experimental literature in economics abounds with studies on individual behaviour in strategic settings. The results of those studies have presented a strong case for the fact that people's behaviour is not always in line with the paradigm of the rational individual who is solely driven by own-payoff concerns.[1] Behavioural economists have engaged in various attempts to change the conception of the representative economic agent, so as to reconcile it with the experimental findings. A particularly popular class

[1] Fehr and Schmidt (2006), and Binmore and Shaked (2010) provide interesting overviews and discussions on the concepts of own-payoff maximisation and selfishness. They also analyse critically the refutations of these concepts in the experimental and behavioural literatures.

of such endeavours involves the concept of other-regarding preferences.

Models of other-regarding preferences do not attack the principle of rationality, according to which an individual strives to maximise her/his utility, but rather focus on the subjective nature of that utility. Specifically, they posit that one's preferences may well be driven by concerns other than the maximisation of one's personal monetary payoff. Further, these concerns may be related to the distribution of payoffs among oneself and other agents, or the actions required to attain those payoffs and the underlying intentions, or both. Equality in the distribution of payoffs (e.g. Bolton, 1991; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), reciprocity (e.g. Rabin, 1993 ; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006), and altruism (e.g. Becker, 1976) are some examples of models of other-regarding preferences.

The empirical validity of these accounts has been the focus of many experimental studies. Methodologically, such studies attempt to evaluate the performance of the models by testing their accuracy or consistency at tracking behaviour. Fehr and Schmidt (2006) detail and review a large body of related evidence from the experimental literature. Bruhin, Fehr, and Schunk (2016) construct a structural model of preferences and conduct an experiment to measure outcome- and reciprocity-based social preferences. They find that social preferences dominate in their sample and that genuinely selfish preferences do not in fact emerge. In their study, all three types of (endogenously emerging) preferences assign higher weights on the payoffs of others when their own payoffs are higher than they do when their own payoffs are lower. Additionally, they report that preferences over the distribution of payoffs generally dominate reciprocity concerns.

A particular account of other-regarding preferences that has received a lot of attention in the experimental literature is the model of inequality aversion proposed by Fehr and Schmidt (1999). According to this, individuals are concerned not only about their own material payoffs, but also about whether these are higher or lower than those of the people they interact with. Specifically, people prefer, to some idiosyncratic extent, their payoffs to be equal to those of others. This model is fairly straightforward and parsimonious (preferences are expressed only over alternative outcome distributions), while it has proved quite powerful in accounting for aggregate behaviour in many classic games (see e.g. Güth, Kliemt and Ockenfels, 2003; Fehr, Naef, and Schmidt, 2006).[2]

Blanco et al. (2011) test the model of Fehr and Schmidt through an experiment involving four different one-shot two-player games. Each of their subjects is called to provide a decision in each of the player-roles in every game. They firstly elicit the subjects' inequality-aversion parameters by asking them to play in a modified dictator and an ultimatum game. Specifically, they determine each subject's values for the model parameters based on the actions (s)he chose as a dictator[3] and as a responder in the ultimatum game. Subsequently, they derive the Fehr-Schmidt model's predictions about behaviour in the other games they deploy, based on the elicited parameter values. Then, they use the decisions made in those other games (the proposer in the ultimatum game, a sequential prisoner's dilemma, and a public-good game) to test for consistency. They find that for the most part the model predicts fairly accurately the shares of people that will choose the different actions (pro-social vs selfish). At the same time, however, the model appears to have little explanatory power with

[2] Blanco et al. do mention that the model of Fehr and Schmidt has been shown to fail at accounting for behaviour in some specific games (e.g. Charness and Rabin, 2002; Engelmann and Strobel, 2004). However, this does not change the fact that it does, in fact, perform well in a wide variety of situations. Therefore, it is still of value to investigate the reasons underlying its performance, so as to obtain an idea of how one can improve upon it.

[3] The authors propose a modified version of the game that allows for the separation of various parameter values - see section 3.2 for a discussion on the limitations of the standard dictator game.

respect to the behaviour of each single individual. They conclude that the Fehr-Schmidt model is able to account for different behavioural motives that are relevant to different games. On the other hand, these motives are not necessarily correlated within each single subject and therefore the model fails at the within-subject level.

What is of particular interest here is that the model's failure to account for people's behaviour in Blanco et al. may be confounded the presence of strategic uncertainty. By 'strategic uncertainty' we mean uncertainty that is related to others' beliefs (of any order) and actions (Morris and Shin, 2002).[4] To see this, note that they elicited their parameter values in situations that do not involve strategic uncertainty.[5] These elicited parameter values, then, appeared to have no explanatory power with respect to behaviour in situations where such uncertainty is present. On the one hand, this may, indeed, indicate a failure of the model to consistently account for people's choices. On the other hand, however, the model's failure may simply be a consequence of the additional uncertainty related to the other player's decision. In this sense, the variation in behaviour (and the consequent failure of the model) may not be an issue of preferences, but rather one related to beliefs about the other player's actions. That is, a pattern of choices that is interpreted as indicative of unstable preferences may instead have resulted from volatile beliefs about what others think and plan to do.[6] Therefore,in such a setting any conclusions about the performance of a model of preferences are sensitive to this confound.

Some evidence in support of this argument is provided by the study

[4] Bradenburger (1993) distinguishes 'strategic' from 'structural' uncertainty, the latter referring to the fundamental causal and statistical structure of the situation at hand.

[5] Both a dictator and an ultimatum responder face no uncertainty with respect to the decisions of the people they are paired with. In the dictator's case the other person makes no decision anyway, while in the ultimatum case the other person's decision has become common knowledge at the time the responder makes her/his own.

[6] Hofstadter's (1985) term 'reverberant doubt' highlights the way in strategic uncertainty expands from the slightest concern, leading to unexpected outcomes.

of Blanco et al. itself. In particular, the single decision which the Fehr-Schmidt model predicted well at the individual level is that of the second mover in the sequential prisoner's dilemma. In this case, people's choices as second movers in the sequential prisoner's dilemma were consistent with the parameters elicited through their behaviour in the dictator game. Thus, the model performed well at the individual level in a situation that removes strategic uncertainty, like the one used to measure the subjects' preferences. In addition, Yang, Onderstal, and Schram (2016) find that the Fehr-Schmidt model performs well at the within-subject level only if reciprocal options are unavailable. They report that the ability to reciprocate others' actions lowers considerably the model's performance. Significant choice-set effects are also being reported by He and Wu (2016), while Dannenberg et al. (2007) also stress the importance of information about the types of one's co-players.

However, the performance of the model across player-roles that do not involve strategic uncertainty has only been evaluated once, for a specific pair of games. In this paper we expand the analysis by focusing exclusively on such situations. To do so, we introduce a series of games in which strategic uncertainty is absent from most player-roles. That is, in these player-roles all uncertainty related to other people's actions has been resolved at the time the decisions are made. Therefore, we can investigate behavioural variations related to preferences isolated from those related to beliefs about others' choices.

Using an approach similar to that of Blanco et al. (2011), we conduct an experimental evaluation of the performance of the Fehr-Schmidt account. Our experimental design features a number of different one-shot pairwise interactions in which each participant is called to engage. Our aim is to determine whether the Fehr-Schmidt model will succeed in tracking the behaviour of our participants across these situations.

For reasons of simplicity and clarity, we focus only on advantageous-inequality aversion, i.e. aversion for one's own payoff being higher than those of others. We measure this aversion by deploying payoff structures that imply either higher or equal payoffs to most player-roles with those of their partners.

We deploy three different games, which are variants of the dictator, the trust, and the lying game. The first two are akin to their traditional versions, while in the latter an agent is asked to report a random outcome truthfully. We describe each one in detail in section 3.3. These games share some important qualitative characteristics. In particular, each person has to make a distributive decision in all but one player-roles. Thus, almost all decisions are about allocations of payoffs and are made in the absence of strategic uncertainty. Our games differ mainly in two important ways.

The first is the process that leads a player to the position of making a distributive choice. In the dictator game the recipient makes no decision and, thus, has no way of influencing the dictator's choice. In the trust game the second mover (who effectively acts as a dictator) only gets to make a choice if the first mover decides to trust her/him. In the lying game the distributive decision of the single active player follows the observation of a random draw.

The second refers to the additional motives that are relevant to each game and is partly a consequence of the differences in process. The dictator simply decides across various alternative payoff allocations. Thus, any motives related to choices of the other player or exogenous outcomes are unlikely to be relevant. In the trust game the second mover only gets to play if the first mover trusts. Therefore, (s)he may be (partly) driven by an additional motive to reciprocate. In the lying game the decision maker is asked to report truthfully an outcome of chance, so lying aversion is likely to be relevant. In sum, each of the three games features a potentially different motive structure.

We investigate whether and how such differences influence behaviour in the absence of strategic uncertainty. Crucially, we can evaluate these influences within a setting of consequentially similar decisions. Our results suggest that the Fehr-Schmidt model fails to account for people's behaviour. It appears that the model's predictions are broadly inconsistent with the actions our subjects choose. We find, however, that the model's failure is not symmetric across all players. Instead, it performs considerably better in accounting for the behaviour of people who exhibit a strong adherence to their respective motives. These are the ones who seem solely concerned about the maximisation of their own payoff and those who exhibit very strong preferences for equality in payoffs. The choices of these people are consistent with the predictions of the Fehr-Schmidt model.

By contrast, people who manifest moderate aversion to payoff-inequality do not do so consistently. Their patterns of choices also appear inconsistent with a range of different pro-social motives. We therefore conclude that there are two main ways to interpret their behaviour. The first is that these people appear (moderately) averse to inequality, while in truth they are driven by different motives (potentially not other-regarding). The second is that the stability of people's preferences depends on how strong these preferences are. Individuals who are strongly motivated adhere to their preferences more consistently than those who do not, irrespectively of whether they are pro-social or entirely selfish. These two lines of reasoning are not mutually exclusive and bear important implications for future research.

3.2 Fehr-Schmidt utility

Consider an interaction among n players. The model of inequality aversion proposed by Fehr and Schmidt (1999) champions the following utility function for a representative player i:

$$
\begin{aligned}
U_i(s_i, s_{-i}) = x_i(s_i, s_{-i}) - \alpha_i \frac{1}{n-1} \sum_{j \neq i} \max\{x_j(s_i, s_{-i}) - x_i(s_i, s_{-i}), 0\} - \\
-\beta_i \frac{1}{n-1} \sum_{j \neq i} \max\{x_i(s_i, s_{-i}) - x_j(s_i, s_{-i}), 0\}
\end{aligned}
$$

Here, s_i represents the strategy deployed by player i, s_{-i} stands for the collection of strategies of all the other players, and $x_k(s_i, s_{-i})$ denotes the payoff accruing to player k from the strategy profile (s_i, s_{-i}). Furthermore, α_i is the parameter that measures i's aversion to disadvantageous inequality, while β_i is the parameter that measures i's aversion to advantageous inequality. The term 'disadvantageous inequality' refers to situations where player i's payoff is lower than those of her/his counterparts. Conversely, 'advantageous inequality' refers to cases where player i receives a payoff that is higher than those of the other players.

In the two-player case and focusing only on aversion to advantageous inequality being experienced by player i, the expression above reduces to:

$$
U_i(s_i, s_j) = x_i(s_i, s_j) - \beta_i \max\{(x_i(s_i, s_j) - x_j(s_i, s_j)), 0\} \qquad (3.2.0.1)
$$

Equation 3.2.0.1 simply states that player i receives positive utility from her/his own material payoff, but also suffers a utility loss equal to the difference between her/his payoff and that of the other player weighted by the idiosyncratic parameter β_i. Note that the omission of α_i from equation 3.2.0.1 is only meant to simplify the exposition given our focus on β_i and is not illustrative of any assumptions on the degree of aversion to disadvantageous inequality experienced by the players.

Fehr and Schmidt (1999) make a number of a priori assumptions regarding the distributions of their model's parameters. The one that is relevant to our investigation is that $0 \leq \beta_i < 1$. The fact that $0 \leq \beta_i$ rules out the possibility of an individual experiencing satisfaction from having obtained a higher payoff than others. The restriction $\beta_i < 1$ postulates that no individual will burn part of her/his own payoff in order to reduce payoff inequality.

Our experimental design is such, that we can obtain a measurement of β_i for each player. We compute these measurements based on our subjects' behaviour in a modified dictator game. We then deploy two more games, for which we form predictions about how individuals with given β_i values will behave. We evaluate the model's performance by checking whether its predictions are in line with people's actual behaviour.

The game we deploy to elicit our subjects' preferences is a variant of the dictator game (Forsythe et al., 1994). As Fehr and Schmidt (1999) pointed out, the traditional version of the game is not suitable for getting a point prediction of the advantageous-inequality parameter. The reason is that due to the linearity of the transfers between the dictator and the recipient, subjects can only be categorised in two broad groups: those with $\beta_i \leq 0.5$, who should choose to keep the whole amount to themselves, and the ones with $\beta_i \geq 0.5$, who should choose the equal split. Such a coarse classification of β does not allow for the formation of sufficiently detailed hypotheses. Therefore, the standard dictator game is unsuitable for our analysis.

For this reason, we deploy a modified version of the game, which affords us greater precision in the measurement of our subjects' β_i values. In our variant each consecutive increase in the payoff of the recipient is progressively more expensive for the dictator. That is, in order to increase the recipient's payoff further, the dictator has to sacrifice an ever-growing amount of her/his own money. With this new payoff structure we are able to characterise many more β_i threshold values and, thus, pinpoint each subject's one more precisely.

To understand the mechanism of this classification, it is helpful to consider an example based on our dictator game. In it the dictator has to decide among ten possible allocations, ranging from the most selfish (keep all the surplus) to the most egalitarian one (divide the surplus equally). Again, the crucial feature of our variant is that the total surplus (the sum of payoffs) varies across the ten actions. This allows for intermediate actions to be optimal given certain β_i values. By choosing one of the ten allocations, the dictator expresses a weak preference for the chosen action over the rest available. Thus, labelling x_o and y_o the payoffs accruing to the dictator and the recipient respectively from action o, we can conclude the following, regarding the immediately previous $(o-1)$ and the immediately next $(o+1)$ action.[7]

$$U_i(x_o) \succeq U_i(x_{o-1}) \Leftrightarrow x_o - \beta_i(x_o - y_o) \geq x_{o-1} - \beta_i(x_{o-1} - y_{o-1}) \quad (3.2.0.2)$$

[7] Notice that since the actions range from the most selfish to the most egalitarian, $x_o \geq y_o$ and, thus, $\max\{(x_a - y_a), 0\} = (x_a - y_a), \ \forall o$.

$$U_i(x_o) \succeq U_i(x_{o+1}) \Leftrightarrow x_o - \beta_i(x_o - y_o) \geq x_{o+1} - \beta_i(x_{o+1} - y_{o+1}) \quad (3.2.0.3)$$

Since β_i is the same across both sides of each of the above inequalities, 3.2.0.2 and 3.2.0.3 imply, respectively, that:

$$\beta_i \geq \frac{x_{o-1} - x_o}{x_{o-1} + y_o - x_o - y_{o-1}} \quad (3.2.0.4)$$

$$\beta_i \leq \frac{x_o - x_{o+1}}{x_o + y_{o+1} - x_{o+1} - y_o} \quad (3.2.0.5)$$

That is, with an appropriate payoff structure β_i can be restricted within these two bounds. Moreover, an appropriate payoff structure will allow for this classification given any choice of the dictator. It is easy to see that the common bound of two consecutive choices is the same, i.e. the supremum β_i corresponding to action o is the infimum β_i corresponding to action $o + 1$.[8] We deploy such payoff structures in our games, presented in the following section.

3.3 Experimental design

Our experiment consists of three one-shot two-player games, a dictator, a trust, and a lying game. The trust game is sequential and involves two

[8] In other words, the greatest value of β_i for which player i will choose action o is the lowest value of β_i for which (s)he will choose action $o + 1$.

player-roles deciding, while in each of the dictator and the lying game only one player-role is making a decision. We deploy a within-subject design. That is, each subject participates in every one of our games. Furthermore, participants are asked to provide decisions in both player-roles in every game prior to learning their actual role (role uncertainty). Finally, in our sequential game we use the strategy method to elicit people's choices as second movers.

As mentioned in section 3.2, the first game we use is a variant of the dictator game. In designing it, we tried to remain as close to the traditional version as possible.[9] Our modified version is described in Table 3.1. It features two players, A (the dictator) and B (the recipient).

Table 3.1
Dictator game - Payoffs and associated β_i threshold values

A's action	A's payoff	B's payoff	β_i-threshold
ONE	£18.00	£0.00	1/10
TWO	£17.80	£1.80	2/10
THREE	£17.40	£3.40	3/10
FOUR	£16.80	£4.80	4/10
FIVE	£16.00	£6.00	5/10
SIX	£15.00	£7.00	6/10
SEVEN	£13.80	£7.80	7/10
EIGHT	£12.40	£8.40	8/10
NINE	£10.80	£8.80	9/10
TEN	£9.00	£9.00	

Table 3.1 contains all actions available to player A and the resulting payoffs for both A and B in the first three columns. Each cell in the last column contains the threshold parameter value of advantageous-inequality

[9] The modified version proposed by Blanco et al. (2011) involves each subject choosing one in each of 21 pairs of allocations. In every pair, the left option always implies £20.00 for the dictator and £0.00 for the recipient. The right option, on the other hand, implies an equal payoff for both participants, ranging from £0.00 to £20.00. Their modified game allows them to distinguish among several parameter values, because the transfers between the dictator and the recipient are no longer linear. However, it differs substantially in form from the standard version of the dictator game. Indeed, the version of Blanco et al. involves an additional element of chance, that can be described as Nature's decision: the dictator does not know which of her/his 21 choices will eventually be implemented.

aversion for which a given player would be indifferent between choosing the action on the same line and the immediately next one. Thus, for example, the threshold value of 3/10 in the dictator game is the one for which a dictator would be indifferent between action THREE and action FOUR.

Our variant is characterised by two appropriate modifications relative to the standard version of the game. The first is that the ten available actions proceed from allocating the whole of the sum to the dictator to distributing it equally between the two players. The second is that every additional amount transferred is more expensive for the dictator. This feature allows us to compute meaningful threshold values for the parameter measuring advantageous-inequality aversion, as discussed above.

It is important to note, however, that our estimates are prone to be biased by concerns about social efficiency. To see this, notice that social efficiency is maximised with actions FIVE and SIX, which correspond to β_i values in $[0.4, 0.5]$ and $[0.5, 0.6]$ respectively. The presence of concerns about efficiency, then, will lead to overestimated β_i values for subjects with 'true' β_i lower than 0.4 and underestimated β_i values for those with 'true' β_i higher than 0.6. This confound is inevitable if linearity is broken, so that more precise parameter thresholds can be computed. The reason is that obtaining more than one parameter thresholds for switching from one action to another requires each successive switch to be more expensive than the previous one at the margin. In the context of our dictator game, this implies that for each additional pound the recipient earns the dictator has to part with a larger sum. Blanco et al. face the same problem in their modification, where efficiency in the egalitarian option increases monotonically across their pairs of choices. Thus, they may have ended up with inflated estimates of all their subjects' β_i values, simply due to inequality aversion and concerns about efficiency being mixed together.[10] We do, however, find some evidence that efficiency-related distortions are small. We discuss this issue in the analysis of our results, in section 4.4.

Recall that we use our dictator game to estimate our subjects' β_i parameters. To evaluate consistency, we then deploy two more games, which feature similar distributive choices. This allows us to provide a basis for comparison with the behaviour in the dictator game, in line with our research focus. Our games are such, that we can draw predictions about the way our subjects will behave, based on our estimates of their β_i parameters. We then compare our predictions with their actual behaviour in each of those games, to evaluate the performance of the Fehr-Schmidt model.

As mentioned previously, our three games share some crucial qualitative characteristics. Specifically, they feature decisions that are defined over alternative allocations of payoffs between the decision-maker and another player. *These decisions do not involve any uncertainty related to the other player's choices.* The differences across our games are, instead, related to the process that leads a player to make a distributive choice and the way this choice is made. Thus, we can focus on the changes of behaviour in response to additional other-regarding motives and procedural changes.

The second game used in our experiment is a version of the trust game (Berg et al., 1995). This game is sequential. It is played by two agents, X (the trustor), who moves first, and Y (the trustee), who moves second. Crucially, the payoff structure corresponding to the choices available to agent Y is similar to that faced by the dictator in our previous game. Specifically, every increase in X's payoff comes at a progressively higher cost to Y.

[10] It is worth pointing out, as Blanco et al. (2011) do in footnote 20 of their paper, that under an alternative utility specification higher concerns might instead lead to a deflated estimate of a subject's true β_i, depending on how it compares with 0.5.

Agent X has to decide between action IN (trust Y) and OUT (do not trust Y). If X chooses OUT, then both agents get a payoff of £4.50. If X chooses IN, then Y gets to choose one of four options, as outlined in Table 3.2.

Table 3.2
Trust game - Payoffs and associated β_i threshold values

X's action	X's payoff	Y's payoff	β_i-threshold
IN	£?	£?	-
OUT	£4.50	£4.50	-
Y's action	Y's payoff	X's payoff	β_i-threshold
ONE	£16.60	£1.10	2/10
TWO	£15.75	£4.50	4/10
THREE	£13.75	£7.50	6/10
FOUR	£10.00	£10.00	

Here, the computation of parameter values corresponding to the choices of X requires additional assumptions regarding her/his expectations about the preferences of Y. Thus, they are not unique within a give payoff structure and therefore we refrain from including them in Table 3.2. The payoff structure following X's choice of IN, however, is such that the actions available to Y can be classified according to threshold values similar to those of the other two games.

It is important to notice that from the point of view of the second mover this game is similar to the dictator game. If X chooses OUT, then Y has no action to take anyway. If X chooses IN, on the other hand, then a Fehr-Schmidt Y acts precisely like a dictator, as her/his concerns are exclusively payoff-related. To the extent, however, that our participants' concerns are not exclusively payoff-related, the different way in which the game proceeds (relative to the dictator game) may have an influence on their behaviour. In particular, Y here does not get to play a part, unless X enables her/him to. Thus, upon deciding, Y knows what X has played.

This feature reveals intentions and allows them to play a role. For example, Y can readily interpret a decision of X to play IN as a kind move and, thus, may want to reciprocate. In this sense, Y's preference for reciprocity complements her/his aversion to advantageous inequality and may, thus, result in a return higher than that predicted by the Fehr-Schmidt model.

Finally, our third game involves a distributive decision in a different context. Its main difference with the dictator game is the way the decision-maker attains each allocation. In particular, all decision-makers are asked to report truthfully an outcome based on chance. On the other hand, they are free to misreport, as there is no control of whether their report is actually truthful. Their report itself is crucial, since it determines the eventual allocation of payoffs between themselves and the persons they are paired with. The aim here is to investigate how a change in the way the decision is to be made (and the additional motives implied) affects behaviour in an otherwise similar allocation problem. Due to this feature of its allocation process, we term this the lying game.

Our lying game features two people, J (the reporter) and K (the dependant). K is entirely passive in this situation. J is confronted with a spinning wheel, divided in three sections of equal size and different colour, namely red, blue, and green. J is asked to spin the wheel (by pressing a 'START' button). Upon activation, the wheel spins for a few seconds and then stops. Subsequently, J is asked to report the outcome of the wheel-spin, i.e. the colour the wheel has landed on. The report that J submits determines the payoff of both J and K, according to the scheme in Table 3.3.

The basic structure of this game is similar to that of a dictator game with a fixed initial surplus (equal to £17.00). The difference is in the setting. Specifically, the report of J in our game is to be determined by a

Table 3.3
Lying game - Payoffs and associated β_i threshold values

J's action	J's payoff	K's payoff	β_i-threshold
RED	£17.00	£0.00	5/10
BLUE	£8.50	£8.50	
GREEN	£0.00	£0.00	-

random draw. Notice that the motives of a pure Fehr-Schmidt person are independent of the random draw. That is, such a person's choice would be exactly the same across all possible draws. Notice, further, that such a person would never choose to report GREEN (hence the absence of a relevant threshold value for the β parameter). Indeed, reporting GREEN is dominated by at least one of the other two options for every value of the parameter. However, again, the setting allows for a variety of other motives to influence people's behaviour. For example, to the extent that our participants are concerned about reporting truthfully, we should expect some deviation from the model's predictions towards the actual outcomes of the wheel-spin.

The experiment was conducted in pen-and-paper format, except for part of the wheel-spin in the lying game, which was conducted in z-Tree (Fischbacher, 2007). In particular, in the lying game subjects received the instructions on paper (and also heard them aloud from the experimenter) and also had to submit their decisions on the relevant decision sheets (i.e. on paper). However, they had to activate and observe the spinning wheel on their computer screens.

We chose this setup in order to address a potential issue related to our lying game. Specifically, we wanted subjects to feel free to lie if they wanted to, unhindered from considerations related to the possibility of them being detected. On the other hand, we did need to know the true outcome of the wheel-spin, as otherwise we would not be able to test for consistency.

Thus, we wanted to downplay the notion that the experimenter would learn the actual outcome of each wheel-spin, in addition to the corresponding report. Therefore, we created a separation between the observation and the reporting of the outcome in this manner. We asked the subjects to spin the wheel on their computer screens (by pressing a start button), observe the outcome, and then write it down on paper.

Each of the three games was presented to the subjects in a separate section of the experimental session. At the beginning of each session the subjects were introduced to the proceedings of the experiment. Afterwards, the instructions for the first task (the dictator game), which had already been distributed, were read aloud. Subsequently, decision sheets were distributed to the subjects. Each was asked to provide her/his decisions in the role of the dictator on them and place them inside one of three envelopes placed on their desk. Once everyone had done so, their envelopes were collected and the instructions for the second task were then distributed. The process remained the same for the following two tasks. The three tasks were followed by a questionnaire.

The order in which the games were presented to the participants was fixed. In every session, the Dictator Game was presented first, followed by the Trust Game. The Lying Game was the third and last one always. We deemed it necessary to maintain this order for two main reasons.

The first one pertains to the dictator game. Specifically, we use the decisions made in it to locate the individual parameter values. Not only it is the simplest one to understand, but also it provides the largest range of actions and, thus, the finest parameter classification affordable in our three games. For this reason, we wanted to eliminate any noise in our measurements owing to them being obtained in later stages of the experiment. This is why we opted for presenting it first to everyone.

The second reason refers to the lying game. In particular, we did not want any subsequent interactions, so that concerns related to further experimental stages would not prevent participants from lying if they wanted to. Thus, we decided to present it after the other two in every session.

Each of our 178 participants played each game once and provided decisions for all player-roles. That is, each subject provided four decisions in total: one in the Dictator Game, two in the Trust Game (as first and second mover), and one in the Lying Game. The participants were made aware that they would be paired up and randomly assigned roles for a randomly chosen game at the end of the experiment. The payoff of each pair would be then determined by the decisions of the players in their assigned roles. The participants received no feedback or payment until the end of the experiment. In our games the joint payoff attainable for each pair ranges between £17.00 and £22.00 within and across the games. Each participant received an additional £2.00 show-up fee. Each session lasted for approximately one hour and thirty minutes. The average payoff per participant was about £10.50.

3.4 Results

In order to conduct our analysis, we first use the decisions made in the dictator game to classify our participants' β_i values in their respective intervals. The distribution of the β_i parameter is summarised in Table 3.4. Overall, the average β_i value in our dictator game lies within the interval $[0.3, 0.5]$, with the median one positioned in $[0.4, 0.5]$.

The distribution of β_i values in our dictator game is quite similar to both that of Blanco et al. and the one derived by Fehr and Schmidt

Table 3.4
Distribution of β - Observations in our data vs Fehr-Schmidt (1999) assumptions and data in Blanco et al.(2011)

Dictator	β_i intervals	Relative frequencies	β_i	Our data	Fehr-Schmidt*	Blanco et al.*
ONE	$\beta_i \leq 0.1$	26%	$\beta_i \leq 0.2$	30%	30%	29%
TWO	$0.1 \leq \beta_i \leq 0.2$	4%				
THREE	$0.2 \leq \beta_i \leq 0.3$	8%	$0.2 \leq \beta_i \leq 0.5$	29%	30%	15%
FOUR	$0.3 \leq \beta_i \leq 0.4$	1%				
FIVE	$0.4 \leq \beta_i \leq 0.5$	20%				
SIX	$0.5 \leq \beta_i \leq 0.6$	19%	$0.5 \leq \beta_i$	41%	40%	56%
SEVEN	$0.6 \leq \beta_i \leq 0.7$	6%				
EIGHT	$0.7 \leq \beta_i \leq 0.8$	6%				
NINE	$0.8 \leq \beta_i \leq 0.9$	2%				
TEN	$0.9 \leq \beta_i$	8%				

* *Note:* Recall that Fehr and Schmidt (1999) and Blanco et al. (2011) use 0.235 as their first threshold value - see footnote 11 for more details.

(1999).[11] This conclusion is supported by χ^2 and Fisher's exact tests. Specifically, our data and the distribution assumed by Fehr and Schmidt appear similar at all levels of statistical significance ($\chi^2 = 0.4409, d.f. = 2, p = 0.802$; Fisher's exact: $p = 0.819$). Regarding the comparison with the distribution in Blanco et al., the differences are more pronounced, but still at most borderline significant ($\chi^2 = 5.5289, d.f. = 2, p = 0.063$; Fisher's exact: $p = 0.062$).

