Essays on Social Preferences and Beliefs in Non-Embedded Social Dilemmas

Ozan Aksoy

Utrecht University

Manuscript committee: Prof. dr. ir. V. Buskens
Prof. dr. D. Gambetta
Prof. dr. H. Hoijtink
Prof. dr. S. Rosenkranz

This research project was funded by the Netherlands Organization for Scientific Research (NWO) under grant 400-08-229.

Ozan Aksoy

Essays on Social Preferences and Beliefs in Non-Embedded Social Dilemmas

Dissertation, Utrecht University, The Netherlands.

Cover design: Ozan Aksoy

Cover illustration: Ian L., "Gear Icons", retrieved from www.stockvault.net

Printed by Wöhrmann Print Service—Zutphen

ISBN 978-90-393-5998-3

This book was composed and typesetted using LaTeX by Ozan Aksoy.

Essays on Social Preferences and Beliefs in Non-Embedded Social Dilemmas

Essays over Sociale Preferenties en Verwachtingen bij Niet-Ingebedde Sociale Dilemma's
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus,
prof. dr. G. J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op
vrijdag 30 augustus 2013 des middags te 12.45 uur

door

Ozan Aksoy

geboren op 4 oktober 1980
te Kayseri, Turkije

Promotor: Prof. dr. W. Raub
Co-promotor: Dr. J. Weesie

In memory of Fikri Alpaltaş,
my first mentor

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis is about *social dilemmas* in *non-embedded* settings. Social dilemmas are situations where individual and collective interests are in conflict (Dawes, 1980; Kollock, 1998). Collective action problems (Olson, 1965), problems of trust (e.g., Coleman, 1990), and cooperation (Taylor, 1987) are examples of social dilemmas. Social dilemmas are often studied using formal rational choice models, particularly game theory. For instance, the Prisoner's Dilemma (PD) game is used extensively to model and study numerous real-world social dilemmas. In the "standard" application of game theory, actors are assumed to be rational and selfish. According to this approach, sustaining cooperation in the PD—when played only once—is not possible. This is because each individual playing the PD would act non-cooperatively because doing so maximizes individual outcomes, irrespective of other's behavior. In game-theoretic language, not cooperating is a dominant strategy, and mutual non-cooperation is the unique Nash equilibrium of the PD game. However, if both actors cooperate, both will be better off. Mutual cooperation, however, is unstable due to individual incentives for defection. More generally, a social dilemma is an interdependent situation represented by a game with a Pareto-inefficient equilibrium (Raub et al., 2014).

In real life, however, we do observe cooperation in many situations that most observers would describe as social dilemmas. In such situations, people often cooperate and trust each other as well as honor placed trust (for an

overview, see Fehr & Gintis, 2007). The question, then, is how one explains cooperation in real life when the standard rational choice solution seemingly suggests otherwise. The answer to this question varies with the type of social dilemma. One can roughly distinguish two types of social dilemmas: *embedded* and *non-embedded*. Standard rational choice models explain cooperation between rational and selfish actors in *embedded* social dilemmas with several mechanisms (see Raub et al., 2014). For example, when the PD is embedded temporally, in the sense of being repeated indefinitely, future individual benefits of cooperation may overcome the immediate benefit of non-cooperative behavior. In that case, conditionally cooperative strategies, such as cooperating as long as the opponent cooperates and defecting otherwise, can be part of the equilibrium (Axelrod, 1984; Fudenberg & Maskin, 1986). Alternatively, when the social dilemma is embedded in a social structure, e.g., in a social network, cooperation can be sustained with mechanisms such as reputation (Raub & Weesie, 1990; Buskens & Raub, 2002). Similar to temporally embedded social dilemmas, in socially embedded social dilemmas conditionally cooperative strategies, such as cooperating if the opponent always cooperated with third parties in the past and defecting otherwise, can be part of an equilibrium. Other forms of embeddedness, such as institutional embeddedness, have similar effects on cooperation. Institutions such as laws and regulations may provide opportunities for actors to voluntarily change the incentive structures of the game. For instance, extending the game with the option to sign contracts or pledge bonds before the dilemma game is played make non-cooperation costly and can sustain cooperation (e.g., Snijders, 1996; Weesie & Raub, 1996). In this thesis we do not examine embedded social dilemmas.

Not all social dilemmas are embedded. In large, open societies, numerous interactions take place among strangers. In such interactions, actors interact only once and will not likely see each other again. Moreover, such encounters often lack opportunities for signing contracts or pledging bonds. We define social dilemmas that take place in such encounters as non-embedded social dilemmas. According to some scholars (e.g., Putnam, 2001; Delhey & Newton, 2005; Gheorghiu et al., 2009), the distinction between embedded and non-embedded social dilemmas corresponds to the distinction between *particular-*

ized or "thick" versus *generalized* or "thin" cooperation and trust. Particularized cooperation and trust take place, e.g., among family or friends, in repeated interactions, whereas generalized cooperation and trust are about interactions between strangers in social encounters. Generalized cooperation and trust in non-embedded settings are important determinants of social capital in societies at large (Putnam, 2001; Delhey & Newton, 2005).[1] Non-embedded social dilemmas, by definition, lack the mechanisms that help sustain cooperation in embedded settings. Thus, the aforementioned mechanisms cannot explain cooperation in non-embedded social dilemmas. Nevertheless, we do observe cooperation in non-embedded settings (see, e.g., Sally, 1995; Fehr et al., 2002; Camerer, 2003, as well as the experiments reported in this thesis). Standard rational choice theory employs two core assumptions mentioned above: the rationality assumption and the selfishness assumption. To explain cooperation in non-embedded settings, one can invoke two sets of mechanisms, each of which relates to one of these two core assumptions of the standard model.

1.1 Cooperation in non-embedded social dilemmas

1.1.1 Revising the rationality assumption

Some explanations of cooperation in non-embedded settings relax the rationality assumption of the standard model. That is, cooperation in non-embedded settings can be explained as the result of confusion, error, or some sort of "irrational" decision making of actors (e.g., Quattrone & Tversky, 1986; Palfrey & Prisbrey, 1997; Binmore, 2010; Burton-Chellew & West, 2013). Note that it is difficult to define what is irrational. A seemingly irrational act, such as self-harm or disclosing one's past crimes, can, in fact, be quite rational if it serves as a way to avoid external harm or help signal one's trustworthiness (Gambetta, 2009). Here, by irrationality, we mean any act that deviates from the standard rationality assumption of game theory, which has two compo-

[1] In fact, limiting people's exposure and motivation to interact with strangers, particularized forms of cooperation and trust can be harmful, as they seem to have a negative impact on generalized cooperation and trust (Yamagishi & Yamagishi, 1994; Ermisch & Gambetta, 2010).

nents, *utility maximization* and *rational beliefs*. Utility maximization implies that actors are goal directed, that is, they choose the act that maximizes utility, whatever the utility function is. The second component, rational beliefs, is more complex and is often neglected in the rational choice literature. The rational beliefs assumption concerns actors' information about each other's actions and utilities and about each other's information in situations with uncertainty. For further information on this conceptualization of rationality, see Harsanyi (1977), Chapter 1. We will discuss the rational beliefs component in detail in the subsection *Beliefs about others' preferences* below, as well as in the upcoming four chapters. Alternative decision theories, such as the theory of metagames (Howard, 1966), regret theory or maximin behavior (Savage, 1951), some bounded rationality models (e.g., Gigerenzer & Todd, 1999; Binmore, 2010), and "mythical thinking" (Quattrone & Tversky, 1986), are approaches that revise the standard rationality assumption.

We do acknowledge the role of error and confusion. Some sort of error is included in the explanatory models presented in several chapters of the thesis. For example, as we will elaborate in Chapters 2 and 3, we assume that people have well-defined utility functions but that they also make unsystematic errors in evaluating the utility of a certain outcome distribution between self and others. Moreover, as we will mention below, we include potential non-rational, ego-centered—consensus type—biases in beliefs in our explanatory models. Including these biases in beliefs relaxes the rational beliefs component of the standard rationality assumption. In this sense, we "stretch" the standard rationality assumption. However, the scope of the thesis does not include dropping the rationality assumption of the standard model altogether. We still assume some form of rationality, that is, actors try to maximize utility, given their potentially biased beliefs, but they are imperfect in accomplishing this, as represented by unsystematic evaluation error. In other words, while we relax the rational beliefs component of the standard rationality assumption, we still keep the utility maximization component nearly intact. We believe that irrationality, in the sense of dropping the utility maximization component, cannot be the main explanation of cooperation in non-embedded settings given behavioral consistencies in many experimental studies, such as subjects'

responsiveness to experimental manipulations and objective outcomes. Andreoni & Miller (2002) explicitly test whether the utility maximization assumption holds empirically. They conclude that, assuming social preferences, the choices of 98% of their subjects are consistent with utility maximization. Moreover, despite its potential drawbacks, the rationality assumption lies at the core of decision theory and game theory, which are quite fruitfully applied to many different domains of action. If one drops the rationality assumption altogether, one should also let go of decision theory and game theory. To our knowledge, there is still no strong alternative for decision theory and game theory that could match them in scope and rigor.

1.1.2 Revising the selfishness assumption

Alternative explanations of cooperation in non-embedded settings replace the selfishness assumption of the standard model. One important stream of research suggests that what people maximize is not solely their own material outcomes and that at least some people have social preferences (e.g., McClintock, 1972; Kelley & Thibaut, 1978; Liebrand & McClintock, 1988; Schulz & May, 1989; Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000; Falk & Fischbacher, 2006). One can distinguish between roughly two types of social preferences: distributional—*outcome-based*—and *process-based* (for an overview, see Fehr & Schmidt, 2006). Distributional social preferences involve how actors value a certain outcome allocation between self and others. Social value orientations, e.g., individualism, cooperativeness, altruism, competitiveness (Schulz & May, 1989), and inequality aversion (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000) are examples of distributional preferences. In process-based social preferences, actors take the history of previous interactions into account. Responding to the kind intentions of others with kindness in the form of more pro-social motives (positive reciprocity) and unkind intentions with more pro-self or competitive motives (negative reciprocity) are examples of process-based social preferences (Falk & Fischbacher, 2006; Vieth, 2009). Assuming such non-selfish preferences, one can indeed explain the behavior of people in many different experimental settings, including non-embedded social dilemmas. When people face a game which is a Prisoner's Dilemma

when only material *outcomes* are considered, they transform these material outcomes into subjective *utilities* due to social preferences (Kelley & Thibaut, 1978). The equilibrium of the effective decision matrix, in turn, may differ from the equilibrium of the PD defined in terms of material outcomes. For instance, cooperation—which cannot be part of the equilibrium of the outcome matrix of the PD—can be part of the equilibrium of the effective decision matrix, e.g., for subjects who are interested in joint outcomes rather than their own outcomes.[2]

The appeal of replacing the selfishness assumption with social preferences but not discarding the rationality assumption completely is that one can still use decision theory and game theory. However, as we will see below, replacing the selfishness assumption with social preferences comes at a cost. First, one needs to address heterogeneity in social preferences. Second, one needs to address beliefs about others' social preferences.

1.1.3 Heterogeneity in social preferences

Research has shown that people are heterogeneous with respect to their social preferences (e.g., McClintock, 1972; Kelley & Thibaut, 1978; Liebrand & McClintock, 1988; Schulz & May, 1989; Fehr & Schmidt, 1999, 2006). For example, some people dislike inequality, others are concerned with social efficiency, and others are selfish. Moreover, the same type of social motive varies in strength, e.g., people who dislike inequality do so in varying degrees. Accounting for this heterogeneity is important in explaining the findings of a wide array of experiments (Fehr & Schmidt, 1999; Simpson, 2004; Erlei, 2008). Thus, it is not enough to simply introduce one single social preference and assume that everybody has the same type and strength of social preference. Deviating from the selfishness assumption, one also needs to account for individual variation in preferences.

Here, we would like to address an ancient debate on explaining social

[2] A curious and quite under-investigated case that follows from this payoff transformation approach is the "altruist's dilemma", a game that is not a social dilemma when material outcomes are considered but that turns into a social dilemma after utility transformation (see Weesie, 1994; Heckathorn, 1996).

phenomena with the (heterogeneous) preferences of actors. Rational choice theorists, particularly economists, have been very reluctant to explain behavior and social phenomena by "adjusting" individual preferences. In their famous article *De Gustibus Non Est Disputandum*, Stigler & Becker (1977) object to introducing interpersonal heterogeneity in preferences, as one can explain virtually any differences in behavior by adjusting preferences *ex post*. Stigler & Becker (1977) advocate explaining differences in behavior not by differences in unobserved subjective preferences but by differences in observable and objective factors that are of importance for classical economics, such as income, production costs, and prices of goods. However, other contemporary economists, especially experimental economists, do proceed with providing explanations via heterogeneous social preferences, so much so that there are literally hundreds of articles that propose some form of social preferences introducing interpersonal heterogeneity in those social preferences (for an overview, see Fehr & Schmidt, 2006). Our stance on this debate is as follows. We do agree that adjusting preferences *ex post* to explain behavior is unscientific, as one can then explain almost everything with preferences. To avoid this fallacy, we start *ex ante* with a well-defined model for social preferences. Assuming *stability* in social preferences across games within a given context, we are able to predict certain patterns of behavior across these games that can be tested empirically (Caplan, 2003). Moreover, we not only give explanations for well-known empirical regularities with this model of social preferences, such as aggregate cooperation rates in non-embedded symmetric PDs, but we also test this model in new settings (Buskens & Raub, 2013). In this thesis, we create such new settings by varying game outcomes, introducing outcome asymmetry in games, and applying social preference models to different games such as Dictator Games and symmetric and asymmetric Prisoner's Dilemmas.

1.1.4 Beliefs about others' preferences

A rational choice analysis not only involves a specification of actions and utilities but also involves information players have about each other's actions and preferences as well as the other players' information. A complication arises when one deviates from the standard selfishness assumption, namely, dealing

with beliefs becomes more complex because of uncertainty with respect to others' social preferences. To be more precise, beliefs are important for the following reason. Social dilemmas are interdependent situations. In such situations, the choices of actors also depend on their beliefs about others' choices.[3] For example, in a trust situation a person may not trust the trustee if she believes that the trustee will abuse her trust, even if the trustor is pro-socially motivated. Beliefs about others' choices, in turn, depend on beliefs about others' (social) preferences. People are, however, heterogeneous with respect to their social preferences, as we discussed above. In other words, selfish as well as pro-social actors coexist. In addition, in a typical non-embedded encounter it is very difficult to observe the social preferences of the interaction partner(s). Thus, there is uncertainty about the social preferences of others. Hence, in addition to social preferences one needs to address beliefs explicitly, otherwise the explanatory model is incomplete.

We would like to discuss another line of research that is relevant here. This line of research concerns external factors that reveal the interaction partners' social preferences. These factors include observable physical cues, such as a blush that signals one's otherwise unobserved social preferences (Frank, 1988). In addition to such physical cues, pre-game play acts could also reveal unobservable social preferences. For example, committing oneself to a cooperative action (Snijders, 1996; Vieth, 2009) or revealing one's pro-sociality with costly signals prior to playing the social dilemma game (Bacharach & Gambetta, 2001) reduce, but likely do not eliminate, incomplete information about others' social preferences. Although interesting, we do not investigate such external factors that may reveal information about others' preferences. The social dilemmas studied in this thesis are truly one-shot without prior interaction[4] and are completely anonymous. That is, actors cannot identify or communicate—explicitly or implicitly—with each other.

There are several approaches to dealing with beliefs. First, one can ignore

[3] Except for special cases such as when the effective decision matrix has dominant strategies.