We deem it particularly important that our results are in agreement with those of the other two aforementioned studies, for a number of reasons. To start with, this comparison constitutes an instrument check. Our success in replicating previous observations is crucial for the significance of our analysis (see Andreoni et al., 2003). Furthermore, even though each of the two previous studies agrees with the distribution observed in our data, Blanco et al. report significant differences between their results and the distribution assumed by Fehr and Schmidt. Our results lie somewhere in

[11] In the categorisation proposed by Fehr and Schmidt (1999) the β_i threshold value between the first and the second group is 0.235. This is also the threshold Blanco et al. (2011) use to compare their own distribution with the Fehr-Schmidt assumptions. To make our distributions comparable, we have to allocate our subjects who chose action THREE in the dictator game either in the $\beta_i \leq 0.235$ or in the $0.235 \leq \beta_i \leq 0.5$ group. We find that even under the most extreme allocations our distribution is statistically similar (at least at the 5% level) to both that observed in Blanco et al. and the one assumed by Fehr and Schmidt.

the middle between the two. Importantly, the distributions are similar despite the fact that our payoff structure is different from both that of a traditional dictator game and the variant deployed by Blanco et al.

On the other hand, the distribution of β_i values in our sample appears quite dissimilar to that found by Yang et al. (2016) and the one in He and Wu (2016). Indeed, χ^2 tests on the proportions of subjects within certain groups of parameter values indicate that our findings differ significantly, at least at the 5% level.[12] It is worth noting that both these studies use elicitation procedures that extend the β-value space to include negative values. Thus, the discrepancies between the distributions may be taken as evidence of choice-set dependency. However, such inferences need to be made cautiously: we are, after all, comparing games that differ in more than one dimensions.

We now turn to the issue of consistency. Given our subjects' parameter values, it is possible to form predictions about their decisions in the other two games. These predictions can then be compared with their actual choices. We focus on each of the two games separately and form specific hypotheses about people's behaviour. We then test these hypotheses to evaluate the explanatory power of the Fehr-Schmidt model. We conduct non-parametric and regression analysis.

3.4.1 Consistency in the trust game

Consider the following hypothesis (payoffs accruing to each action in parentheses, payoff of Y first):

Hypothesis 1. Agent i as a second mover in the trust game will choose:

[12] Our tests return $\chi^2 = 42.304, d.f. = 2, p = 0.000$ for the comparison with Yang et al. and $\chi^2 = 7.4169, d.f. = 2, p = 0.025$ for the comparison with He and Wu.

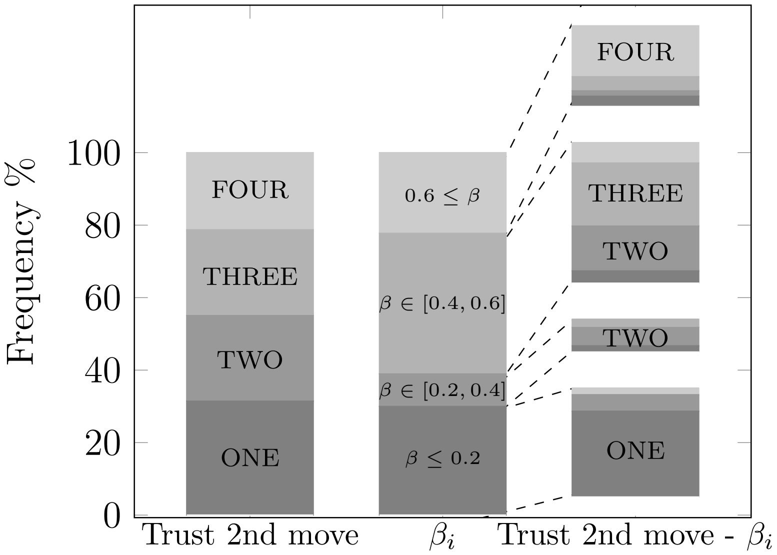


Figure 3.1: 2nd-move responses and parameter values in the trust game

- Action ONE (£16.60 , £1.10), iff $\beta_i \leq 0.2$
- Action TWO (£15.75 , £4.50), iff $0.2 \leq \beta_i \leq 0.4$
- Action THREE (£13.75 , £7.50), iff $0.4 \leq \beta_i \leq 0.6$
- Action FOUR (£10.00 , £10.00), iff $0.6 \leq \beta_i$

Our findings are summarised in Figure 3.1. The first column outlines people's choices, while the shares of the corresponding β_i values are depicted in the second one. The third column presents the choices of each β-group separately. Overall, we observe 31.5% of our participants choosing action ONE, 23.6% choosing TWO, 23.6% choosing THREE, and 21.3% choosing FOUR as second movers in the trust game. At the same time, about 30% of them are characterised by $\beta_i \leq 0.2$, 9% by $0.2 \leq \beta_i \leq 0.4$, 39% by $0.4 \leq \beta_i \leq 0.6$, and 22% by $0.6 \leq \beta_i$. Thus, there is a substantial discrepancy between the distribution of choices expected according to the Fehr-Schmidt model and that we actually observe. Simply put, we document a significantly higher share of action TWO and a lower share of action THREE than what the model has predicted. Our statistical tests

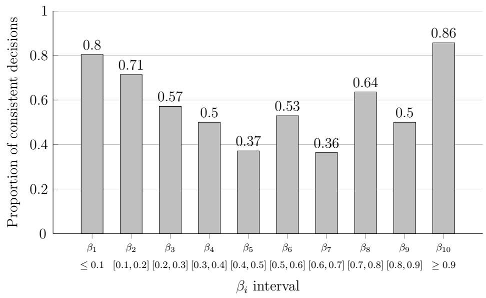


Figure 3.2: Trust game - Proportions of 2nd-mover decisions consistent with model's predictions across all β_i intervals

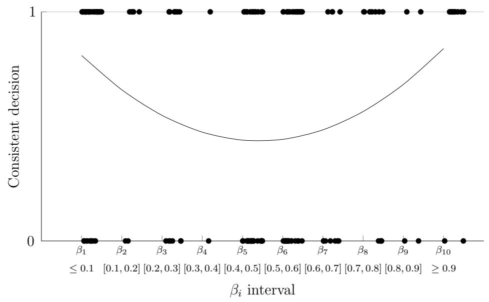


Figure 3.3: Trust game - Decisions and predicted probabilities of consistency across all β_i intervals

confirm this discrepancy. The difference between the two distributions is statistically significant ($\chi^2 = 18.357, d.f. = 3, p = 0.000$; Fisher's exact: $p = 0.000$). Thus, the model fails to predict our subjects' behaviour.

A feature of interest in the model's failure is the fact that it is not symmetric across the β_i values. At a first level this is obvious in Figure 3.1. Specifically, among the people with β_i values lower than 0.2 or higher than 0.6 we observe substantially high rates of consistency with the model's predictions (79% and 63% respectively). By contrast, people with $\beta_i \in [0.2, 0.4]$ exhibit a consistency rate of 56%, while those with $\beta_i \in [0.4, 0.6]$ an even lower one (45%). However, a direct comparison across those four groups in terms of their consistency rates would be confounded by the fact that they are asymmetric in size. That is, almost half of the parameter-value intervals are concentrated in the same group ($\beta_i \geq 0.5$). In order to take this into account, we proceed to a finer classification of the β_i values, namely the one afforded to us by the dictator game.

Figure 3.2 presents the proportions of consistent subjects for all dictator actions. Recall that each of these ten actions is associated with a specific β-value interval. One can easily see that the deviation rate forms a U-shaped pattern. That is, people with very low (≤ 0.1) or very high (≥ 0.9) degrees of aversion to advantageous inequality tend to be more consistent with their preferences than people with intermediate such concerns. This finding is illustrated in Figure 3.3.

To test this hypothesis, we run a logistic regression. Recall that hypothesis 1 outlines all patterns of choices that are consistent with the F-S model. We consider the shares of consistent and inconsistent choices in the trust game across all β groups (as defined in the dictator game) and try to identify the model that provides the best fit to our data. That is, we evaluate the relationship between the probability of one's trust-game choice being inconsistent with one's β_i value and the β_i value itself, through a maximum-likelihood estimation.

Our results, presented in Figure 3.4, exhibit two main features of interest. To start with, we observe a statistically significant association between the magnitude of β_i and the probability of behaving in accordance with the F-S model. That is, the extent to which people are consistently inequality-averse appears, indeed, related to the degree of inequality aversion they manifest. Furthermore, a non-linear specification provides a significantly better fit than a linear one does ($\chi^2(1) = 16.06$, $p = 0.000$). Thus, it turns out that the change in the degree of consistency is not constant, but varies with β. In addition, this variation is non-monotonic. In particular, consider the probability that the trust-game choice will be inconsistent with the F-S model, as estimated by the logit regression. This probability attains its smallest values for the $\beta_i \leq 0.1$ and $\beta_i \geq 0.9$ groups and is maximised for $\beta \in [0.4, 0.6]$.[13]

Our analysis so far yields a number of conclusions. To start with, it is clear that the F-S model leaves at least some part of the behavioural variation unexplained, even when it performs well. This is evident by the

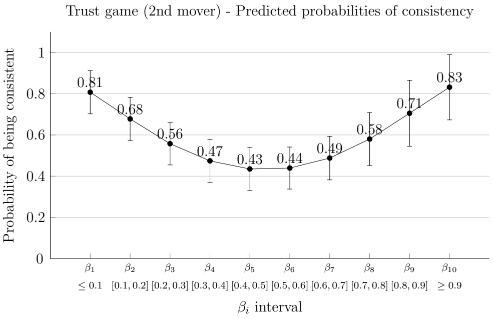
[13] The Table outlines the regression output, while the Figure depicts the estimated relationship.

Trust game (2nd mover) - Non-linear logistic regression of consistent behaviour on the β_i groups

Dependent variable:	Consistent β_i	Number of obs. LR χ^2:	178 21.22
Log likelihood:	-109.10215	Prob. $> \chi^2$: Pseudo R^2:	0.000 0.0886

	Coeff.	Std. Err.	z	$Pr[>\lvert z\rvert]$	[95% Conf. Interval]	
β_i	-0.954	0.222	-4.30	0.000	-1.388	-0.519
β_i^2	0.088	0.022	4.01	0.000	0.045	0.132
constant	2.297	0.507	4.53	0.000	1.304	3.290

The regression estimates the relationship between one's β_i value and the probability that one's behaviour is consistent with the Fehr-Schmidt model. The quadratic term is taken as continuous.



Trust game (2nd mover) - Predicted probabilities of consistency

Figure 3.4: Trust game - Estimated relationship between one's β_i value and the probability that one's decision is consistent with the Fehr-Schmidt model

fact that all estimated probabilities of consistent decisions are significantly different from one.[14] Moreover, the performance of the model varies significantly with the degree of advantageous-inequality aversion exhibited. The pattern of this variation is in line with our initial hypothesis. That is, people who exhibit very high ($\beta_i \geq 0.9$) or very low ($\beta_i \leq 0.1$) aversion to advantageous inequality behave are indeed significantly more consistent with their preferences. The variation in the degrees of consistency is substantial and significant, as it is evident in our regression. To provide additional support for this claim, we compare the distributions of consistent/inconsistent trust-game decisions across all dictator-game choices. We do so in a pairwise manner, using Fisher's exact tests. Our results, which can be found in Table 3.5, reaffirm our previous findings.

To summarise, the Fehr-Schmidt model broadly fails to account for our subjects' behaviour. This is primarily due to the fact that people with intermediate β_i values do not follow their preferences. That is, the model's performance varies with the strength of people's aversion to payoff inequality. According to our results from the trust game, people who exhibit very strong or very weak concerns for payoff inequality are doing so much more consistently than people who are moderately so averse.

[14] This is true at least at the 5% level, see Figure 3.4.

Table 3.5
Trust game - Statistical comparisons of differences in deviation rates across Dictator choices

	ONE	TWO	THREE	FOUR	FIVE	SIX	SEVEN	EIGHT	NINE	TEN
ONE	-	0.626	0.092	0.377	0.000	0.014	0.007	0.251	0.206	1.000
TWO		-	0.656	1.000	0.018	0.438	0.335	1.000	0.576	0.574
THREE			-	1.000	0.222	1.000	0.428	1.000	1.000	0.209
FOUR				-	1.000	1.000	1.000	1.000	1.000	0.350
FIVE					-	0.230	1.000	0.169	0.631	0.004
SIX						-	0.491	0.729	1.000	0.049
SEVEN							-	0.395	1.000	0.017
EIGHT								-	1.000	0.350
NINE									-	0.350
TEN										-

Each action in the first column is compared to every action in the first row, in terms of number of consistent and inconsistent participants. The null hypothesis is that the share of consistent participants is the same across all actions. We test this hypothesis using Fisher's exact tests. The p-value corresponding to the test between the action in row i and that in column j is reported in cell c_{ij}. Darker shades correspond to more significant differences. Thus, for example, the share of participants who are consistent with the Fehr-Schmidt model differs significantly between actions ONE and FIVE.

How can we interpret the fact that people with moderate β_i values deviate from the model's predictions at higher proportions? Perhaps the most straightforward inference is that in reality the proportions of inconsistent people do not really differ. Rather, it may be the case that deviations from each prediction are randomly dispersed around that prediction. Then, people in the middle of our β_i-value space can deviate from the model's predictions towards both directions, whereas people at the extremes can only do so in one way. To see this, recall that there are four actions available to the second mover in our trust game. People, then, with $\beta_i \in [0.2, 0.6]$ can violate the model's predictions by choosing either actions that are more egalitarian than what the model has predicted or ones that are less so. By contrast, people with very low ($\beta_i \leq 0.1$) or very high ($\beta_i \geq 0.9$) degrees of aversion to inequality can only invalidate the model by changing their behaviour in a single direction. In this sense, the pattern of deviations we observe may simply be due to the fact that our action space is bounded.

However, our data are not in line with this interpretation. For such a pattern to arise, people's violations of the model's predictions would need to be random. What we find, instead, is that these deviations are indicative of some patterns. In total, we observe substantially more deviations towards more selfish options than towards more egalitarian ones. It is true that the deviations of people with $\beta_i \in [0.5, 0.6]$ appear randomly spread around the mode's prediction. However, most of the people with $\beta_i \in [0.4, 0.5]$ behaved more selfishly than predicted. When we consider these two groups jointly, we see that only 26% of those deviating did so towards action FOUR, while 58% of them went for action TWO instead. Given that these two groups together contain about 39% of our subject pool (and are by far the most populated among the interior β_i groups), we conclude that people's deviations from the F-S model's predictions can not be rationalised as randomly occurring.

Exploring the potential for different other-regarding preferences, we argue that the observed pattern cannot be accounted for by reciprocity. If subjects were motivated by such concerns, we should observe disproportionately high frequencies of actions THREE and FOUR being chosen relative to what the β_i values suggest.[15] Instead, we find the opposite to be the case: the highest share of inconsistent subjects is to be found among those who should have chosen action THREE, many of whom opt for TWO instead. That is, the pattern of inconsistencies is the opposite of what one should expect if our subjects were reciprocal.

Remarkably, efficiency concerns do not help here either. Notice that the most socially efficient option in our trust game is, in fact, action THREE. Thus, efficiency concerns would result in action THREE being chosen more frequently that expected according to the Fehr-Schmidt model. Again, what we observe is exactly the opposite pattern. Action THREE is in fact chosen less frequently than expected. Therefore, the pattern we observe in our data goes directly against both reciprocity and efficiency

[15] These are agent Y's actions that would endow agent X with a payoff higher than what (s)he would have obtained had (s)he opted for OUT instead of IN in the first stage. Thus, these are the actions that a reciprocal agent would choose in response to a kind move by her/his partner.

considerations.

What may be the case is that people who are classified as moderately averse to payoff inequality are not in fact driven by such concerns. Instead, they may be driven by a desire to maintain a certain self-image, while trying to achieve as high a personal payoff as possible. People with such concerns will opt for the payoff-maximising action, so long as that action is not too 'morally deplorable', that is, it does not result in a too negative view of oneself. If this is the case, then their threshold for what constitutes an excessively deplorable action can be sensitive to many factors. For example, the degree of 'immorality' they assign to a specific action may vary with the other actions available. Additionally, their perception may be partly shaped by the social norms that are prevalent in their environment. In this sense, their views may be rank-dependent or socially determined (or both). That would result, for example, in them choosing a particular payoff allocation in one setting, but discarding a similar one in another.

At any rate, the Fehs-Schmidt model's performance is notably better with respect to extreme β_i values. The shares of people with very weak and very strong aversion to inequality are consistently predicted. In addition, the behaviour of those people is consistently accounted for by the model with a higher degree of accuracy relative to the behaviour of the moderately averse ones. We therefore conclude that the model predicts well the behaviour of some, but not all participants in our trust game.

3.4.2 Consistency in the lying game

We now turn our attention to behaviour in the lying game. We proceed to form the following two hypotheses (payoffs accruing to each action in parentheses, payoff of J first):

Hypothesis 2. No player will ever report GREEN (£0.00 , £0.00), irrespectively of the value of her/his β_i parameter.

Hypothesis 3. Irrespectively of the outcome of the wheel, player i as individual J in the lying game will report:

- RED (£17.00 , £0.00) iff $\beta_i \leq 0.5$
- BLUE (£8.50 , £8.50) iff $\beta_i \geq 0.5$

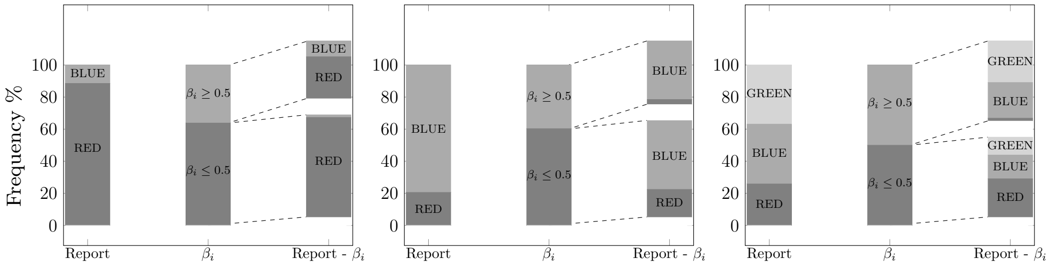
We begin with hypothesis 2. We observe that 11.25% of our participants reported GREEN. Thus, if we consider the model's predictions as deterministic, this hypothesis is confidently rejected (binomial test: $p = 0.000$). If, on the other hand, we allow for a probability of error in decision-making, then the model's prediction can be salvaged. Consider such a trembling-hand specification, according to which individual i will choose optimally given β_i with probability $p_{-\epsilon} = 1 - \epsilon$ and make a random erroneous decision with probability $p_\epsilon = \epsilon$.[16] Then, the focus turns to the magnitude of ϵ necessary to confirm hypothesis 2. It turns out that for our results to be rationalisable as random errors at the 10% level of significance, ϵ would need to be at least equal to about 0.12. That is, in order for our findings to be accountable for by this interpretation, our subjects would have to err about 12% of the time. We view this minimum error threshold as rather restrictive and implausible. Furthermore, the fact that all GREEN reports were provided in cases where the true outcome of the wheel-spin was GREEN (as we discuss below) indicates that they are not

[16] By random erroneous decision we mean here that the two non-optimal reports are chosen with equal probabilities. Due to the fact that the options in the lying game are not uniquely well-ordered, a uniformly random specification is the most reasonable option. In addition, by restricting the random decision to the two non-optimal reports, we provide a test that is favourable to the model: Had we allowed randomness across all three of them, an even higher probability of error would be needed to account for the number of GREEN reports we observe. We demonstrate that even under these more favourable conditions ϵ needs to be implausibly large.

really random errors. We conclude that the model does not manage to correctly predict the occurrences of GREEN reports.

We now proceed to test hypothesis 3. To be precise, we perform our tests given each actual outcome of the wheel-spin separately. The reason is that, as the hypothesis points out, the true colour the wheel has landed on should have no effect on the report of a Fehr-Schmidt agent. Had we tested across all wheel-spin outcomes, any effects of concerns about truthful reporting could have been misinterpreted as evidence for inequality aversion. By focusing on each outcome in isolation we can control for such effects.

We refer to each actual outcome of the spinning wheel as a state, in the sense that it is a product of chance that our participants find themselves in. Figure 3.5 presents the distributions of responses, parameter values, and responses conditional on parameter values within each state. We focus first on the participants whose wheels landed on red. Among them, 64% are characterised by $\beta_i \leq 0.5$, while 88.5% report RED. The distribution of parameter values is significantly different from that of reports ($\chi^2 = 10.178, d.f. = 1, p = 0.001$, Fisher's exact: $p = 0.003$). The same is true for those whose wheels landed on blue. Among the people who found themselves in that state 39.7% feature $\beta_i \geq 0.5$, while 79.4% reported BLUE. Again, the distributions differ significantly ($\chi^2 = 20.588, d.f. = 1, p = 0.000$, Fisher's exact: $p = 0.000$). In addition, the absence of any GREEN reports in these two states constitutes evidence against erroneous decision-making by our subjects. Finally, looking at the subjects whose wheels ended up on green, we observe 26% reporting RED, 37% reporting BLUE, and 37% reporting GREEN, while exactly 50% of them exhibit $\beta_i \geq 0.5$. We can confidently reject the hypothesis that the distributions are similar in this state, too ($\chi^2 = 25.164, d.f. = 2, p = 0.000$, Fisher's



(a) State RED **(b)** State BLUE **(c)** State GREEN

Figure 3.5: State-specific reports and parameter values in the lying game

exact: $p = 0.000$).

We can, thus, conclude that the Fehr-Schmidt model fails to account for the behaviour of our subjects in the lying game. It seems that the actual outcome of the wheel-spin exerts a strong influence on people's behaviour. The model appears unable to capture this influence, even when no one reports GREEN (as is the case in states RED and BLUE). As a result, it does not perform well in our within-state evaluations.

Given the model's poor performance in accounting for our subjects' choices, we examine again whether its predictive power varies across different parameter values. To do so, we evaluate the consistency of behaviour with the model's predictions across the same β-groups we used for the trust game. Thus, our results are readily comparable. Note that in state RED we focus our attention to people with $\beta_i \geq 0.5$, as those are the ones who would lie by reporting the Fehr-Schmidt prediction. For the same reason, we investigate the behaviour of people with $\beta_i \leq 0.5$ in state BLUE. In state GREEN all agents with Fehr-Schmidt type preferences would lie, so we can use our β classification in its entirety. As before, we conduct pairwise comparisons and logistic regressions across our β-intervals within each state. The estimated probabilities are reported in Table 3.6.

Starting from state RED, presented in Figure 3.5a, we find that about

Table 3.6
Logit estimates of variation in degree of consistency with model's predictions across β-groups

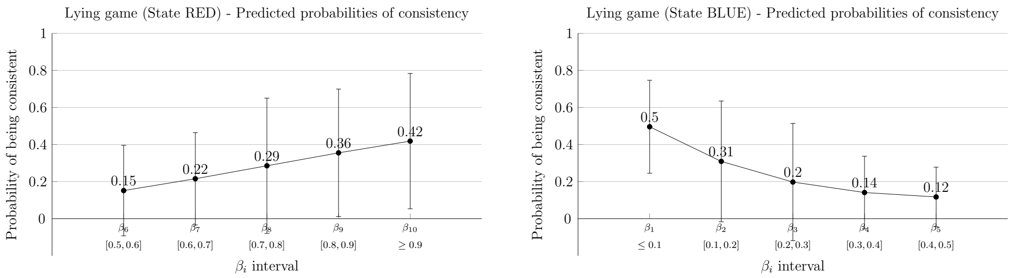
	State RED	State BLUE	State GREEN	
	Report BLUE	Report RED	Report RED	Report BLUE
$\beta_i \leq 0.1$	-	0.496 (0.128)	0.774 (0.110)	-
$\beta_\in[0.1, 0.2]$	-	0.309 (0.166)	0.585 (0.111)	-
$\beta_\in[0.2, 0.3]$	-	0.198 (0.161)	0.420 (0.101)	-
$\beta_\in[0.3, 0.4]$	-	0.141 (0.100)	0.318 (0.092)	-
$\beta_\in[0.4, 0.5]$	-	0.117 (0.082)	0.273 (0.084)	-
$\beta_\in[0.5, 0.6]$	0.152 (0.124)	-	-	0.274 (0.080)
$\beta_\in[0.6, 0.7]$	0.215 (0.127)	-	-	0.323 (0.085)
$\beta_\in[0.7, 0.8]$	0.285 (0.186)	-	-	0.428 (0.115)
$\beta_\in[0.8, 0.9]$	0.355 (0.175)	-	-	0.596 (0.167)
$0.9 \leq \beta_i$	0.418 (0.186)	-	-	0.784 (0.177)

Estimated probabilities of reports being consistent with the F-S model are reported next to the relevant β_i intervals and under their respective states (with standard errors in parentheses). The differences among the probabilities are statistically insignificant in states RED and BLUE. In state GREEN, on the other hand, they are highly significant (at the 1% level).

36% of the people whose wheels landed on this colour exhibit $\beta_i \geq 0.5$. Among them a non-negligible portion (27%) chose to report BLUE. Half of those who did so belong in the $\beta_i \geq 0.9$ group, while the rest are uniformly distributed across the remaining ones. We find that the probability of being consistent with the Fehr-Schmidt model is 2.75 times higher for the people in the $\beta_i \geq 0.9$ group relative to those in the $\beta_i \in [0.5, 0.6]$ one. However, this result is statistically insignificant, likely due to the low number of people with such parameter values in this state and the dominance of other concerns.[17]

Repeating the exercise in state BLUE, we find the same pattern, this

[17] Truthful reporting and self-serving bias are good candidates to be considered as dominant concerns here.



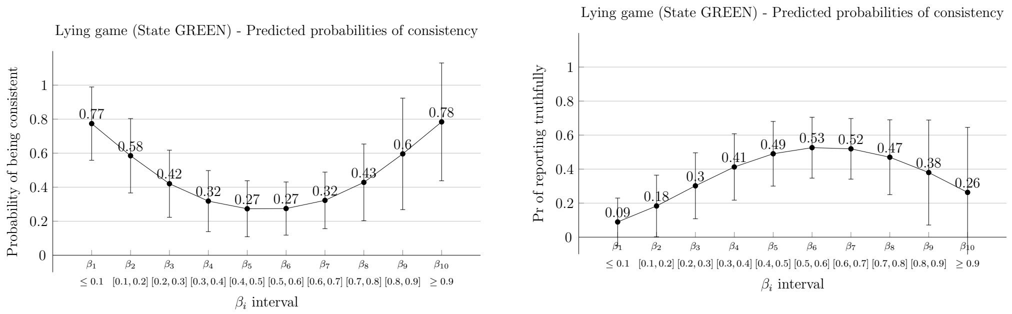
(a) Average predicted probabilities of consistency in state RED

(b) Average predicted probabilities of consistency in state BLUE

Figure 3.6: Lying game - Average predicted degrees of consistency across the β groups in states RED and BLUE

time among those with $\beta_i \leq 0.5$. Figure 3.5b depicts the situation. Here we observe about 60% of the people exhibiting $\beta_i \leq 0.5$. Of these, 29% chose to report RED. Thus, the model failed to account for the behaviour of those in the low-β_i group, most of whom behaved in the exact opposite way. On the other hand, people with $\beta \leq 0.1$ appear more consistent with the Fehr-Schmidt model than people with $\beta_i \in [0.4, 0.5]$. The estimated probability of being consistent with the model is 4.24 times higher for people in the $\beta_i \leq 0.1$ relative to those in the $\beta_i \in [0.4, 0.5]$ one. However, this result is, again, statistically insignificant.

Lastly, we turn to state GREEN, the only one in which we observe GREEN reports. In contrast to the model's predictions, the share of these reports is non-negligible. Figure 3.5c summarises our results. As mentioned before, 37% of the people in this state chose to report GREEN. The rest are divided between another 37%, who reported BLUE, and the remaining 26%, who reported RED. With respect to their parameter values, 50% of our subjects exhibit $\beta_i \leq 0.5$ and the other 50% feature $\beta_i \geq 0.5$. About 48% of the subjects in the $\beta_i \leq 0.5$ group opted for reporting RED. The vast majority of these (69%) are in the $\beta_i \leq 0.1$ group. Among those with $\beta_i \geq 0.5$ almost 44.5% reported BLUE. Interestingly, we observe a substantially high proportion of inconsistent decisions among the people with $\beta_i \in [0.5, 0.6]$, with only 29% of them reporting BLUE. Indicatively,



(a) Average predicted probabilities of consistency across the β groups

(b) Average predicted probabilities of reporting truthfully across the β groups

Figure 3.7: Average predicted degrees of consistency and truthfulness in state GREEN of the lying game

about 61% of those with $\beta_i \geq 0.6$ did the same (and thus were consistent). In total, the people with extreme β_i values are again significantly more consistent with the Fehr-Schmidt model than the moderate ones. The highest estimated probabilities of consistent decisions correspond to those with $\beta_i \leq 0.1$ and $\beta_i \geq 0.9$ (0.774 and 0.784, respectively). People with β_i in the $[0.4, 0.6]$ range are again the least likely to behave consistently (with an estimated probability equal to 0.273). This result, depicted in Figure 3.7a, is highly significant.