[4] Except the sequential social dilemmas analyzed in Chapter 5, where the second player observes the first player's action. However, the observed action of the first player still takes place within the social dilemma game.

the uncertainty about others' preferences discussed above by assuming that social preferences are common knowledge. That is, actors are assumed to know each other's social preferences. This assumption greatly simplifies the analysis. The second way to model beliefs is indeed to model them as uncertain and to turn to games with incomplete information. In behavioral game theory, uncertainty about others' utilities is dealt with via the Bayesian-Nash equilibrium framework (Harsanyi, 1968). In this framework, beliefs are assumed to be rational. One aspect of this condition is that beliefs are independent of one's own preferences, and actors know the true distribution of preferences in the "population". Unfortunately, while being mathematically elegant, this rational belief assumption is not empirically supported. Social psychological studies often demonstrate clear biases in beliefs, such as the "false" consensus, the triangle, or the cone effects (for an overview see Aksoy & Weesie, 2012b). These social psychological effects indicate that actors' own social preferences and beliefs are not independent. Such biases in beliefs are often ignored in the behavioral game theory literature (see Aksoy & Weesie, 2013a). In this thesis, we compare and test several theories on beliefs.

We need to make a further important theoretical distinction. We differentiate between two types of beliefs: beliefs about others' *preferences* and beliefs about others' *behavior*. In this thesis, a micro theory, in the form of a social preference model, is the theoretical basis of the analysis of both actors' own behavior and their beliefs about others' behavior. If we are explaining the actors' own behavior with a utility model, we also want to explain their beliefs about others' behavior with the same utility model. In other words, we endogenize beliefs about others' behavior in the explanatory model and explain beliefs about others' behavior by beliefs about others' preferences. In principle, we would have to address the infinite series of beliefs about beliefs, about, i.e., higher-order beliefs. In this thesis, however, we mostly model first-order beliefs, i.e., beliefs about others preferences. We address higher-order beliefs only in Chapter 5.

1.2 Overview of research questions and chapters

As discussed above, this thesis tries to explain cooperation in non-embedded settings. The next four chapters (Chapters 2 to 5) develop a formal game-theoretic approach that "stretches" the rationality assumption and revises the selfishness assumption of standard rational choice theory. In these four chapters, we tackle the complications that come about when one relaxes these standard assumptions, such as dealing with heterogeneity in preferences and dealing with beliefs about others' preferences. Chapters 6 and 7 include two applications of the game-theoretic framework developed in Chapters 2 to 5. Below, we briefly discuss how individual chapters fit into the general framework. We first start with the research questions of Chapters 2 to 5 and then discuss those of Chapters 6 and 7.

1.2.1 Game-theoretic analysis of cooperation in non-embedded social dilemmas

Chapters 2 and 3 develop statistical and methodological tools to measure outcome-based social preferences and beliefs about others' outcome-based social preferences. The main research questions of these two chapters are how social preferences are distributed and how beliefs about others' social preferences are related to one's own social preferences. These two chapters directly test several models of beliefs, such as rational beliefs assumed within the Bayesian-Nash equilibrium concept, and three social psychological hypotheses that incorporate ego-centered biases in beliefs. Chapters 2 and 3, as well as a statistical appendix to Chapter 2, complement each other by addressing the same substantive question with different statistical methodologies. Chapter 2 uses a two-step estimation method. In the first step, individual social preference parameters are estimated. In the second step, the individual social preference parameter estimates obtained in the first step are used to predict beliefs about others' social preferences. This two-step estimation procedure is potentially biased because measurement error in individual social preference estimates obtained in the first step is ignored in the second step. Moreover, in this two-step procedure the variation in beliefs, conditional on one's own

social preferences, also has to be ignored. We address these statistical issues in an appendix to Chapter 2. Exploiting structural equation modeling (SEM) tools on latent interactions, we provide a one-step estimation method for an approximation of our model using Mplus (Muthén & Muthén, 1998–2010). This SEM approach alleviates the problems of two-step estimation. However, this works for only very simple models, i.e., models with only one unknown social preference parameter. With models with two unknown social preference parameters, e.g., Fehr & Schmidt (1999), this SEM approach fails numerically. Chapter 3 solves this problem by proposing a one-step hierarchical Bayesian estimator using OpenBugs (Lunn et al., 2009). This Bayesian approach proves to be very flexible and works well with relatively complex models.

Chapter 4 brings in process-based social preferences, such as negative and positive reciprocity, in addition to outcome-based social preferences. Thus, in addition to the questions studied in Chapters 2 and 3, Chapter 4 investigates possible "history" effects (Snijders, 1996; Gautschi, 2000; Vieth, 2009). To clarify what is meant by "history", consider a sequential social dilemma. In a sequential social dilemma, the situation where the first player cooperated represents a different history than the situation where the first player defected. This chapter investigates whether and how social preferences, beliefs about others' social preferences, and the relationship between the two depend on history. In addition, the behavior of the first mover (player 1) in the sequential dilemma game depends on the first mover's beliefs about the social preferences of the second mover (player 2). Thus, in this chapter we study the behavioral consequences of alternative models of beliefs for the behavior of the first mover (player 1).

Chapter 5 extends the analysis by focusing on PDs where players make their decisions *simultaneously*, that is, neither player knows each other's choices prior to making their own choices. The complication in the simultaneous case is that infinite series of higher-order beliefs enter in the game-theoretic analysis. Chapter 5 includes a rigorous game-theoretic analysis on how different types of social preferences and beliefs about others' preferences influence cooperation and beliefs about others' cooperation in simultaneously played non-embedded PDs. Another important feature of Chapter 5 is that it analyzes outcome-

based social preferences but does not include process-based social preferences. It also analyzes what we call *normative* social preferences, which cannot be classified as outcome or process-based. As in all other chapters, Chapter 5 tests the game-theoretic predictions using experimental data.

1.2.2 Where do heterogeneous social preferences come from?

A question one may ask reading this thesis is where social preferences and heterogeneity in social preferences come from. There is a heated discussion on why people have social preferences, especially toward strangers, and why there is heterogeneity in social preferences. The discussion involves various disciplines, including evolutionary psychology, behavioral game theory, sociology, and social psychology (see Fehr & Gintis, 2007). Situational factors, such as the "framing" of the decision situation, also seem to influence social preferences (Hertel & Fiedler, 1998; Liberman et al., 2004; Lindenberg & Steg, 2007). A thorough investigation of the origins of social preferences and heterogeneity in social preferences is beyond the scope of this thesis. The thesis does not include a "regression" of social preferences on sociological variables such as gender, age, education, or religion (e.g., Bekkers & Wiepking, 2011), or on socialization variables (e.g., Gattig, 2002, Chapter 5). Nor does the thesis include an evolutionary model that explains the evolution of social preferences during the Pleistocene period or among our primate ancestors (e.g., Bowles & Gintis, 2004). We also do not study systematically how situational factors influence social preferences; yet, the thesis is not completely silent on the issue. As summarized below, the thesis investigates three factors, namely, history, procedural justice, and social identity, that may influence social preferences.

Chapters 2, 3, and 5 treat social preferences as given without saying much about the origins of these preferences. Bringing in process-based social preferences, Chapter 4 investigates a contextual determinant of social preferences, that is, the history between the decision maker and her interaction partner. For instance, Chapter 4 studies whether and how the social preferences of an actor change after the interaction partner defects compared with the situation where the interaction partner cooperates or where there was no prior interaction between the actor and the partner. Chapters 6 and 7 investigate two

additional sociologically important factors that may influence social preferences.

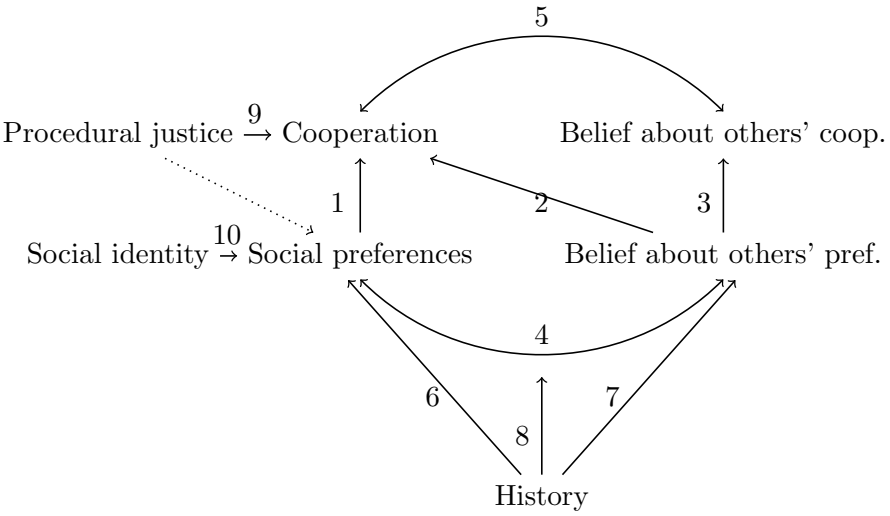
Chapter 6 studies how the process through which resources are allocated, viz., procedural justice, influences cooperation. In all other chapters, subjects play games where the material outcomes in those games are simply given. Chapter 6 investigates whether and how the process through which these outcomes are allocated influences cooperation. In particular, Chapter 6 experimentally compares the effect of three allocation procedures on cooperation: allocating resources randomly, via merit, and via ascription. To be clear, Chapter 6 does not study directly how procedural justice influences social preferences. This chapter investigates the *effects* of procedural justice on cooperation, which works theoretically via social preferences. A detailed test of the assumption that social preferences are indeed altered by procedural justice is left for future research.

The final empirical chapter, *Chapter 7*, studies another factor that may help understand where social preferences come from. In particular, Chapter 7 investigates inter-cultural differences in social preferences. With an experimental design with Turkish, Dutch, and Turkish-Dutch subjects, Chapter 7 studies how the social identities of the decision maker and her interaction partner influence the social preferences of the decision maker. Systematic differences between cultures with respect to social preferences, if any, will hint at cultural roots for social preferences. In addition, the results of Chapter 7 will show whether the social distance between actors influences social preferences.

1.2.3 A schematic overview

Figure 1.1 provides a schematic representation of the aforementioned research questions. Arrows 1 and 2 represent the effects of social preferences and beliefs about others' social preferences on cooperation. As discussed above, beliefs about the cooperative *behavior* of others are explained by beliefs about others' social *preferences*. This relationship is symbolized by Arrow 3. Possible relationships between social preferences and beliefs about others' preferences are symbolized by Arrow 4. For example, the false consensus effect predicts a high positive correlation between social preferences and beliefs about preferences

Figure 1.1: Schematic overview of the research questions of the thesis.



whereas the Bayesian-Nash equilibrium concept assumes that correlation to be zero. Arrow 5 symbolizes the relationship between cooperation and beliefs about others' cooperation. To be precise, as we will elaborate in Chapter 5, our theoretical model explains the relationship between cooperation and beliefs about others' cooperation (Arrow 5) via the relationship between social preferences and beliefs about others' preferences (Arrow 4) as well as via Arrows 1, 2, and 3. Arrows 6, 7, and 8 represent the effects of history on social preferences, beliefs about others' preferences, and the relationship between the two, respectively. Arrow 9 represents the effect of procedural justice on cooperation, which theoretically works via social preferences (via the untested dotted arrow).[5] Finally, Arrow 10 represents the effects of the social identities of the decision maker and her interaction partner on the social preferences of the decision maker. Table 1.1 summarizes which chapter investigates which research question, as represented by the arrows in Figure 1.1.

[5]We depict the influence of procedural justice on social preferences with a dotted arrow because we do not study empirically how procedural justice influences social preferences. Procedural justice influences behavior theoretically via social preferences.

Table 1.1: Research questions of the chapters and corresponding arrows in Figure 1.1. x = chapter includes research questions on the relation indicated by the respective arrow.

	Arrow in Figure 1.1									
	1	2	3	4	5	6	7	8	9	10
Chapter 2				x						
Chapter 3				x						
Chapter 4	x	x	x	x	x	x	x	x		
Chapter 5	x	x	x	x	x					
Chapter 6									x	
Chapter 7										x

1.3 Research strategy

An important feature of any thesis is its research strategy. This thesis attempts to unify theoretical and statistical models. Coleman (1981) distinguishes two alternative ways of performing empirical research: *classical hypothesis testing* and *model calibration* (also see Hedström, 2005, Chapter 5). In classical hypothesis testing, one starts with a theoretical model, derives several hypotheses about relationships between several variables from the theoretical model and tests these hypotheses about parameters of convenient, relatively simple, statistical models. Most sociological research falls under this classical hypothesis testing paradigm. Classical hypothesis testing in game-theoretic analysis refers to testing the comparative statics. For example, using a game-theoretic model with a particular type of social preferences representing, say, inequality aversion, one may derive the prediction that the level of cooperation in game X will be greater than in game Y due to the differences in material outcomes in games X and Y. One then tests the implication of the game-theoretic model by comparing the cooperation rates in games X and Y, without estimating how much people dislike inequality. We discuss this approach at length and provide examples in Chapter 5. Studies that employ classical hypothesis testing within the field of behavioral game theory include Nikiforakis & Normann (2008) and Croson (2007).

Model calibration, on the other hand, involves beginning with a theoret-

ical model with some primitives—prior parameters—that could, in principle, *generate* data. In other words, one lays out a mathematical model that mirrors the data generation process, where the theoretical and statistical models overlap. After data collection, one calibrates the primitives, that is, estimates the prior parameters of the theoretical model. In our case, continuing with the above example, model calibration would imply specifying a game-theoretic model in which inequality aversion is clearly defined with a parameter. Given this inequality aversion parameter, this game-theoretic model can generate choice data in games X and Y. After data collection, one directly estimates the inequality aversion parameter of each subject or the distribution of inequality aversion in the population, using the choices of people in games X and Y. The magnitude of the estimates as well as model fit will reveal whether the theoretical model is supported by the data. Unreasonable—e.g., extremely high or low—parameter values or a bad model fit will undermine the theoretical model. Examples of model calibration include McKelvey & Palfrey (1992); Palfrey & Prisbrey (1997) and Bellemare et al. (2008).

Coleman (1981) and Hedström (2005) argue that model calibration has some advantages over classical hypothesis testing. The main advantage is the following. In model calibration, the estimated parameters have substantive theoretical meaning, as they are a direct part of the theoretical model. For example, a parameter that describes how much, on average, people dislike inequality, or a variation in how much people dislike inequality, has an obvious substantive interpretation. The quantities in classical hypothesis testing, such as regression coefficients, however, often have no meaning aside from describing the relationships between variables of importance. The distinction between classical hypothesis testing and model calibration refers also to the distinction between testing the implications of the model and testing the assumptions of the model. The instrumentalism of classical rational choice theory suggests that incorrect assumptions are fine as long as the model's predictions are correct. In the more extreme version of instrumentalism, having correct predictions at the aggregate level is enough for a theory to be supported, even if predictions at the micro level are falsified and the assumptions of the theory are wrong (Friedman, 1953; Blanco et al., 2011). This instrumentalism is, how-

ever, often criticized as reasoning on false premises often results in incorrect conclusions with only occasional exceptions (for an overview, see Hedström, 2005, Chapter 3). Classical hypothesis testing can be seen as an application of instrumentalism, as it tests a model based solely on its predictions. Model calibration, on the other hand, works directly on the assumptions of the model by estimating the primitives of the model. In model calibration, the magnitudes of parameter estimates and the fit of the model not only provide an assessment of the accuracy of the predictions of the model but also of the appropriateness of the assumptions.

Each of the empirical chapters of this thesis includes a (formal) theoretical and an empirical analysis. In these chapters, we employ both classical hypothesis testing and model calibration. In Chapters 6 and 7, we use classical hypothesis testing. All other empirical chapters employ model calibration. In Chapters 2 to 5, we build mathematical models that are thought to generate data, and we calibrate the parameters of this mathematical model using experimental data. We can make the theoretical and statistical models overlap by a "random utility approach" (Thurstone, 1927; McFadden, 1974). In a random utility approach, social preferences are included in formal-theoretical models. Predictions are not deterministic because the model also includes stochastic terms capturing various forms of decision noise. This stochastic aspect of our theoretical models makes the theoretical models also statistical models that can be fitted to data.[6] Estimates for decision noise can be interpreted as the misfit of our theoretical models as well as lapses of rationality. In classical hypothesis testing, assessment of the misfit of theoretical models is not directly available. Another practical advantage of model calibration in our case is that the designs of the experiments are guided by the formal theoretical model that is supposed to generate data; this improves the statistical precision of the parameter estimates.

In our theoretical models, we relax the classical assumptions of rational choice theory. For example, we introduce heterogeneous social preferences

[6] Quantal response equilibrium (QRE) (McKelvey & Palfrey, 1995) is a random utility approach applied to situations with strategic interdependence, such as social dilemmas. Our game-theoretic models can be seen as applications of QRE. For another application of QRE to public good games, see Anderson et al. (1998).

and relax the rational beliefs assumption. Thus, we are making the theory more complex than the standard model. However, we are still able to connect with data, that is, we can fit our models. This is because we take advantage of modern statistical techniques such as developments in structural equation modeling on latent interactions (Muthén & Muthén, 1998–2010) and hierarchical Bayesian modeling (Gelman et al., 2004). Finally, incorporating Bayesian tools within both theoretical and statistical models also provides methodological unification. The theoretical models we develop involve subjects' decisions under uncertainty. For example, our models include subjects' prior beliefs about the social preferences of others and their conditional choices given their prior beliefs. This corresponds to Bayesian decision making (Raiffa, 1970). As mentioned above, we also use Bayesian methods as statistical tools to fit our models. Moreover, because we use vague, almost uninformative, prior distributions for the unknown parameters of our statistical models, some Bayesian estimates can also be interpreted as frequentist results (Gelman et al., 2004; Snijders & Bosker, 2012). These properties of Bayesian and frequentist approaches, as well as using Bayesian decision making as a theoretical tool, provide a further step toward the unification of statistical and theoretical modeling.

1.4 A remark on terminology

The empirical chapters in this thesis are published in outlets from three disciplines: sociology, social psychology, and economics. Chapters 5 and 6 are published in the *Journal of Mathematical Sociology* (Aksoy & Weesie, 2013b, 2009). A Dutch language translation of Chapter 7 is published in a book edition of the Dutch sociology journal *Mens en Maatschappij* (Aksoy & Weesie, 2012a). Chapter 2 is published in the *Journal of Experimental Social Psychology* (Aksoy & Weesie, 2012b). Chapter 3 is published in *Games* (Aksoy & Weesie, 2013a). Chapter 4 is currently under review. The terminology used in and the style of a chapter is adapted to the general jargon and style of the discipline and the requirements of the journal. We thus use social preferences, social motives, other-regarding preferences, and other-regarding

motives interchangeably in this thesis. We also use beliefs and expectations interchangeably.

Finally, as its title implies, this thesis is not a monograph but a collection of independently written articles. Thus, while tackling parts of an overarching research theme, namely, cooperation in non-embedded settings, the chapters are written as independent papers, and each chapter can be read independently.

Chapter 2

Beliefs about the social orientations of others: A parametric test of the triangle, false consensus, and cone hypotheses*

Abstract

This study tests a number of hypotheses proposed in the literature concerning the relationship between an actor's social orientation and her beliefs about the social orientations of others. In contrast to the existing literature, this study employs a parametric approach with an innovative methodology. First, the social orientation parameters of actors are estimated: the weights respondents add to (1) the outcomes of Alter and to (2) the absolute difference between the outcomes for Ego and Alter. Then, the mean and the variance of the distribution of beliefs about the social orientation parameters of others are estimated,

*This chapter is written in collaboration with Jeroen Weesie and published in *Journal of Experimental Social Psychology* (Aksoy & Weesie, 2012b). We thank Werner Raub, Vincent Buskens, Rense Corten, Rene Torenvlied, Michael Mäs, and the participants of the 2009 ESA Innsbruck Conference for comments on the earlier drafts of this chapter.

conditional on the actor's social orientation parameters. The results show that (1) there is a positive association between an actor's social orientation and her belief about the *mean* of the social orientations of others and (2) those who have approximately zero social orientation parameter values (individualists) expect the *variation* of others' social orientations to be lower than those with smaller (competitors) or larger (cooperators/egalitarians) social orientation parameter values. These results support the cone model, which models the "false" consensus effect where the "false" consensus is highest for individualists.

2.1 Introduction

The social orientation of an actor describes how she values a particular distribution of outcomes for herself and for others. Numerous social orientations have been distinguished in the literature. A person who has a cooperative orientation maximizes the sum of her own and other's outcomes. A person who has an individualist orientation maximizes only her own outcomes. A person who has a competitive orientation maximizes outcomes for herself and minimizes others' outcomes. A person who has an equality orientation minimizes the difference between her own and other's outcomes. A person who has a maximin orientation maximizes outcomes of the person who receives the lowest outcomes (McClintock, 1972; Griesinger & Livingston, 1977; Schulz & May, 1989). Next to identifying and distinguishing different types of social orientations, the relationship between one's social orientation and one's beliefs about others' social orientations is another important research area in the social orientation literature. Three hypotheses have been proposed on this relationship: the triangle hypothesis (Kelley & Stahelski, 1970), the structural assumed similarity bias (SASB) (Kuhlman et al., 1992), and the cone model of Iedema (1993). In this paper, we test these three hypotheses with a sophisticated approach and an innovative methodology. After briefly explaining the three hypotheses, we address some methodological shortcomings in the social orientation literature. The approach that we propose ameliorates these methodological shortcomings.

Triangle Hypothesis: Classifying individuals into two categories, competitors and cooperators, the triangle hypothesis postulates that competitors expect others to be "mono-typically" competitive whereas cooperators expect more variation in the population; that is, they expect others to be either competitors or cooperators. The reason for these heterogeneous beliefs, as stated by Kelley & Stahelski (1970), is as follows. People with competitive and cooperative orientations interact randomly. When a competitor and a cooperator interact, the competitor behaves non-cooperatively. In turn, this non-cooperative behavior makes the cooperator behave non-cooperatively. Competitors do not acknowledge the behavioral assimilation of cooperators, so competitors believe that everyone is a competitor. In other words, competitors observe mostly non-cooperative *behavior* from their interaction partners, and they infer from these behaviors that most people are of a competitive "*type*". Cooperators have more diverse experiences. When matched with another cooperator, mutual cooperation takes place. When a cooperator is matched with a competitor, however, the cooperator also behaves non-cooperatively, yielding mutual non-cooperative behavior. Thus, cooperators observe both cooperative and competitive behaviors and develop more diverse beliefs about the social orientations of others. The triangle hypothesis is usually tested within the Prisoner's Dilemma game (PD) framework. The PD is a paradigmatic 2-person game where actors have a cooperative option and an option to defect (non-cooperative). The non-cooperative option yields a higher outcome for the self irrespective of the other actor's choice. However, if each actor chooses to defect, the actors end up with a mutually worse situation than if they both had cooperated. Kelley & Stahelski (1970); Schenkler & Goldman (1978); Van Lange (1992) suggest that actors who consistently defect in the PD expect most others to defect as well, and actors who cooperate in the PD expect that the population includes both cooperators and defectors. This supports the triangle hypothesis.

Structural Assumed Similarity Bias (SASB): Some scholars (e.g., Kuhlman et al., 1992, 1986; Kuhlman & Wimberley, 1976) measure social orientations and expected social orientations using Decomposed Games (DGs)

Table 2.1: Example of a Decomposed Game. Option B maximizes the outcome of the other and thus is a cooperative option; Option A minimizes the outcome for the other and is a competitive option.

	You get	**Other gets**
Option A	540 Points	500 Points
Option B	540 Points	555 Points

instead of the PD. DGs comprise two or more choice alternatives, each of which yields an (hypothetical) outcome pair for the self and an anonymous other (see Table 2.1). Using subjects' choices in a series of DGs, subjects are classified in one of the social orientation categories. Similarly, subjects' beliefs about others' orientations are assessed in relation to their beliefs about others' choices in the DGs. These DG studies allow a more detailed classification of subjects by introducing additional social orientations, such as the 'individualist' and sometimes the 'equality' or 'maximin' orientations. A consistent finding in these studies is that actors classified in one of the social orientation categories expect others to belong predominantly to their own category. This finding is in line with the false consensus effect (Ross et al., 1977), which states that without specific information, people expect others to be similar to themselves in terms of tastes, preferences, problem-solving strategies, and other attributes.[1] Proponents of the SASB argue that the false consensus effect is applicable to all social orientation types to comparable degrees, and cooperators do not have more diverse beliefs. They also suggest that the triangle effect is found mainly in PD-like DGs in which one choice maximizes both individual and relative gains (defective choice) and another choice maximizes joint gains (cooperative choice). Thus, the triangle effect observed in a PD or in a DG representing PD emerges because individualists' and competitors' responses are merged. Kuhlman et al. (1986) coined the term for these findings: Structural Assumed Similarity Bias (SASB).

[1] Dawes (1989) indicates that the consensus effect is not necessarily *false*, that is, people would also use information on others' choices if this information were made available. For the purpose of this paper, this difference is not of great importance.

Cone Model: Iedema & Poppe (1999, 1994a) and Iedema (1993) find that, in line with SASB, actors in all social orientation categories expect others to be similar to the self. However, they discover that consensus is highest among individualists. Other types, such as cooperators and competitors, expect greater variation in the population. Iedema (1993) describes these findings as the cone model because if actors are placed on a continuum from competitors to cooperators with the individualists in the middle, false consensus is highest in the middle. Iedema (1993) finds that other orientations that cannot be placed within this competitive-cooperative orientation spectrum, such as the equality and maximin orientations, also have more diverse expectations than individualists. According to Iedema (1993), the cone effect emerges as an outcome of a combination of multiple phenomena. First, all orientation types expect their own type to be more frequent in the population, in line with the false consensus effect. In addition, Iedema (1993) argues that individualism is expected to a high degree by everyone because it is a common stereotype about other people, at least in the Western world. This adds up to individualists' consensus expectations. Thus, individualists expect the occurrence of their own orientation more than other types expect the occurrence of their own orientation.

2.1.1 Methodological shortcomings in the social orientation literature

In most studies in the social orientation literature, the distinction between motivation and behavior is unclear. The triangle hypothesis distinguishes the general 'types' of cooperators and competitors and beliefs about others' 'types'. However, these types are assessed by the behavior of actors and their beliefs about the behaviors of others. For example, the triangle hypothesis is tested using the PD game, which combines motives and behaviors. Inferring social motives and beliefs about the social motives of others from behaviors in such interdependent games is problematic. In interdependent situations, observed behavior is an outcome of the actor's social orientation, her beliefs about the social orientation of her interaction partner and her beliefs about her partner's behavior (e.g., see: Aksoy & Weesie, 2013b). For example, in the

PD, an actor may not cooperate even if she has a cooperative orientation if she expects that the interaction partner is a competitor.

This objection is less stringent for the studies that use DGs instead of PDs. However, the measurement of social orientations with DGs can still be improved methodologically. First, an unguided descriptive methodology is used, that is, subjects are classified in distinct categories as cooperative, individualist, or competitive depending on their choice behavior in DGs.[2] For example, those who choose the cooperative option in 60% of the DGs are placed in the cooperative orientation category irrespective of the remaining 40% of their choices. Thus, this qualitative categorization ignores variation in a subject's choices in different DGs once she is classified in one of the social orientation categories. Second, it is theoretically proposed that there is an underlying continuous distribution of social orientations, for example, from competition to cooperation (Kelley & Stahelski, 1970; Kuhlman & Wimberley, 1976). Thus, these types are particular intervals on a continuum in which cut points for the social orientation categories are rather arbitrary. If one chooses different cut points, different results will be obtained. Likewise, subjects' beliefs about others' orientations are assessed categorically and qualitatively.

An additional drawback is the following. Typically, the relationship between one's social orientation and one's beliefs about others' social orientations is restricted to the degree of consensus between one's own and expected orientations. Iedema & Poppe (1999, 1994a); Iedema (1993) test the triangle hypothesis by comparing cooperators' and competitors' perceived consensus of their social orientation. In addition, Iedema (1993, Chapters 2 and 4) and Iedema & Poppe (1994a) ask subjects only to state what choice they expect predominantly from others. This is problematic because one may expect others to be, on average, the same as one's self in terms of social orientation but may still expect much variation in the population. Alternatively, one may expect others to be uniformly the same as the self without any variation. These two different beliefs, however, cannot be differentiated by this method. Studying

[2]One exception is the Ring method (Liebrand & McClintock, 1988), which measures social orientations as angles. In practice, however, estimates from the Ring method are used to place subjects in distinct groups (Liebrand & McClintock, 1988).

only the predominantly expected orientation or the perceived consensus is not sufficient to describe beliefs about others' social orientations nor is it sufficient to understand the relationship between one's own and expected social orientations. Thus, as an alternative, we propose to differentiate and investigate simultaneously measures of the central tendency (the mean) and dispersion (the variance) of expected social orientations.

In summary, in this study, we empirically investigate the association between actors' social orientations and their beliefs about the orientations of others. We contribute to the social orientation research by overcoming some of its drawbacks using a formal parametric approach. We estimate social orientations more precisely by obtaining direct social orientation parameters per subject using choice data on DGs, whereas past social orientation research relies on a relatively simple categorization of competitors, individualists, cooperators etc. Instead of describing beliefs about others' social orientations qualitatively, we estimate the means and variances of the distribution of expected social orientations of others conditional on one's own social orientation. Using this parametric approach, we test the triangle hypothesis, the SASB, and the cone model. This approach enables us to test more refined hypotheses about the specific shape of the association between one's own social orientations and the mean and variance of expected social orientations, which was not possible with previously available methods.

We further assert that our results are relevant for the experimental economics and rational choice literatures. Recently, those disciplines shifted from the classical assumption of self-interested individuals to the recognition that individuals are interested in the welfare of their interaction partners (Fehr & Schmidt, 2006; Gaechter & Herrmann, 2009; Chaudhuri, 2011). Moreover, those fields have also acknowledged heterogeneity in terms of people's non-selfish preferences, that is, some actors are selfish and other actors are non-selfish to varying extents (Simpson, 2004; Fehr & Schmidt, 2006; Fischbacher & Gaechter, 2010). Deviations from the standard selfishness assumption and the acknowledgment of heterogeneity in non-selfish motives, however, open up a Pandora's Box of new complications for formal theory. When heterogeneous, non-selfish motives are introduced in formal models, it is implausible

to assume that actors' social motives are visible/known to everyone (see: Orbell & Dawes, 1993). This is a problem because, in interdependent situations, an actor's behavior is usually dependent on her beliefs about the social motives of other(s) (and, in some cases, even higher order beliefs, e.g., beliefs about others' beliefs about the social motives of the self). Without modeling these beliefs explicitly, behavioral predictions are incomplete. Despite its importance, the experimental economics and rational choice literatures rarely address this issue. In these disciplines, beliefs about others' motives are dealt with by either making overly simplified assumptions such as common knowledge of social motives (Chaudhuri, 2011) or by applying the Bayesian-Nash equilibrium concept. In Bayesian-Nash equilibrium, actors are assumed to know the actual non-selfish preference distribution, and this distribution is assumed to be independent of one's own type. (e.g., Fehr & Schmidt, 1999). This application of Bayesian-Nash equilibrium is not empirically grounded because it ignores the complex relationships between one's social motives and one's beliefs about others' social motives, as captured by the triangle, SASB and cone hypotheses. In this paper, we describe the relationship between preferences and expected preferences empirically to provide an empirical basis for future formal modeling.[3]

2.2 Theory and hypotheses

2.2.1 Social orientations

We start with a widely applied one-parameter social orientation utility model, according to which an actor assigns a weight θ to the other's outcome (McClintock, 1972; Griesinger & Livingston, 1977). The utility of an actor i with

[3]In the experimental economics literature, there are studies that elicit beliefs (e.g., Croson, 2007; Neugebauer et al., 2009; Gaechter & Herrmann, 2009; Fischbacher & Gaechter, 2010) to analyze the association between *behavior* and *expected behavior*, that is, between the level of personal contribution to the public good and the expected average contribution from others. In contrast, the triangle, SASB and cone hypotheses yield refined predictions about the relationship between *preferences* (measured as social orientations) and *expected preference* distributions.