Overall, we find a strong propensity towards reporting truthfully. This propensity is particularly pronounced in states RED and BLUE. In state GREEN, predictably, truthful reports are less common, but still far from scarce. Interestingly, preferences for truthful reporting appear to be strongest for people with moderate β_i values (in the range $[0.4, 0.6]$), i.e. those who are the least consistent with the Fehr-Schmidt model.[18] Figure 3.7b depicts the change in the estimated probability of reporting truthfully across the

[18] We observe this pattern in the GREEN state. This is the only state that allows us to evaluate this hypothesis, as in it the two motives are always in conflict with each other. The probability of a truthful report is estimated at 0.09 (0.26) for people with $\beta_i \leq 0.1$ ($\beta_i \geq 0.9$) and at about 0.52 for those in the $\beta_i \in [0.4, 0.6]$ group. Additionally, when we focus on the other two states, we see that people with higher β_i values tend to lie significantly more in the RED state and significantly less in the BLUE one. Thus, we indeed observe the pattern of conflict between truthful reporting and inequality aversion that we wanted to generate through this game

different β_i-value intervals we obtain from the dictator game.

So, what do our results actually mean? It appears that behaviour is sensitive to a host of factors that the Fehr-Schmidt model cannot account for. Blanco et al. (2011) argue that the apparent success of the model is to be attributed to its ability to account for a variety of different behavioural motives. This, in turn, implies that these motives are well-aligned, i.e. they prescribe actions that yield similar payoff distributions. Indeed, when we focus on situations where there is conflict among different motives, the model appears unable to account for people's behaviour in general.

In addition, concerns about efficiency do not appear to exert a significant influence here, either. If anything, people who chose the most socially efficient actions (FIVE and SIX) in the dictator game exhibit stronger preferences for truthful reporting. That is, the share of subjects who reported GREEN truthfully is significantly larger among those who had chosen FIVE or SIX as dictators. Given that GREEN is the least efficient option, preferences for efficiency do not appear to cause much distortion in our measurements.

However, it is still the case that people with particularly high or low degrees of inequality aversion are significantly more consistent with the Fehr-Schmidt model than people with moderate such preferences. We observe a positive correlation between the degree of extremity in people's β_i values and their rates of consistency with the model's predictions in all three states of our lying game. In state GREEN this correlation is highly significant.

To summarise, our results indicate that in the lying game the state generally dominates responses. We find strong propensities towards reporting truthfully, even when doing so is counter-efficient and detrimental for one's personal payoff. Given this finding, however, we also observe that some people remain consistent to their β_i values. These people are primarily the ones who exhibit very low (≤ 0.1) or very high (≥ 0.9) such values. This latter result is in line with what we found to be the case in our trust game. Thus, it lends support to the idea that the Fehr-Schmidt model performs consistently well only for subjects with extreme β_i values.

3.5 Conclusion

Our evaluation of the Fehr-Schmidt model yields two main findings. The first is that the model appears generally unable to account for people's behaviour in a consistent manner. This is true even in the absence of strategic uncertainty (and the consequent strategic considerations). The second finding is that the rates of deviation differ across different types of individuals. People who manifest strong preferences either for their own material payoff or for equality in the distribution of payoffs do so throughout the experiment. Remarkably, these people are also more likely to lie to achieve their preferred payoff distributions. By contrast, those who exhibit moderate aversion to advantageous inequality (according to our parameter-value space) appear significantly less committed to their preferences.

With respect to the model's performance, our results differ from those of previous studies. Specifically, we find that the model fails to account for our subjects' behaviour in both the trust and the lying game. By contrast, Blanco et al. report that it has considerable predictive power at the aggregate level (apart from the case of the sequential prisoner's dilemma), while Yang et al. even find considerable individual-level performance. We attribute the differences in our findings to two main design features. Firstly, the second mover in our trust game has four available options. Accordingly, Hypothesis 1 provides a stricter test for the model than the binary-prediction hypotheses in Blanco et al. Secondly, our lying game may induce a conflict of motives. For example, individuals that exhibit concerns about truthful reporting in addition to inequality aversion may appear inconsistent with the model's predictions if these two motives prescribe different actions.

Fehr and Schmidt (1999) do point out that 'positive α_i's and β_i's can be interpreted as a direct concern for equality as well as a reduced-form concern for intentions. [...] As a consequence, our preference parameters are compatible with the interpretation of intentions-driven reciprocity.' Interestingly, we find that the model's failure cannot be accounted for by concerns about reciprocity. This finding also contrasts those of Blanco et al. and Yang et al., who conclude that reciprocal motives can account for at least some of the behavioural variation. Furthermore, the patterns of behaviour we observe appear incompatible with efficiency concerns, too.

Regarding the substantially higher rates of consistency among people with extreme (high or low) degrees of inequality aversion, our results point towards the existence of types (similar to Fischbacher and Gächter, 2006). We can thus discern between people who are very selfish or very egalitarian, who are consistently so throughout, and people who are 'in the middle' (in the model's sense), whose behaviour is inconsistent with inequality aversion. One interpretation is that this latter group is driven by different motives, which are not always correlated with inequality aversion. Another one is that they are unsure about their preferences and, thus, sensitive to environmental cues that sway them towards one direction or another. Both these interpretations are consistent with the observation that people with moderate β_i values ($\beta_i \in [0.4, 0.6]$) exhibit the highest degree of truthful reports in our lying game.

3.6 References

Andreoni, J., Castillo, M., & Petrie, R. (2003). What do bargainers' preferences look like? Experiments with a convex ultimatum game. *The American Economic Review, 93*(3), 672-685.

Andreoni, J., & Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica, 70*(2), 737-753.

Becker, G. S. (1976). Altruism, egoism, and genetic fitness: Economics and sociobiology. *Journal of Economic Literature, 14*(3), 817-826.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*(1), 122-142.

Binmore, K. (2005). Economic man - or straw man?. *Behavioral and Brain Sciences, 28*(06), 817-818.

Binmore, K., & Shaked, A. (2010). Experimental economics: Where next?. *Journal of Economic Behavior & Organization, 73*(1), 87-100.

Blanco, M., Engelmann, D., & Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior, 72*(2), 321-338.

Bolton, G. E. (1991). A comparative model of bargaining: Theory and evidence. *The American Economic Review*, 1096-1136.

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

Brandenburger, A. (1993). *Strategic and structural uncertainty in* *games*. Division of Research, Harvard Business School.

Bruhin, A., Fehr, E., & Schunk, D. (2016). The Many Faces of Human Sociality: Uncovering the Distribution and Stability of Social Preferences.

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

Dannenberg, A., Riechmann, T., Sturm, B., & Vogt, C. (2007). *Inequity Aversion and Individual Behavior in Public Good Games: An Experimental Investigation* (No. 07-034). ZEW-Zentrum für Europäische Wirtschaftsforschung/Center for European Economic Research.

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior, 47*(2), 268-298.

Engelmann, D., & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *The American Economic Review*, 857-869.

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior, 54*(2), 293-315.

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

Fehr, E., & Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism - experimental evidence and new theories. *Handbook of the Economics of Giving, Altruism and Reciprocity, 1*, 615-691.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1998). Gift exchange and reciprocity in competitive experimental markets. *European Economic Review, 42*(1), 1-34.

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

Fischbacher, U., & Gächter, S. (2006). Heterogeneous social preferences and the dynamics of free riding in public goods.

Fisman, R., Kariv, S., & Markovits, D. (2007). Individual preferences for giving. *The American Economic Review, 97*(5), 1858-1876.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic behavior, 6*(3), 347-369.

Goeree, J. K., & Holt, C. A. (2000). Asymmetric inequality aversion and noisy behavior in alternating-offer bargaining games. *European Economic Review, 44*(4), 1079-1089.

Güth, W., Kliemt, H., & Ockenfels, A. (2003). Fairness versus efficiency: an experimental study of (mutual) gift giving. *Journal of Economic Behavior & Organization, 50*(4), 465-475.

He, H., & Wu, K. (2016). Choice set, relative income, and inequity aversion: An experimental investigation. *Journal of Economic Psychology, 54*, 177-193.

Douglas R. Hofstadter. (1985). *Metamagical themas: Questing for the essence of mind and pattern.* Basic Books.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the assumptions of economics. *Journal of Business*, S285-S300.

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.

Levin, J. (2006). Fairness and Reciprocity.

López-Pérez, R., & Spiegelman, E. (2013). Why do people tell the truth? Experimental evidence for pure lie aversion. *Experimental Economics, 16*(3), 233-247.

Morris, S., & Shin, H. S. (2002). Measuring strategic uncertainty. *Princeton University. www. princeton. edu/~hsshin/www/barcelona. pdf.*

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 1281-1302.

Yang, Y., Onderstal, S., & Schram, A. (2016). Inequity aversion revisited. *Journal of Economic Psychology, 54*, 1-16.

Chapter 4

Consistency of pro-social preferences - The case of compliance with social norms

4.1 Introduction

Economic experiments have generated a large number of findings that contradict the assumption that individuals are solely concerned about their own material gain. Participants in dictator games (Forsythe et al, 1994) transfer some of their endowments to anonymous recipients. They trust others and honour others' trust in them (Berg et al, 1995) They co-operate in social dilemmas, when defecting would increase their material payoffs (Ledyard, 1995). And they punish free-riders, even at a personal cost (Gächter et al, 2008).

An interesting feature of such pro-social behaviour is that it appears sensitive to the setting in which it is expressed. Dictator-game transfers vary with the degree of anonymity, the action space available, and the way the dictator is appointed (see e.g. Cherry et al, 2002; Krupka and Weber, 2013). Punishment can be pro- or anti-social, depending on the culture in which it is administered (Hermann et al, 2008). A view that attempts to account for such contextual effects posits that people are intrinsically motivated to comply with social norms. A social norm is a collectively held perception of a rule that prescribes appropriate behaviour and is applicable to a particular situation.[1] Individuals experience normative (dis)utility when they act in ways that are socially (in)appropriate. The degree to which a norm applies varies with the characteristics of the particular situation at hand. Therefore, differences in behaviour across different settings can be accounted for by different norms being relevant to those settings.[2]

The influence of social norms on behaviour has been analysed and documented in various settings. Akerlof and Kranton (2000, 2005, 2010) investigate how perceptions of social identity can influence economic behaviour. Chang et al. (2015) apply their framework in examining the effects of political-identity priming on redistributive behaviour. They find that the primed political identities (democratic or republican) of US citizens determine the degree to which they perceive redistribution as socially appropriate and can account for differences in their redistributive behaviour. Using the same analytical framework on the topic of discrimination, Barr et al. (2015) show that discriminatory behaviour is significantly weaker in contexts where it is less socially appropriate.

[1] For extensive discussions on the concept of social norms see e.g. Axelrod (1986), Bernheim (1994), Elster (2000), Bendor and Swistak (2001), Hechter and Opp (2001), Bicchieri (2006), Young (2015). See also Kanazawa (2001) on the compatibility of social-norm accounts with evolutionary psychology.

[2] A large body of literature in psychology (see e.g. Schwartz, 1973; Cialdini et al., 1990) distinguishes among injunctive norms, which are prescriptions of what should be done, descriptive norms, which describe what is regularly done, and personal norms, which are one's own ethical opinions. In the field of business organisation, Burks and Krupka (2012) demonstrate that a misalignment between one's own ethical opinion and the normative perceptions of one's peers or managers is associated with job dissatisfaction and dishonesty for personal monetary gain. Our focus in this study lies on social injunctive norms, i.e. shared beliefs about what one ought to do in specific situations.

Conceptually, the idea of adherence to social norms offers a plausible and testable interpretation of behaviour. Krupka and Weber (2013) report that a simple specification of normative utility can track the differences in transfers across different versions of the dictator game. In their model people care about the extent to which their actions are socially appropriate. They elicit the degree of social appropriateness pertaining to each action by using incentivised coordination games. Krupka et al (2016) use the same norm-elicitation protocol and report similar findings with respect to dictator and Bertrand games. On the other hand, Gächter et al (2013) compare the model of social norms with a model of distributional pro-social preferences and find that the latter outperforms in terms of explanatory power. When considered together, the findings of these studies seem to suggest that the normative model is better at capturing behaviour in some games than it is in others.

A related aspect of the model's performance is the degree to which it is consistent in correctly anticipating people's behaviour. What is meant by 'consistent' here is that the model should (in principle) be able to predict correctly the choices of each individual across different games, given the relevant information on the social appropriateness of the actions in each game.[3] That is necessary even if we restrict our attention to games in which the model has been shown to perform well in accounting for aggregate patterns of behaviour. This is important, because according to the model's narrative, although social norms may change across different situations,

[3] Consistency of performance can be defined in different ways. Blanco et al. (2011) refer to aggregate and individual-specific patterns of behaviour. In their setup, a model performs well at the aggregate level if it is able to predict correctly the proportions of the different types of agents in each game. To perform well at the individual level, on the other hand, the model has to predict the choices of each specific individual with a sufficient degree of accuracy. As the authors state, given these definitions, consistency at the aggregate level does not (in general) imply consistency at the individual one, and vice versa. In our setup aggregate-level consistency follows more readily from consistency at the individual level and we thus focus mostly on the latter, without explicitly referring to it as such.

each individual's propensity to adhere to what is socially appropriate is the same. It is due to people's stable preferences that the model is able to generate predictions about their behaviour in the first place.

In this study we investigate this aspect of the normative model's performance through an experiment. Our basic setup involves three games, variants of the dictator, the trust, and a lying game. Using the norm-elicitation method developed by Krupka and Weber (2013) we measure the degree of social appropriateness pertaining to each of the actions involved in each of these games. We then ask people in a different subject pool to actually play all three games. Each participant is called to provide a decision in every node of every player role in each game. We evaluate the model's predictive power using the elicited measures of social appropriateness and our participants' revealed propensities to comply with social norms. Conceptually, our study is akin to that by Kimbrough and Vostroknutov (2016 a), who use a similar experimental setup and conclude that pro-social behaviour is the result of the rule-following propensities of individuals.[4] We depart from their framework mainly in three ways. Firstly, we only consider decisions that do not involve strategic uncertainty.[5] Secondly, we implement a more fine-tuned analysis, linked directly to the norm-elicitation method of Krupka and Weber (2013). Finally, we observe the decisions of the same subjects in three different games.

We have already argued about our choice of games in the previous chapter of this thesis. The same arguments apply here, with the only difference that they are now to be viewed through the perspective of adherence to social norms (rather than distributional preferences per se). Furthermore,

[4] See also Kimbrough and Vostroknutov (2016 b), for a complementary discussion on eliciting desires for rule-following.

[5] By 'strategic uncertainty' we mean uncertainty that is related to others' beliefs (of any order) and actions - see Morris and Shin (2002), as well as the previous chapter of this thesis.

the Krupka-Weber model has been shown to perform well in accounting for the behavioural variation in similar games (see e.g. Krupka et al, 2016). An additional argument that can be made with respect to normative preferences relates to our lying game. Specifically, this game involves the potential for normative reinforcement or conflict. For example, to the extent that norms of payoff-equality and truthfulness are relevant to it, they may prescribe the same action or different ones. This is a very appealing feature, since it allows us to separate preferences for compliance with social norms from preferences for adherence to specific ideals. It also allows us to investigate the effectiveness of the Krupka-Weber norm-elicitation method.

Our results indicate that, in general, the Krupka-Weber model does not perform well in anticipating the behaviour of our subjects. We find that the proportions of people who manifest given preferences for norm-compliance are not stable across our games. We also find that a lot of our subjects do not exhibit stable norm-following propensities. Some of our subjects appear to be motivated by strong selfish preferences. Some others seem to firmly adhere to a principle of payoff equality, even when the actions it prescribes are not judged as the most socially appropriate. The behaviour of these two types is largely consistent with their respective motives throughout the three games. Of them, only the selfish types can be accounted for adequately by the Krupka-Weber model. Furthermore, the behaviour of the rest is not suggestive of stable norm-following preferences. In this sense, our findings contradict those of Kimbrough and Vostroknutov (2016). In addition, we show that the model's predictions are not always determinate and that it sometimes ex ante precludes choices that prove to be quite popular among our subjects.

The remainder of the paper proceeds as follows. In section 4.2 we illustrate the Krupka-Weber model and our process of eliciting the individual-specific propensities for norm-compliance. In section 4.3 we present our experimental design and discuss its features. In section 4.4 we describe our results and assess the model's performance. In section 4.5 we conclude with a critical evaluation of our findings.

4.2 Norm-dependent utility

Krupka and Weber (2013) model adherence to norms as an individual-specific feature of deep preferences. They argue that in choice environments every option available to an individual is characterised by a certain degree of social appropriateness. Then, all options can be ordered and compared in terms of their degrees of social appropriateness, in the same manner that they are in terms of the payoffs they lead to.

Consider game $\mathcal{G}$ featuring action set $A = \{a_1, a_2, ..., a_K\}$. The game is played by $I \geq 2$ individuals. Each individual $i \in I$ is concerned both about her/his resulting monetary payoff and the social appropriateness of the action (s)he chooses. Function $\pi : A^I \rightarrow \mathbb{R}$ ascribes a monetary payoff to each action chosen by player i, depending on the choices of the other players. The degree of social appropriateness that corresponds to each action is given by a function $N : A \rightarrow \mathbb{R}$. An individual-specific parameter γ_i measures player i's sensitivity to concerns about social appropriateness. Effectively, $\gamma_i \in \mathbb{R}^+$ can be thought of as the weight that normative concerns have on i's utility function relative to concerns about her/his material payoff.[6] Thus, player i's utility from action a can be written as:

[6] One can think of individuals with $\gamma_i < 0$ as anti-social (i.e. people who receive satisfaction from violating social norms). In most of the situations we examine here the behaviour of such individuals cannot be distinguished from that of people with $\gamma = 0$. In those cases where we can discern between the two, we find that the proportion of people with $\gamma_i < 0$ is very small.

$$U_i(a^i, a^{-i}) = \pi(a^i, a^{-i}) + \gamma_i N(a^i, a^{-i}) \tag{4.2.0.1}$$

It is useful to bear in mind what this specification implies. In particular, one may notice that both the payoff and the social-appropriateness function are universal. In other words, all I individuals enjoy the same level of utility from a given amount of material payoff and are in agreement about the degree of appropriateness ascribed to each action. The fact that they experience the latter differently is only due to it being weighted by the individual-specific parameter γ_i.

For our experimental investigation we assume linear functional forms for both the material payoff resulting from action a_k and its degree of social appropriateness. This postulate confers analytical simplicity without rendering the account itself more restrictive.[7] As far as our study is concerned, we show in section 4.4 that, assuming accurate judgements, non-linear functional forms for $U(\pi(.))$ cannot account for our results.[8]

Of course, this does not imply that a linear utility function precisely describes people's preferences. However, in our experiment the differences in payoffs across the games are relatively small. Therefore, utility over money can be reasonably expected to be linear.[9]

With these in mind and given a^{-i}, individual i chooses an action $\hat{a}^i$

[7] The reason is that any non-linear effects of social-norm adherence on utility can be captured by function $N(.)$. In fact, such a conception of function $N(.)$ is very useful, because it can also account for the existence of interior maxima in the strategy space. Consider a generic set of options with elements in the interval $[\underline{a}, \overline{a}]$. Even if the payoff function is linear, an interior maximum can still exist for $0 < \gamma < \infty$, so long as there is conflict between personal payoff and social appropriateness. This is because perceptions about social appropriateness need not be linear in the available options. And since perceptions are elicited (using the Krupka-Weber, 2013 method), they can be applied to generate an interior maximum with a linear normative-utility function.

[8] It is worth noting here that in their original paper Krupka and Weber use a more general value function, $V(\pi(.))$, to capture people's material utility.

[9] If anything, we observe people for whom payoff variations should matter most exhibiting the highest degrees of consistency.

according to:

$$\hat{a}^i \in \arg\max_{a}^{i} U_i(a^i, a^{-i}) = \pi(a^i, a^{-i}) + \gamma_i N(a^i, a^{-i}) \qquad (4.2.0.2)$$

In a discrete choice set equation 4.2.0.2 implies that individual i will choose the option that delivers the highest level of utility, given γ_i. Thus, inferences about γ_i can be made based on the individual's choice, taking into account all alternative choices. Specifically, a rational player will choose action a_1 over action a_2 iff the former confers a (weakly) higher utility than the latter does:

$$\begin{aligned} a_1 \succeq a_2 \Rightarrow U_i(a_1^i, a^{-i}) \geq U_i(a_2^i, a^{-i}) \Rightarrow \\ \Rightarrow \pi(a_1^i, a^{-i}) + \gamma_i N(a_1^i, a^{-i}) \geq \pi(a_2^i, a^{-i}) + \gamma_i N(a_2^i, a^{-i}) \Rightarrow \\ \Rightarrow \gamma_i \underset{<}{\geq} \frac{\pi(a_1^i, a^{-i}) - \pi(a_2^i, a^{-i})}{N(a_2^i, a^{-i}) - N(a_1^i, a^{-i})} \end{aligned} \qquad (4.2.0.3)$$

Provided that social appropriateness is in conflict with personal payoff in the comparison between a_1 and a_2, a positive threshold value for γ_i can be computed. This is the lower or upper bound of the set of possible γ_i values player i can have, given her/his choice. We can extend this reasoning to discrete choice sets with more than two actions. Consider the set $\Gamma_1 \subset \mathbb{R}^+$, which contains the threshold γ_i values that result from the binary comparisons of the chosen action, a_1, with all alternatives. Furhter, let Γ_i be partitioned into Γ_1^L and Γ_1^H. Γ_1^L contains all γ_i thresholds that have resulted from the comparisons between a_1 and each of the actions that yield a lower degree of social appropriateness (and a higher personal payoff). Conversely, Γ_1^H contains all γ_i thresholds that have resulted from the comparisons between a_1 and each of the actions that yield a higher degree of social appropriateness (and a lower personal payoff). Then, the lower and upper bound of the set of admissible γ_i values following the choice of a_1 are simply the maximum and minimum elements of Γ_1^L and Γ_1^H, respectively. Then, the consistency of the account can be tested by generating predictions about future behaviour given the elicited γ_i values.

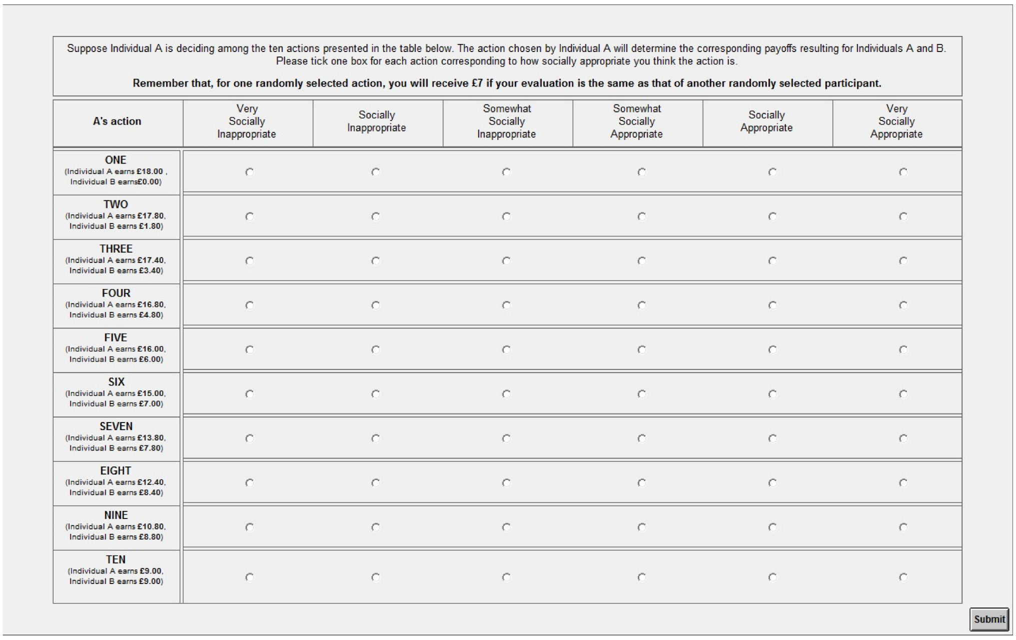
4.3 Experimental design

Our study consists of two separate experiments. In the first one, labelled 'behavioural', we asked each of our 178 subjects to take part in three games. These games are the ones we also used for the evaluation of the Fehr-Schmidt (1999) model of inequality aversion and are described in Chapter 2 (see section 3, pages 6-10). As a brief reminder, they are modified versions of the Dictator Game , the Trust Game, and a Lying Game.

For our evaluation of the Krupka-Weber (2013) account of norm compliance, we additionally needed to have a measure of the social appropriateness of each action featuring in our games. It is worth repeating that for our purposes this is ultimately an empirical matter. That is, we did not try to derive the form of the $N(.)$ function analytically, based on certain priors. Instead, we used the norm-elicitation task developed by Krupka and Weber (2013) to obtain people's judgements about how socially appropriate each action is.

We used a separate subject pool to elicit perceptions of social appropriateness. This second experiment, labelled 'normative', was conducted using z-Tree (Fischbacher, 2007). Our 100 participants in the Krupka-Weber task initially saw a description of each of our games on their computer screens. Every game was presented as an interaction between two generic agents.

Thus, the dictator game was played by Individuals A and B, the trust game by Individuals X and Y, and the lying game by Individuals J and K. Our participants in the normative experiment obtained information about each game that matched the information available to those who engaged in it in the behavioural experiment. This included the payoff structure, the order of moves, the actions available to each player, and their starting positions. After each description, they were asked to evaluate the degree of social appropriateness of each available action on a discrete ordinal scale. This scale is the one used by Krupka et al (2016). They could judge any one of them as *Very Socially Inappropriate*, *Socially Inappropriate*, *Somewhat Socially Inappropriate*, *Somewhat Socially Appropriate*, *Socially Appropriate*, or *Very Socially Appropriate*. They provided their judgements on their computer screens and in private. A screen-shot of the assessment table for the dictator game is provided in Figure 4.1.



Suppose Individual A is deciding among the ten actions presented in the table below. The action chosen by Individual A will determine the corresponding payoffs resulting for Individuals A and B. Please tick one box for each action corresponding to how socially appropriate you think the action is.

Remember that, for one randomly selected action, you will receive £7 if your evaluation is the same as that of another randomly selected participant.

A's action	Very Socially Inappropriate	Socially Inappropriate	Somewhat Socially Inappropriate	Somewhat Socially Appropriate	Socially Appropriate	Very Socially Appropriate
ONE (Individual A earns £18.00 , Individual B earns£0.00)	○	○	○	○	○	○
TWO (Individual A earns £17.80, Individual B earns £1.80)	○	○	○	○	○	○
THREE (Individual A earns £17.40, Individual B earns £3.40)	○	○	○	○	○	○
FOUR (Individual A earns £16.80, Individual B earns £4.80)	○	○	○	○	○	○
FIVE (Individual A earns £16.00, Individual B earns £6.00)	○	○	○	○	○	○
SIX (Individual A earns £15.00, Individual B earns £7.00)	○	○	○	○	○	○
SEVEN (Individual A earns £13.80, Individual B earns £7.80)	○	○	○	○	○	○
EIGHT (Individual A earns £12.40, Individual B earns £8.40)	○	○	○	○	○	○
NINE (Individual A earns £10.80, Individual B earns £8.80)	○	○	○	○	○	○
TEN (Individual A earns £9.00, Individual B earns £9.00)	○	○	○	○	○	○

Figure 4.1: Table of normative assessments - Dictator game

It was made explicit to the participants that what is meant by *'socially (in)appropriate'* is *behaviour that most people agree is the '(in)correct' or '(un)ethical' thing to do.* That is, they were informed that it is not their personal views that are relevant for the task, but rather their perceptions about the views of society as a whole. To ensure adherence to this rule, they were incentivised to try to match the social view. Specifically, they were informed that they would have the opportunity to receive a substantial bonus in addition to their show-up fee. For this purpose, we would pair them up and choose a game and an action randomly at the end of the session. If a participant's assessment for that action matched the one of the person (s)he was paired with, they would both earn an additional £7.00 bonus. Otherwise, they would only receive the show-up fee, which was £5.00.

In the trust game, for completeness, we elicit ratings of social appropriateness for both player-roles. However, we mostly focus on those corresponding to the second mover's options to evaluate the model. The reason is that the behaviour of the first mover can be partly attributed to the presence of strategic uncertainty about the second mover's choice. Therefore, his propensity for behaving in a socially appropriate way cannot be straightforwardly elicited.

In the lying game we elicit ratings for each report in every state (every true colour the wheel has landed on). Thus, each participant has to assess how appropriate it is to report RED given that the true colour is RED, how appropriate it is to report RED given that the true colour is BLUE, and how appropriate it is to report RED given that the true colour is GREEN. We then ask them to assess the social appropriateness of reporting BLUE and that of reporting GREEN in the same way. We do this to capture the state-specific effect on the degree of appropriateness of each report.

The order in which the games were presented remained fixed throughout the normative experiment. We chose this option to maintain a close correspondence with the behavioural experiment.[10] We think that this correspondence is important, because it allows variations in behaviour to be aligned with variations in the normative assessments. That is, even though the people in our behavioural experiment are different from those in the normative one, a Krupka-Weber agent would make a choice based on her/his own perceptions about social appropriateness. Thus, if that agent had to face the three games in our fixed order, (s)he would also judge how appropriate each of the actions is in the same order.

We ran ten experimental sessions in total, six for the behavioural and four for the normative experiment. All our sessions took place at the University of Nottingham, in 2015. Of our behavioural sessions, three were run on the 5^{th} of June in the CRIBS laboratory and three on the 4^{th} of December in the CeDEx laboratory. Each of these sessions lasted for approximately one hour and 15 minutes and yielded an average payoff of £10.50 for each of our subjects. Our normative sessions were all run in the CeDEx laboratory, on the 28^{th} and 29^{th} of May, and on the 9^{th} of December. Their average duration was 50 minutes and each participant received approximately £7.00 on average.

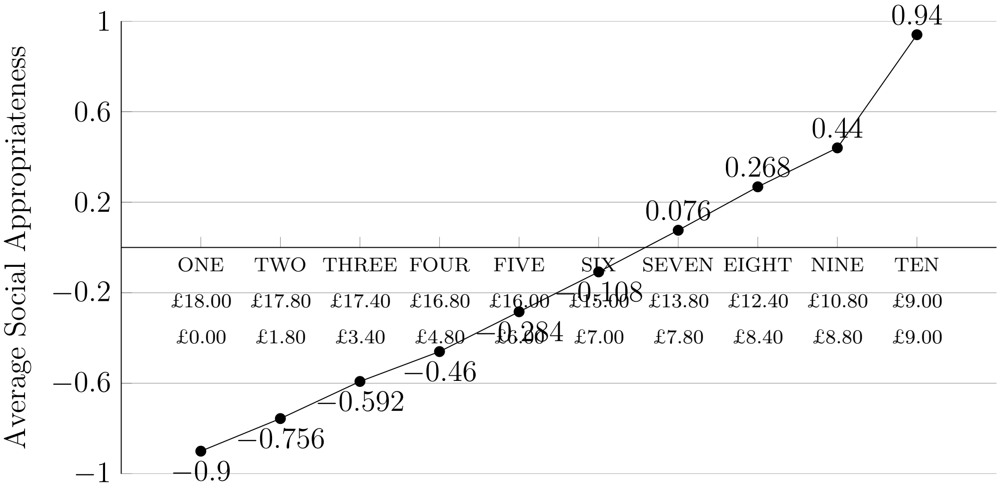
4.4 Results

4.4.1 Parameter estimation

In order to categorise the participants in terms of their norm-following propensities, we use the results from our normative experiment. We start by assigning numerical values to the ratings of social appropriateness. Our

[10] Recall that in the behavioural experiment the order had to be fixed, in order to minimise the amount of noise in our measurements.

Figure 4.2: Average Normative Assessments - Dictator Game



assignment is similar to that of Krupka et al (2016). That is, we ascribe the following scores to the ratings: -1 to *Very Socially Inappropriate*, -0.6 to *Socially Inappropriate*, -0.2 to *Somewhat Socially Inappropriate*, 0.2 to *Somewhat Socially Appropriate*, 0.6 to *Socially Appropriate*, and 1 to *Very Socially Appropriate.* We then compute the average degree of social appropriateness pertaining to each of the actions in our games, based on these values.[11] We interpret each degree as the value assigned by the normative function, $N(.)$, to the corresponding action. We then proceed to compute the γ thresholds associated with that action, that is, the values of γ_i for which someone is indifferent between that action and its closest alternatives (in terms of payoff/appropriateness correspondence).

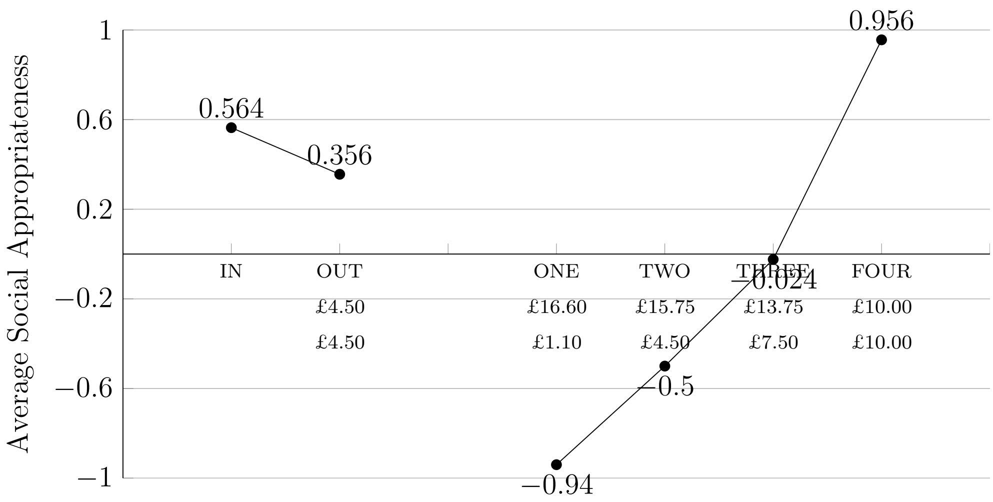
Figures 4.2, 4.3, and 4.4 depict the average assessments of social appropriateness based on the responses in our normative experiment. In each Figure the actions available in the corresponding game are listed on the horizontal axis, followed by the implied payoffs to the participants. The

[11] It is important to remember that in assigning numerical values to the qualitative statements used in the Krupka-Weber framework we assume equal distance between the various categories. For instance, for our purposes the distance between the ratings 'Very Socially Inappropriate' and 'Socially Inappropriate' is exactly equal to the distance between the ratings 'Somewhat Socially Inappropriate' and 'Somewhat Socially Appropriate'. This is also in line with Krupka and Weber (2013).

climax on the vertical axis measures the degree of social appropriateness, spanning from -1 (Very Socially Inappropriate) to +1 (Very Socially Appropriate).

An immediate feature to notice is that for every player-role the distribution of appropriateness ratings pertaining to the actions available is single-peaked. Importantly, this does not preclude the potential for conflicting norms to arise (a situation we are particularly interested in), as the assessments in the Lying Game indicate. Additionally, the average ratings in the dictator and the trust game reveal an interesting relationship between changes in the payoff distribution implied by the different actions and changes in their perceived appropriateness. In particular, the data seem to suggest that a deviation from the most appropriate action is more costly (in a normative sense) from a same-step deviation from a less appropriate one (towards the same direction). This is consistent with the notion of discontinuity around the norm, widely observed in experimental settings (see e.g. Andreoni and Bernheim, 2009). To see this, let the norm in both the decision node of the dictator and that of the second mover in the trust game be the equal split. This is conferred by action TEN in the dictator game and action FOUR in the trust game. In the former, the difference between the average assessment of action TEN and that of action NINE is significantly higher that that of any other two neighbouring actions (all Wilcoxon's sign-rank tests yield p-values lower than 0.01). The same is true for the difference between actions FOUR and THREE in the trust game (again, all Wilcoxon's sign-rank p-values are lower than 0.01). This pattern is all the more significant when examined in light of the fact that our payoff structure in both these games points towards the exact opposite direction: Increasing deviations from the most socially appropriate action return progressively lower marginal material benefits. Regarding the Lying Game, the average report ratings across states seem to point towards

Figure 4.3: Average Normative Assessments - Trust Game



conflicting norms, as will become clearer below.

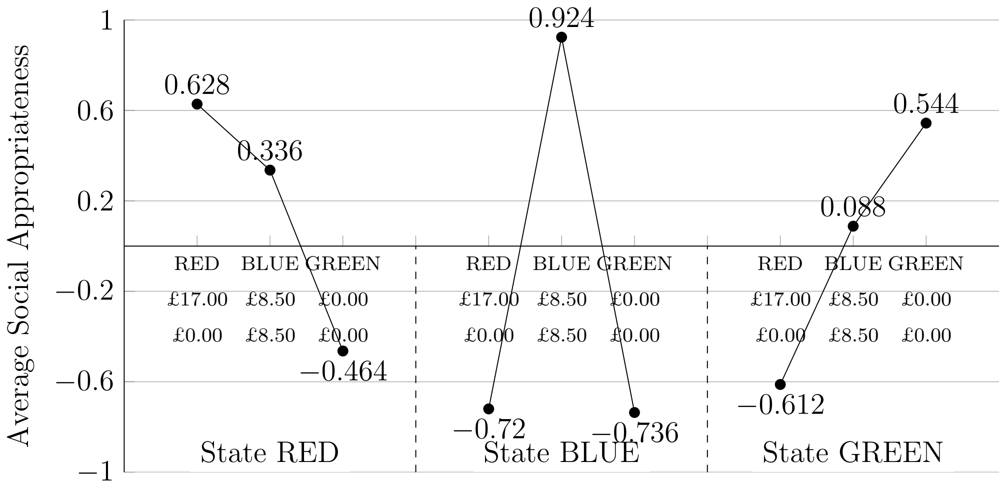
The pattern of appropriateness ratings in the dictator game is in line with those in the various versions of the game analysed in Krupka and Weber (2013). Note, again, that in our variant the most egalitarian outcome results from the last action (action TEN). Thus, the qualitative properties of the normative function appear to be maintained across the different games.[12]

It is important to note here that social-efficiency concerns do not seem to have a substantial effect on appropriateness.[13] Indeed, we find no evidence that the difference in average degrees of social appropriateness (the slope of the line in Figure 4.2) changes as we move towards and away from the most efficient options (FIVE and SIX). This finding can be interpreted in different ways. For example, it may be the case that people didn't actually realise that social efficiency varies across actions. On the other hand, maybe people do indeed disregard it and focus on equality when

[12] This qualitative similarity can be viewed as evidence in support of the robustness and suitability of the Krupka-Weber elicitation method, given that changes in the ratings do, in fact, occur across different versions.

[13] By social-efficiency concerns we mean considerations about the maximisation of the joint payoff.

Figure 4.4: Average Normative Assessments - Lying Game



assessing social appropriateness. These two scenarios are not conflicting with the normative account, since what we are interested in are people's judgements, irrespectively of their rationale.

The assessments of the actions available to the second mover in the trust game are similar to those of the dictator's options. In particular, we observe that social appropriateness is increasing monotonically as one moves from the most selfish to the most egalitarian action. Interestingly, people deemed the OUT option, available to the first mover, as significantly less socially appropriate than its alternative, IN, which gives the second mover the opportunity to play (Wilcoxon's sign-rank p-value: 0.002).

With respect to the lying Game, we observe that the state generally dominates in the assessment of social appropriateness. That is, people seem to consider truthful reports as the most socially appropriate in every instance, irrespective of the ensuing degree of efficiency or distribution of the resulting total payoff. It is also true, however, as is evident in Figure 4.4, that the pro-social option of reporting BLUE is always characterised by positive degrees of appropriateness on average. This is not true for the other two options, of which the one (RED) dominates in terms of own-payoff considerations, while the other (GREEN) exhibits the same degree of payoff equality with BLUE, but is less efficient.

We may thus argue that normative conflict can be generated in our lying game primarily between two principles, egalitarianism and truthfulness. We find evidence that both are at play from two indicative patterns. On the one hand, there is the fact that reporting BLUE is never socially inappropriate (on average) irrespectively of the true colour of the wheel. This constitutes evidence that a norm of payoff-equality is present. On the other hand, the report that is dominant (on average) in terms of social appropriateness is the one that matches the state. This points towards a norm of truthful reporting, which dominates concerns about payoff-equality.

Notice that payoff equality does not appear to be important purely in itself. If it were, then reporting GREEN would be just as appropriate as BLUE is. Instead, the comparison between the average ratings of these two options indicates that, to some extent, efficiency matters to our subjects. At the same time, however, social efficiency does not appear to matter much in the absence of payoff equality. To see this, notice that the average degree of social appropriateness of reporting GREEN[14] when the state is GREEN is not statistically different from that of reporting RED in state RED (Wilcoxon signed-rank test: $p = 0.154$). On the other hand, reporting RED in state GREEN appears less appropriate than reporting GREEN in state RED. Their difference is significant at the 10% level (Wilcoxon signed-rank test: $p = 0.054$). Thus, there is some tentative evidence that truthfulness notwithstanding, payoff equality is perceived as a more important norm by our subjects than social efficiency. This result is all the more striking when one takes into account the large difference between the aggregate payoff induced by reporting RED and that which

[14] Recall that a GREEN report is the socially inefficient option in this game.

follows a GREEN report.

An important element in the assessments that we obtained is that the social consensus is not equally strong across the actions in each game. That is, the degree of dispersion of the normative assessments varies across actions. As we will see in sub-section 4.4.3, this variation raises a crucial issue for the validity of our conclusions. We examine it in detail in that section and argue that it cannot be the sole justification for our findings.

4.4.2 Evaluation of norm-following behaviour

So far we have analysed the perceptions of the participants in our normative experiment on how socially appropriate each action is in our games. We now use these normative assessments to characterise the behaviour of the participants in our behavioural experiment.[15] To do so, we start by taking the average assessment pertaining to each action to denote its degree of social appropriateness (the value of the $N(.)$ function for that action). Using these averages, we then compute the γ threshold values as described in section 4.2. Each of these thresholds is a value for γ_i that renders a normative agent i indifferent between two options (see, e.g., Table 4.1). In each game, these thresholds define a set of parameter groups. Each parameter group is a collection of all γ_i values that prescribe the same choice. Finally, we classify our participants in the behavioural experiment in those groups, based on their decisions in the dictator game. This classification allows us to draw predictions about their behaviour in the other two games. We test these predictions to evaluate the model's performance in accounting for the behavioural variation in our sample.

[15] Recall that the two experiments involve different participants.

4.4.2.1 Choices in the dictator game

Out of the 178 people that participated in our behavioural experiment, around 26% chose action ONE in the Dictator game (giving nothing to their Recipients). On the opposite side, the equal split (action TEN) was chosen by about 8% of them. What is worth noting is that actions FIVE and SIX, which are the ones that maximise social efficiency, were chosen quite often (namely, by 20% and 19% of the people respectively).

We allocate our participants into different categories, based on the relative strength of their normative preferences. In this way, we have a profile of 'types' of players, defined according to their γ_i values. Table 4.1 summarises our results. Out of our 178 subjects 30% are characterised by $\gamma_i \leq 2.439$, while 41% exhibit $\gamma_i \geq 5.725$ and the remaining 29% are spread in the middle. The average γ_i value in our dictator game lies in the interval $[4.545, 5.682]$.

Table 4.1 Threshold values for the γ parameter - Dictator game

Action	Own Payoff	Av. N.A.	γ-threshold	Data
ONE	£18.00	-0.9	1.389	26%
TWO	£17.80	-0.756	2.439	4%
THREE	£17.40	-0.592	4.545	8%
FOUR	£16.80	-0.46	4.545	1%
FIVE	£16.00	-0.284	5.682	20%
SIX	£15.00	-0.108	5.725	19%
SEVEN	£13.80	0.076	-	6%
EIGHT	£12.40	0.268	-	6%
NINE	£10.80	0.44	-	2%
TEN	£9.00	0.94		8%

Threshold values for γ based on the average normative assessment scores. Each threshold is the value of γ at which a player is indifferent between the action at the same line and the immediately next undominated one (hence why there is no threshold value in the same line with action TEN). Where threshold values are missing, the corresponding actions are strictly dominated based on the theory and the observed average degrees of social appropriateness.

An intriguing feature of our dictator game depicted in Table 4.1 is that not every action is optimally chosen by people with some γ_i. Specifically, there are three actions (SEVEN, EIGHT, and NINE) that should never be opted for by a norm-following agent, irrespectively of her/his parameter value. To see why this is so, consider the case of equality in relation 4.2.0.3. Given any two actions, the threshold value for the γ_i parameter is the one for which player i will be indifferent between them. A higher (lower) γ_i value would render one action or the other more attractive. At some point the marginal gain from the next more socially appropriate choice is so small, that only agents with substantially higher γ_i values would be willing to incur the payoff loss. But these agents may find it optimal to choose an even more appropriate action, if what they gain in terms of social appropriateness exceeds the additional payoff they have to forgo.

This feature of the model may at first seem counter-intuitive, but it is rather appealing. To see this, consider a choice environment with more than two options. Effectively, given the structure of payoffs and degrees of social appropriateness some options may confer a trade-off between the two that is never optimal: At least one other option confers a better trade-off. In our dictator game this is easy to see. Notice that from action SIX every subsequent option decreases the dictator's payoff exponentially. Their social appropriateness, however, increases in a linear fashion up to action TEN, at which point it jumps up substantially (see Figure 4.2). As a result, a norm-following agent that is willing to incur a payoff-cost to opt for an action higher than SIX will find it optimal to go all the way to the most socially appropriate action.[16] One of the implications is that the

[16] In our dictator game the threshold for being indifferent between actions SIX and TEN is lower than that between actions SIX and SEVEN. That is, every person with a γ_i high enough to prefer action SEVEN from SIX will also prefer action TEN from SEVEN. Additionally, for some γ_i values a person will prefer action TEN from SIX, but not action SEVEN. Conversely, the threshold for being indifferent between actions SIX and TEN is higher than that between actions NINE and TEN. That is, every person with a γ_i low enough to prefer action NINE from TEN will also prefer action SIX from

empirical pattern of discontinuity around the norm acquires a theoretical rationale, at least with respect to our dictator game.[17] Another one is that the restriction of the set of admissible choices constitutes itself a testable proposition. Consider the following hypothesis:

Hypothesis 4. No player should choose action SEVEN, EIGHT, or NINE in the dictator game, irrespectively of the value of her/his γ_i parameter.

We find that more than 14% of our participants opt for one of the actions precluded by hypothesis 4. This frequency is not easily reconcilable with the model's prediction. We might be tempted, for instance, to assert that people's behaviour is largely consistent with the model, but with some positive probability they make random errors. But in this case, the probability required to generate the frequencies of dominated choices we observe would need to be substantially high. To see this, note that the 14% rate is a lower bound for the probability of error necessary to generate this pattern of choices (assuming that people are equally likely to err in either direction). Thus, the model's prediction that actions SEVEN, EIGHT, and NINE will never be chosen is falsified by our data. The pattern of discontinuity around the norm, however, seems to be present in our data. Action TEN, which is the most socially appropriate one, is also substantially more popular than action NINE. Notice that this is all the more striking given our payoff structure, where a deviation from action NINE to TEN is the most expensive in terms of personal monetary payoff.

The falsification of hypothesis 4 notwithstanding, the behaviour of people who choose dominated actions cannot be rationalised by the model.

NINE. Additionally, for some γ_i values a person will prefer action SIX from TEN, but not action NINE. The same is true for action EIGHT. As a result, there is no parameter-value interval such, that any of actions SEVEN, EIGHT, and NINE is preferred against all alternatives. For every (weakly positive) γ_i they are dominated by at least one other option.

[17] An analysis of dictator games with different payoff structures would be necessary to assess the robustness of this claim.

Therefore, we do not consider them in our evaluation of its consistency in correctly predicting people's choices. A distinction needs to be made here between two different indicators of performance. On the one hand, there may be people whose choices cannot be accounted for by the model. These are the people for whom no admissible γ_i value exists. Consequently, they cannot be accommodated within the framework and no predictions can be made about them. On the other hand, there may be people whose choices can be accounted for by the model, but not in a consistent manner. These are the ones for whom γ_i values exist, but are not stable across different games. These two different aspects of performance are equally important in the model's evaluation.

4.4.2.2 Consistency in the trust game

We proceed to examine whether the Krupka-Weber account can describe the behavioural regularities in our data in a way that is consistent across the different games. We focus firstly on the relationship between parameter values elicited in the dictator game and second-mover choices in the trust game. Recall that for our analysis of consistency we exclude all participants who chose the dominated actions (SEVEN, EIGHT, and NINE) in the dictator game. Consider the following two hypotheses:

Hypothesis 5. No player should choose action THREE as second mover in the trust game, irrespectively of the value of her/his γ_i parameter.

Hypothesis 6. Player i as second mover in the trust game should choose:

- Action ONE (£16.60 , £1.10), iff $\gamma_i \leq 1.932$
- Action TWO (£15.75 , £4.50), iff $1.932 \leq \gamma_i \leq 3.949$
- Action FOUR (£10.00 , £10.00), iff $3.949 \leq \gamma_i$

There are four alternative actions available to the second mover in our trust game. Table 4.2 contains the γ thresholds associated with these actions and the proportions of our sample that opted for them.[18] We see that action THREE is also dominated according to the model and the elicited ratings of social appropriateness. This instance is harder to rationalise as discontinuity around the norm. It is true that action FOUR is the most appropriate on average and can, thus, be considered the normatively prescribed option. However, the action space is considerably smaller than that in our dictator game and, as a result, choices are much more concentrated. On the empirical side, it turns out that more subjects chose action THREE than action FOUR. If we focus our attention only to those who did not choose dominated options in the dictator game (152 people), 23% of them opted for THREE and only about 16% chose FOUR.[19] Thus, hypothesis 5 can be confidently rejected.

Table 4.2 Threshold values for the γ parameter - Trust game (2nd mover)

Action	Own Payoff	Av. N.A.	γ-threshold	Data
ONE	£16.60	-0.94	1.932	35%
TWO	£15.75	-0.5	3.949	26%
THREE	£13.75	-0.024	-	23%
FOUR	£10.00	0.956		16%

The table presents the threshold values for γ based on the average normative-assessment scores. Each threshold is the value of γ at which a player is indifferent between the action at the same line and the immediately next undominated one. Where threshold values are missing, the corresponding actions are strictly dominated based on the theory and the observed average degrees of social appropriateness. The proportions refer to the 152 people who made consistent choices in the dictator game.

Regarding hypothesis 6, we also find little support for the model. Remember that to test for consistency we can only consider those participants that did not make a dominated action in either the dictator or the trust game. There are 117 participants that satisfy this criterion and are thus

[18] These proportions are relative to the total of 152 participants who chose non-dominated choices in the dictator game.

[19] Out of the total sample of 178 people, 24% chose THREE and 21% chose FOUR.

Table 4.3 Parameter values and returns in the trust game - Total of subjects with no dominated choices in the dictator and trust game

γ_i	Frequency %	Predicted choice	Choices observed %		
			ONE	TWO	FOUR
$\gamma_i \leq 1.389$	39.3%	ONE	**80.4%**	13%	6.6%
$\gamma_i \in [1.389, 2.439]$	6%	ONE/TWO	**71.4%**	28.6%	0%
$\gamma_i \in [2.439, 4.545]$	9.3%	TWO/FOUR	27.3%	**72.7%**	0%
$\gamma_i \in [4.545, 4.545]$	0.9%	FOUR	0%	**100%**	0%
$\gamma_i \in [4.545, 5.682]$	18.8%	FOUR	13.6%	**77.3%**	9.1%
$\gamma_i \in [5.682, 5.725]$	13.7%	FOUR	18.7%	31.3%	**50%**
$\gamma_i \geq 5.725$	12%	FOUR	14.3%	0%	**85.7%**

The table lists the relative frequencies of actual choices in the trust game (2nd mover), as well as the prediction(s) of the Krupka-Weber model for each γ_i group. The percentage reported in each cell of the last three columns is relative to the corresponding γ_i group. The table includes our 117 subjects that made non-dominated choices in both the dictator and the trust game. The modal choice within each group is highlighted in boldface.

eligible for this analysis. Of them, about 39% would be expected to choose action ONE and 46% action FOUR. There is also a 6% that may consistently pick either ONE or TWO, while the remaining 9% would be expected to choose between TWO and FOUR. For these two latter groups there is some overlap between the parameter intervals suggested by the dictator game and those relevant to the trust game. This is why their behaviour would be consistent with more than one choices.

What we observe, instead, is that a higher proportion of our sample opted for action ONE and a lower one for action FOUR than what the model had anticipated. Specifically, more than 45% of our subjects chose ONE, less than 22% selected FOUR, and about 33% went for TWO. Even if we assume that the 16% whose choices are not completely determinate opted for action TWO,[20] the difference between the expected and the observed distribution is significant at the 1% level ($\chi^2(2) = 15.649$; $p = 0.000$, Fisher's exact: $p = 0.000$).

Our results are depicted in Table 4.3. Consider our 117 subjects that

[20] This assumption is the most forgiving for the model.

did not choose a dominated option in any of the two games. We see that the choices of people in the first three parameter groups are to a large extent consistent with their γ_i values. However, a large proportion of those in the last group, who were expected to choose action FOUR, deviated to action TWO instead. As a result, we find a much higher proportion of choices of TWO and a much lower one of choices of FOUR than we would theoretically anticipate. The difference between the observed distribution of choices and the one expected by the model is significant at the 1% level ($\chi^2(2) = 18.283$, $p = 0.000$; Fisher's exact: $p = 0.000$). We conclude that the Krupka-Weber model is not able to capture accurately the behavioural variability across our dictator and trust game. In addition, the model is unable to account for 34% of our subjects (61 out of our initial sample of 178 participants), each of whom made at least one dominated choice.

An interesting feature of the model's failure to predict our subjects' choices accurately is that it is not uniform across their γ_i values. This is evident in Table 4.3. To start with, most of those who should have chosen action ONE actually did so. Similarly, among those with $\gamma_i \geq 5.725$ about 86% chose action FOUR, in line with the model's prediction. By contrast, among those with $\gamma_i \in [4.545, 5.682]$ only about 9% went for action FOUR (more than 77% opted for TWO instead). Furthermore, half of the people with $\gamma_i \in [5.682, 5.725]$ preferred actions ONE and TWO.

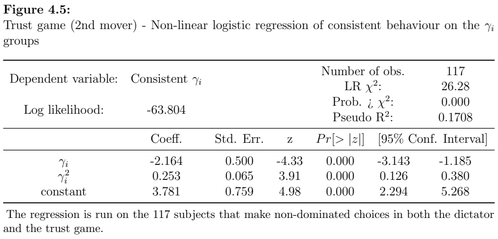


Figure 4.5:
Trust game (2nd mover) - Non-linear logistic regression of consistent behaviour on the γ_i groups

Dependent variable:	Consistent γ_i			Number of obs. LR χ^2:		117 26.28
Log likelihood:	-63.804			Prob. ¿ χ^2: Pseudo R^2:		0.000 0.1708
	Coeff.	Std. Err.	z	$Pr[>\|z\|]$	[95% Conf.	Interval]
γ_i	-2.164	0.500	-4.33	0.000	-3.143	-1.185
γ_i^2	0.253	0.065	3.91	0.000	0.126	0.380
constant	3.781	0.759	4.98	0.000	2.294	5.268

The regression is run on the 117 subjects that make non-dominated choices in both the dictator and the trust game.

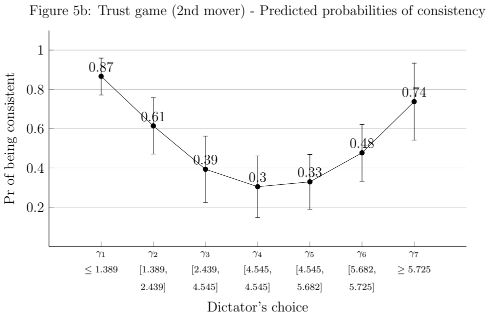


Figure 5b: Trust game (2nd mover) - Predicted probabilities of consistency

Figure 4.6: Trust game - Estimated relationship between one's γ_i value and the probability that one's decision is consistent with the Krupka-Weber model

Thus, our results are suggestive of a non-monotonic relationship between the strength of one's commitment to do what is socially appropriate and one's consistency in doing so. A logistic regression confirms this. The predicted probability of making a decision that is consistent with the Krupka-Weber model depends on one's γ_i value in a U-shaped manner. The quadratic specification fits the pattern of our data significantly better than a linear one ($\chi^2(1) = 15.26$, $p = 0.000$). The two parts of Figure 4.6 detail and depict this relationship between the predicted probability that one is consistent with the model and one's choice in the dictator game.

4.4.2.3 Consistency in the lying game

We now turn to our participants' behaviour in the lying game. Table 4.4 presents the parameter thresholds associated with the available options in every state and the proportions of our participants that chose them.[21] The social appropriateness of each action in this game is heavily dependent on the state, i.e. the actual outcome of the wheel-spin. In state RED the most socially appropriate action is to report RED. Thus, since there is no conflict between one's own material payoff and social appropriateness, a Krupka-Weber agent will always report RED, irrespectively of her/his γ_i value. Every other report is dominated. In state BLUE, where BLUE is the most socially appropriate report to give, RED is still the option of agents with sufficiently low γ_i values. Here the only completely dominated option is GREEN (as it both implies the lowest payoff and has a very low degree of social appropriateness). Finally, in state GREEN it is the GREEN report that dominates in terms of social appropriateness, with BLUE being the second most appropriate choice. As a result, no option is entirely dominated: For every report R there exists an interval $[\underline{\gamma}_i^R, \overline{\gamma}_i^R]$

[21] These proportions are, again, relative to the total of 152 participants who chose non-dominated choices in the dictator game.

such, that agent i will report R iff $\gamma_i \in [\underline{\gamma}_i^R, \overline{\gamma}_i^R]$.