θ_i for an outcome allocation x (for the self) and y (for the other) is defined as:[4]

$$U_i(x, y) \equiv U_i^*(x, y; \theta_i) = x + \theta_i y. \tag{2.1}$$

The parameter θ_i, the weight that actor i attaches to the outcome of the other, is called i's social orientation. If $\theta_i > 0$, i gains utility from the outcome of the other, she is called a cooperator. If $\theta_i \approx 0$, i is an individualist because she is interested mainly in the outcome for the self. If $\theta_i < 0$, i is a competitor because she derives negative utility from the outcome of the other.

The model in (2.1) is deterministic and does not allow any incorrect predictions of behavior. To use the social orientation utility function for statistical analyses, we also introduce a stochastic component. Resembling McFadden's discrete choice approach (McFadden, 1974), or equivalently Thurstone (1927), we introduce an additive disturbance ϵ into the utility specification with further abuse of notation:

$$\begin{aligned} U_i(x, y) &\equiv U_i^*(x, y; \theta_i) + \epsilon \\ &= x + \theta_i y + \epsilon \qquad \epsilon \sim N(0, \tau^2). \end{aligned} \tag{2.2}$$

ϵ captures random utility for evaluation error. If people make choices by comparing the random utilities of alternatives, their choices are probabilistic. As we model it, the probability of choosing the wrong option in a DG decreases as the utility difference between options increases. As in most standard regression models, we assume that error is non-systematic: ϵ is normally distributed with zero mean and variance $\tau^2 > 0$. Here, we assume that ϵ does not depend on subjects, θ_i or outcomes in the decision situations.[5]

[4] An alternative representation is: $U(x, y; w_i) = w_{i1}x + w_{i2}y$. This alternative representation is equivalent to (2.1) with $\theta_i = \frac{w_{i2}}{w_{i1}}$ provided $w_{i1} > 0$, that is, actors prefer more for themselves rather than less. This is because utility is generally seen as cardinal, that is, defined only up to increasing affine transformations. We have checked our data, and for none of the subjects $w_{i1} < 0$.

[5] There are alternative ways to model decision error (see: Palfrey & Prisbrey, 1997). For example, a random term can be added in the decision weight assigned to the other's outcome representing random tastes: $U^b(x, y; \theta) = x + (\theta + \epsilon)y$, with $\epsilon \sim \mathrm{N}(0, \tau^2)$. In our

A potential problem with the model in (2.2) is that research has shown that the social orientations of a minority of subjects cannot be captured by (2.2) (Schulz & May, 1989). In particular, some respondents consider the absolute difference between the outcomes for self and others, as in the equality or maximin orientations (Schulz & May, 1989; Grzelak et al., 1977). To account for these orientations, Schulz & May (1989) extend the model in (2.2) by adding the weighted absolute difference between the outcomes for the self and other. To avoid biasing our results by ignoring subjects whose preferences are not captured by (2.2), we also provide analyses for the extended two-parameter model:

$$\begin{aligned} U_i(x,y) &\equiv U_i^*(x,y;\theta_i,\beta_i)+\epsilon \\ &= x+\theta_i y-\beta_i|x-y|+\epsilon \qquad \epsilon\sim N(0,\tau^2). \end{aligned} \tag{2.3}$$

The model in (2.3) contains competitive, individualist, cooperative, equality, and maximin orientations as special cases (Schulz & May, 1989). The competitive, individualist, and cooperative orientations are captured with the variation in the θ parameter in case $\beta \approx 0$. Those with $\theta \approx 0$ and large β are of an equality orientation. Those with $\beta = -\theta \to 1$ are of maximin orientation. We do not expect many large negative βs, controlled for θ. Subjects of the competitive orientation are captured by a large negative θ. For convenience in statistical analyses, we assume that (θ, β) has a bivariate normal distribution in the subject pool. Instead of classifying subjects in categorical groups, we define a subject i's social orientation by her θ_i and β_i scores.

Equality concerns have also received attention from economists and sociologists. The inequality aversion models of Fehr & Schmidt (1999) and Bolton & Ockenfels (2000) are extremely popular in experimental economics. Ignoring the random utility term, ϵ, (3) can be written as $U_i^*(x,y;\theta_i,\beta_i) =$

datasets, this alternative representation yielded a worse fit than (2.2) and is not used. One may also model decision error as a fixed probability of choosing the wrong option due to a tremble in response, such as pushing the wrong button. In this alternative representation, the probability of making a decision error does not depend on the utility difference between alternatives. Palfrey & Prisbrey (1997) suggests that this representation of decision error predicts data worse than the random utility representation.

$\begin{cases} (1-\beta)x + (\theta+\beta)y & \text{if } x \geq y \\ (1+\beta)x + (\theta-\beta)y & \text{if } x < y \end{cases}$, which is mathematically equivalent to the inequality aversion model of Fehr & Schmidt (1999) and a special case of Charness & Rabin (2002) and Tutic & Liebe (2009). Thus, the model in (2.3) is a fairly general model used extensively in various disciplines.

2.2.2 Beliefs about others' social orientations

First, consider the one-parameter social orientation model in (2.2). In our model of beliefs, every actor has a subjective belief about the distribution of θ in the population. Compared with other actors, some actors believe that people are, in general, more cooperative. This can be captured as the belief that the *mean* of θ in the population is higher for those actors. Similarly, some actors believe that people vary considerably with respect to social orientations, whereas other actors believe that most people are of similar social orientations. This can be captured by stating that the belief about the *variance* of θ may vary among actors. The hypotheses discussed in the introduction yields predictions on how beliefs about the mean and variance of social orientations depend on one's orientation.

We rewrite this more formally, also extending to the two-parameter utility specification in (2.3). Let us define $\mu(\widetilde{\theta}_i)$ and $\sigma^2(\widetilde{\theta}_i)$ as actor i's beliefs about the mean and the variance of θ in the population, respectively. Similarly, $\mu(\widetilde{\beta}_i)$ and $\sigma^2(\widetilde{\beta}_i)$ are i's beliefs about the mean and the variance of β, respectively. We also define $\rho(\widetilde{\theta}_i, \widetilde{\beta}_i)$ as i's belief about the correlation between θ and β. In line with the hypothesis discussed in the introduction, these beliefs depend on i's social orientation, namely θ_i and β_i. To formulate these hypotheses conveniently, we use the polynomial functions below.

$$
\begin{aligned}
\mu(\widetilde{\theta}_i) &= b_{a0} + b_{a1}\theta_i \\
\mu(\widetilde{\beta}_i) &= b_{b0} + b_{b1}\beta_i \\
\ln \sigma^2(\widetilde{\theta}_i) &= b_{sa0} + b_{sa1}\theta_i + b_{sa2}\theta_i^2 \\
\ln \sigma^2(\widetilde{\beta}_i) &= b_{sb0} + b_{sb1}\beta_i + b_{sb2}\beta_i^2 \\
\rho(\widetilde{\theta}_i, \widetilde{\beta}_i) &= b_0. \qquad (2.4)
\end{aligned}
$$

In (2.4), we use the logarithmic transformation for beliefs about the variance of the social orientation parameters because the variance can never be negative. For statistical convenience, we assume a particular *shape* for actors' subjective beliefs. A subject's beliefs about the distribution of θ and β in the population follow a bivariate normal distribution. As will be explained shortly, we estimate the $\widehat{b}$ parameters in (2.4).

2.2.3 Hypotheses

Defining beliefs as parameterized probability distributions allows us to formulate hypotheses about the relationship between one's social orientations and the parameters of expected social orientations more clearly and precisely.

We begin with the one-parameter social orientation model in (2.2). The *triangle hypothesis* postulates that cooperative people, that is, people with a larger θ will expect more variation in the population. This hypothesis can be formulated mathematically as a positive (linear) relationship between θ and $\ln \sigma^2(\widetilde{\theta})$. In addition, according to the triangle hypothesis, competitors (people with negative θs) expect most others to be competitive whereas cooperators expect some people to be competitive and some people to be cooperative. This yields the prediction that compared with competitors, cooperators expect others to be, on average, more cooperative, thus yielding a positive (linear) relationship between θ and $\mu(\widetilde{\theta})$. Taking these two predictions together, the relationship between beliefs about others' orientations and the actors' orientations can be depicted as in Figure 2.1a (H1a). In Figure 2.1, the grey-shaded regions represent the relationship between the actors' orientations and beliefs

about the variance in others' orientations (uncertainty). The thick lines within the grey regions represent the relationship between the actors' orientations and beliefs about the mean of others' social orientations.

The *SASB* predicts that subjects of all social orientations will believe that others are, on average, of the same social orientation as the self. This is, again, a positive (linear) relationship between θ and $\mu(\widetilde{\theta})$. However, in contrast to the triangle hypothesis, the SASB does not provide clear predictions about the relationship between θ and $\sigma(\widetilde{\theta})$. The SASB mainly claims that the false consensus effect is expected to be stable across different social orientation types. This can be interpreted as the prediction that uncertainty about the social orientations of others will be evenly distributed across social orientation types. See Iedema & Poppe (1999, p. 1444) for such an interpretation of the SASB. The predictions of the SASB is given in H2a in Figure 2.1a.

The *cone model* of Iedema (1993) predicts yet another pattern. The cone model extends the false consensus effect, indicating that the degree of false consensus will be highest for individualists. This means that people expect others to be similar to the self but the expected variance will be lowest for individualists, that is, those with $\theta \approx 0$. The expected variance will increase as θ deviates from individualism ($\theta \approx 0$). Translated to our terminology, this is a positive (linear) relationship between θ and $\mu(\widetilde{\theta})$ and a u-shape relationship between θ and $\ln \sigma^2(\widetilde{\theta})$ (see H3a in Figure 2.1a).

Deriving predictions from the triangle hypothesis on the equality dimension, i.e., the β parameter in (2.3), is less straightforward and should be treated as exploratory. The triangle hypothesis was originally proposed for the competition-cooperation dimension, thus for θ. Equality and maximin orientations captured by the β parameter are ignored by the original triangle formulation. Iedema (1993) also notes this difficulty. One solution is simply to not test the triangle hypothesis by ignoring the results for the β parameter. Another solution is to interpret large β values as reflecting more cooperative orientations (Iedema, 1993). This yields the prediction that those with large β will also expect more variation in the β parameter (H1b in Figure 2.1b).[6]

[6]Interpreting larger β values as more cooperative orientations is far from trivial. Aksoy & Weesie (2013b) suggest that the probability of cooperation in the PD increases in β but

The SASB yields the same predictions for the β parameter as for the θ parameter. This is because the SASB predicts the level of false consensus to be invariant across all social orientation categories (H2b in Figure 2.1b).

The cone model can also easily be extended to the β parameter. It predicts that the level of false consensus will decrease and that the expected variance in social orientations will increase as social orientations move away from individualism. Iedema (1993) shows that, as compared to individualists, consensus expectations are lower for equality and maximin orientations. The only difference for the β parameter compared with the θ parameter is that we do not expect many large negative βs. People do not *like* inequality *per se* on top of the competitive orientation captured by the θ parameter. We also do not observe large negative βs in our subject pool. This yields the truncated cone shape for the β parameter (H3b in Figure 2.1b).[7]

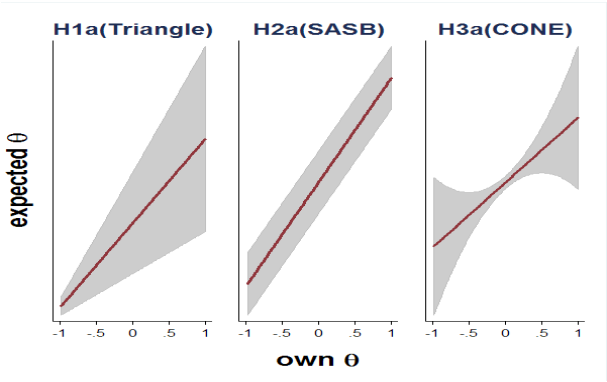
We want to stress an additional important point. In our hypotheses we do *not* claim that the distribution of expected orientations is *caused* by one's own orientations or vice versa. Each of the three explanatory models that we consider in the paper (triangle, SASB, and cone) involves several mechanisms. In some of these mechanisms, actors' orientations influence beliefs; in others, beliefs influence actors' orientations; in still other cases, beliefs and orientations co-influence each other. However, these three explanatory models predict three distinct *shapes* of the relationship between actors' social orientations and uncertainty, that is, their beliefs about the variance in others' orientations. We test these predictions on the shape of the relationship between actors' social orientations and beliefs.

only under *some* conditions, such as in games where payoff asymmetry is low.

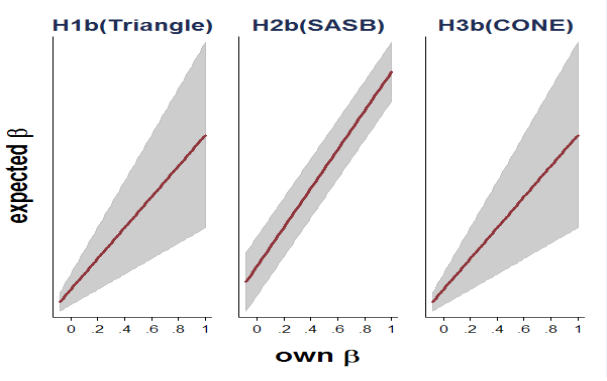
[7] A strictly truncated version of H3a would be curved. H3a is curved because it predicts a non-monotonous, i.e, first decreasing then increasing relationship. H3b predicts a monotonous, i.e., increasing relationship. Thus, we simplify the cone effect with a linear relationship for the β parameter.

Figure 2.1: Hypotheses. (a) The grey regions represent the relationship between uncertainty ($\sigma^2(\widetilde{\theta})$) and θ. The thick lines within the grey regions represent the relationship between $\mu(\widetilde{\theta})$ and θ. θ is the weight that an actor attaches to the outcome of others. $\theta < 0$ captures competitive, $\theta \approx 0$ captures individualist, and $\theta > 0$ captures cooperative orientations. (b) The same representation as in (a) but for the β parameter. β is the weight that a subject attaches to the absolute difference between the outcomes for self and other.

(a) Hypotheses on θ



(b) Hypotheses on β



2.3 Method

2.3.1 Subjects

A total of 155 subjects were recruited using the Online Recruitment System for Economic Experiments (ORSEE; Greiner (2004)). These students are from various study fields including economics (23%), psychology (7%) and numerous other fields. Subjects received an average of 16 euro for participating in the experiment, depending on their choices throughout the experiment.

2.3.2 Procedure

Subjects participated in one of eight sessions in the fall 2009. They were seated randomly in one of the cubicles in the Experimental Lab for Sociology and Economics (ELSE) at Utrecht University so that they could not see each other or the experimenter. Before social orientations and beliefs about others' social orientations were measured, the subjects played eight incentivized PDs, always with a different partner, but they did not receive feedback about others' decisions in these games or about how much they earned during the games. After the eight PDs were played without feedback, the social orientations and beliefs about others' social orientations were measured as explained below. The eight PDs were included because Kelley & Stahelski (1970) explicitly relates the triangle effect to the PD framework. Kelley & Stahelski (1970) differentiates only two types, competitors and cooperators, and the two options in the PD correspond to these two types. Playing PDs before the measurement of social orientations and beliefs about other's social orientations makes our design similar to that of Kelley & Stahelski (1970) and invokes the competitor-cooperator decision frame. However, we measure social orientations based not on PD response but on DGs to avoid the shortcomings of using the PD response to measure social orientations.

Subjects recorded their choices in an experiment booklet. Subjects' decisions involved other *real* subjects in the Alter role. The matching of subjects with other subjects and the calculation of their final earnings is done after all subjects completed the experiment. Thus, payment took place within two weeks after the last session. Subjects were given the choice to receive the pay-

ment on their bank accounts or to collect the payment in person using a code slip. Subjects who choose the first payment option, provided us with their bank account numbers and names in a closed and sealed envelope, given separately from the experiment booklet which ensured anonymity. After matching subjects' decisions with other subjects using a unique subject identifier number printed on experiment booklets, overall earnings of the subjects were calculated. Subjects were fully informed on this matching and payment procedure during the experiment. We received no indication that the procedure was unclear or that subjects distrusted that the payments would really be made.

Measurement of social orientations: Subjects stated their preferences in 18 DGs with two options. Table 2.1 includes an example of these 18 DGs. Appendix A.1 includes the full list of the DGs used in the experiment. The order in which subjects receive these DGs was randomized individually.

Measurement of beliefs about others' social orientations: Besides stating their own preferences in the DGs, each subject stated the percentages of other participants in the experiment who they thought would prefer option A over option B in each of these DGs. The accuracy of their beliefs was rewarded with 500 points for a perfect hit, that is, they received 500 points when the guessed percentage was the same as the actual percentage. Subjects earned 20 points less for each percentage point deviation from the actual percentage. If the guess was off by more than 25%, the subject earned no points, and no money.

2.4 Results

2.4.1 Social orientations

We explain the statistical analysis using the one-parameter social orientation model in (2.2). Except for estimating additional parameters, the procedure is the same for the extended two-parameter social orientation model in (2.3).

The outcomes in the two options of a given DG can be transformed into utilities using the function in (2.2). The strength of the preference for option

A over option B for subject i in a DG is the utility difference between the two options, that is:

$$\begin{aligned} U_{AB}(\theta_i) &= \{U^*(x_A, y_A; \theta_i) + \epsilon_A\} - \{U^*(x_B, y_B; \theta_i) + \epsilon_B\} \\ &= (x_A - x_B) + \theta_i(y_A - y_B) + (\epsilon_A - \epsilon_B) \quad \epsilon_A - \epsilon_B \sim N(0, 2\tau^2). \end{aligned} \tag{2.5}$$

where x_A is the outcome for the self in option A and y_B is the outcome for the other in option B in a DG. A subject is expected to prefer option A over option B in a DG if $U_{AB}(\theta_i) > 0$. Treating θ as a normally distributed variate in the subject population implies a multilevel probit model with a random coefficient θ. The 18 DG choices are nested in subjects. The (latent) dependent variable is a subject's preferences between options A and B, $U_{AB}(\theta)$, and the independent variables are the differences between the outcomes for the self and for the other in options A and B in each of the DGs as shown in (2.5). We estimate the parameters of this multilevel probit model with maximum likelihood using the Stata software GLLAMM (Rabe-Hesketh et al., 2002).[8] In addition to the mean and variance of θ, the variance of $\epsilon_A - \epsilon_B$ is estimated directly. In line with (2.5), the coefficient of $x_A - x_B$ is constrained to be 1 in the regression model. After fitting this probit multilevel model, we obtain the posterior means (empirical Bayes predictions) of θ per subject (Rabe-Hesketh et al., 2002). These posterior means are individual θs estimated conditionally on the overall model.

The results of this multilevel probit model for the one-parameter social orientation model are shown in Figure 2.2a (Model 1). Both the mean and the variance of θ are statistically significantly different from zero. The mean and standard deviation of θ are estimated as 0.11 and 0.16 ($\sqrt{0.026}$), respectively. The range of θ estimates is from -0.4 to 0.45.[9]

For the extended two-parameter social orientation model, the multilevel

[8]The Stata scripts used to analyse social orientations and beliefs about others' social orientations are available with our data upon request.

[9]The number of decisions is 2788 instead of 2790 (18×155) in Models 1 and 2 because a subject failed to respond in two DGs.

probit model additionally includes the mean and variance of β. Because θ and β may be correlated, we also include the covariance between θ and β. The results are shown in Figure 2.2a (Model 2). Compared with Model 1, Model 2 fits the data better: the log-likelihood is significantly smaller ($\chi^2(3)$=185.10, p<0.01). Related to this, the variance of the error term ($\epsilon_A - \epsilon_B$) is also smaller in Model 2. Both the mean and the variance of β are statistically significantly different from zero. The mean and the standard deviation of β are estimated as 0.05 and 0.13, respectively. As we expected, there are not many large negative βs. Individual β estimates are between -0.12 and 0.46. Figure 2.2b includes the kernel density plots of individual β and θ estimates and the scatter plot of the β and θ obtained from Model 2. Another important observation is that compared with the results of Model 1, the θ estimates are only slightly different. In Model 2, the mean of θ is slightly lower and the variance of θ is somewhat larger. Now, the range of θ estimates is from -0.55 to 0.52. Most importantly, individual θs estimated using Model 1 and Model 2 are almost perfectly correlated (ρ=0.99). Introducing the equality dimension hardly changes the θ estimates. There is also a low negative correlation between θ and β ($\rho = -.34$).

2.4.2 Beliefs about others' social orientations

The method of our analysis of beliefs is the following. Each subject stated her guess about the percentage of other subjects choosing option A in each of the 18 DGs. We begin with the one-parameter social orientation model. A subject's guess about the percentage of others who choose option A in a given DG can be represented as the subject's belief about the probability that the utility difference (2.5) is positive.[10] The probability that (2.5) is positive depends on the subject's belief about the mean and the variance of θ in the population. In addition, the subject's belief about the mean and the variance of θ depend on the subject's own θ as given in (2.4). Assuming a certain shape for the subjective belief distribution of θ, which, in this case, is the normal distribution, the b parameters in (2.4) can be estimated using

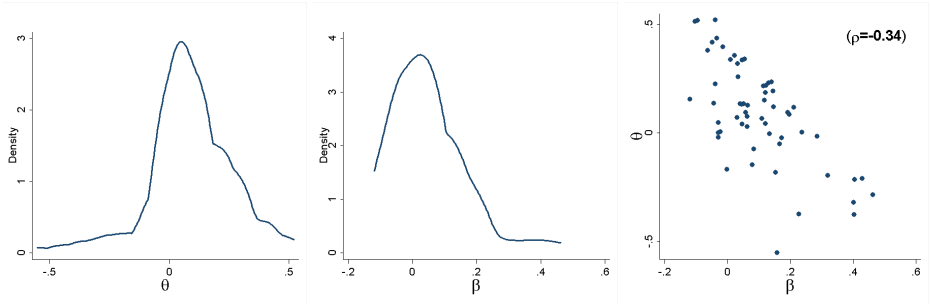
[10] In addition, in our method of analyzing beliefs, subjects also make unsystematic errors in stating their guesses. We assume that the expected value of this error is zero. We do not make further constraining assumptions about the shape of the distribution of this error.

Figure 2.2: (a) Results of the multilevel probit regression with random coefficients for the one-parameter (Model 1) and for the extended two-parameter (Model 2) social orientation models. In (a), the p-values for the means are computed using the Wald test; for the variances of the random effect and decision error, p-values are derived from the boundary tests using the mixture distribution $\overline{\chi}^2(01)$ of $\chi^2(1)$ and $\chi^2(0)$ (see Self & Liang, 1987). (b) Kernel density plots of empirical Bayes estimates of θs and βs and the scatter plot of the θs and βs obtained from Model 2.

(a)

	Model 1		Model 2	
Parameter	Coef.	Std. Err.	Coef.	Std. Err.
Random effect				
mean(θ_i)	.112**	.015	.089**	.018
var(θ_i)	.026**	.005	.039**	.007
mean(β_i)			.052**	.013
var(β_i)			.017**	.004
cov(θ_i,β_i)			-.009**	.004
var($\epsilon_A - \epsilon_B$)	.063**	.005	.035**	.003
log-likelihood	-1088.565		-996.017	

(b)



a subject's guess as the dependent variable. The independent variable is a nonlinear combination of a subject's θ score and outcomes in the DG. The same method is extended to the two-parameter social orientation model. In the two-parameter case, a subject's subjective belief is defined by a bivariate normal distribution for θ and β. The means and variances of the belief distribution depend on a subject's own θ and β as given in (2.4). We leave the details of the estimation method and the computations to Appendix A.2.

Figure 2.3a presents our results for the one-parameter social orientation model (Model 3). The estimated function for $\mu(\widetilde{\theta})$ is $\mu(\widetilde{\theta}) = -0.071 + 1.138\theta$. $\ln\sigma^2(\widetilde{\theta})$ varies non-linearly with θ, and the quadratic term is positive and highly significant. The estimated function is $\ln\sigma^2(\widetilde{\theta}) = -1.94 - 1.66\theta + 16.12\theta^2$. The minimum value of $\ln\sigma^2(\widetilde{\theta})$ is attained for $\theta = 0.04$, supporting the cone model.

Figure 2.3a also presents our results for the two-parameter social orientation model where the equality dimension is added (Model 4). Model 4 fits the data better than Model 3 because R^2 increased about 3% points. Now, the estimated function for $\mu(\widetilde{\theta})$ is $\mu(\widetilde{\theta}) = -0.076 + 1.528\theta$. The estimated function for $\ln\sigma^2(\widetilde{\theta})$ is $\ln\sigma^2(\widetilde{\theta}) = -2.593 + 1.164\theta + 16.049\theta^2$. Again, the quadratic term is highly statistically significant. The minimum value of $\ln\sigma^2(\widetilde{\theta})$ is observed when $\theta = -0.04$. In short, the results for θ are qualitatively the same as the results in Model 3. The left panel of Figure 2.3b depicts the relationship between beliefs about others' θ and θ in a graphical form. The results strongly support the cone model. The estimated function for $\mu(\widetilde{\beta})$ is $\mu(\widetilde{\beta}) = 0.052 + 1.455\beta$. The estimated function for $\ln\sigma^2(\widetilde{\beta})$ is $\ln\sigma^2(\widetilde{\beta}) = -3.430 - 20.448\beta - 24.104\beta^2$. The linear term is highly significant whereas the quadratic term is not. The right panel of Figure 2.3b shows the relationship between beliefs about others' β and β in a graphical form. The results for the β parameter are in line with both the triangle hypothesis and the cone model.[11]

To summarize, there is a strong positive association between an actor's social orientation and her beliefs about the *mean* of others' social orientations. In other words, on average, the more cooperative an actor is, the more she

[11] Although not reported in Figure 2.3b, the belief about the correlation between θ and β, $\rho(\theta_i, \beta_i)$ is estimated as -.47 (z=4.64, p<0.01).

expects others to be cooperative. Similarly, on average, the more an actor dislikes inequality, the more she expects others to dislike inequality. This finding is in line with the triangle hypothesis, the structural assumed similarity bias (SASB), and the cone model. Thus, when the belief about the mean of orientations is considered, all of these hypotheses are supported. When the relationship between an actor's social orientation and her beliefs about the *variance* of the social orientations of others is considered, the results strongly support the cone model. Individualists, that is, those with $\theta \approx 0$ and $\beta \approx 0$ expect less variation in the population, and expected variance increases as θ and β deviate from zero.

2.5 Additional analyses and discussion

We now discuss specific issues and report some additional analyses. The first discussion is on the construct validity of our measure of social orientation. Our subjects played 8 one-shot PDs without feedback before the DGs. The question is if θ and β estimates of subjects predict their PD responses. The answer to this question is non-trivial. Behavior in a PD depends not only on an actor's social orientations but also on other factors, such as her beliefs about others' social orientations. In addition, social orientations are preferences for outcome distributions between self and others. Thus, an actor's social orientation interacts with outcomes in the PD (Aksoy & Weesie, 2013b). We leave a detailed analysis of the PD behavior and social orientations to a future paper. As preliminary results, however, we report some tests. We regress the 8 binary PD decisions of subjects (1 = cooperation, 0 = defection) on θ and β estimates using a multilevel probit regression model with a random intercept for subjects. Both θ and β estimates significantly predict cooperation in a PD. Controlling for β, an increase of 0.1 in θ increases the odds of cooperating by a factor of 2.2 (b=7.88, Wald $\chi^2(1)$=50.05, p<0.01). Controlling for θ, an increase of 0.1 in β increases the odds of cooperating by a factor of 2.1 (b=7.50, Wald $\chi^2(1)$=18.91, p<0.01). These preliminary results are affirmative in demonstrating the construct validity of our social orientation measures.

We also demonstrate the effects of categorization of social orientations. Using the one-parameter social orientation model, we categorize subjects as cooperative if $\theta > 0.1$, and as competitive otherwise, i.e., a mean-split. The results following this categorization supports the triangle hypothesis. Cooperators ($\sigma^2(\widetilde{\theta}) = 0.277$) expect more variation with respect to social orientations than competitors ($\sigma^2(\widetilde{\theta}) = 0.135$), and the difference is statistically significant with a one-sided test. By increasing the number of categories, it is possible to capture the cone shape. For example, when we classify subjects into three groups using arbitrary cut points as competitors ($\theta < 0$), individualists ($0 \leq \theta \leq 0.1$), and cooperators ($\theta > 0.1$), we observe the cone pattern. Individualists are significantly less uncertain ($\sigma^2(\widetilde{\theta})$ is lower) than competitors or cooperators. However, if the arbitrary cut points are too low or too high, the cone effect disappears. The advantage of the continuous parametric approach that we adopt in this paper is that it does not rely on arbitrary cut points.

Our analyses so far ignore differences in beliefs among actors with the same social orientation. In other words, under the one-parameter social orientation, there is no other source of variation in $\mu(\widetilde{\theta})$ and $\sigma^2(\widetilde{\theta})$ besides θ. To investigate possible effects of ignoring subject-level heterogeneity on our results, we estimate $\mu(\widetilde{\theta})$ and $\sigma^2(\widetilde{\theta})$ separately for each subject using the 18 observations per subject. In this manner, we ensure that $\mu(\widetilde{\theta})$ and $\sigma^2(\widetilde{\theta})$ are estimated independently per subject without restricting subject-level heterogeneity at the cost of losing statistical power. Then, we regress $\mu(\widetilde{\theta})$ and $\sigma^2(\widetilde{\theta})$ on the actor's θ. The results of these regressions are qualitatively the same as those of Model 3 reported in Figure 2.3a. $\mu(\widetilde{\theta})$ is linearly related to θ, and $\sigma^2(\widetilde{\theta})$ is quadratically related to θ, with a u-shape.

The two-parameter social orientation model in (2.3) captures the social orientations of most subjects (Schulz & May, 1989). However, it is always possible to improve the fit of the social orientation model by including additional parameters. For example, a utility model that includes weights for y^2, $|x-y|^2$ and $y \times |x-y|$ fits the choice data better. There is a trade-off between better fit on the one hand and interpretability of coefficients and parsimony on the other hand. We fitted this complex model with these additional quadratic and interaction terms. The individual θ and β estimates obtained from this

more complex model correlates almost perfectly ($\rho > 0.99$) for both θ and β with those obtained using the two-parameter social orientation model in (2.3). Thus, there is little added value to fitting such complex models.

The random utility component we include in the two social orientation models, ϵ, is assumed to be unsystematic and therefore independent of other variables. In reality, this component might depend on some variables. For example, there might be individual heterogeneity in ϵ, or the variance of ϵ might decrease as subjects make more decisions due to learning. Dealing with individual heterogeneity in ϵ is computationally very demanding and not feasible with a relatively small dataset. However, the method that we use can easily be extended to make the variance of the error dependent on certain variables, such as the number of previous DG decisions in the experiment. We fit an extended version of Model 2 by making the variance of ϵ depend on the number of previous DG decisions in the experiment. Indeed, we find a small and statistically significant effect: the variance of the error decreases by approximately 2% per period. Individual θs and βs estimated while controlling for this period effect correlates almost perfectly ($\rho > .99$) for both θ and β parameters with the estimates obtained from Model 2. Thus, we conjecture that ignoring factors that might influence decision error does not affect our results substantially.