Table 4.4 Threshold values for the γ parameter - Lying game

State	Report	Own Payoff	Av. N.A.	γ-threshold	Data
RED	RED	£17.00	0.628		32%
	BLUE	£8.50	0.336	-	3%
	GREEN	£0.00	-0.464	-	0%
BLUE	RED	£17.00	-0.72	5.170	9%
	BLUE	£8.50	0.924		27%
	GREEN	£0.00	-0.736	-	0%
GREEN	RED	£17.00	-0.612	12.143	9%
	BLUE	£8.50	0.088	18.640	9%
	GREEN	£0.00	0.544		11%

Threshold values for γ based on the average normative assessment scores. Each threshold is the value of γ at which a player is indifferent between the action at the same line and the immediately next undominated one. Where threshold values are missing, the corresponding actions are strictly dominated based on the theory and the observed average degrees of social appropriateness. The proportions refer to the 152 people who made consistent choices in the dictator game.

We begin our analysis of the lying game by examining whether choices that are deemed dominated by the model are indeed avoided by our participants. Afterwards, we exclude the participants who have made such choices and evaluate the performance of the model in anticipating the behaviour of the rest. We initially focus on each state separately and then consider the game as a whole.

We start with state RED. Here there is no conflict between personal payoff and social appropriateness. Consider, thus, the following hypothesis:

Hypothesis 7. In state RED of the lying game every player should report RED, irrespectively of the value of her/his γ_i parameter.

Out of the 61 people who found themselves in this state 54 are eligible for our analysis (based on their dictator-game choices). Most of these 54 (91%) reported RED. Those who deviated (9%) reported BLUE instead. This degree of deviation from the model's prediction may be the result of random errors in decision-making. A Fisher's exact test between the distribution we observe and the one we theoretically anticipate indicates that the two are not statistically different at the 5% level (Fisher's exact: $p = 0.057$). Thus, the Krupka-Weber model performs well in accounting for the behaviour of our subjects in the RED state of the lying game.

We next turn to state BLUE, where the most socially appropriate report is also an equitable one. We can form the following hypothesis:

Hypothesis 8. In state BLUE of the lying game no player should report GREEN, irrespectively of the value of her/his γ_i parameter.

We can draw predictions for 54 out of the 63 people that found themselves in this state. Out of these 54 subjects, 24% reported RED and 76% reported BLUE. Thus, hypothesis 8 is confidently confirmed. None of our subjects reported GREEN in this state.

It is particularly encouraging that no one opted for GREEN, unless it was the truthful report. This is a further indication that people understood how the game works and were not choosing randomly. Additionally, the very low frequency of BLUE reports in the RED state is an indication that the Krupka-Weber model is, to some extent, able to account for the way in which people make choices.

We now proceed to test the model's consistency in accounting for our participants' behaviour. As before, we exclude the people who reported BLUE when the true outcome was RED, as well as those who made dominated choices in the dictator game, from the analysis. This brings our initial sample size of 178 down to 147 subjects. Table 4.5 outlines the distributions of choices and parameter values in each state. Notice that it only include subjects with non-dominated dictator-game choices.

With respect to state RED, the performance of the model is tested by hypothesis 7 itself. Given that this hypothesis cannot be rejected, the

Table 4.5 Parameter values and returns in the trust game - Total of subjects with no dominated choices in the dictator and lying game

State	γ_i	Frequency %	Predicted choice	Frequency % of observed choice			Total frequency % of choices per colour
				RED	BLUE	GREEN	
RED	$\gamma_i \leq 1.389$	36.7%	RED	**100%**	0%	0%	RED: 100%
	$\gamma_i \in [1.389, 2.439]$	2%	RED	**100%**	0%	0%	
	$\gamma_i \in [2.439, 4.545]$	18.4%	RED	**100%**	0%	0%	BLUE: 0%
	$\gamma_i \in [4.545, 4.545]$	0%	RED	0%	0%	0%	
	$\gamma_i \in [4.545, 5.682]$	20.4%	RED	**100%**	0%	0%	GREEN: 0%
	$\gamma_i \in [5.682, 5.725]$	14.3%	RED	**100%**	0%	0%	
	$\gamma_i \geq 5.725$	8.2%	RED	**100%**	0%	0%	
BLUE	$\gamma_i \leq 1.389$	27.8%	RED	46.7%	**53.3%**	0%	RED: 24.1%
	$\gamma_i \in [1.389, 2.439]$	5.6%	RED	**66.7%**	33.3%	0%	
	$\gamma_i \in [2.439, 4.545]$	5.6%	RED	0%	**100%**	0%	BLUE: 75.9%
	$\gamma_i \in [4.545, 4.545]$	3.7%	RED	0%	**100%**	0%	
	$\gamma_i \in [4.545, 5.682]$	27.8%	RED/BLUE	13.3%	**86.7%**	0%	GREEN: 0%
	$\gamma_i \in [5.682, 5.725]$	22.2%	BLUE	8.3%	**91.7%**	0%	
	$\gamma_i \geq 5.725$	7.4%	BLUE	25%	**75%**	0%	
GREEN	$\gamma_i \leq 1.389$	27.3%	RED	**75%**	16.7%	8.3%	RED: 29.6%
	$\gamma_i \in [1.389, 2.439]$	6.8%	RED	**66.7%**	0%	33.3%	
	$\gamma_i \in [2.439, 4.545]$	4.6%	RED	**100%**	0%	0%	BLUE: 31.8%
	$\gamma_i \in [4.545, 4.545]$	0%	RED	0%	0%	0%	
	$\gamma_i \in [4.545, 5.682]$	22.8%	RED	0%	**60%**	40%	GREEN: 38.6%
	$\gamma_i \in [5.682, 5.725]$	31.8%	RED	0%	28.6%	**71.4%**	
	$\gamma_i \geq 5.725$	6.8%	RED/BLUE/GREEN	0%	**66.7%**	33.3%	

The table lists the relative frequencies of actual choices in the lying game, as well as the prediction(s) of the Krupka-Weber model for each γ_i group, per state. The percentage frequency of each group is relative to the total within the respective state. The percentage reported in each cell of the last three columns is relative to the corresponding γ_i group. The table includes our 147 subjects that made non-dominated choices in both the dictator and the lying game. The modal choice within each group is highlighted in boldface.

model performs quite well in predicting our subjects' choices. Notice that this is not a case of relative social appropriateness. That is, it is not that reporting RED is so much more appropriate than all other options that even agents will low γ_i values are compelled to choose it. Instead, RED is unequivocally the optimal report, irrespectively of one's γ_i value. Thus, the total conformity predicted by the Krupka-Weber model is due to the alignment of incentives. Indeed, our data support this prediction.

In state BLUE the situation is different. Here own-payoff considerations and concerns about social appropriateness point towards different directions. As a result, we can form the following hypothesis:

Hypothesis 9. Player i in state BLUE of the lying game should report:

- RED (£17.00 , £0.00), iff $\gamma_i \leq 5.17$
- BLUE (£8.50 , £8.50), iff $\gamma_i \geq 5.17$

Recall that we can form predictions for 54 subjects in this state. With respect to their parameter values, 42% of them exhibit $\gamma_i \leq 4.545$ and 30%

feature $\gamma_i \leq 5.682$. The former group is expected to have reported RED and the latter BLUE. These expectations are definite. The remaining 28% is characterised by γ_i in the range $[4.545, 5.682]$. In the absence of a finer classification, we have to consider both RED and BLUE reports as consistent with their parameter values. However, even under the assumption that these people should all report BLUE,[22] the observed distribution of reports appears significantly different from the expected one ($\chi^2(1) = 4.167$, $p = 0.041$, Fisher's exact: $p = 0.065$). Additionally, the distribution of RED and BLUE reports does not seem to differ significantly across the parameter groups ($\chi^2(2) = 4.972$, $p = 0.083$, Fisher's exact: $p = 0.101$). That is, the proportions of reports appear more or less stable across the relevant γ_i intervals. We conclude that in spite of correctly predicting the absence of GREEN reports, the Krupka-Weber model cannot account for the behavioural variation we observe in this state of the lying game.

Lastly, we turn to state GREEN. This was the true colour of the wheel-spin for 54 of our participants, 44 of whom are eligible for our analysis based on their decisions as dictators. The truthful report is the most socially appropriate option in this state too. In addition, every report is explicable by some γ_i values. Consider the following hypothesis:

Hypothesis 10. Player *i* in state GREEN of the lying game should report:

- RED (£17.00 , £0.00), iff $\gamma_i \leq 12.143$
- BLUE (£8.50 , £8.50), iff $12.143 \leq \gamma_i \leq 18.64$
- GREEN (£0.00 , £0.00), iff $18.640 \leq \gamma_i$

Of our 44 participants here, 29% reported RED, 32% BLUE, and 39% GREEN. Yet, 93% exhibited $\gamma_i \leq 5.725$ and should, according to the model,

[22] This is the most favourable interpretation for the model. It involves characterising all people with γ_i in the interval $[4.545, 5.682]$ as featuring $\gamma_i \geq 5.17$.

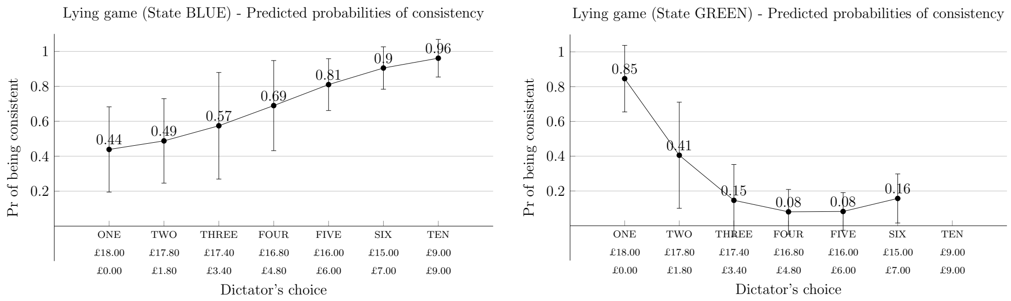


Figure 4.7: Lying game - States BLUE and GREEN - Estimated relationship between one's γ_i value and the probability that one's decision is consistent with the Krupka-Weber model

report RED. Moreover, we cannot make any more specific statements about the people that comprise the other 7%. That is, they have manifested γ_i values larger than 5.725, but we have no way of knowing whether their values are such, that they should have reported RED, BLUE, or GREEN. Therefore, in principle any report is consistent with the parameter values of those people.

Nevertheless, the model does not perform well, even given this indeterminacy. To start with, the distribution of the actual reports is markedly different from that predicted by the model ($\chi^2(2) = 37.627$, $p = 0.000$, Fisher's exact: $p = 0.000$). Furthermore, if we focus on the group of people who were expected to have reported RED,[23] we see that only 32% of them did so. In fact, most of them (39%) opted for truthfully reporting GREEN instead. Thus, once more, the Krupka-Weber model appears unable to account for the behaviour of our participants.

Are the consistency rates in the lying game suggestive of a pattern similar to that we found in the trust game? From what we can infer, this does not appear to be the case. We ran logistic regressions within each state[24] to try and estimate how the probability of being consistent with

[23] Here we refer to all participants with $\gamma_i \leq 5.725$, for whom the model gives a definite prediction.

[24] This excludes state RED, where all eligible subjects are consistent by construction: Reporting RED is the only non-dominated option.

the model's prediction varies with one's γ_i value. Our results, depicted in Figure 4.7, indicate that the relationship between degrees of consistency and parameter values differs markedly across states.

Interestingly, in state BLUE people appear more compelled to report truthfully than in state GREEN (as indicated by the aggregate percentage frequencies of reports in Table 4.5). The difference in the proportions of true reports across the two states is highly significant ($\chi^2(1) = 13.956$, $p = 0.000$, Fisher's exact: $p = 0.000$). This finding is consistent with the pattern of average ratings that we get from our other subject pool in the normative experiment. Specifically, reporting BLUE seems to be always socially appropriate to some extent, at least on average, irrespectively of the true outcome of the wheel-spin.[25] By contrast, reporting GREEN (and, interestingly, RED) is only appropriate when it is truthful. We can therefore argue that, in general, there is more normative support for reporting BLUE than there is for any of the other two options. The behaviour of our subjects in the BLUE and GREEN state of the behavioural experiment is in line with this observation.

Notice that the Krupka-Weber model does not perform badly in accounting for people's behaviour. Indeed, 70% of the 147 eligible subjects make choices that are consistent with the γ_i groups they are classified in by the model. This rate of 'success', however, comes at the expense of determinacy: The hypotheses we test are such, that quite often the model's predictions are, to a varying extent, vague. As an example, consider the subjects with $\gamma_i \geq 5.725$ in the GREEN state. From the point of view of the model, any report is, in principle, consistent with the parameter value each of these subjects may have. But then the model affords us no new insight about the way they make their decisions. Another example is the

[25] What we mean here is that reporting BLUE has a positive average score in terms of social appropriateness in all states of the lying game.

case of those with $\gamma_i \in [4.545, 5.682]$ in the BLUE state. Here the model informs us that none of them will report GREEN, which is true, but not very interesting. On the proportions of choices of RED and BLUE reports, however, the model is silent. It cannot distinguish between the two types in that parameter group.

Finally, note that there are 31 people the behaviour of whom is inexplicable by the model. These are the ones who made dominated decisions. If we consider their cases too, then the proportion of consistent subjects falls to 58%.

4.4.3 Normative disagreement and inconsistent behaviour

So far we have taken each action's average score in the Krupka-Weber task as an accurate indicator of how socially appropriate that action is. There are, however, substantial differences in the dispersion of valuations across actions. That is, there are variations in the degree of normative disagreement: For some actions we observe a much higher percentage of ratings in our normative experiment favouring the same option than for others. One can see this by noticing the varying degrees of standard deviation around the mean assessments. Tables 4.6, 4.7, and 4.8 present these differences for our three games. Each table contains the distribution of assessments across the six judgement categories for each action in the corresponding game.

This variation is important, because it can have a direct bearing on behaviour. The reason is that apart from one's propensity to choose what is socially appropriate, an equally important determinant of behaviour is one's judgement of how social appropriateness varies across the different actions. In the language of the model, knowing γ_i is not enough; to determine i's

Table 4.6 Distributions of assessments on social appropriateness - Dictator game

	Very Socially Inappropriate	Socially Inappropriate	Somewhat Socially Inappropriate	Somewhat Socially Appropriate	Socially Appropriate	Very Socially Appropriate	Mean
ONE	87%	6%	5%	0%	1%	1%	-0.9
TWO	57%	34%	4%	2%	2%	1%	-0.756
THREE	31%	49%	13%	2%	4%	1%	-0.592
FOUR	17%	45%	30%	3%	4%	1%	-0.46
FIVE	14%	24%	37%	20%	4%	1%	-0.284
SIX	8%	19%	27%	36%	8%	2%	-0.108
SEVEN	4%	12%	23%	36%	22%	3%	0.076
EIGHT	3%	4%	21%	21%	47%	4%	0.268
NINE	3%	2%	11%	17%	50%	17%	0.44
TEN	0%	0%	3%	1%	4%	92%	0.94

Note: The table presents the relative frequencies of the degrees of social appropriateness and the mean score for each action. The modal response is shaded.

Table 4.7 Distributions of assessments on social appropriateness - Trust game

	Very Socially Inappropriate	Socially Inappropriate	Somewhat Socially Inappropriate	Somewhat Socially Appropriate	Socially Appropriate	Very Socially Appropriate	Mean
IN	1%	1%	6%	21%	40%	31%	0.564
OUT	1%	9%	17%	22%	25%	26%	0.356
ONE	91%	7%	0%	1%	0%	1%	-0.94
TWO	19%	47%	27%	5%	1%	1%	-0.5
THREE	9%	10%	29%	33%	18%	1%	-0.024
FOUR	0%	0%	0%	2%	7%	91%	0.956

Note: The table presents the relative frequencies of the degrees of social appropriateness and the mean score for each action. The modal response is shaded.

Table 4.8 Distributions of assessments on social appropriateness - Lying game

		Very Socially Inappropriate	Socially Inappropriate	Somewhat Socially Inappropriate	Somewhat Socially Appropriate	Socially Appropriate	Very Socially Appropriate	Mean
State RED	RED	2%	3%	9%	11%	22%	53%	0.628
	BLUE	14%	4%	9%	14%	25%	34%	0.336
	GREEN	45%	19%	11%	13%	6%	6%	-0.464
State BLUE	RED	63%	22%	5%	3%	6%	1%	-0.72
	BLUE	1%	0%	0%	0%	14%	85%	0.924
	GREEN	65%	17%	11%	3%	2%	2%	-0.736
State GREEN	RED	53%	18%	16%	7%	4%	2%	-0.612
	BLUE	19%	11%	11%	17%	22%	20%	0.088
	GREEN	6%	2%	9%	14%	21%	48%	0.544

Note: The table presents the relative frequencies of the degrees of social appropriateness and the mean score for each action. The modal response is shaded.

behaviour, we also need to know the exact value of the $N(.)$ function (s)he assigns to every action.

In this respect, a problem arises for our testing approach for actions that exhibit high variances in their normative assessments (high degrees of normative disagreement). For such actions we cannot immediately determine whether a person's behaviour disputes the model's prediction or it should rather be attributed to their perception of how appropriate each action is. In addition, some actions that are deemed dominated based on the average assessments may not be so based on the assessments of each particular individual. In other words, if there is a large variance in how people assess an action, focusing on a test of consistency based on deviations from the average may be misleading. The reason is that a person's behaviour may appear to invalidate the model's prediction simply because their normative evaluation of a relevant action is far away from the average.

To see the problem raised by normative disagreement for our analysis, consider the case of reporting BLUE when the true outcome is GREEN in the lying game (Table 4.8).This is one of the extreme cases in our sample, where different norms appear to be in conflict. Notice, in particular, that 20% of our subjects in the normative experiment view this report as *Very Socially Appropriate*, while 19% of them judge it as *Very Socially Inappropriate*. The normative disagreement between these two groups is not merely a quantitative mis-coordination. Their assessments are qualitatively different: The first group seems to be strongly driven by a norm of pro-social payoff-equality,[26] while the second one disregards it completely (perhaps in the name of honesty).

It is thus obvious that in the presence of normative disagreement our

[26] By *pro-social* payoff-equality we mean the principle of attaining the most *socially efficient* payoff-equality. The characterisation is necessary, because, strictly speaking, reporting GREEN also achieves payoff-equality, albeit in a very inefficient way.

arguments about the performance of the Krupka-Weber model are compromised: An agent may be behaving according to the model's rationale, but based on normative judgements that are substantially different from those of our average subject. Of course, we do not know the appropriateness judgements of each of our participants in the behavioural experiment. However, we take the normative ratings elicited through the Krupka-Weber task to be representative of people's judgements in aggregate. Thus, we can try to account (at least in part) for the problem of normative disagreement by examining the *patterns* of deviation from the model's predictions.

To start with, we can immediately point out that normative disagreement should have no effect on the behaviour of people with very low γ_i values (in our behavioural experiment, those with $\gamma_i \leq 1.389$). These people should choose the payoff-maximising option in every game. Of course, people are classified based on the normative assessments in the first place. However, the assessments pertaining to actions ONE and TWO in the dictator game were relatively decisive: 87% of our normative subjects judged ONE as *Very Socially Inappropriate* and a total of 91% evaluated TWO as either *Very Socially Inappropriate* or *Socially Inappropriate* (see Table 4.6). It is, thus, highly unlikely that people in our behavioural experiment who chose that action did so thinking that it was socially appropriate to do so. Therefore, their deviations from the model's predictions can be seen as indicative of its inability to account for their behaviour. Out of the 46 people who chose action ONE in the dictator game, 15 made an inconsistent choice in at least one of the other two games, while one made a dominated choice (reported BLUE in state RED of the lying game).[27] That is, we can confidently state that around a third of our subjects with $\gamma_i \leq 1.389$ behaved in ways that were not anticipated by the model.

[27] Although the extent to which this choice is dominated can be disputed (the social appropriateness of reporting BLUE is difficult to determine due to the pull of the payoff-equality norm), this only affects one of the 46 subjects.

On the other side of the parameter-value spectrum, people with $\gamma_i \geq 5.725$ can hardly be erroneously classified as avid norm followers. As Table 4.6 indicates, action TEN in the dictator game was perceived as *Very Socially Appropriate* by 92% of our normative subjects. Thus, we can at least claim that people in our behavioural experiment who chose that action are characterised by relatively high γ_i values. Normative disagreement should affect the behaviour of these people in predictable ways. For example, they should not be expected to report RED or GREEN in the BLUE state of the lying game.[28] We can form predictions for most of these people[29] and, indeed, as we have already pointed out, their behaviour is largely consistent with the Krupka-Weber model (81% of those eligible confirm the model's predictions in both the trust and the lying game).

Let us now turn to the people with γ_i values that are most likely to have been estimated imprecisely. To what extent can we expect their observed deviations from the predictions we have generated to be due to their differing opinions on social appropriateness? Consider actions FIVE and SIX, which exhibit the highest degrees of normative disagreement in our dictator game.[30] As table 4.6 informs us, in our normative experiment most of the assessments (81% for action FIVE and 82% for action SIX) are concentrated between *Socially Inappropriate* and *Somewhat Socially Appropriate*. In our behavioural experiment 35 people chose action FIVE and 34 chose action SIX. For simplicity, ignore those who made a choice that we previously deemed dominated in either the trust or the lying game.[31]

[28] One can see in Table 4.8 that the highest degree of normative agreement in the lying game is attained in state BLUE.

[29] Specifically, 11 out of the 14 people with $\gamma_i \geq 5.725$ made non-dominated choices in all games.

[30] The social appropriateness of action SEVEN is equally ambiguous, but recall that it is dominated based on the average assessments (and, thus, we cannot form predictions for those who chose it).

[31] With respect to action THREE in the trust game, notice that, according to Table 4.7, it is highly likely to be deemed a dominated choice by a given individual. The reason is that most of the assessments that deviate from the modal one are lower. Thus, the marginal gain in terms of social appropriateness one receives by switching from action

We are left with 38 subjects whose choices we can evaluate in terms of their consistency with what the model had predicted. Of them, 84% have violated at least one of those predictions.

How can this finding be interpreted? Consider the possibility that the apparent violations of the model's predictions are in fact instances of normative disagreement. That is, that the behaviour of these subjects is actually consistent with the model and it is their normative assessments that are different from the average ones. We do not have data on the normative assessments of the participants in our behavioural experiment (as they are the ones that actually played the games). However, we can get an idea of how the distribution of those assessments would look like by looking at the responses in our normative experiment. In other words, we can attempt to infer the aggregate distribution of assessments our participants in the behavioural experiment are likely to exhibit by examining the aggregate distribution of the assessments in our normative experiment.

In Appendix C.2 we offer a full analysis of this approach and outline of our results. Here we present our findings in a rather intuitive way. Our analysis focuses on the subjects who chose action FIVE or SIX in the dictator game, as they are the ones who exhibit the highest rates of deviation from the model's predictions and their number is adequate for this type of analysis. Recall that the prediction of the Krupka-Weber model for those people is that they will choose action FOUR in the trust game. We combine the assessment patterns in the normative experiment with the choices made in the behavioural experiment. This allows us to infer that among those who chose FIVE or SIX in the dictator game about 44% could have been expected to choose actions ONE or TWO in the trust game, on the grounds of normative disagreement. In our behavioural experiment,

TWO to action THREE is likely to be small.

on the other hand, this proportion is equal to 74%. Simply put, we find significantly higher rates of deviation that what normative disagreement alone can account for.

We repeat the analysis for the same group of people in state GREEN of the lying game. Here too we find that the rates of deviation do not match the pattern of assessments in Table 4.8. Intuitively, there are far fewer RED, and more BLUE and, especially, GREEN reports than what should be expected solely due to normative disagreement. To see this, note that virtually none of the people in this state reported RED (the model's prediction). One could argue that this is because they perceived reporting RED as too inappropriate compared to the other two options. However, this argument is problematic: Given the range of γ_i values these people can have based on their dictator-game choices, their perceived difference between the social appropriateness of RED and that of BLUE or GREEN reports would have to be very high. It is immediately obvious in Table 4.8, though, that this can hardly be the case for all these people. The assessments on the social appropriateness of each option in state GREEN are quite dispersed. As a result, it should have been the case that some of them had not judged RED as too inappropriate to be chosen. Thus, even given the imprecision in estimating people's perceptions on the social appropriateness of each of the available actions, the model cannot account for the behavioural patterns we observe.

We conclude that the Krupka-Weber model of social conformity does not perform well in anticipating the behaviour of our subjects, even when normative disagreement is taken into account. Moreover, it falls short in different dimensions. Specifically, it sometimes precludes actions that turn out to be quite popular, and generates predictions that are often inaccurate and/or inconsistent with people's actual behaviour.

4.5 Concluding remarks

In this paper we evaluate the ability of the Krupka-Weber model of normative preferences to consistently account for the behaviour of our participants. For this purpose, we use an experiment involving three games that share some important qualitative characteristics: They do not involve strategic uncertainty and they are likely to be relevant to this model (based on previous studies with similar games). We find little support for the model's predictive power. The proportions of people characterised by certain degrees of sensitivity towards norm-compliance are not stable across our games. In addition, individual sensitivities also tend to vary from one game to the other.

In contrast to the narrative of the normative model, we find that most of the people who are either very selfish or very egalitarian exhibit stable preferences. Those who are very selfish try to maximise their personal material payoffs. The egalitarian ones, on the other hand, seek to achieve payoff equality. Their common element is that they are consistent in pursuing their respective goals across the three games. If we take the average normative assessments as indicative of how socially appropriate the actions in our games are, the behaviour of the egalitarian group is hard to reconcile with the Krupka-Weber model. In the face of normative conflict, where payoff equality is at odds with honesty, people who strongly value the former choose to lie, in order to stick to their principle. Due to the dominance of the honesty norm, however, they appear as norm-violators.

It is this feature that may highlight the main problem with the normative account. It may be true that people adhere to particular principles or behavioural rules due to a desire for conformity, in the fist place. However, this still does not imply that they hold all principles equally. To the extent that their perceptions about which ones dominate are different, some narrowly-focused normative agents will appear as deviants. Additionally, if we are to view them as moral agents instead, we may be able to account for their behaviour, in these games at least, more efficiently, with a more parsimonious model. Moreover, subjects who do not appear strongly motivated by selfish or pro-social principles are also unpredictable. Their propensities to abide by norms vary substantially across the three games. The behaviour of these people does not lend itself to straightforward classification. What we can conclude is that it cannot be consistently tracked by the Krupka-Weber model. Perhaps it will be fruitful for future research to use different games, in order to try and explain these people's preferences.

It can be argued that the shortcomings of the model may have resulted from differences in perceptions about social appropriateness. Overall, our analysis indicates that this is unlikely to be the case. That is, the behavioural variation is not sufficiently explicable by differences in the individual-specific values of the $N(.)$ function. Therefore, assuming that our testing procedure is valid, we can interpret the model's failure in two ways. The first is that the people in our sample are not driven by concerns about social appropriateness, but by some other motives. Each of these motives may (or may not) be specific to some, but not all of our games.[32] According to this narrative, in order to explain behaviour, we need a completely different framework altogether. The second interpretation is that people's preferences for complying with what is socially appropriate are not stable across games. In other words, each person i is characterised by different γ_i values, depending on the situation (s)he faces. In this case, if behaviour is still explicable, then there must be a way in which we can associate the various situations with these different γ_i values. However,

[32] For example, concerns about payoff equality are relevant to all games, while a desire for maximising social efficiency is only relevant to the dictator and the trust game.

in doing so, we may end up with a model that is isomorphic to the one based on separate motives. We will then have to consider the epistemic conditions imposed by each to determine which one is superior.

Finally, it may simply be the case that the norm-elicitation task does not work very well for an analysis of this level of detail. For example, our assumption that people perceive the rating 'Very Socially Inappropriate' in the same way in all three games may not be valid. If people interpret 'Very Socially Inappropriate' differently across our three games, then we have no basis for computing the relevant parameter thresholds. In this case, it is the shortcomings of the elicitation mechanism that are causing the problem. To the extent that this is true, however, we are at an impasse, since it is hard to imagine how this mechanism can be improved.

Our data suggest that some people are strongly motivated by notions that are irrelevant to social appropriateness. A very high proportion of those who choose the options that maximise their own material payoffs do so consistently across our games. The same is true for those who choose the most egalitarian options. On the other hand, some people do not manifest strong concerns of this kind. It might be tempting to assume that these are the ones who are driven by concerns about social appropriateness, but this does not seem to be the case.

4.6 References

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

Akerlof, G. A., & Kranton, R. E. (2005). Identity and the Economics of Organizations. *The Journal of Economic Perspectives, 19*(1), 9-32.

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

Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review, 80*(04), 1095-1111.

Barr, A., Lane, T., & Nosenzo, D. (2015). *On the social appropriateness of discrimination* (No. 2015-25). CeDEx Discussion Paper Series.

Bendor, J., & Swistak, P. (2001). The evolution of Norms. *American Journal of Sociology, 106*(6), 1493-1545.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*(1), 122-142.

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

Bicchieri, C. (2006). *The grammar of society: The emergence and dynamics of social norms.* Cambridge University Press.

Blanco, M., Engelmann, D., & Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior, 72*(2), 321-338.

Burks, S. V., & Krupka, E. L. (2012). A multimethod approach to identifying norms and normative expectations within a corporate hierarchy: Evidence from the financial services industry. *Management Science, 58*(1), 203-217.