Finally, the method used to analyze beliefs about other's social orientations is a two-stage procedure. In the first stage, we estimate the subjects' θ and β. In the second stage, we use those estimates to predict the subjects' beliefs about the means and variances of others' θ and β. In this second stage, the measurement error in individual θ and β scores is ignored. This means that the standard errors of the parameters in the second stage, in Model 3 and Model 4, are underestimated. In principle, this problem could be solved by calculating the standard errors using bootstrapping. However, this calculation requires extensive computing time because fitting a single multilevel probit model for actors' choices and a nonlinear regression model for beliefs is already computationally very demanding. Thus, we refrain from using the bootstrapping method. Moreover, because the rejected null hypotheses have very low p-values in our sample, we believe that this issue does not substan-

tially affect our results.

2.6 Summary and conclusions

In this paper, we investigate the relationship between actors' social orientations and their beliefs about the social orientations of others by testing hypotheses proposed in the social orientation literature. In contrast to previous research, we employ a formal parametric approach. First, actors' social orientation parameters are estimated using a new method. With the new method, rather than classifying subjects into distinct social orientation categories, social orientations are modeled as a continuous distribution. Second, beliefs about the mean and variance of the social orientations of others, conditional on one's own social orientation, are estimated, again with a newly introduced method. We analyze a one-parameter social orientation model and a two-parameter social orientation model that includes an equality dimension.

The formal parametric approach allows us to test more refined hypotheses about the specific shape of the association between actors' social orientations and their beliefs about the mean and variance of others' social orientations. We find a strong, positive association between an actor's social orientation and her belief about the mean of others' social orientations. This finding is in line with the three hypotheses tested in this paper: the triangle hypothesis, the SASB hypothesis, and the cone model. In addition, we find that individuals with social orientation parameter values of approximately zero - applicable to respondents who have the individualist social orientation - expect less variation in others' social orientations than those with smaller or larger social orientation parameter values. This finding supports the cone model.

Because previous studies rely on a relatively simple categorization of competitors, individualists, cooperators, egalitarians, etc., such elaborate tests are not possible with the methods used in past research. In past research, hypotheses were usually tested via the degree of perceived consensus between one's orientations and one's beliefs about others' orientations. For example, Iedema (1993, Chapters 2 and 4) and Iedema & Poppe (1994a) ask subjects to state what an average person would choose in a given Decomposed Game.

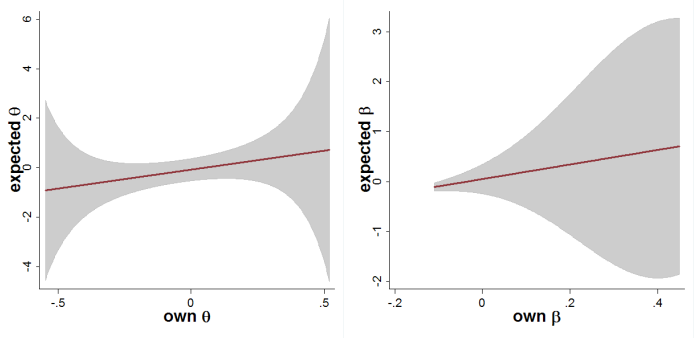
The answer provided reflects only the respondent's belief about the mean of social orientations, and asking only about one aspect of beliefs, i.e., the mean, obscures variation in respondent's beliefs. One may expect others to be similar to one's self on average, but still expect variation in other's beliefs. Alternatively, one may expect others to be similar to the self without any variation. These two different beliefs cannot be distinguished if only the degree of perceived consensus is considered. We also check the effect of the categorization of social orientations used in past research on our results. When we classify our subjects using arbitrary cut points to identify competitors and cooperators, the results misleadingly support the triangle hypothesis. Uncertainty (beliefs about the variance of others' social orientations) is lower for competitors when this method is used. By introducing additional categories to the competitor-cooperator dichotomy, such as the third category of individualists, one can capture the cone effect that we find. However, if one chooses higher or lower cut points in their categorization, the cone effect disappears. We conclude that simplifying the underlying continuous distribution of social orientations into distinct categories can yield misleading results. Our method of measuring one's social orientation and beliefs about others' social orientations does not rely on such categorization.

Figure 2.3: (a) the fitted models on the relationship between actors' beliefs about the mean/variance of social orientations and actors' social orientations for the one-parameter (Model 3) and two-parameter (Model 4) social orientation models. (b) Relationship between actors' social orientations and beliefs about other's social orientations in a graphical form, obtained from Model 4. Grey-shaded regions represent the relationship between beliefs about the variance (uncertainty) of others' social orientations and one's own social orientations: the boundaries of the areas are $\mu(\widetilde{\theta}) \pm 1.65\sigma(\widetilde{\theta})$ for the left panel and $\mu(\widetilde{\beta}) \pm 1.65\sigma(\widetilde{\beta})$ for the right panel. The thick lines in the grey regions show the relationship between beliefs about the mean of social orientations and actors' social orientations.

(a)

	Model 3		Model 4			
Var	$\mu(\widetilde{\theta})$	$\ln\sigma^2(\widetilde{\theta})$	$\mu(\widetilde{\theta})$	$\ln\sigma^2(\widetilde{\theta})$	$\mu(\widetilde{\beta})$	$\ln\sigma^2(\widetilde{\beta})$
	Coef. (S.E.)	Coef. (S.E.)	Coef. (S.E.)	Coef. (S.E.)	Coef. (S.E.)	Coef. (S.E.)
Cons	-.071**	-1.938**	-.076**	-2.594**	.052**	-3.430**
	(.020)	(0.216)	(.016)	(.322)	(.011)	(.715)
θ	1.138**	-1.657	1.529**	1.164		
	(.137)	(1.82)	(.149)	(.883)		
θ^2		16.117**		16.049**		
		(4.456)		(2.674)		
β					1.455**	20.448**
					(.163)	(7.142)
β^2						-24.104
						(14.044)
R^2	0.705		0.728			

(b)



Chapter 3

Hierarchical Bayesian analysis of biased beliefs and distributional other-regarding preferences*

Abstract

This study investigates the relationship between an actor's beliefs about others' other-regarding (social) preferences and her own other-regarding preferences, using an *"avant-garde"* hierarchical Bayesian method. We estimate two distributional other-regarding preference parameters, α and β, of actors using incentivized choice data in binary Dictator Games. Simultaneously, we estimate the distribution of actors' beliefs about others' α and β, conditional on actors' own α and β, with incentivized belief elicitation. We demonstrate the benefits of the Bayesian method compared to it's hierarchical frequentist counterparts. Results show a positive association between an actor's own (α, β) and

*This chapter is written in collaboration with Jeroen Weesie and published in *Games* (Aksoy & Weesie, 2013a). We thank Werner Raub for his extensive comments on earlier drafts and Vincent Buskens and Rense Corten for their input in conducting the experiment. We also thank Axel Ockenfels, Siegfried Berninghaus, and the participants of the June 2010 Maastricht Behavioral and Experimental Economics Symposium and of the September 2010 IAREP/SABE/ICABEEP conference.

her beliefs about *average* (α, β) in the population. The association between own preferences and the *variance* in beliefs about others' preferences in the population, however, is curvilinear for α and insignificant for β. These results are partially consistent with the cone effect Iedema (1993), (Aksoy & Weesie, 2012b) which is described in detail below. Because in the Bayesian-Nash equilibrium concept, beliefs and own preferences are assumed to be independent, these results cast doubt on the application of the Bayesian-Nash equilibrium concept to experimental data.

3.1 Introduction

Experimental evidence shows that utility models incorporating other- regarding preferences often explain choice data better than the classical economic model with selfish actors. Consequently, numerous social preference models have been proposed and many such model have received great attention in the literature. (For example, according to our most recent search of Google Scholar on November 21, 2012, Fehr & Schmidt (1999) inequality aversion model is cited 5299 times, approaching the 7454 citations of Adam Smith's "The Theory of Moral Sentiments"). It is also acknowledged that there is heterogeneity in preferences, i.e., some actors are selfish, whereas some actors have other-regarding preferences to varying degrees (e.g., Fehr & Schmidt, 1999; Kohler, 2012). This heterogeneity is often invoked as an important factor to explain why, under some experimental conditions, results seem to converge to the classical economic model's predictions, while under other conditions, results deviate significantly from the classical model.

Introducing other-regarding preferences in micro-economic models and acknowledging heterogeneity in preferences, however, creates the need to also model actors' beliefs about others' preferences. In many cases (without dominant equilibria), predictions of micro-economic models depend strongly on actors' beliefs about others' behavior. Beliefs about others' behavior, in turn, depend on actors' beliefs about others' preferences. Thus, modeling actors' own preferences is necessary but not sufficient to obtain behavioral predictions. In addition to actors' own preferences, actors' beliefs about others' preferences

should also be dealt with. Otherwise, empirical tests of utility models are incomplete because the observed behavior could be the result of not only own preferences, but also beliefs about others' preferences. In the behavioral economics literature uncertainty about others' utilities, thus beliefs about others' preferences are typically dealt with via the application of the Bayesian-Nash equilibrium (Harsanyi, 1968) with "rational beliefs" (e.g., Fehr & Schmidt, 1999). Beliefs are assumed to be independent of own preferences, and actors are assumed to know the actual distribution of preferences in the "population". While making sense theoretically, this concept may not be empirically solid. Social psychology studies often demonstrate clear biases in expectations, as summarized by the "false" consensus hypothesis (Ross et al., 1977), the triangle hypothesis (Kelley & Stahelski, 1970), or the Cone model of Iedema (1993) indicating that actors' own social preferences and expectations about others' social preferences may not be independent. See also Blanco et al. (2009, 2011) for a critic of Bayesian-Nash equilibrium within the experimental economics context. We investigate actors' beliefs about others' preferences empirically. Particularly, we focus on the relationship between own preferences and beliefs about others' preferences.

There is a rich literature on belief formation and learning (recent examples include Danz et al., 2012; Hyndman et al., 2012) which are not among the topics of the current paper. Comparable existing studies on beliefs in the micro-economics literature focus mainly on the relationship between the actor's *behavior* and the actor's beliefs about others' *behavior* (e.g., Dufwenberg & Gneezy, 2000; Croson, 2000; Gaechter & Renner, 2010; Ellingsen et al., 2010), or between the actor's *preferences* and the actor's beliefs about others' *behavior* (e.g., Bellemare et al., 2008; Offerman et al., 1996; Blanco et al., 2009). In our view, a micro theory, such as a social utility model, should be the theoretical basis of the analysis of both the actor's own behavior and the actor's beliefs about others' behavior. If we are to explain actors' own behavior with a utility model, we should also explain actors' beliefs with the same utility model. Thus, the distinctive feature of our study is that we explicitly analyze the relationship between actors' own *preferences* and their beliefs about others' *preferences*, nested in the same utility model. We esti-

mate two other-regarding preference parameters for a variant of the Charness & Rabin (2002), an extension of Fehr & Schmidt (1999) utility specification, using choice data in binary Dictator Games. Simultaneously, we estimate the moments of the distribution of actors' beliefs about others' other-regarding preferences, conditional on own other-regarding preferences, with incentivized belief elicitation. To be clear, analyzing the relationship between preferences and beliefs about others' preferences, we deviate from rational beliefs—yet rational beliefs remain a special case of our model where the aforementioned relationship is absent—. However, we do assume that people's preferences are described by the utility model we use, *and* people think that other people's preferences are given by the same utility model.

Because the main focus of this study is on the relationship between beliefs and preferences, we restrict our attention to a single utility model, namely a variant of Charness & Rabin (2002), which is an extension of the inequality aversion model of Fehr & Schmidt (1999). Although we provide an assessment of empirical fit of the model, the primary aim of this paper is *not* testing the utility model used in this paper or finding the utility model that best explains our data. This has been done extensively in the literature. We are aware of other relevant social motives and other successful models of other-regarding preferences (e.g., Bolton & Ockenfels, 2000; Engelmann & Strobel, 2004; Falk & Fischbacher, 2006; Rodriguez-Lara & Moreno-Garrido, 2012). Yet, because the model of Charness & Rabin (2002), and of Fehr & Schmidt (1999) which is contained in Charness and Rabin are among the most cited and applied social utility models, describing the relationship between actors' own motives and beliefs within this framework is useful. In addition, in our study, we use data on simple Dictator Games. In such games, preferences given by the model used here are in line with other potentially relevant types of motives, e.g., maximin preferences and inequality aversion. Finally, variants of other-regarding preferences given by Charness & Rabin (2002) are a common theme in other disciplines such as social psychology (e.g., Schulz & May, 1989), and rational choice sociology (e.g., Aksoy & Weesie, 2013b; Tutic & Liebe, 2009). Thus, our results would be relevant for a variety of disciplines. In addition, the simple shape of the function we use yields substantial convenience in statistical

analyses.

An important contribution of this paper, however, is it's statistical methodology. We use a hierarchical Bayesian method to estimate the other-regarding preferences and the moments of the distribution of actors' beliefs about others' other-regarding preferences. Hierarchical Bayesian methods have some practical advantages over their hierarchical frequentist counterparts (Gelman et al., 2004). First and foremost, incorporating strong numerical estimation procedures, Bayesian methods are very flexible. They could be relatively easily applied to many complex statistical models. Relying on the maximum likelihood approach, however, frequentist methods are limited in this sense. For example, the simultaneous analysis of own other-regarding preferences and beliefs that we perform in this paper is nearly impossible within the frequentist framework. In addition, the Bayesian approach provides a strong and flexible tool to assess model fit via posterior predictive sampling (Gelman et al., 1996), which we exploit in the current paper. Assessing model fit for relatively complex models as ours within the frequentist framework is, again, very difficult. In addition to those practical advantages of Bayesian methods, given that one uses fairly uninformative priors the results of the Bayesian methods converge to "would be" maximum likelihood estimates (Snijders & Bosker, 2012; Gelman et al., 2004). In other words, (some of) the Bayesian results obtained using uninformative priors may be interpreted as frequentist estimates. Our statistical analysis is implemented in OpenBugs (Lunn et al., 2009). We provide the estimation routine as a supplementary file in Appendix B.3 so that other scholars could replicate, modify, and apply the routine.

The main results of the hierarchical Bayesian analysis are the following:

- *Result 1*: There is a strong positive relationship between own other-regarding preferences and beliefs about *average* other-regarding preferences.

- *Result 2*: There is a U-shaped association between own other-regarding preferences and beliefs about the *variance* in others' other-regarding preferences for one of the two parameters in the utility model, namely α. For the other β parameter, the same relationship is insignificant.

- *Result 3*: The utility model that we use and the model for beliefs that we develop below explain the choices and beliefs of subjects in binary Dictator Games adequately.

We want to elaborate on how this current paper improves on a previous study of the authors (Aksoy & Weesie, 2012b). In the current paper we use a richer dataset, combining data from two experiments. Thus statistical power is improved. Second, Aksoy & Weesie (2012b) employs a *different* other-regarding preference model, namely the social orientation model, a model predominantly used in social psychological literature. In the social orientation model in Aksoy & Weesie (2012b), actors are interested in absolute inequality as they are assumed to not differentiate between advantageous and disadvantageous forms of inequality. Despite using a different utility model in this paper, we still find some support for the cone effect. This further reinforces the existence of the cone pattern, at least for some type of other-regarding preferences. Thirdly, and most importantly, Aksoy & Weesie (2012b) uses a two step estimation method that predicts individual social preference parameters in the first step, and in the second step uses these individual estimates to model beliefs. As we acknowledge in Aksoy & Weesie (2012b), this two step estimation procedure is flawed as in the second step it does not take into account measurement error in the first step. In this paper, the Bayesian estimation routine solves this problem as simultaneous estimation of preferences and beliefs becomes possible.

The organization of the paper is as follows. After describing the experimental procedure, we describe the model for other-regarding preferences and beliefs in detail. We then move on to the details of the hierarchical Bayesian analysis where we briefly compare the Bayesian method with a frequentist alternative. A final discussion of the main results concludes the paper.

3.2 Method

3.2.1 Subjects

The data come from two independent experiments. The only differences between the experiments are: (1) the first experiment reported in (Aksoy & Weesie, 2012b) was conducted in January 2010; the second experiment in May 2010, and (2) the second experiment was embedded in a larger set of experiments which included additional treatments. In this paper, we use a subset of data from the second experiment. This subset is collected using the identical procedure as in the first experiment. We performed careful statistical analyses to compare the data from the two experiments. These analyses—available from the authors—showed that when the parameters of the statistical models in this paper are considered, both for other-regarding preferences and beliefs, the hypothesis that the two samples come from the same population could not be rejected. Thus, we collapsed the data of the two experiments, and report below the results for this combined dataset. This yielded a sample of 187 subjects (155 from the 1st and 32 from the 2nd experiment) recruited using the Online Recruitment System for Economic Experiments (ORSEE) (Greiner, 2004). The two experiments comprised 10 sessions in total, each of which was run with 16 to 20 subjects who were seated randomly in one of the cubicles in the Experimental Laboratory for Sociology and Economics (ELSE) of Utrecht University. Since the experiment was conducted in the English language a good command of English was a prerequisite for participation.

3.2.2 Procedure

We followed the standard procedures of experimental economics, e.g., anonymity, real and anonymous partners, incentive compatibility, etc. (Friedman & Cassar, 2004). Social preferences were measured with 18 different binary Dictator Games. Subjects were informed that for each Dictator Game, the recipient was a randomly selected participant. The outcomes in these 18 Dictator Games were chosen to (approximately) optimize the precision of statistical estimates of the other-regarding preference parameters, building on the previous evidence on the empirical distribution of such parameters. In 16 of

the games, one option included an equal distribution, and the other option included inequity of varying degrees (see Appendix B.1 for the parameters of these 18 Dictator Games and the rationale behind choosing these particular games). To measure beliefs, subjects were asked to guess the percentage of other participants taking part in the experiment who would choose each option in each of these 18 Dictator Games. Subjects earned points from their own decisions and were rewarded for the accuracy of their guesses regarding others' behavior. More precisely, for each of the 18 Dictator Games, if a subject's guess of how many other subjects would choose each option was equal to the actual percentage, she earned 500 points. For each percentage point of deviation from the actual percentage, the subject earned 20 fewer points. If the guess was off by more than 25%, the subject earned no points.[1] In addition, next to their decisions as "senders" in the Dictator Games, as a subsequent step, subjects also passively earned points as "recipients" of other randomly selected participants. No feedback about the accuracy of the guesses or about the choices of others was given until all 18 games were complete. The order in which a subject received these 18 games was randomized.

Here we want to discuss three methodological issues that may concern the reader. First of all, we paid for both game outcomes and beliefs. One may argue that subjects might hedge by stating beliefs to insure against bad outcomes in the game. We do not think this is a major concern as Blanco et al. (2008) study this problem explicitly and show clearly that hedging in designs as ours is not a major problem. Secondly, we pay for all games instead of paying for a randomly selected Dictator Game. Our data analysis assumes that subjects treat each Dictator Game as a one-shot game independent from

[1] In our analyses, we assume that subjects report their *average* beliefs, which we think is a natural response for most subjects. Theoretically, incentivising beliefs with a quadratic loss function would ensure reporting average beliefs (Costa-Gomes & Weizsäcker, 2008; Rey-Biel, 2009; Palfrey & Wang, 2009). We opted for a "spline-like" loss fuction rather than a quadratic loss function to make the incentive structure more accessible to the subjects. We do not think this is a major issue as the exact incentive function does not seem to influence the distribution of beliefs to a great extent. For example, Gaechter & Renner (2010) compares the distributions of incentivized and non-incentivized beliefs and reports some but not substantial differences. Palfrey & Wang (2009) report that compared to a linear loss function, a quadratic loss function improves the accuracy of beliefs to some extent. But even a quadratic loss function is used, biases in beliefs remain (also see Rey-Biel, 2009).

other games, rather than all games as one big game. Sophisticated readers might claim that if subjects are also sophisticated enough they may consider distributional outcomes not game by game but across the set of 18 games (see Oechssler, Forthcoming). To stress the (assumed) one-shot nature of each decision, in the experiment the recipient was a randomly selected new participant in each Dictator Game, i.e., stranger matching. Alternatively we could pay for a randomly selected game, but that procedure is not free of problems either (e.g., Stahl & Haruvy, 2006). Finally, in the experiment a subject decided as the sender in the Dictator Game, but also was a recipient of another person. This was done not to reduce the sample size by half. This feature of the design does not make the game strictly one-person and may potentially have some consequences. Yet, we believe that this does not influence our results in a substantial way. First of all, the recipient and the sender for whom the subject was a recipient were both randomly selected subjects, i.e., they were not necessarily the same person. Moreover, our parameter estimates are quite similar to those found in other studies and as we discuss below the statistical model fits data quite well. An otherwise finding would make us suspect such potential design effects. We will revisit these methodological issues in the discussion.

3.3 Theoretical model: other-regarding preferences and beliefs

3.3.1 Other-regarding preferences

We use the following utility function for distributional other-regarding preferences. For an outcome allocation for the self (x) and the other (y), the (random) utility for actor i is:

$$\begin{aligned} U^r(x, y; \alpha_i, \beta_i) &= U(x, y; \alpha_i, \beta_i) + \epsilon \\ &= x - \alpha_i \max(0, y - x) - \beta_i \max(0, x - y) + \epsilon, \\ &\quad \epsilon \sim N(0, \tau^2). \end{aligned} \tag{3.1}$$

Following Charness & Rabin (2002), for subjects with $0 < \beta < \alpha$, this function reduces to the inequity aversion model of Fehr & Schmidt (1999), where α and β capture the utility losses due to disadvantageous and advantageous inequity, respectively. Subjects with $\beta < 0$ and $\alpha > 0$ are competitive, that is, they prefer to receive higher outcomes than the other. Subjects with $\beta > 0$ and $\alpha < 0$ are motivated by so called "social welfare", that is, their utility increases in the outcomes of the other, irrespective of whether the other earns more or less than self, yet with different weights. Finally, those with $\beta = \alpha = 0$ are selfish. To be able to use this utility function for statistical purposes, we included a stochastic term ϵ (McFadden, 1974) in the function, which is assumed to be normally distributed with zero mean, independent across subjects, evaluations, and alternatives in the games. Then, the probability that actor i chooses option A over option B in a binary Dictator Game j will be,

$$\begin{aligned}
\Pr(U^r_{ijA} > U^r_{ijB}) &= \\
\Pr\left(U^r_{jAB}(\alpha_i, \beta_i) > 0\right) &= \Pr\left(U(x_{jA}, y_{jA}; \alpha_i, \beta_i) - U(x_{jB}, y_{jB}; \alpha_i, \beta_i) + \Delta_{ij\epsilon} > 0\right) \\
&= \Pr\left(\Delta_{jx} - \alpha_i \cdot \Delta_{jyx} - \beta_i \cdot \Delta_{jxy} + \Delta_{ij\epsilon} > 0\right) \\
&= \Phi\left(\frac{\Delta_{jx} - \alpha_i \cdot \Delta_{jyx} - \beta_i \cdot \Delta_{jxy}}{\sqrt{2}\tau}\right). \qquad (3.2)
\end{aligned}$$

where x_{jA} is the outcome for the self in option A, y_{jB} is the outcome for the other in option B, $\Delta_{jyx} = \max(0, y_{jA} - x_{jA}) - \max(0, y_{jB} - x_{jB})$, $\Delta_{jxy} = \max(0, x_{jA} - y_{jA}) - \max(0, x_{jB} - y_{jB})$, $\Delta_{ij\epsilon} = \epsilon_{jA} - \epsilon_{jB} \sim N(0, 2\tau^2)$ for subject i in game j, and Φ is the cumulative standard normal distribution. Note that this probability is equivalent to the Quantal Response Equilibrium prediction with normal distributed evaluation errors (McKelvey & Palfrey, 1995).

3.3.2 Beliefs

For each of the 18 Dictator Games, subjects guessed the percentage of other participants who preferred option A over option B. Let p_{ij} be subject i's stated belief about the percentage of others choosing option A in Dictator Game j. We define π_{ij} as i's belief about the probability that (3.2) holds.

We assume that actor i's beliefs about others' α and β, $(\widetilde{\alpha}_i, \widetilde{\beta}_i)$, are rep-

resented by a multivariate normal distribution $\begin{pmatrix} \widetilde{\alpha}_i \\ \widetilde{\beta}_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{\widetilde{\alpha}_i}(\alpha_i) \\ \mu_{\widetilde{\beta}_i}(\beta_i) \end{pmatrix}, \begin{pmatrix} \sigma^2_{\widetilde{\alpha}_i}(\alpha_i) & \rho_{\widetilde{\alpha}_i,\widetilde{\beta}_i} \\ \rho_{\widetilde{\alpha}_i,\widetilde{\beta}_i} & \sigma^2_{\widetilde{\beta}_i}(\beta_i) \end{pmatrix}\right)$, where $\mu_{\widetilde{\alpha}_i}(\alpha_i)$ and $\mu_{\widetilde{\beta}_i}(\beta_i)$ are i's beliefs about the means of α and β, $\sigma^2_{\widetilde{\alpha}_i}(\alpha_i)$ and $\sigma^2_{\widetilde{\beta}_i}(\beta_i)$ are i's beliefs about the variance of α and β, and $\rho_{\widetilde{\alpha}_i,\widetilde{\beta}_i}$ is i's belief about the correlation between α and β.[2] Then π_{ij} satisfies

$$\begin{aligned} \pi_{ij} &= \Pr\left(U^r_{AB}(\widetilde{\alpha},\widetilde{\beta}) > 0 \,\middle|\, \begin{pmatrix} \widetilde{\alpha} \\ \widetilde{\beta} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{\widetilde{\alpha}_i}(\alpha_i) \\ \mu_{\widetilde{\beta}_i}(\beta_i) \end{pmatrix}, \begin{pmatrix} \sigma^2_{\widetilde{\alpha}_i}(\alpha_i) & \rho_{\widetilde{\alpha}_i,\widetilde{\beta}_i} \\ \rho_{\widetilde{\alpha}_i,\widetilde{\beta}_i} & \sigma^2_{\widetilde{\beta}_i}(\beta_i) \end{pmatrix}\right)\right) \\ \Phi^{-1}(\pi_{ij}) &= \frac{\Delta_{jx} - \mu_{\widetilde{\alpha}_i}(\alpha_i)\cdot\Delta_{jyx} - \mu_{\widetilde{\beta}_i}(\beta_i)\cdot\Delta_{jxy}}{\sqrt{2\widetilde{\tau}^2 + \sigma^2_{\widetilde{\alpha}_i}(\alpha_i)\cdot\Delta^2_{jyx} + \sigma^2_{\widetilde{\beta}_i}(\beta_i)\cdot\Delta^2_{jxy} + 2\rho_{\widetilde{\alpha}_i,\widetilde{\beta}_i}\sigma(\widetilde{\alpha}_i)\sigma(\widetilde{\beta}_i)\cdot\Delta_{jyx}\Delta_{jxy}}} \end{aligned} \tag{3.3}$$

where $2\widetilde{\tau}^2$ is the variance of $\widetilde{\epsilon_A} - \widetilde{\epsilon_B}$, the variance of the difference of random disturbances and Φ^{-1} is the inverse cumulative normal distribution. Ignoring boundary cases, we model $\Phi^{-1}(p_{ij}) = \Phi^{-1}(\pi_{ij}) + \varepsilon_{ij}$ and assume that ε_{ij} is normally distributed with zero mean, thus

$$\Phi^{-1}(p_{ij}) = \Phi^{-1}(\pi_{ij}) + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0,\tau_p^2) \tag{3.4}$$

Moreover, we model the moments of the subjective belief distribution given in (3.3) as polynomial functions of own social preference parameters. More precisely:

$$\begin{aligned} \mu_{\widetilde{\alpha}_i}(\alpha_i) &= b_{a0} + b_{a1}\alpha_i + \varepsilon_{\alpha i} \qquad \varepsilon_{\alpha i} \sim N(0,\tau^2_\alpha) \\ \mu_{\widetilde{\beta}_i}(\beta_i) &= b_{b0} + b_{b1}\beta_i + \varepsilon_{\beta i} \qquad \varepsilon_{\beta i} \sim N(0,\tau^2_\beta) \\ \ln\sigma^2_{\widetilde{\alpha}_i}(\alpha_i) &= b_{sa0} + b_{sa1}\alpha_i + b_{sa2}\alpha_i^2 \\ \ln\sigma^2_{\widetilde{\beta}_i}(\beta_i) &= b_{sb0} + b_{sb1}\beta_i + b_{sb2}\beta_i^2 \\ \rho_{\widetilde{\alpha}_i,\widetilde{\beta}_i} &= b_{c0}. \end{aligned} \tag{3.5}$$

[2] In our theoretical model for own preferences, we assume that (α,β) is multivariate normal. We ascertained that this normality assumption for own (α,β) is reasonable by estimating (α,β) with fixed effects, that is, without assuming normality. As own (α,β) is normal, we see no reason to assume a different distribution for beliefs.

In (3.5), $\mu_{\widetilde{\alpha}_i}(\alpha_i)$ and $\mu_{\widetilde{\beta}_i}(\beta_i)$ depend on α and β in a linear form, whereas $\ln \sigma^2_{\widetilde{\alpha}_i}(\alpha_i)$ and $\ln \sigma^2_{\widetilde{\beta}_i}(\beta_i)$ depend on α and β in a curvilinear form. We also performed analyses with higher order polynomial terms, but the functional forms above in (3.5) are sufficient to describe our main results. We will discuss this issue below in the results section.

Also note that there are stochastic error terms $\varepsilon_{\alpha i}$ and $\varepsilon_{\beta i}$ in the equation. For convenience, we assume that $\varepsilon_{\alpha i}$ and $\varepsilon_{\beta i}$ are independent. These error terms capture the possibility that two subjects who have the same values for (α, β) may have different beliefs about average (α, β) in the population.[3] It is in principle possible to add similar error terms for $\ln \sigma^2_{\widetilde{\alpha}_i}(\alpha_i)$ and $\ln \sigma^2_{\widetilde{\beta}_i}(\beta_i)$. We tried to do so. However adding so many random terms complicate analysis yielding estimation and convergence problems. Thus we included error terms only for $\mu_{\widetilde{\alpha}_i}(\alpha_i)$ and $\mu_{\widetilde{\beta}_i}(\beta_i)$.

3.4 Analyses and results

We will first discuss the analyses and the results for other-regarding preferences without introducing beliefs. In a subsequent step, we will provide the simultaneous analysis of other-regarding preferences and beliefs.

3.4.1 Bayesian and frequentist analysis of other-regarding preferences

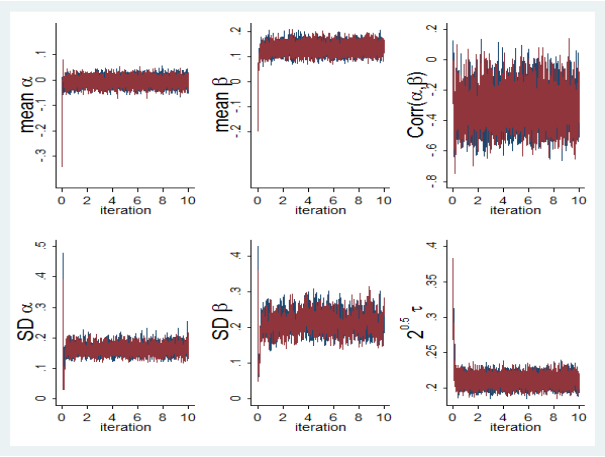
We assume multivariate normality for β and α in our subject pool. Consequently, (3.2) yields a multilevel probit model with random coefficients for α and β. The dependent variable is a subject's choice between options A and B, and the independent variables are the differences between the outcomes for the self and the two terms for outcome differences in options A and B as given in (3.2). Note that the coefficient of Δ_{jx} is 1. The parameters of (3.2) to be estimated are the means and (co)variances of (α_i, β_i) and the variance of evaluation error τ^2. Since (3.2) yields a fairly standard multilevel probit model with

[3] Besides these error terms, the means of the belief distribution depend only on (α_i, β_i). However, the model could be easily adapted to include other subject level covariates, such as age, gender, study field etc.

random coefficients, it can be fitted using the frequentist maximum likelihood approach as well as the Bayesian framework. To facilitate comparison and show indeed that the Bayesian approach with uninformative priors yield very similar results as the frequentists maximum likelihood approach, we present the results of both estimation procedures. When we introduce beliefs later on, however, estimation using the frequentist approach becomes infeasible, thus will provide only the Bayesian solution.