Chang, D., Chen, R., & Krupka, E. (2015). Social norms and identity dependent preferences. *Univ Mich Work Pap.*

Cherry, T. L., Frykblom, P., & Shogren, J. F. (2002). Hardnose the dictator. *The American Economic Review, 92*(4), 1218-1221.

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.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology, 51*(3), 629.

Elster, J. (2000). Social norms and economic theory. In *Culture and Politics* (pp. 363-380). Palgrave Macmillan US.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior, 6*(3), 347-369.

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science, 322*(5907), 1510-1510.

Gächter, S., Nosenzo, D., & Sefton, M. (2013). Peer effects in pro-social behavior: Social norms or social preferences?. *Journal of the European Economic Association, 11*(3), 548-573.

Hechter, M., & Opp, K. D. (Eds.). (2001). *Social norms.* Russell Sage

Foundation.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362-1367.

Kanazawa, S. (2001). De gustibus est disputandum. *Social Forces, 79*(3), 1131-1162.

Kimbrough, E. O., & Vostroknutov, A. (2016a). Norms make preferences social. *Journal of the European Economic Association, 14*(3), 608-638.

Kimbrough, E. O., & Vostroknutov, A. (2016b). *Eliciting respect for social norms.* working paper

Krupka, E. L., Leider, S., & Jiang, M. (2016). A meeting of the minds: Informal agreements and social norms. *Management Science, 63*(6), 1708-1729.

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.

Ledyard, J. O. (1995). *Public Goods. A Survey of Experimental Research.* S. 111-194 in: Kagel, J. H., & Roth, A. E. (Eds.). (2016). *The Handbook of Experimental Economics, Volume 2: The Handbook of Experimental Economics.* Princeton university press.

Schwartz, S. H. (1973). Normative explanations of helping behavior: A critique, proposal, and empirical test. *Journal of Experimental Social Psychology, 9*(4), 349-364.

Young, H. P. (2015). The evolution of social norms. *Economics, 7*(1), 359-387.

Chapter 5

Conclusions

The three studies reported in this thesis investigate different aspects of prosocial behaviour, using game-theoretic tools and experimental methods. The first of these studies (reported in chapter 2) explores the potential for moral preferences to arise as optimal correction mechanisms that eliminate the effects of time-inconsistency in preferences. The other two (reported in chapters 3 and 4) examine whether social behaviour as it is manifested in the experimental laboratory can be consistently accounted for by two exemplary models of social preferences.

Chapter 2 proposes a game-theoretic account of the emergence of moral preferences. Adopting a consequentialist viewpoint, the study links the inculcation of such preferences into an individual's utility function to their potential in improving her/his material situation. At a first glance, this might appear a very narrow conception of morality. However, it can be useful in at least three distinct ways. To start with, the issue of how moral rules may arise and survive selection processes based on material outcomes (which is how we think that evolution operates) is a very important one, indeed. If anything, the moral content of an action or a particular type of behaviour may turn out to be complementary to its material consequences, and if this is the case, it is worth being pointed out. The same analysis may also be used to caution against cases where these two aspects are (at least seemingly) contradictory and in those cases it is not clear why one should prevail over the other (see e.g. Dawkins, 2016 on the discrepancy between genes and humans). On a separate, methodological note, accounting for a variety of behavioural determinants within a single framework may prove to be a useful exercise (see Dietrich and List, 2013 for a larger-scale classification). Finally, the fact that morality can be examined from a materialist perspective in the first place is worth being put forward in its own right, because it highlights an aspect of the notion that is worth considering, particularly with respect to how it is construed in the first place. These three lines of inquiry are areas where the research reported in chapter 2 can be expanded. Another project of great interest is an expansion of the existing model that considers the interplay among the agents in a more comprehensive manner.

Chapters 3 and 4 report studies that investigate the potential of two seminal accounts of social preferences to accurately and consistently track behaviour in an experimental setting. The first of these two accounts is the model of inequality aversion, proposed by Fehr and Schmidt (1999). The second is the model of adherence to social norms, advanced by Krupka and Weber (2013). While the separate examination of the performance of each model is interesting and informative in its own right, a comparative view of their relative strengths and shortcomings affords some profound insights on the determinants of behaviour.

The evaluation of the inequality-aversion model is reported in chapter 3. The setup involves three one-shot, two-player games, aimed at distinguishing among people with varying degrees of aversion to advantageous payoff inequality (also referred to as guilt). Each participant is asked to provide a decision for every decision node of every player in each of these games. The results indicate that the behaviour of the participants in general cannot be consistently accounted for by the model. However, most of those who exhibit either particularly selfish or very egalitarian preferences are also consistently doing so. That is, while the model appears unable to account for the whole of the behavioural variation that we observe in our data, it does seem to be able to capture some of it. That is, it does seem to be able to account for the behaviour of certain *types* of people. A useful test of the robustness of this claim would be an experimental investigation involving games with payoff structures that expand progressively towards both directions. Such a setup would allow for a finer distinction between, e.g., selfish and status-quo preferences, as well as between egalitarian and altruistic ones.

Chapter 4 reports the evaluation of the norm-adherence model. The experimental setup is the same one used for the study in chapter 3, with the addition of an experiment aimed at determining the degree of social appropriateness pertaining to each action available in the games. The model's performance is not supported by the results. It appears that the only people the behaviour of whom can be explained by it are those who choose the payoff-maximising option every time, whether it is socially appropriate to do so or not. Thus, it does not appear to afford any additional interpretative power relative to the standard materialist account, while it imposes further epistemic requirements. An element of interest in the fallibility of some of the model's predictions is that they appear to be driven by people's tendency to adhere to their particular motives even when the actions they prescribe are not the most socially appropriate.

When comparing the two models in terms of their performance this particular feature stands out, not least because it is the opposite of what one might expect. The dependency of choices on the context within which they are expressed is a well-documented regularity in the wider experimental literature (see e.g. Krupka and Weber, 2013 for a review). The norm-adherence model aims to account for this dependency by providing a more nuanced and flexible concept for the determinants of behaviour, which is also more demanding: In addition to the individual-specific propensities for norm-following, one also needs to know the ratings of social appropriateness that pertain to all the actions involved in the choice problem at hand in order to draw predictions. What this comparative evaluation suggests, however, is that the more nuanced account may in fact be worse at anticipating people's behaviour. At the very least, it appears that there are some individuals who are strongly motivated by particular principles (here payoff equality) and they tend to adhere to them even when they compare unfavourably, in society's view, to other principles. This finding suggests that at least some individuals can be classified as conforming to certain types, defined according to some principles that are independent of context. It thus reinforces the conclusions reached by Fischbacher and Gächter (2006), who find clear evidence in support of the existence of heterogeneous types in public-good games.

In this sense, the findings of the two experimental studies provide some empirical support for the way moral preferences are construed in the game-theoretic model proposed in this thesis. Given that some agents appear to exhibit preferences for particular moral rules, an interesting avenue for further research involves the analysis of interactions of distinct moral doctrines. To some extent, this 'battle of ideas' scenario already features in many formal conceptualisations of social preferences, since the principle of maximisation of one's own material standing can itself be thought of as one such doctrine. Thus, for example, the analyses in Bisin and Verdier (2001), Adriani and Sonderegger (2009), and Alger and Weibull

(2012, 2013) already incorporate this feature. However, the consideration of a wider variety of moral rules, with different behavioural prescriptions, offers a richer structure that can afford deeper insights about the ways in which societies determine their moral codes. This potential is enhanced by the incorporation of elements of network theory in the game-theorist's tool-kit.

Finally, it is worth bearing in mind that attempts to theoretically expand and empirically evaluate notions of individual preferences can contribute towards the integration of viewpoints prevalent in different social science disciplines. This integration, if at all feasible, may provide the scientific community not only with a unified account of social behaviour, but also with a more profound understanding of the factors that determine it.

5.1 References

Adriani, F., & Sonderegger, S. (2009). Why do parents socialize their children to behave pro-socially? An information-based theory. *Journal of Public Economics, 93*(11-12), 1119-1124.

Alger, I., & Weibull, J. W. (2012). A generalization of Hamilton's rule - Love others how much?. *Journal of Theoretical Biology, 299*, 42-54.

Alger, I., & Weibull, J. W. (2013). Homo moralis - Preference evolution under incomplete information and assortative matching. *Econometrica, 81*(6), 2269-2302.

Bisin, A., & Verdier, T. (2001). The economics of cultural transmission and the dynamics of preferences. *Journal of Economic Theory, 97*(2), 298-319.

Dawkins, R. (2016). *The selfish gene.* Oxford university press.

Dietrich, F., & List, C. (2013). A Reason-Based Theory of Rational Choice. *Nous, 47*(1), 104-134.

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

Fischbacher, U., & Gächter, S. (2006). *Heterogeneous social preferences and the dynamics of free riding in public goods* (No. 2011). IZA Discussion Papers.

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.

Appendix A

A.1 P's problem

Consider the objective function of P as defined by 2.2.4.2. Upon satisfaction of the first-order condition:

$$
\begin{aligned}
F.O.C.: -\frac{1}{\beta\delta}\left(b_1 - \delta\frac{b_1 - n}{\beta\delta}\right) f\left(\frac{b_1 - n}{\beta\delta}\right) - \frac{C^{'}(n)}{\delta} &= 0 \Rightarrow \\
\Rightarrow \big[(1-\beta)b_1 - n\big] &= \frac{\beta^2}{f(\frac{b_1-n}{\beta\delta})}C^{'}(n) \Rightarrow \\
\Rightarrow n^* = (1-\beta)b_1 - \frac{\beta^2}{f(\frac{b_1-n^*}{\beta\delta})}C^{'}(n^*) & \qquad \text{(A.1.0.1)}
\end{aligned}
$$

The second-order condition for a strict maximum suggests that:

$$
-\frac{1}{\beta\delta}\big[(1-\beta)b_1 - n^*\big]f^{'}\left(\frac{b_1 - n^*}{\beta\delta}\right) - f\left(\frac{b_1 - n^*}{\beta\delta}\right) - \beta^2 C^{''}(n^*) < 0 \qquad \text{(A.1.0.2)}
$$

Notice that inequality A.1.0.2 is not necessarily satisfied for every n^* that satisfies A.1.0.1. Assumption 2.2.6 ensures that the n^* that satisfies A.1.0.1 maximises P's utility function. Assumption 2.2.5 guarantees that this maximum point is unique.

A.2 Parameter variations

We now investigate how variations in the parameters of the model affect the level of n and the rate of Y's adherence to P's preference. We focus firstly on b_1. Recall that according to A.1.0.1:

$$\big[(1-\beta)b_1 - n\big] f\left(\frac{b_1 - n}{\beta\delta}\right) - \beta^2 C^{'}(n) = 0 \tag{A.2.0.1}$$

Deriving A.2.0.1 with respect to b_1 we find:

$$\frac{\partial F.O.C.}{\partial b_1} = \frac{1}{\beta\delta}[(1-\beta)b_1 - n]f^{'}\left(\frac{b_1 - n}{\beta\delta}\right) + (1-\beta)f\left(\frac{b_1 - n}{\beta\delta}\right) \tag{A.2.0.2}$$

Deriving A.2.0.1 with respect to n we find:

$$\frac{\partial F.O.C.}{\partial n} = -\frac{1}{\beta\delta}[(1-\beta)b_1 - n]f^{'}\left(\frac{b_1 - n}{\beta\delta}\right) - f\left(\frac{b_1 - n}{\beta\delta}\right) - \beta^2 C^{''}(n) \tag{A.2.0.3}$$

Consider an exogenous shift from $\bar{b}_1$ to $\hat{b}_1$, where $|\bar{b}_1| < |\hat{b}_1|$. Let $db_1 \equiv |\hat{b}_1| - |\bar{b}_1|$ and $dn^* \equiv \hat{n}^* - \bar{n}^*$. Note that P will respond to the change in b_1 by adjusting n according to A.2.0.1. Therefore, A.2.0.2 and A.2.0.3 together add up to:

$$\left[\frac{1}{\beta\delta}[(1-\beta)b_1 - n]f'\left(\frac{b_1-n}{\beta\delta}\right) + (1-\beta)f\left(\frac{b_1-n}{\beta\delta}\right)\right]db_1 -$$
$$-\left[\frac{1}{\beta\delta}[(1-\beta)b_1 - n]f'\left(\frac{b_1-n}{\beta\delta}\right) + f\left(\frac{b_1-n}{\beta\delta}\right) + \beta^2 C''(n)\right]dn = 0$$

Thus, it is true that:

$$\frac{dn^*}{db_1} = \frac{\frac{1}{\beta\delta}[(1-\beta)b_1 - n^*]f'\left(\frac{b_1-n^*}{\beta\delta}\right) + (1-\beta)f\left(\frac{b_1-n^*}{\beta\delta}\right)}{\frac{1}{\beta\delta}[(1-\beta)b_1 - n^*]f'\left(\frac{b_1-n^*}{\beta\delta}\right) + f\left(\frac{b_1-n^*}{\beta\delta}\right) + \beta^2 C''(n^*)} \tag{A.2.0.4}$$

The sign of $\frac{dn^*}{db_1}$ depends on the sign and magnitude of $f'\left(\frac{b_1-n}{\beta\delta}\right)$. To see this, recall firstly that from equation A.1.0.2 $f'\left(\frac{b_1-n^*}{\beta\delta}\right)$ has a lower bound:

$$f'\left(\frac{b_1-n^*}{\beta\delta}\right) > -\frac{\beta\delta f\left(\frac{b_1-n^*}{\beta\delta}\right) + \beta^3\delta C''(n^*)}{(1-\beta)b_1 - n^*}$$

This means that the denominator of the fraction on the right-hand side of equation A.2.0.4 is positive for every n^* that constitutes a maximum. The numerator, on the other hand, will be negative if:

$$f'\left(\frac{b_1-n^*}{\beta\delta}\right) < -\frac{\beta\delta(1-\beta) f\left(\frac{b_1-n^*}{\beta\delta}\right)}{(1-\beta)b_1 - n^*}$$

Taking the above into account, we can discern the following cases:

$$\frac{dn^*}{db_1} = \begin{cases} y > 0, & if\ \frac{f'\left(\frac{b_1-n^*}{\beta\delta}\right)}{f\left(\frac{b_1-n^*}{\beta\delta}\right)} > -\frac{\beta\delta(1-\beta)}{(1-\beta)b_1-n^*} \\ y < 0, & if\ \frac{f'\left(\frac{b_1-n^*}{\beta\delta}\right)}{f\left(\frac{b_1-n^*}{\beta\delta}\right)} < -\frac{\beta\delta(1-\beta)}{(1-\beta)b_1-n^*} \\ y = 0, & if\ \frac{f'\left(\frac{b_1-n^*}{\beta\delta}\right)}{f\left(\frac{b_1-n^*}{\beta\delta}\right)} = -\frac{\beta\delta(1-\beta)}{(1-\beta)b_1-n^*} \end{cases} \tag{A.2.0.5}$$

It is worth noting that when $\frac{dn^*}{db_1} < 0$, n^* will fall to zero following an increase in b_1. The reason is that from 2.2.6 it can be seen that $f\left(\frac{b_1-n^*}{\beta\delta}\right)$ is decreasing more rapidly than $C(n)$. Thus, as is the case in our baseline scenario, in response to an increase in b_1 P will either increase n^*, or eliminate it altogether. We can describe the relationship between changes in b_1 and changes in n^* in a general proposition. Consider game $\mathcal{G}$ with $\bar{b}_1$, $\bar{b}_2 \sim \mathcal{F}(\bar{b}_2, \bar{\sigma}^2)$, and $\bar{n}^*$. Suppose that $\bar{b}_1$ is replaced with $\hat{b}_1$, where $|\hat{b}_1| > |\bar{b}_1|$. Such a change will, ceteris paribus, lead to:

- $\hat{n}^* > \bar{n}^*$, if $\frac{f'\left(\frac{b_1-n^*}{\beta\delta}\right)}{f\left(\frac{b_1-n^*}{\beta\delta}\right)} > -\frac{\beta\delta(1-\beta)}{(1-\beta)b_1-n^*}$.
- $\hat{n}^* < \bar{n}^*$, if $\frac{f'\left(\frac{b_1-n^*}{\beta\delta}\right)}{f\left(\frac{b_1-n^*}{\beta\delta}\right)} < -\frac{\beta\delta(1-\beta)}{(1-\beta)b_1-n^*}$.
- $\hat{n}^* = \bar{n}^*$, if $\frac{f'\left(\frac{b_1-n^*}{\beta\delta}\right)}{f\left(\frac{b_1-n^*}{\beta\delta}\right)} = -\frac{\beta\delta(1-\beta)}{(1-\beta)b_1-n^*}$.

The first of these cases corresponds to corollary 2.2.8. It suggests that so long as the percentage change in the frequency of the cut-off point is above a certain threshold, P will have an incentive to increase n^* in response to increases in b_1.

Changes in b_1 also have a bearing on compliance, which, according to corollary 2.2.8, may be negative. Following our definition of compliance (2.2.2), we can measure its variations as changes in the cumulative probability that Y's choice wil *not* conform with P's preference. As this probability dwindles, the degree of compliance increases.

Let NC be the cumulative probability that the choice of Y will be different from P's preference. Then, $NC = \mathcal{C}^{\mathcal{F}}\left(\frac{b_1-n^*}{\beta\delta}\right) - \mathcal{C}^{\mathcal{F}}\left(\frac{b_1}{\delta}\right)$, where $\mathcal{C}^{\mathcal{F}}(.)$ is the cumulative distribution function of distribution $\mathcal{F}(.)$. Consider, then, the change in this difference in response to a change in b_1.

$$\frac{\partial\big(\mathcal{C}^{\mathcal{F}}\big(\frac{b_1-n^*}{\beta\delta}\big)-\mathcal{C}^{\mathcal{F}}\big(\frac{b_1}{\delta}\big)\big)}{\partial b_1}=\left(1-\frac{\partial n^*}{\partial b_1}\right)\frac{1}{\beta\delta}f\left(\frac{b_1-n^*}{\beta\delta}\right)-\frac{1}{\delta}f\left(\frac{b_1}{\delta}\right) \tag{A.2.0.6}$$

Given that $\frac{\partial n^*}{\partial b_1}<1$ (from equation A.2.0.4), the first term of the right-hand side of A.2.0.6 is always positive. Therefore, for a sufficiently low $f\big(\frac{b_1}{\delta}\big)$ an increase in b_1 will lead to a lower degree of compliance.

To clarify this argument further, we also provide a numerical example. Consider game $\mathcal{G}$ with $\bar{b}_1=4,\hat{b}_1=6,C(n)=4n,\delta=1,\beta=0.25,\mathcal{F}(b_2,\sigma^2)=\mathcal{N}(14,2)$, where $\mathcal{N}(\mu,\sigma^2)$ is the normal distribution with mean μ, variance σ^2, and probability density function $g(x\mid\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}exp\big(-\frac{(x-\mu)^2}{2\sigma^2}\big)$. Let $\bar{n}^*$ be the equilibrium level of n corresponding to $\bar{b}_1$ and $\hat{n}^*$ the one corresponding to $\hat{b}_1$. Then, from A.2.0.1, solving for $\bar{n}^*$:

$$\begin{aligned}\bar{n}^*&=(1-\beta)\bar{b}_1-\frac{\beta^2}{f\big(\frac{\bar{b}_1-\bar{n}^*}{\beta\delta}\big)}C^{'}(\bar{n}^*)=\\&=3-\frac{0.0625}{f\big(\frac{4-\bar{n}^*}{0.25}\big)}4=\\&\approx 1\end{aligned}$$

On the other hand, solving for $\hat{n}^*$:

$$
\begin{aligned}
\hat{n}^* &= (1-\beta)\hat{b}_1 - \frac{\beta^2}{f\left(\frac{\hat{b}_1-\hat{n}^*}{\beta\delta}\right)}C^{'}(\hat{n}^*) = \\
&= 4.5 - \frac{0.0625}{j\left(\frac{4-\hat{n}^*}{0.25}\right)}4 = \\
&\approx 2.875
\end{aligned}
$$

We see that P has increased n^* in response to the rise in b_1. With respect to compliance, it is easy to see that the cumulative probability of disagreement between the two agents has increased. In particular, under $\bar{b}_1$ this probability is equal to:

$$
\mathcal{C}^{\mathcal{F}}\left(\frac{\bar{b}_1-\bar{n}^*}{\beta\delta}\right) - \mathcal{C}^{\mathcal{F}}\left(\frac{\bar{b}_1}{\delta}\right) = \mathcal{C}^{\mathcal{F}}(12) - \mathcal{C}^{\mathcal{F}}(4) \approx 0.159 \qquad \text{(A.2.0.7)}
$$

Under $\hat{b}_2$, on the other hand, it becomes:

$$
\mathcal{C}^{\mathcal{F}}\left(\frac{\hat{b}_1-\hat{n}^*}{\beta\delta}\right) - \mathcal{C}^{\mathcal{F}}\left(\frac{\hat{b}_1}{\delta}\right) = \mathcal{C}^{\mathcal{F}}(12.5) - \mathcal{C}^{\mathcal{F}}(6) \approx 0.227 \qquad \text{(A.2.0.8)}
$$

Thus, compliance decreases following the increase of b_1 from $\bar{b}_1$ to $\hat{b}_1$.

We now turn to variations in b_2 and their effect on the equilibrium level of morality. In what follows, $b_1 = \bar{b}_1$. In accordance with assumptions 2.2.5 and 2.2.6, let $b_2 \sim \mathcal{F}(\bar{b}_2, \bar{\sigma}^2)$, where $\mathcal{F}(.)$ is quasi-concave. To start with, suppose that a variance-preserving shift occurs, from $\mathcal{F}(\bar{b}_2, \bar{\sigma}^2)$ to $\mathcal{H}(\hat{b}_2, \bar{\sigma}^2)$, where $\hat{b}_2 > \bar{b}_2 > 0$. In other words, the distribution shifts towards higher values of b_2, making option B less appealing than before. Let $\bar{n}^*$ denote the equilibrium n under $\mathcal{F}(.)$ and $\hat{n}^*$ that under $\mathcal{H}(.)$. In addition, let $f(.)$ denote the probability density function of distribution $\mathcal{F}(.)$ and $h(.)$ that of distribution $\mathcal{H}(.)$. It is straightforward to verify from equation A.2.0.1 that an increase (decrease) of the density of the cut-off point that has resulted from a change in the distribution will lead to an increase (decrease) in n^*. The reason is that such a change adjusts the importance of $C^{'}(n^*)$ in the determination of n^*. In other words, it is true that if $h\left(\frac{\bar{b}_1-\bar{n}^*}{\beta\delta}\right) > f\left(\frac{\bar{b}_1-\bar{n}^*}{\beta\delta}\right)$, then $\hat{n}^* > \bar{n}^*$, while if $h\left(\frac{\bar{b}_1-\bar{n}^*}{\beta\delta}\right) < f\left(\frac{\bar{b}_1-\bar{n}^*}{\beta\delta}\right)$, then $\hat{n}^* < \bar{n}^*$. In addition, from A.2.0.1, the following two equations are true.

$$(1-\beta)\bar{b}_1 - \bar{n}^* - \frac{\beta^2}{f\left(\frac{\bar{b}_1-\bar{n}^*}{\beta\delta}\right)}C^{'}(\bar{n}^*) = 0$$

$$(1-\beta)\bar{b}_1 - \hat{n}^* - \frac{\beta^2}{h\left(\frac{\bar{b}_1-\hat{n}^*}{\beta\delta}\right)}C^{'}(\hat{n}^*) = 0$$

Therefore, it follows that:

$$\bar{n}^* - \hat{n}^* = \frac{\beta^2}{h\left(\frac{\bar{b}_1-\hat{n}^*}{\beta\delta}\right)}C^{'}(\hat{n}^*) - \frac{\beta^2}{f\left(\frac{\bar{b}_1-\bar{n}^*}{\beta\delta}\right)}C^{'}(\bar{n}^*) \qquad \text{(A.2.0.9)}$$

Then, comparing $\bar{n}^*$ with $\hat{n}^*$, one can see that:

$$\bar{n}^* > \hat{n}^* \Rightarrow f\left(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\right) > h\left(\frac{\bar{b}_1 - \hat{n}^*}{\beta\delta}\right)\frac{C^{'}(\bar{n}^*)}{C^{'}(\hat{n}^*)}$$

The converse is also true:

$$\bar{n}^* < \hat{n}^* \Rightarrow f\left(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\right) < h\left(\frac{\bar{b}_1 - \hat{n}^*}{\beta\delta}\right)\frac{C^{'}(\bar{n}^*)}{C^{'}(\hat{n}^*)}$$

Suppose, now, that $\frac{d^2C(n)}{dn^2} \geq 0$, that is, that the cost function is either convex or linear in n. In this case $\bar{n}^* > \hat{n}^* \Rightarrow \frac{C^{'}(\bar{n}^*)}{C^{'}(\hat{n}^*)} > 1$ and vice-versa. Thus:

$$\bar{n}^* > \hat{n}^* \Rightarrow f\left(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\right) > h\left(\frac{\bar{b}_1 - \hat{n}^*}{\beta\delta}\right)$$

$$\bar{n}^* < \hat{n}^* \Rightarrow f\left(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\right) < h\left(\frac{\bar{b}_1 - \hat{n}^*}{\beta\delta}\right)$$

It, thus, becomes apparent that equation A.2.0.9 implies an upper and a lower bound for the density of the new cut-off point, $h\left(\frac{\bar{b}_1 - \hat{n}^*}{\beta\delta}\right)$. That is, the density of a cut-off point that has resulted from an increase in n can never be lower than that of the initial cut-off point. Conversely, the density of a cut-off point that has resulted from a reduction in n will never surpass that of the initial cut-off point. This result implies that a parallel (variance-preserving) shift in the distribution of b_2, such as the one described above, always enhances compliance.

To see why this is the case, consider such a change, whereby rule $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is replaced by $h : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ such, that $h(b_2) = f(b_2 - \Delta)\ \ \forall b_2$, where $\Delta > 0$. Recall that $\bar{n}^*$ is the equilibrium level of morality under $f(.)$ and $\hat{n}^*$ that under $h(.)$. Then, it is true that:

$$\frac{\bar{b}_1 - \hat{n}^*}{\beta\delta} \leq \frac{\bar{b}_1 - \bar{n}^*}{\beta\delta} + \Delta \tag{A.2.0.10}$$

To see this, one can start from $\hat{n} = \bar{n}^* - \beta\delta\Delta$ and show that this is, in fact, not equal to $\hat{n}^*$.[1] Recall that if $\hat{n}$ was an equilibrium level under $h(.)$,

[1] If it were, the cut-off point $\frac{\bar{b}_1 - \hat{n}^*}{\beta\delta}$ would be in the same relative position given $h(.)$ with that of $\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}$ under $f(.)$.

then it would need to satisfy A.2.0.1. However:

$$\begin{aligned}
&\big[(1-\beta)\bar{b}_1 - \hat{n}\big]h\bigg(\frac{\bar{b}_1 - \hat{n}}{\beta\delta}\bigg) - \beta^2 C^{'}(\hat{n}) = \\
&\qquad = \big[(1-\beta)\bar{b}_1 - \bar{n}^* + \beta\delta\Delta\big]h\bigg(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta} + \Delta\bigg) - \beta^2 C^{'}(\bar{n}^* - \beta\delta\Delta) = \\
&\qquad = \big[(1-\beta)\bar{b}_1 - \bar{n}^* + \beta\delta\Delta\big]f\bigg(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\bigg) - \beta^2 C^{'}(\bar{n}^* - \beta\delta\Delta) = \\
&\qquad = \beta^2\big[C^{'}(\bar{n}^*) - C^{'}(\bar{n}^* - \beta\delta\Delta)\big] + \beta\delta\Delta f\bigg(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\bigg) > 0
\end{aligned}$$

It is obvious that A.2.0.1 is not satisfied by $\hat{n} = \bar{n}^* - \beta\delta\Delta$. Therefore, P has an incentive to further increase n, thereby increasing the probability that Y will make her preferred choice in the next period.

We can organise our findings with respect to variance-preserving distributional shifts in another general proposition. Consider game $\mathcal{G}$ with $\bar{b}_1$, $\bar{b}_2 \sim \mathcal{F}(\bar{b}_2, \bar{\sigma}^2)$, and $\bar{n}^*$. Consider a shift from $\mathcal{F}(\bar{b}_2, \bar{\sigma}^2)$ to $\mathcal{H}(\hat{b}_2, \bar{\sigma}^2)$, where $0 < \bar{b}_2 < \hat{b}_2$. Let $f(.)$ denote the probabilty density function of distribution $\mathcal{F}(.)$ and $h(.)$ denote the probability density function of distribution $\mathcal{H}(.)$. Then, such a change will, ceteris paribus, lead to:

- $\hat{n}^* < \bar{n}^*$, if $h(\frac{b_1 - \bar{n}^*}{\beta\delta}) < f(\frac{b_1 - \bar{n}^*}{\beta\delta})$.
- $\hat{n}^* > \bar{n}^*$, if $h(\frac{b_1 - \bar{n}^*}{\beta\delta}) > f(\frac{b_1 - \bar{n}^*}{\beta\delta})$.
- $\hat{n}^* = \bar{n}^*$, if $h(\frac{b_1 - \bar{n}^*}{\beta\delta}) = f(\frac{b_1 - \bar{n}^*}{\beta\delta})$.

If additionally $\frac{b_1}{\delta} < \bar{b}_2$, then, ceteris paribus, the probability that Y's choice will comply with P's preference increases as $E[b_2]$ grows larger.

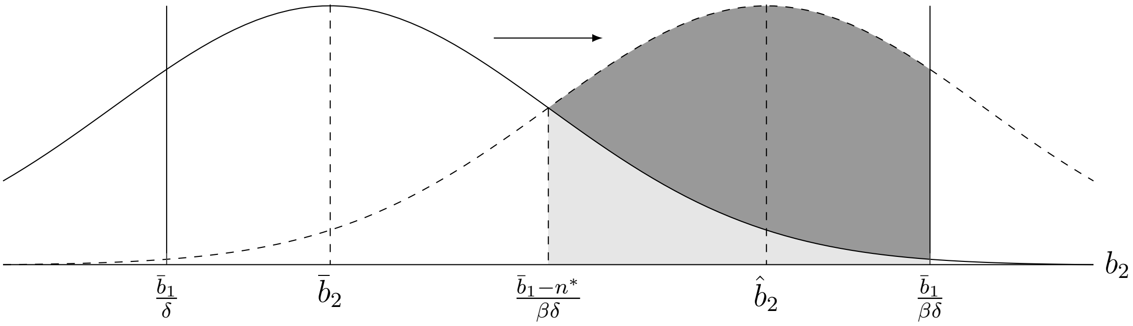

$\frac{\bar{b}_1}{\delta}$
$\bar{b}_2$
$\frac{\bar{b}_1-n^*}{\beta\delta}$
$\hat{b}_2$
$\frac{\bar{b}_1}{\beta\delta}$
b_2

Figure A.1: $\hat{b}_2 > \bar{b}_2$: The expected future consequence is relatively larger, but the level of n^* is the same.

A.3 Example of a distributional shift

What if both the mean and the variance increase as a result of the distributional shift? This is the case pertaining to proposition 2.2.10, which states that such a change may decrease both n and compliance. We show how this can be the case through an example situation. Consider game $\mathcal{G}$ with $b_1 = 4, C(n) = 2n, \delta = 1, \beta = 0.5, \mathcal{F}(\bar{b}_2, \bar{\sigma}^2) = \mathcal{N}(5.5, 0.4), \mathcal{J}(\hat{b}_2, \hat{\sigma}^2) = \mathcal{N}(7, 1)$, where $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with mean μ, variance σ^2, and probability density function $g(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\big(-\frac{(x-\mu)^2}{2\sigma^2}\big)$. Let $\bar{n}^*$ be the equilibrium level of n under distribution $\mathcal{F}(.)$ and $\hat{n}^*$ that under $\mathcal{J}(.)$. Then, from A.2.0.1, solving for $\bar{n}^*$:

$$
\begin{aligned}
\bar{n}^* &= (1-\beta)\bar{b}_1 - \frac{\beta^2}{f\big(\frac{\bar{b}_1-\bar{n}^*}{\beta\delta}\big)} C^{'}(\bar{n}^*) = \\
&= 2 - \frac{0.25}{f\big(\frac{4-\bar{n}^*}{0.5}\big)} 2 = \\
&\approx 1.38
\end{aligned}
$$

On the other hand, solving for $\hat{n}^*$:

$$
\begin{aligned}
\hat{n}^* &= (1-\beta)\bar{b}_1 - \frac{\beta^2}{j\left(\frac{\bar{b}_1-\hat{n}^*}{\beta\delta}\right)} C^{'}(\hat{n}^*) = \\
&= 4 - \frac{0.25}{j\left(\frac{4-\hat{n}^*}{0.5}\right)} 2 = \\
&\approx 0.67
\end{aligned}
$$

Thus, the level of n^* has decreased as a result of the distributional shift. Regarding compliance, let $\mathcal{C}^{\mathcal{N}}(.)$ denote the cumulative distribution function of $\mathcal{N}(.)$. Then, under $\mathcal{F}(5.5, 0.4)$ the share of b_2 values for which Y would conform with P's preference in the case of conflict was:

$$
\mathcal{C}^{\mathcal{F}}\left(\frac{\bar{b}_1}{\beta\delta}\right) - \mathcal{C}^{\mathcal{F}}\left(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\right) \approx 1 - 0.258 = 0.742
$$

Under $\mathcal{J}(7, 1)$ the share of b_2 values for which Y will conform with P's preference in the case of conflict becomes:

$$
\mathcal{C}^{\mathcal{J}}\left(\frac{\bar{b}_1}{\beta\delta}\right) - \mathcal{C}^{\mathcal{J}}\left(\frac{\bar{b}_1 - \hat{n}^*}{\beta\delta}\right) \approx 0.841 - 0.345 = 0.496
$$

Thus, both morality and compliance decrease as a result of the distributional shift. The results are illustrated in figure 2.10, in sub-section 2.2.4. In general terms, the proposition may be stated as follows. Consider game $\mathcal{G}$ with $\bar{b}_1$, $\bar{b}_2 \sim \mathcal{F}(\bar{b}_2, \bar{\sigma}^2)$, and $\bar{n}^*$. Consider a shift from $\mathcal{F}(\bar{b}_2, \bar{\sigma}^2)$ to $\mathcal{J}(\hat{b}_2, \hat{\sigma}^2)$, where $\frac{1}{\delta} < \bar{b}_2 < \hat{b_2}$ and $\bar{\sigma}^2 < \hat{\sigma}^2$. Let $f(.)$ denote the probability density function of distribution $\mathcal{F}(.)$ and $j(.)$ denote the probability density function of distribution $\mathcal{J}(.)$. Then, $\exists f, j : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ such, that $\hat{n}^* < \bar{n}^*$

and the degree of compliance is lower.

Appendix B

B.1 Experimental instructions

Instructions

Welcome and thank you for taking part in this experiment on decision making. This experiment is run by the “Centre for Decision Research and Experimental Economics” and has been financed by various research foundations. For your participation you will receive a **show-up fee of £2**. In addition, you may receive some more money, based on your choices and the choices of others.

There are other people in this room, who are also participating in this experiment. Everyone is participating for the first time, and all participants are reading the same instructions. During the experiment, we request that you **turn off your mobile phone, remain quiet, and do not attempt to communicate with other participants.** If you have a question at any time, please raise your hand and wait for the experimenter to come to your desk to answer it. Participants not following this request may be asked to leave without receiving payment.

There will be **three tasks** for all participants to perform in this experiment. In each task you will be asked to make one or more decisions, and will have a chance to earn money. You will not receive feedback on the outcome of any task until the end of the experiment, and decisions that will be made in one task will not affect decisions or earnings in the other tasks. You will not receive any instructions for or information about a task until you have completed all previous tasks. After the third task, there will also be a questionnaire. The anonymity of your responses to all parts of all tasks and questions is guaranteed.

Only **one task will be used for determining your earnings** from the experiment. At the end of the experiment, we will roll a fair six-sided die. If we roll a 1 or a 2, all participants in this experiment will be paid according to their earnings from Task 1 only. If we roll a 3 or a 4, all participants will be paid according to their earnings from Task 2 only. And if we roll a 5 or a 6, all participants will be paid according to their earnings from Task 3 only. As you will not know until the end of the experiment which task you will receive payment for, please make your decisions in each task carefully. Your earnings will be paid out to you in private and in cash at the end of the experiment.

Shortly, you will receive detailed instructions about Task 1. You will receive detailed instructions about Task 2 once everyone in the room has completed Task 1, and instructions about Task 3 once everyone in the room has completed Task 2.

If you have a question now, please raise your hand and the experimenter will come to your desk to answer it.

Task 1 - Instructions

In this task you will be randomly paired with another person in this room. At the end of this task the pair will be dissolved, and you will not be paired with this person again during this experiment.

Each person in the pair will be randomly assigned a role: "Individual A" or "Individual B", with equal probability. Individual A must choose one of ten possible actions, while Individual B has no action to take. The action taken by Individual A determines the final earnings for Individual A and Individual B in Task 1 of the experiment.

The ten possible actions that Individual A can take are listed in the table below. For each action, the table shows the corresponding earnings for Individual A and Individual B.

Individual A's action	**Individual A's earnings**	**Individual B's earnings**
ONE	£18.00	£0.00
TWO	£17.80	£1.80
THREE	£17.40	£3.40
FOUR	£16.80	£4.80
FIVE	£16.00	£6.00
SIX	£15.00	£7.00
SEVEN	£13.80	£7.80
EIGHT	£12.40	£8.40
NINE	£10.80	£8.80
TEN	£9.00	£9.00

For instance, suppose that Individual A chooses action FOUR. Then, Individual A's final earnings from Task 1 are £16.80 and Individual B's final earnings are £4.80.

Exactly who takes the role of Individual A in your pair will not be revealed until the end of the experiment. In the meantime, we ask you to make a decision **as if** you are Individual A.

At the end of the experiment, if this task is selected for payment, we will toss a fair coin to determine whether you or the person you are paired with take the role of Individual A.

- If you are selected as Individual A, then your choice will be implemented, and you and the other person will be paid according to your decision.
- If the other person in the pair is selected as Individual A, then his or her choice will be implemented, and you and the other person will be paid according to his or her decision.

Before we continue with the experiment, in order to make sure that each participant understands how their earnings from Task 1 are calculated, we ask you to answer the questions below. The experimenter will check your answers in a few minutes. Once everyone has answered all questions, we will continue with the experiment.

1. *Which of the following statements is true (circle your answer):*
   a. *You will decide who takes the role of Individual A in your pair.*
   b. *The experimenter will toss a coin to decide who takes the role of Individual A in your pair. You and the other person will be informed of the outcome of the coin toss before you make any decision in the task.*
   c. *The experimenter will toss a coin to decide who takes the role of Individual A in your pair. You and the other person will only be informed of the outcome of the coin toss at the end of the experiment.*

2. *Suppose that you choose action THREE and the other person in your pair chooses action SIX. If this task is selected for payment, and you are selected as Individual A:*
   a. *What are your earnings?* ____________
   b. *What are the other person's earnings?* ____________

3. *Suppose that you choose action TEN and the other person in your pair chooses action SEVEN. If this task is selected for payment, and the other person is selected as Individual A:*
   a. *What are your earnings?* ____________
   b. *What are the other person's earnings?* ____________

Task 1 – Decision Sheet

Please make a decision in the role of Individual A. Please choose one of the ten actions below and indicate your choice in the space provided below.

Individual A's Action	**Individual A's earnings**	**Individual B's earnings**
ONE	£18.00	£0.00
TWO	£17.80	£1.80
THREE	£17.40	£3.40
FOUR	£16.80	£4.80
FIVE	£16.00	£6.00
SIX	£15.00	£7.00
SEVEN	£13.80	£7.80
EIGHT	£12.40	£8.40
NINE	£10.80	£8.80
TEN	£9.00	£9.00

I choose action □

Once you have made your decision, fold the paper in half and put it in one of the envelopes that are placed on your desk. Shortly, the experimenter will come around to collect your envelope.

Task 2 - Instructions

In this task you will be randomly paired with another person in this room. At the end of this task the pair will be dissolved, and you will not be matched with this person again during this experiment.

Each person in the pair will be randomly assigned a role: "Individual X" or "Individual Y", with equal probability. Individual X can choose between two actions: “IN” or “OUT”.

If Individual X chooses OUT, Individual Y has no action to take, and both Individual X and Individual Y earn £4.50 each.

If Individual X chooses IN, then Individual Y must choose one of four possible actions, listed in the table below. For each action, the table shows the corresponding earnings for Individual X and Individual Y.

Individual Y’s action	**Individual Y’s earnings**	**Individual X’s earnings**
ONE	£16.60	£1.10
TWO	£15.75	£4.50
THREE	£13.75	£7.50
FOUR	£10.00	£10.00

For instance, suppose that Individual X chooses IN and Individual Y chooses action TWO. Then, Individual Y's final earnings from Task 2 are £15.75 and Individual X's final earnings are £4.50.

Exactly who in your pair takes the role of Individual X or Individual Y will not be revealed until the end of the experiment. In the meantime, we ask you to make **a decision for each role**. That is, we ask you to make two decisions: one decision as if you are Individual X, and one decision as if you are Individual Y.

At the end of the experiment, if this task is selected for payment, we will toss a fair coin to determine whether you take the role of Individual X (and thus the other person in your pair takes the role of Individual Y), or Individual Y (and thus the other person in your pair takes the role of Individual X).

- If you take the role of Individual X, then your decision in the role of Individual X and the other person’s decision in the role of Individual Y will be used to compute earnings.
- If you take the role of Individual Y, then your decision in the role of Individual Y and the other person’s decision in the role of Individual X will be used to compute earnings.

Before we continue with the experiment, in order to make sure that each participant understands how their earnings from Task 2 are calculated, we ask you to answer the questions below. The experimenter will check your answers in a few minutes. Once everyone has answered all questions, we will continue with the experiment.

1. *Which of the following statements is true (circle your answer):*
   a. *You are paired with another person in this task. You do not know whether you will be assigned the role of Individual X or Individual Y until the end of the experiment. Therefore, you are asked to make two decisions: one in the role of Individual X and one in the role of Individual Y.*
   b. *You have been assigned the role of Individual Y in this task.*
   c. *You are paired with another person in this task. You do not know whether you will be assigned the role of Individual X or Individual Y until the end of the experiment. You are asked to make a decision in the role of Individual X and the other person is asked to make a decision in the role of Individual Y.*

2. *Suppose that you choose IN in the role of Individual X, and action THREE in the role of Individual Y. Suppose that the other person in your pair chooses IN in the role of Individual X, and action ONE in the role of Individual Y. If this task is selected for payment, and you are selected as Individual X:*
   a. *What are your earnings?* ____________
   b. *What are the other person's earnings?* ____________

3. *Suppose that you choose IN in the role of Individual X, and action ONE in the role of Individual Y. Suppose that the other person in your pair chooses OUT in the role of Individual X, and action FOUR in the role of Individual Y. If this task is selected for payment, and you are selected as Individual Y:*
   a. *What are your earnings?* ____________
   b. *What are the other person's earnings?* ____________

Task 2 – Decision Sheet

Please make a decision in the role of Individual X, and a decision in the role of Individual Y. Please indicate your choices in the spaces provided below.

Individual X:

Please choose between IN and OUT.

As Individual X, I choose □

Individual Y:

Please choose one of the four actions below.

Individual Y's action	Individual Y's earnings	Individual X's earnings
ONE	£16.60	£1.10
TWO	£15.75	£4.50
THREE	£13.75	£7.50
FOUR	£10.00	£10.00

As Individual Y, I choose action □

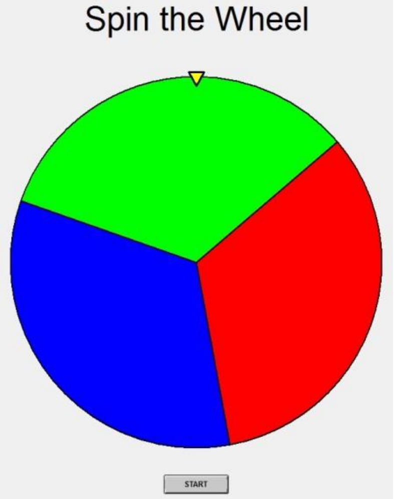
Once you have made your decision, fold the paper in half and put it in one of the envelopes that are placed on your desk. Shortly, the experimenter will come around to collect your envelope.

Task 3 - Instructions

In this task you will be randomly paired with another person in this room. At the end of this task the pair will be dissolved, and you will not be paired with this person again during this experiment.

Each person in the pair will be randomly assigned a role: "Individual J" or "Individual K", with equal probability. Only Individual J has an action to take in this task, and this action will determine the final earnings for Individual J and Individual K in Task 3 of the experiment.

On his or her computer screen, Individual J will see a “spinning wheel”, divided in three different-coloured sections of equal size: RED, BLUE, and GREEN. A screenshot of the spinning wheel is provided below.



Individual J activates the wheel by clicking the START button. The wheel will spin for a few seconds and then will stop. The screen will then go blank. Individual J must report on the decision sheet the outcome of the wheel spin. The final earnings from Task 3 for Individual J and Individual K depend on this outcome. In particular:

- If the wheel stops on RED, Individual J earns £17.00 and Individual K earns £0.00
- If the wheel stops on BLUE, both Individual J and Individual K earn £8.50 each
- If the wheel stops on GREEN, both Individual J and Individual K earn £0.00 each

Note that each colour has an equal chance of being selected.

Exactly who takes the role of Individual J in your pair will not be revealed until the end of the experiment. In the meantime, we ask you to spin the wheel **as if** you are Individual J. At the end of the experiment, if this task is selected for payment, we will toss a fair coin to determine whether you or the person you are paired with take the role of Individual J.

- If you are selected as Individual J, then we will use the outcome of the wheel spin reported on your decision sheet to compute your earnings for Task 3.
- If the other person is selected as Individual J, then we will use the outcome of the wheel spin reported on his or her decision sheet to compute your earnings for Task 3.

Before we continue with the experiment, in order to make sure that each participant understands how their earnings from Task 3 are calculated, we ask you to answer the questions below. The experimenter will check your answers in a few minutes. Once everyone has answered all questions, we will continue with the experiment.

1. *If you are Individual J and the wheel stops on GREEN:*
   1. *What are your earnings?* ____________
   2. *What are the other person's earnings?* ____________

2. *Which of the following statements is true (circle your answer):*
   1. *Your report of the outcome of the wheel spin will certainly not be used to compute earnings in this task.*
   2. *Your report of the outcome of the wheel spin will certainly be used to compute earnings in this task.*
   3. *Your report of the outcome of the wheel spin will be used to compute earnings in this task only if you are randomly assigned the role of Individual J at the end of the experiment.*

Task 3 – Decision Sheet

Please report the outcome of the wheel spin that you saw on your computer screen.

The outcome of the wheel spin was □

Once you have made your decision, fold the paper in half and put it in one of the envelopes that are placed on your desk. Shortly, the experimenter will come around to collect your envelope.

Appendix C

C.1 Experimental instructions

C.1.1 Behavioural experiment

Instructions

Welcome and thank you for taking part in this experiment on decision making. This experiment is run by the “Centre for Decision Research and Experimental Economics” and has been financed by various research foundations. For your participation you will receive a **show-up fee of £2**. In addition, you may receive some more money, based on your choices and the choices of others.

There are other people in this room, who are also participating in this experiment. Everyone is participating for the first time, and all participants are reading the same instructions. During the experiment, we request that you **turn off your mobile phone, remain quiet, and do not attempt to communicate with other participants.** If you have a question at any time, please raise your hand and wait for the experimenter to come to your desk to answer it. Participants not following this request may be asked to leave without receiving payment.

There will be **three tasks** for all participants to perform in this experiment. In each task you will be asked to make one or more decisions, and will have a chance to earn money. You will not receive feedback on the outcome of any task until the end of the experiment, and decisions that will be made in one task will not affect decisions or earnings in the other tasks. You will not receive any instructions for or information about a task until you have completed all previous tasks. After the third task, there will also be a questionnaire. The anonymity of your responses to all parts of all tasks and questions is guaranteed.

Only **one task will be used for determining your earnings** from the experiment. At the end of the experiment, we will roll a fair six-sided die. If we roll a 1 or a 2, all participants in this experiment will be paid according to their earnings from Task 1 only. If we roll a 3 or a 4, all participants will be paid according to their earnings from Task 2 only. And if we roll a 5 or a 6, all participants will be paid according to their earnings from Task 3 only. As you will not know until the end of the experiment which task you will receive payment for, please make your decisions in each task carefully. Your earnings will be paid out to you in private and in cash at the end of the experiment.

Shortly, you will receive detailed instructions about Task 1. You will receive detailed instructions about Task 2 once everyone in the room has completed Task 1, and instructions about Task 3 once everyone in the room has completed Task 2.

If you have a question now, please raise your hand and the experimenter will come to your desk to answer it.

Task 1 - Instructions

In this task you will be randomly paired with another person in this room. At the end of this task the pair will be dissolved, and you will not be paired with this person again during this experiment.

Each person in the pair will be randomly assigned a role: "Individual A" or "Individual B", with equal probability. Individual A must choose one of ten possible actions, while Individual B has no action to take. The action taken by Individual A determines the final earnings for Individual A and Individual B in Task 1 of the experiment.

The ten possible actions that Individual A can take are listed in the table below. For each action, the table shows the corresponding earnings for Individual A and Individual B.

Individual A's action	**Individual A's earnings**	**Individual B's earnings**
ONE	£18.00	£0.00
TWO	£17.80	£1.80
THREE	£17.40	£3.40
FOUR	£16.80	£4.80
FIVE	£16.00	£6.00
SIX	£15.00	£7.00
SEVEN	£13.80	£7.80
EIGHT	£12.40	£8.40
NINE	£10.80	£8.80
TEN	£9.00	£9.00

For instance, suppose that Individual A chooses action FOUR. Then, Individual A's final earnings from Task 1 are £16.80 and Individual B's final earnings are £4.80.

Exactly who takes the role of Individual A in your pair will not be revealed until the end of the experiment. In the meantime, we ask you to make a decision **as if** you are Individual A.

At the end of the experiment, if this task is selected for payment, we will toss a fair coin to determine whether you or the person you are paired with take the role of Individual A.

- If you are selected as Individual A, then your choice will be implemented, and you and the other person will be paid according to your decision.
- If the other person in the pair is selected as Individual A, then his or her choice will be implemented, and you and the other person will be paid according to his or her decision.

Before we continue with the experiment, in order to make sure that each participant understands how their earnings from Task 1 are calculated, we ask you to answer the questions below. The experimenter will check your answers in a few minutes. Once everyone has answered all questions, we will continue with the experiment.

1. *Which of the following statements is true (circle your answer):*
   a. *You will decide who takes the role of Individual A in your pair.*
   b. *The experimenter will toss a coin to decide who takes the role of Individual A in your pair. You and the other person will be informed of the outcome of the coin toss before you make any decision in the task.*
   c. *The experimenter will toss a coin to decide who takes the role of Individual A in your pair. You and the other person will only be informed of the outcome of the coin toss at the end of the experiment.*

2. *Suppose that you choose action THREE and the other person in your pair chooses action SIX. If this task is selected for payment, and you are selected as Individual A:*
   a. *What are your earnings?* ____________
   b. *What are the other person's earnings?* ____________

3. *Suppose that you choose action TEN and the other person in your pair chooses action SEVEN. If this task is selected for payment, and the other person is selected as Individual A:*
   a. *What are your earnings?* ____________
   b. *What are the other person's earnings?* ____________

Task 1 – Decision Sheet

Please make a decision in the role of Individual A. Please choose one of the ten actions below and indicate your choice in the space provided below.

Individual A's Action	Individual A's earnings	Individual B's earnings
ONE	£18.00	£0.00
TWO	£17.80	£1.80
THREE	£17.40	£3.40
FOUR	£16.80	£4.80
FIVE	£16.00	£6.00
SIX	£15.00	£7.00
SEVEN	£13.80	£7.80
EIGHT	£12.40	£8.40
NINE	£10.80	£8.80
TEN	£9.00	£9.00

I choose action

Once you have made your decision, fold the paper in half and put it in one of the envelopes that are placed on your desk. Shortly, the experimenter will come around to collect your envelope.

Task 2 - Instructions

In this task you will be randomly paired with another person in this room. At the end of this task the pair will be dissolved, and you will not be matched with this person again during this experiment.

Each person in the pair will be randomly assigned a role: "Individual X" or "Individual Y", with equal probability. Individual X can choose between two actions: “IN” or “OUT”.

If Individual X chooses OUT, Individual Y has no action to take, and both Individual X and Individual Y earn £4.50 each.

If Individual X chooses IN, then Individual Y must choose one of four possible actions, listed in the table below. For each action, the table shows the corresponding earnings for Individual X and Individual Y.

Individual Y’s action	**Individual Y’s earnings**	**Individual X’s earnings**
ONE	£16.60	£1.10
TWO	£15.75	£4.50
THREE	£13.75	£7.50
FOUR	£10.00	£10.00

For instance, suppose that Individual X chooses IN and Individual Y chooses action TWO. Then, Individual Y's final earnings from Task 2 are £15.75 and Individual X's final earnings are £4.50.

Exactly who in your pair takes the role of Individual X or Individual Y will not be revealed until the end of the experiment. In the meantime, we ask you to make **a decision for each role**. That is, we ask you to make two decisions: one decision as if you are Individual X, and one decision as if you are Individual Y.

At the end of the experiment, if this task is selected for payment, we will toss a fair coin to determine whether you take the role of Individual X (and thus the other person in your pair takes the role of Individual Y), or Individual Y (and thus the other person in your pair takes the role of Individual X).

- If you take the role of Individual X, then your decision in the role of Individual X and the other person’s decision in the role of Individual Y will be used to compute earnings.
- If you take the role of Individual Y, then your decision in the role of Individual Y and the other person’s decision in the role of Individual X will be used to compute earnings.

Before we continue with the experiment, in order to make sure that each participant understands how their earnings from Task 2 are calculated, we ask you to answer the questions below. The experimenter will check your answers in a few minutes. Once everyone has answered all questions, we will continue with the experiment.

1. *Which of the following statements is true (circle your answer):*
   a. *You are paired with another person in this task. You do not know whether you will be assigned the role of Individual X or Individual Y until the end of the experiment. Therefore, you are asked to make two decisions: one in the role of Individual X and one in the role of Individual Y.*
   b. *You have been assigned the role of Individual Y in this task.*
   c. *You are paired with another person in this task. You do not know whether you will be assigned the role of Individual X or Individual Y until the end of the experiment. You are asked to make a decision in the role of Individual X and the other person is asked to make a decision in the role of Individual Y.*

2. *Suppose that you choose IN in the role of Individual X, and action THREE in the role of Individual Y. Suppose that the other person in your pair chooses IN in the role of Individual X, and action ONE in the role of Individual Y. If this task is selected for payment, and you are selected as Individual X:*
   a. *What are your earnings?* ____________
   b. *What are the other person's earnings?* ____________

3. *Suppose that you choose IN in the role of Individual X, and action ONE in the role of Individual Y. Suppose that the other person in your pair chooses OUT in the role of Individual X, and action FOUR in the role of Individual Y. If this task is selected for payment, and you are selected as Individual Y:*
   a. *What are your earnings?* ____________
   b. *What are the other person's earnings?* ____________

Task 2 – Decision Sheet

Please make a decision in the role of Individual X, and a decision in the role of Individual Y. Please indicate your choices in the spaces provided below.

Individual X:
Please choose between IN and OUT.

As Individual X, I choose □

Individual Y:
Please choose one of the four actions below.

Individual Y's action	**Individual Y's earnings**	**Individual X's earnings**
ONE	£16.60	£1.10
TWO	£15.75	£4.50
THREE	£13.75	£7.50
FOUR	£10.00	£10.00

As Individual Y, I choose action □

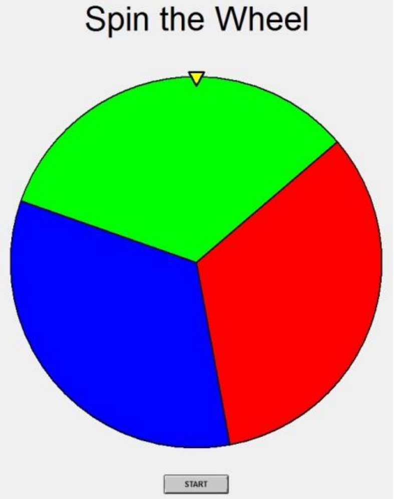
Once you have made your decision, fold the paper in half and put it in one of the envelopes that are placed on your desk. Shortly, the experimenter will come around to collect your envelope.

Task 3 - Instructions

In this task you will be randomly paired with another person in this room. At the end of this task the pair will be dissolved, and you will not be paired with this person again during this experiment.

Each person in the pair will be randomly assigned a role: "Individual J" or "Individual K", with equal probability. Only Individual J has an action to take in this task, and this action will determine the final earnings for Individual J and Individual K in Task 3 of the experiment.

On his or her computer screen, Individual J will see a “spinning wheel”, divided in three different-coloured sections of equal size: RED, BLUE, and GREEN. A screenshot of the spinning wheel is provided below.



Individual J activates the wheel by clicking the START button. The wheel will spin for a few seconds and then will stop. The screen will then go blank. Individual J must report on the decision sheet the outcome of the wheel spin. The final earnings from Task 3 for Individual J and Individual K depend on this outcome. In particular:

- If the wheel stops on RED, Individual J earns £17.00 and Individual K earns £0.00
- If the wheel stops on BLUE, both Individual J and Individual K earn £8.50 each
- If the wheel stops on GREEN, both Individual J and Individual K earn £0.00 each

Note that each colour has an equal chance of being selected.

Exactly who takes the role of Individual J in your pair will not be revealed until the end of the experiment. In the meantime, we ask you to spin the wheel **as if** you are Individual J. At the end of the experiment, if this task is selected for payment, we will toss a fair coin to determine whether you or the person you are paired with take the role of Individual J.

- If you are selected as Individual J, then we will use the outcome of the wheel spin reported on your decision sheet to compute your earnings for Task 3.
- If the other person is selected as Individual J, then we will use the outcome of the wheel spin reported on his or her decision sheet to compute your earnings for Task 3.

Before we continue with the experiment, in order to make sure that each participant understands how their earnings from Task 3 are calculated, we ask you to answer the questions below. The experimenter will check your answers in a few minutes. Once everyone has answered all questions, we will continue with the experiment.

1. *If you are Individual J and the wheel stops on GREEN:*
    a. *What are your earnings?* ____________
    b. *What are the other person's earnings?* ____________

2. *Which of the following statements is true (circle your answer):*
    a. *Your report of the outcome of the wheel spin will certainly not be used to compute earnings in this task.*
    b. *Your report of the outcome of the wheel spin will certainly be used to compute earnings in this task.*
    c. *Your report of the outcome of the wheel spin will be used to compute earnings in this task only if you are randomly assigned the role of Individual J at the end of the experiment.*

Task 3 – Decision Sheet

Please report the outcome of the wheel spin that you saw on your computer screen.

The outcome of the wheel spin was □

Once you have made your decision, fold the paper in half and put it in one of the envelopes that are placed on your desk. Shortly, the experimenter will come around to collect your envelope.

C.1.2 Normative experiment

Instructions

Welcome and thank you for taking part in this experiment on decision making. This experiment is run by the "Centre for Decision Research and Experimental Economics" and has been financed by various research foundations. For your participation you will receive a **show-up fee of £5.** In addition, you may receive some more money, based on your choices and the choices of others.

There are other people in this room, who are also participating in this experiment. Everyone is participating for the first time, and all participants are reading the same instructions. During the experiment, we request that you **turn off your mobile phone, remain quiet, and do not attempt to communicate with other participants.** If you have a question at any time, please raise your hand and wait for the experimenter to come to your desk to answer it. Participants not following this request may be asked to leave without receiving payment.

In this experiment, you will read descriptions of **three situations**. In these situations one or two person(s) must decide how to act. For each situation, you will be given a description of the various possible actions that each person can choose to take.

After you read the description of each situation, you will be asked to evaluate the various possible actions that each person can take. You must indicate, for each of the possible actions, whether taking that action would be "**socially appropriate**" and "**consistent with moral or proper social behaviour**", or "**socially inappropriate**" and "**inconsistent with moral or proper social behaviour**". By socially appropriate, we mean behaviour that most people agree is the "correct" or "ethical" thing to do. Another way to think about it is that if a person were to select a socially inappropriate action, then someone else might be angry at him or her for having done so.

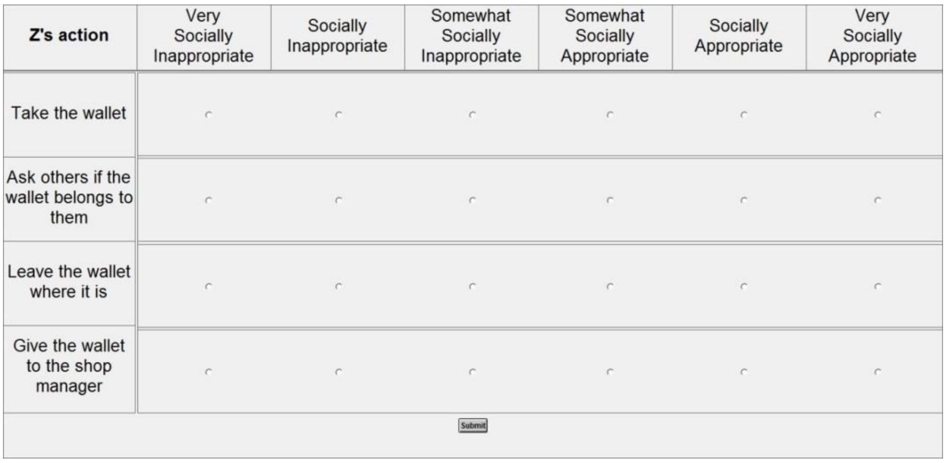
In each of your responses, we would like you to answer as truthfully as possible, based on your opinions of what constitutes socially appropriate or socially inappropriate behaviour.

To give you an idea of how the experiment will proceed, on the next pages we will go through an example situation and show you how you will indicate your responses.

Example situation

Individual Z is at a local coffee shop near campus. While there, Individual Z notices that someone has left a wallet at one of the tables. Individual Z must decide what to do and can choose one of four possible actions: take the wallet; ask others nearby if the wallet belongs to them; leave the wallet where it is; or give the wallet to the shop manager.

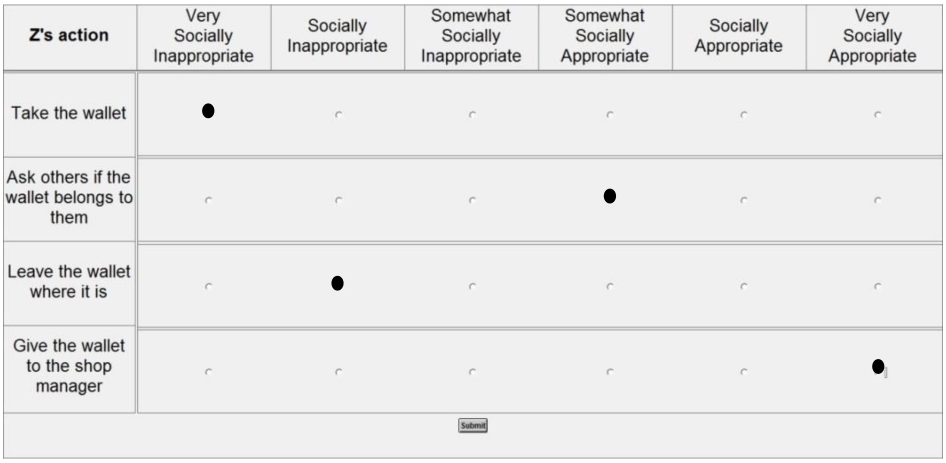
The table below presents the list of the possible actions that Individual Z can choose. For each of the actions, you would be asked to indicate whether you believe choosing that action is very socially inappropriate, socially inappropriate, somewhat socially inappropriate, somewhat socially appropriate, socially appropriate, or very socially appropriate. To indicate your response, you would click on the corresponding button.



Z's action	Very Socially Inappropriate	Socially Inappropriate	Somewhat Socially Inappropriate	Somewhat Socially Appropriate	Socially Appropriate	Very Socially Appropriate
Take the wallet	○	○	○	○	○	○
Ask others if the wallet belongs to them	○	○	○	○	○	○
Leave the wallet where it is	○	○	○	○	○	○
Give the wallet to the shop manager	○	○	○	○	○	○

If this was one of the situations for this study, you would consider each of the possible actions and, for that action, indicate the extent to which you believe taking that action would be “socially appropriate” or “socially inappropriate”. Recall that by socially appropriate we mean behaviour that most people agree is the “correct” or “ethical” thing to do.

For example, suppose you thought that taking the wallet was very socially inappropriate, asking others nearby if the wallet belongs to them was somewhat socially appropriate, leaving the wallet where it is was socially inappropriate, and giving the wallet to the shop manager was very socially appropriate. Then you would indicate your responses as follows:



Z's action	Very Socially Inappropriate	Socially Inappropriate	Somewhat Socially Inappropriate	Somewhat Socially Appropriate	Socially Appropriate	Very Socially Appropriate
Take the wallet	●	○	○	○	○	○
Ask others if the wallet belongs to them	○	○	○	●	○	○
Leave the wallet where it is	○	●	○	○	○	○
Give the wallet to the shop manager	○	○	○	○	○	●

If you have any questions about this example situation or about how to indicate your responses, please raise your hand now.

Your task in today's experiment

You will next be given a description of three situations where one or two participants in an experiment have to choose among various possible actions. After you read each description, you must consider the possible actions and indicate on your computer screen how socially appropriate these are in tables similar to the one shown above for the example situation.

How your earnings are determined

At the end of the experiment, the computer will randomly select one of the three situations. For this situation, the computer will also randomly select one of the persons involved in the situation (if applicable) and one of the possible actions that this person could choose.

The computer will then pair you randomly with another person participating in the experiment here today. Your evaluation of the selected action will be compared with that of this randomly selected participant. **If your evaluation is the same as theirs, you will receive £7 for this task; otherwise you will receive zero**.

For instance, imagine the example situation above was the actual situation and the possible action "Leave the wallet where it is" was selected by the computer. If your evaluation had been "somewhat socially inappropriate" then your task earnings would be £7 if the person you are paired with also evaluated the action as “somewhat socially inappropriate”, and zero otherwise.

Before we continue with the experiment we want to check that each participant understands how their earnings will be calculated. To do this we ask you to answer the questions below. In a couple of minutes the experimenter will check your answers. When each participant has answered all questions correctly we will continue with the experiment.

If you have a question at any time, raise your hand and the experimenter will come to your desk to answer it.

Questions

- For the action selected for payment, if your rating is "Very socially appropriate" and the rating of the person who is randomly matched with you is "Very socially appropriate", your earning is: _________________

- For the action selected for payment, if your rating is "Very socially appropriate" and the rating of the person who is randomly matched with you is "Socially inappropriate", your earning is: _________________

Situation 1

Description of the situation

Suppose that Individual A, a participant in an experiment, is randomly paired with another participant, Individual B. The pairing is anonymous, meaning that neither individual will ever know the identity of the other individual with whom he or she is paired.

In the experiment, Individual A must choose one of ten possible actions, while Individual B has no action to take. The action taken by Individual A determines the final earnings for Individual A and Individual B in the experiment.

The ten possible actions that Individual A can take are listed in the table below. For each action, the table shows the corresponding earnings for Individual A and Individual B.

Individual A's action	Individual A's earnings	Individual B's earnings
ONE	£18.00	£0.00
TWO	£17.80	£1.80
THREE	£17.40	£3.40
FOUR	£16.80	£4.80
FIVE	£16.00	£6.00
SIX	£15.00	£7.00
SEVEN	£13.80	£7.80
EIGHT	£12.40	£8.40
NINE	£10.80	£8.80
TEN	£9.00	£9.00

For instance, suppose that Individual A chooses action FOUR. Then, Individual A's final earnings from the experiment are £16.80 and Individual B's final earnings are £4.80.

After Individual A has chosen an action, both participants are informed of the action chosen and are paid accordingly in private and in cash.

Before we continue with the experiment, in order to make sure that each participant understands how Situation 1 works, we ask you to answer the questions below. The experimenter will check your answers in a few minutes. Once everyone has answered all questions, we will continue with the experiment.

1. *Suppose that Individual A would choose action THREE:*
   a. *What would be the earnings of Individual A?* ____________
   b. *What would be the earnings of Individual B?* ____________

Your task in today's experiment

On your computer screen you will see a table where you must indicate, for each of the ten possible actions available to Individual A, whether you believe that choosing that action is very socially inappropriate, socially inappropriate, somewhat socially inappropriate, somewhat socially appropriate, socially appropriate, or very socially appropriate. Recall that by socially appropriate we mean behaviour that most people agree is the ‘’correct’’ or ‘’ethical’’ thing to do. To indicate your response, please choose one option in each row.

At the end of the experiment, if Situation 1 is selected for payment, the computer will select one possible action by Individual A at random. If your response matches the response of another randomly selected participant, you will receive £7; otherwise you will receive zero.

Please now look at your computer screen and indicate your responses.

Situation 2

Description of the situation

Suppose that Individual X, a participant in an experiment, is randomly paired with another participant, Individual Y. The pairing is anonymous, meaning that neither individual will ever know the identity of the other individual with whom he or she is paired.

Individual X can choose between two actions: "IN" or "OUT".

If Individual X chooses OUT, Individual Y has no action to take, and both Individual X and Individual Y earn £4.50 each.

If Individual X chooses IN, then Individual Y must choose one of four possible actions, listed in the table below. For each action, the table shows the corresponding earnings for Individual X and Individual Y.

Individual Y's action	**Individual Y's earnings**	**Individual X's earnings**
ONE	£16.60	£1.10
TWO	£15.75	£4.50
THREE	£13.75	£7.50
FOUR	£10.00	£10.00

For instance, suppose that Individual X chooses IN and Individual Y chooses action TWO. Then, Individual Y's final earnings from the experiment are £15.75 and Individual X's final earnings are £4.50.

After Individual X and Individual Y have chosen their actions, both participants are informed of the actions chosen and are paid accordingly in private and in cash.

Before we continue with the experiment, in order to make sure that each participant understands how Situation 2 works, we ask you to answer the questions below. The experimenter will check your answers in a few minutes. Once everyone has answered all questions, we will continue with the experiment.

1. *Suppose that Individual X would choose IN and Individual Y would choose action THREE:*
   a. *What would be the earnings of Individual X?* ____________
   b. *What would be the earnings of Individual Y?* ____________

2. *Suppose that Individual X would choose OUT and Individual Y would choose action FOUR:*
   a. *What would be the earnings of Individual X?* ____________
   b. *What would be the earnings of Individual Y?* ____________

Your task in today's experiment

On your computer screen you will see two tables, one listing the actions available to Individual X, and another listing the actions available to Individual Y. For each table, and for each action, you must indicate whether you believe that choosing that action is very socially inappropriate, socially inappropriate, somewhat socially inappropriate, somewhat socially appropriate, socially appropriate, or very socially appropriate. Recall that by socially appropriate we mean behaviour that most people agree is the ‘’correct’’ or ‘’ethical’’ thing to do. To indicate your response, please choose one option in each row.

At the end of the experiment, if Situation 2 is selected for payment, the computer will randomly select one of the two tables. For the selected table, the computer will also randomly select one action. If your response matches the response of another randomly selected participant, you will receive £7; otherwise you will receive zero.

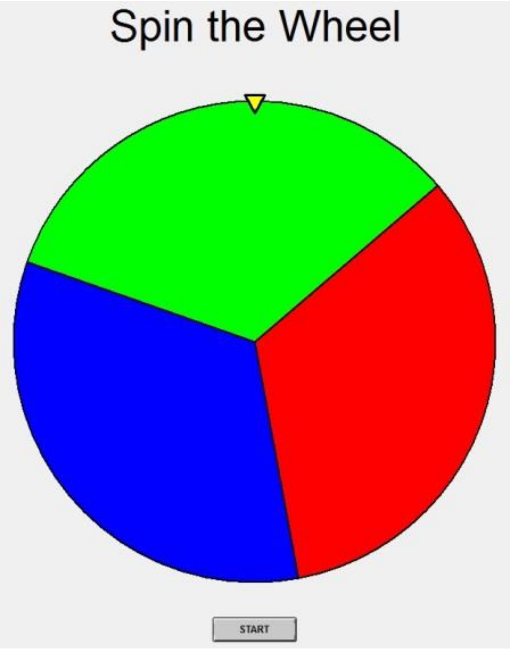
Please now look at your computer screen and indicate your responses.

Situation 3

Description of the situation

Suppose that Individual J, a participant in an experiment, is randomly paired with another participant, Individual K. The pairing is anonymous, meaning that neither individual will ever know the identity of the other individual with whom he or she is paired.

Individual J is seated at a visually separated computer workstation, like yours. On his or her computer screen, Individual J sees a "spinning wheel", divided in three different-coloured sections of equal size: RED, BLUE, and GREEN. A screenshot of the spinning wheel is provided below.



Individual J activates the wheel by clicking the START button. The wheel spins for a few seconds and then stops. The screen then goes blank. Individual J must report on a decision sheet the outcome of the wheel spin. The final earnings from the experiment for Individual J and Individual K depend on this outcome. In particular:

- If the wheel stops on RED, Individual J earns £17.00 and Individual K earns £0.00
- If the wheel stops on BLUE, both Individual J and Individual K earn £8.50 each
- If the wheel stops on GREEN, both Individual J and Individual K earn £0.00 each

After Individual J has reported the outcome of the spin, the experimenter collects the decision sheet and pays Individual J and Individual K in private and in cash according to Individual J's report.

Your task in today's experiment

On your computer screen you will see a list of the actions available to Individual J. These will be presented in three tables, each containing three rows:

1. In the first table you will evaluate Individual J’s choice to report RED after having observed the wheel stopping on RED, BLUE, or GREEN.
2. In the second table you will evaluate Individual J’s choice to report BLUE after having observed the wheel stopping on RED, BLUE, or GREEN.
3. In the third table you will evaluate Individual J’s choice to report GREEN after having observed the wheel stopping on RED, BLUE, or GREEN.

For each table and each row, you must indicate whether you believe that choosing that action is very socially inappropriate, socially inappropriate, somewhat socially inappropriate, somewhat socially appropriate, socially appropriate, or very socially appropriate. Recall that by socially appropriate we mean behaviour that most people agree is the ‘’correct’’ or ‘’ethical’’ thing to do. To indicate your response, please choose one option in each row.

At the end of the experiment, if Situation 3 is selected for payment, the computer will randomly select one of the three tables. For the selected table, the computer will also randomly select one row. If your response matches the response of another randomly selected participant, you will receive £7; otherwise you will receive zero.

Please now look at your computer screen and indicate your responses.

C.2 Normative disagreement and preference consistency

In this section we examine whether what we interpret as inconsistent behaviour on the part of our subjects in the behavioural experiment can be simply due to the fact that their judgements about social appropriateness are different from the average ones in the normative treatment. To start with, notice that what determines the behaviour of the a normative agent are not the normative scores that they assign to the available actions themselves, but rather the differences between these scores (along with the actions' material payoffs and the agent's parameter value). Suppose, then, that for a given agent, i, we do not know her/his parameter value and her/his perceptions on the normative appropriateness of each of the available actions. What we do know (assume, to be exact), however, is that the agent is characterised by a stable parameter value and that (s)he is consistently rational. That is, for any pairs of actions, α_1 and α_2 in game 1, and α_3 and α_4 in game 2, it is true that:

$$\alpha_1 \succeq \alpha_2 \Rightarrow \gamma_i \leq \frac{\pi(\alpha_1) - \pi(\alpha_2)}{N_i(\alpha_2) - N_i(\alpha_1)}$$
$$\alpha_3 \succeq \alpha_4 \Rightarrow \gamma_i \leq \frac{\pi(\alpha_3) - \pi(\alpha_4)}{N_i(\alpha_3) - N_i(\alpha_4)}$$

Stability of γ_i implies that if i exhibits the following preferences:

$$\alpha_1 \succeq \alpha_2 \quad \alpha_4 \succeq \alpha_3,$$

then it must be the case that:

$$\frac{\pi(\alpha_1) - \pi(\alpha_2)}{N_i(\alpha_2) - N_i(\alpha_1)} \geq \gamma_i \geq \frac{\pi(\alpha_3) - \pi(\alpha_4)}{N_i(\alpha_4) - N_i(\alpha_3)}$$

Notice that this expression (in particular, the $N_i(.)$ function) is person-specific and does not depend on aggregate measures. We can investigate, then, whether the patterns of deviation from the model's predictions that we observe in our behavioural treatment are similar to the structures of deviations from the mean responses in our normative treatment. That is, consider the ratio of the people within each group in our *behavioural* experiment that deviate from the model's predictions. We can examine if this ratio is statistically similar to the ratio of the people in our *normative* experiment whose assessments are consistent with such behaviour, in the manner described above.

For our analysis, we focus on the people in our behavioural experiment who chose actions FIVE and SIX in the dictator game. These groups exhibit the highest rates of deviation from the model's predictions and are sufficiently populated for such an analysis. Consider, firstly, their behaviour in the trust game. The prediction of the Krupka-Weber model, based on the average assessments in the normative experiment, is that they will all choose action FOUR as second movers (recall that action THREE is dominated). Those that do not can switch either to action ONE or to action TWO. There are, thus, four possible combinations of choices, given these deviations:

1.1.a. $\{FIVE, ONE\}$: $\frac{\pi(ONE)-\pi(FOUR)}{N_i(FOUR)-N_i(ONE)} \geq \gamma_i \geq \frac{\pi(FOUR)-\pi(FIVE)}{N_i(FIVE)-N_i(FOUR)}$

1.2.a. $\{FIVE, TWO\}$: $\frac{\pi(TWO)-\pi(FOUR)}{N_i(FOUR)-N_i(TWO)} \geq \gamma_i \geq \frac{\pi(FOUR)-\pi(FIVE)}{N_i(FIVE)-N_i(FOUR)}$

1.3.a. $\{SIX, ONE\}$: $\frac{\pi(ONE)-\pi(FOUR)}{N_i(FOUR)-N_i(ONE)} \geq \gamma_i \geq \frac{\pi(FIVE)-\pi(SIX)}{N_i(SIX)-N_i(FIVE)}$

1.4.a. $\{SIX, TWO\}$: $\frac{\pi(TWO)-\pi(FOUR)}{N_i(FOUR)-N_i(TWO)} \geq \gamma_i \geq \frac{\pi(FIVE)-\pi(SIX)}{N_i(SIX)-N_i(FIVE)}$

Given the payoff structures in the two games, cases 1.1.a-1.4.a. imply, respectively, that:

1.1.b. $N_i(FOUR) - N_i(ONE) \leq 8.25[N_i(FIVE) - N_i(FOUR)]$

1.2.b. $N_i(FOUR) - N_i(TWO) \leq 7.1875[N_i(FIVE) - N_i(FOUR)]$

1.3.b. $N_i(FOUR) - N_i(ONE) \leq 6.6[N_i(SIX) - N_i(FIVE)]$

1.4.b. $N_i(FOUR) - N_i(TWO) \leq 5.75[N_i(SIX) - N_i(FIVE)]$

That is, given these relations in the relevant pairs of normative assessments, the related choices can be rationalised within the context of the Krupka-Weber model. In our normative experiment 46% of the participants satisfy 1.1.b. and 1.2.b. (no one satisfies one and not the other) and 40% of them satisfy 1.3.b. and 1.4.b. (again, no one satisfies one and not the other). In our behavioural experiment there are 38 people who chose FIVE or SIX in the dictator game and something other than THREE in the trust game (and, thus, we can form predictions about their behaviour). Of them, 22 chose FIVE and 16 SIX in the dictator game. Based on this composition and the patterns of ratings in our normative experiment, we would expect 43.5% of them to choose either ONE or TWO as second-movers in the trust game. What we find instead is that 73.7% of them made such choices. The difference between the expected and the observed absolute frequencies of choices is significant at the 1% level ($\chi^2(1) = 6.592$, $p = 0.000$, Fisher's exact: $p = 0.009$).

To replicate the analysis in the lying game, we focus on each state separately. We star with state GREEN and consider, again, only those who chose action FIVE or SIX in the dictator game. We can infer, at the very least, that if the Krupka-Weber model is to track behaviour consistently, then the following four conditions need to hold (noting that the prediction of the model for all these people is that they will choose to report RED in this state):

2.1.a. $\{FIVE, BLUE\} \quad : \quad \frac{\pi(FIVE)-\pi(SIX)}{N_i(SIX)-N_i(FIVE)} \geq \gamma_i \geq \frac{\pi(RED)-\pi(BLUE)}{N_i(BLUE|GREEN)-N_i(RED|GREEN)}$

2.2.a. $\{FIVE, GREEN\} \quad : \quad \frac{\pi(FIVE)-\pi(SIX)}{N_i(SIX)-N_i(FIVE)} \geq \gamma_i \geq \frac{\pi(RED)-\pi(GREEN)}{N_i(GREEN|GREEN)-N_i(RED|GREEN)}$

2.3.a. $\{SIX, BLUE\} \quad : \quad \frac{\pi(SIX)-\pi(SEVEN)}{N_i(SEVEN)-N_i(SIX)} \geq \gamma_i \geq \frac{\pi(RED)-\pi(BLUE)}{N_i(BLUE|GREEN)-N_i(RED|GREEN)}$

2.4.a. $\{SIX, GREEN\} \quad : \quad \frac{\pi(SIX)-\pi(SEVEN)}{N_i(SEVEN)-N_i(SIX)} \geq \gamma_i \geq \frac{\pi(RED)-\pi(GREEN)}{N_i(GREEN|GREEN)-N_i(RED|GREEN)}$

Recall that the rating of social appropriateness attached to each report depends on the actual state. Thus, $N_i(R|S)$ represents agent i's judgement about the social appropriateness of report R conditional on state S having occurred. Notice that here action SIX is compared to action SEVEN in terms of social appropriateness. It can equally be compared to action TEN (given that SEVEN is one of those dismissed as dominated). That may render the test more or less favourable to the model, depending on the convexity of the $N_i(.)$ function relative to that of the payoff structure. Since this is ultimately an empirical matter, we examine both versions. Given the payoff structures in the two games, cases 2.1.a.-2.4.a. imply, respectively, that:

2.1.b. $N_i(BLUE|GREEN) - N_i(RED|GREEN) \geq 8.5[N_i(SIX) - N_i(FIVE)]$

2.2.b. $N_i(GREEN|GREEN) - N_i(RED|GREEN) \geq 17[N_i(SIX) - N_i(FIVE)]$

2.3.b. $N_i(BLUE|GREEN) - N_i(RED|GREEN) \geq 7.083[N_i(SEVEN) - N_i(SIX)]$

2.4.b. $N_i(GREEN|GREEN) - N_i(RED|GREEN) \geq 14.167[N_i(SEVEN) - N_i(SIX)]$

Our normative assessments indicate that 59% of our normative group satisfy 2.1.b. and 57% of them satisfy 2.2.b. (there is a negligible proportion that satisfies each one and not the other). In addition, 49% of these participants satisfy 2.3.b. and 45% of them satisfy 2.4.b. (again, there is a negligible proportion that satisfies each one and not the other). There are 24 people in our behavioural experiment who chose either FIVE or SIX in the dictator game and then found themselves in state GREEN of the lying game. Of them, 10 chose FIVE and 14 chose SIX in the dictator game. Based on this composition of choices in our behavioural experiment and the pattern of ratings in our normative experiment, we would expect at most 55% of them to choose either BLUE or GREEN in the GREEN state of the lying game. What we find instead is that all of them in fact chose one of these two options. The difference between the expected and the observed absolute frequencies of choices is significant at the 1% level

($\chi^2(1) = 12.632$, $p = 0.000$, Fisher's exact: $p = 0.001$).

If we instead compare action SIX to action TEN in the dictator game:

3.1.a. $\{FIVE, BLUE\} \quad : \quad \frac{\pi(FIVE)-\pi(SIX)}{N_i(SIX)-N_i(FIVE)} \geq \gamma_i \geq \frac{\pi(RED)-\pi(BLUE)}{N_i(BLUE|GREEN)-N_i(RED|GREEN)}$

3.2.a. $\{FIVE, GREEN\} \quad : \quad \frac{\pi(FIVE)-\pi(SIX)}{N_i(SIX)-N_i(FIVE)} \geq \gamma_i \geq \frac{\pi(RED)-\pi(GREEN)}{N_i(GREEN|GREEN)-N_i(RED|GREEN)}$

3.3.a. $\{SIX, BLUE\} \quad : \quad \frac{\pi(SIX)-\pi(TEN)}{N_i(TEN)-N_i(SIX)} \geq \gamma_i \geq \frac{\pi(RED)-\pi(BLUE)}{N_i(BLUE|GREEN)-N_i(RED|GREEN)}$

3.4.a. $\{SIX, GREEN\} \quad : \quad \frac{\pi(SIX)-\pi(TEN)}{N_i(TEN)-N_i(SIX)} \geq \gamma_i \geq \frac{\pi(RED)-\pi(GREEN)}{N_i(GREEN|GREEN)-N_i(RED|GREEN)}$

Cases 3.1.1.-3.1.4. in turn imply, respectively, that:

3.1.b. $N_i(BLUE|GREEN) - N_i(RED|GREEN) \geq 8.5[N_i(SIX) - N_i(FIVE)]$

3.2.b. $N_i(GREEN|GREEN) - N_i(RED|GREEN) \geq 17[N_i(SIX) - N_i(FIVE)]$

3.3.b. $N_i(BLUE|GREEN) - N_i(RED|GREEN) \geq 1.4167[N_i(SEVEN) - N_i(SIX)]$

3.4.b. $N_i(GREEN|GREEN) - N_i(RED|GREEN) \geq 2.833[N_i(SEVEN) - N_i(SIX)]$

Our normative experiment indicates that 59% of the participants satisfy 3.1.b. and 57% of them satisfy 3.2.b. (noting, again, that a small proportion satisfies each one and not the other). We also find that 18% satisfy 3.3.b. and 10% satisfy 3.4.b. (here the proportions of those that satisfy one and not the other are even smaller). Based on these proportions and the distribution of participants across actions FIVE and SIX of the dictator game in our behavioural experiment, we would expect at most 40% of them to choose either BLUE or GREEN in the GREEN state of the lying game. As it has already been mentioned, all of them actually reported either BLUE or GREEN. The difference between the expected and and the observed absolute frequencies of choices is even more pronounced than before($\chi^2(1) = 19.765$, $p = 0.000$, Fisher's exact: $p = 0.000$).

The above evidence suggests that the discrepancy between the behaviour we observe in the laboratory and that which is predicted by the Krupka-Weber model cannot be solely attributed to a disagreement among the subjects about how socially appropriate each option is. It may well be the case that some confusion of this kind is present, but that alone can only account for part of the behavioural variation we observe in our subjects' choices. Our results instead appear to be in favour of the argument that people's preferences cannot be consistently accounted for by the Krupka-Weber model.