We fitted (3.2) with maximum likelihood using the Stata software GLLAMM (Rabe-Hesketh et al., 2002) (also see: Aksoy & Weesie, 2012b). The hierarchical Bayesian estimation requires a more elaborate description, firstly because scholars are less familiar with it, and secondly it involves three main elements each of which should be described clearly. Thus, we spare somewhat more space for the Bayesian procedure. The three main elements of a Bayesian analysis are the prior distributions of the parameters, the likelihood of data, and the MCMC method to obtain the posterior distribution of the parameters. We used the following standard uninformative priors for the parameters of (3.2). The prior for the covariance matrix of (α, β) is an inverse Wishart with 2 degrees of freedom with a scale matrix of $10^2 \cdot I_2$ where I_2 is the 2×2 identity matrix (Kunreuther et al., 2009). For the mean vector of (α, β), we used a multivariate normal prior with zero means and $10 \cdot I_2$ variances. For the variance of the evaluation error τ^2, we used a Gamma(10,10) prior. The likelihood of data is implied by (3.2). We used OpenBugs (Lunn et al., 2009), a freely available program, to obtain (draws from) the posterior distribution of the parameters. The estimated posterior distribution comprised 22.000 draws from two chains after a burn-in of 30.000 draws and recording every 10th draw in each chain (Nilsson et al., 2011). Convergence was confirmed by visual inspection of the history of the draws from the two chains as well as using the Gelman & Rubin (1992) $\widehat{R}$ statistics. If the model has converged, the chains started from different values should mix and the posterior distribution should stabilize. This can be seen from Figure 3.1, which includes the history of the draws for the parameters of (3.2) until the 10.000th iteration. In fact convergence is achieved as early as after a few hundred iterations. In addition, values of Gelman & Rubin (1992) $\widehat{R}$ statistics for the parameters after excluding the

Figure 3.1: History of draws from the posterior distribution of parameters per 10^3 iterations to assess convergence.



burn-in iterations are all smaller than 1.05, in fact all are virtually 1.

Table 3.1 includes the results of the frequentist and the Bayesian estimations, as well as the bivariate scatter plot of individual (α_i, β_i) estimates obtained as empirical posterior means. For the Bayesian procedure Table 3.1-a includes the posterior means (P.M.) as point estimates of the parameters and posterior standard deviations (P.SD.) as standard errors of those point estimates (Snijders & Bosker, 2012). The first thing to note is the extreme similarity of the maximum likelihood and Bayesian results. This is not surprising, given that we used uninformative priors. When uninformative priors are used, the results of Bayesian procedures converge to the frequentist results for relatively large samples (Gelman et al., 2004). Note that the Bayesian procedure yields an entire posterior distribution, not only posterior means or standard deviations. This is demonstrated in Figure 3.2, which presents the posterior density plots of the parameters.

Both the frequentist and Bayesian results show that the estimated mean of β is roughly .14. Also, there is significant variation among subjects with respect to β. The mean of α_i is estimated as -.01, which is statistically "insignificant", that is the estimate is not significantly different from zero (ML

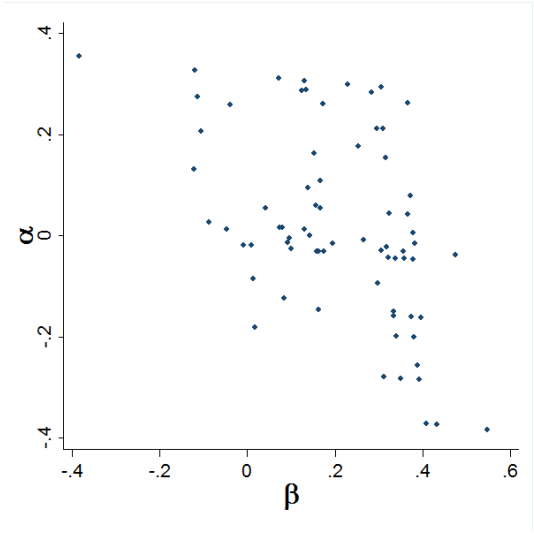
Table 3.1: (a) Hierarchical maximum likelihood (ML) and hierarchical Bayesian estimates and their standard errors for the means and standard deviations of the other-regarding preference parameters (α, β) and the standard deviation of the evaluation error (τ). (b) Bivariate scatter plot of empirical Bayes estimates of (α, β).

(a)

	ML		Bayesian	
parameter	Coef.	S.E.	P.M.	P.SD.
mean(β_i)	.140	.022	.134	.019
mean(α_i)	-.007	.016	-.010	.015
sd(β_i)	.208	.023	.211	.024
sd(α_i)	.163	.015	.166	.013
corr(α_i, β_i)	-.303	.010	-.304	.093
τ	.211	.013	.212	.007

N(decision) = 3366, N(subject) = 187

(b)

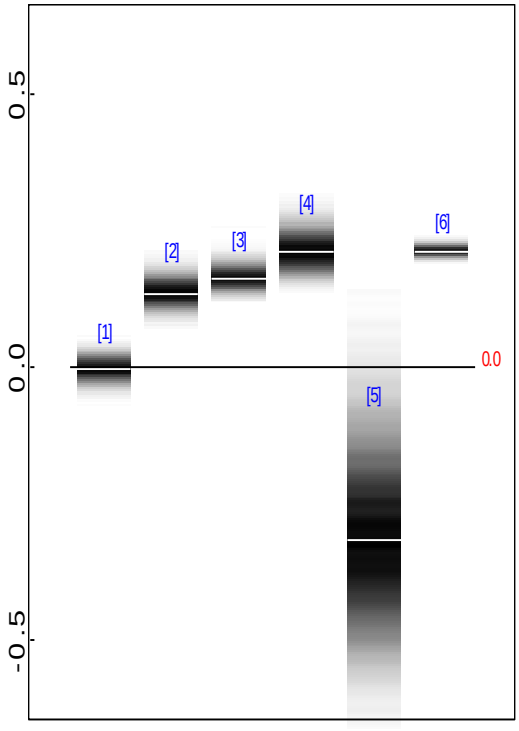


result), and the posterior density of the parameter is centered around zero (Bayesian result). However, albeit small, the estimated variance of α_i is statistically significant. Thus, although, on average, α is about zero, subjects differ significantly in terms of their α values. We also find a moderate negative correlation between α and β. Note, however, that the uncertainty/precision of this correlation parameter is somewhat high, reflected as a wider density strip (see Figure 3.2), and a larger posterior standard deviation and standard error (see Table 3.1).

The utility function that we use is an extension of Fehr and Schmidt's inequity aversion model, thus, we can assess to what extent the assumptions of Fehr and Schmidt hold. Our estimates show that the typical assumptions of Fehr & Schmidt on the distributions of α and β, e.g., $\alpha > 0$, $\beta > 0$, and $\alpha > \beta$, do not hold in the Dictator Games (see also Table 3.1-b). For example, a substantial portion of our sample has negative α estimates, and for quite some of our subjects $\alpha < \beta$. Although our results point to such deviations from the original assumptions of Fehr & Schmidt, they are in fact very close to the estimates obtained for the same utility function in a study that used a representative Dutch sample (Bellemare et al., 2008). Bellemare et al. (2008), for instance, also report negative α values for a large portion of their sample. Another recent study that estimates (α_i, β_i) with maximum likelihood in a Swiss sample also yields a very similar (α, β) distribution as ours where for many subjects $\beta > \alpha$ and for quite some $\alpha < 0$ (Morishima et al., 2012). Blanco et al. (2011) also reports that the assumptions of Fehr and Schmidt on the distribution of α, β are violated in a battery of games. This shows the usefulness of Charness and Rabin extension of the Fehr and Schmidt model.

As a side, we find no evidence of quadratic/nonlinear preferences; i.e., when added to the model, the coefficients of Δ^2_{xy} and Δ^2_{yx} were not significant. Additionally, remember that each subject received the 18 Dictator Games in random order. We checked whether the order that a subject received these games mattered and concluded that the game order did not influence the social preference parameters. The only, albeit small, order effect was that the variance of evaluation error, τ, decreased in the number of previous decisions. Yet, including this order effect on evaluation error in our models did not

Figure 3.2: Density strips of the posterior distribution of parameters: [1] = μ_α, [2] = μ_β, [3] = σ_α, [4] = σ_β, [5] = Corr(α, β), [6] = $\sqrt{2}\tau$.



change our results in any substantial way. We will discuss below the fit of the other-regarding preference model in detail.

3.4.2 Bayesian analysis of other-regarding preferences and beliefs

When the beliefs are entered in the model, statistical analysis becomes significantly difficult. First of all, now the analysis involves fitting four simultaneous equations, namely (3.2), (3.3), (3.4), and (3.5). Secondly, these four equations include highly nonlinear forms. If one wants to stay within the frequentist approach, options are very limited and those limited options are problematic, too (Aksoy & Weesie, 2012b). The first option is a two-step procedure: in the first step fitting (3.2) and obtaining individual (α, β) estimates as posterior

means (Rabe-Hesketh et al., 2002); in the second step fitting (3.3), (3.4), and (3.5) with the nonlinear least squares method, feeding in individual estimates of (α, β) obtained in the first stage as observed variables. This approach, however, would yield incorrect estimates, because in the second step the measurement error in (α, β) is ignored. Moreover, in this two stage approach it is impossible to include the two error terms in (3.5), namely $\varepsilon_{\alpha i}$ and $\varepsilon_{\beta i}$. In other words, individual variation in beliefs has to be ignored, yielding potentially important biases. A second frequentist alternative is turning to (multilevel) structural equation modeling (SEM), which is a natural tool to solve for simultaneous equations within the frequentist framework (Aksoy & Weesie, 2012c). Yet, it proves to be impossible to obtain convergence due to high number of unknowns and highly non-linear forms in equations (3.2), (3.3), (3.4), and (3.5). Thus, we do not discuss further the frequentist solutions for the simultaneous analysis of other-regarding preferences and beliefs.

The hierarchical Bayesian approach, on the other hand, is very flexible. The likelihood of data is given by (3.2), (3.3), (3.4), and (3.5). One still needs to assign priors to the unknowns in those equations and the strong tools incorporated in the Bayesian framework, i.e., MCMC and the Gibbs sampler take care of obtaining (draws from) the posterior distribution of the parameters. We now describe this procedure. As above, we used rather uninformative priors. The priors for the parameters of (3.2) are given in the previous subsection. The priors for the $b_\bullet$ parameters in (3.5) are univariate normals with zero mean and 10^2 variance, except b_{c0} to which a Uniform[-1,1] prior is assigned. For the variances of error terms, $\tau_\bullet$, in (3.5) and (3.4) we assigned Gamma(10,10) priors.[4] As above after 30.000 burn-in draws from two MCMC chains and recording every 10th draw in each chain, a final posterior distribution with 56.000 draws is obtained using OpenBugs. As explained above, convergence is ascertained by visual inspection of the history of the draws from the two chains as well as using the Gelman & Rubin (1992) $\widehat{R}$ statistics. Table 3.2 shows the posterior means (P.M.) and posterior standard deviations (P.SD.)

[4]We fix $2\widetilde{\tau}^2$ in (3.3) to zero, otherwise the MCMC procedure failed to converge. We performed a sensitivity analysis and observed that assigning different fixed values for $2\widetilde{\tau}^2$ hardly influenced the parameters of interest.

Table 3.2: Beliefs about others' other-regarding preferences: Points estimates (posterior means) of the parameters of equation (3.5). Posterior standard deviations are given in parentheses below as the standard errors of posterior means. N(decision) = 3366, N(subject) = 187.

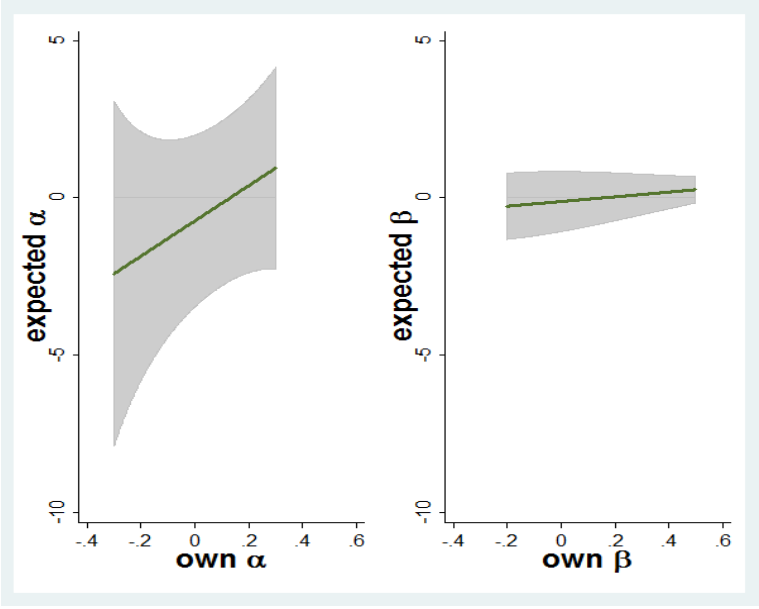
$\mu_{\widetilde{\alpha}_i}(\alpha_i) =$	$-.737$	$+5.596\alpha_i$	$+\varepsilon_{\alpha i} \sim N(0, .770^2)$
	(.105)	(.991)	(.107)
$\mu_{\widetilde{\beta}_i}(\beta_i) =$	$-.117$	$+ .747\beta_i$	$+\varepsilon_{\beta i} \sim N(0, .075^2)$
	(.024)	(.143)	(.018)
$\ln \sigma^2_{\widetilde{\alpha}_i}(\alpha_i) =$	1.020	$-1.788\alpha_i$	$+9.516\alpha_i^2$
	(.115)	(.732)	(4.203)
$\ln \sigma^2_{\widetilde{\beta}_i}(\beta_i) =$	-1.076	$-1.622\beta_i$	$-3.325\beta_i^2$
	(.103)	(2.406)	(7.103)
$\rho_{\widetilde{\alpha}_i, \widetilde{\beta}_i} =$	.650		
	(.106)		

of the parameters of equation 3.5.[5]

The strong positive relationship between own other-regarding preferences and expected *average* other-regarding preferences is apparent in Table 3.2 for both α and β. Those with higher other-regarding preferences parameters also believe that the average other-regarding preferences in the population is higher. The variances of individual errors in average other-regarding preferences, τ_α and τ_β, are estimated as .77 and .08, respectively with relatively low standard errors/high precision. Thus, although beliefs depend on own preferences, there is significant variance in beliefs that cannot be totally explained by own preferences. One finding to be noted is the slope of α_i on $\mu_{\widetilde{\alpha}_i}(\alpha_i)$, which is 5.6. Although this slope seems higher than what one would expect, in fact the variation in α_i is small. The difference between the two standard deviations below and above the average α_i, that is the difference between the

[5]Note again that the Bayesian results include not only P.M.s and P.SD.s but the entire posterior distributions of parameters. Thus, as in Figure 3.2, it is possible to obtain posterior density strips of all parameters which we omit for brevity.

Figure 3.3: Expected other-regarding preferences versus own other- regarding preferences based on the posterior means of the parameters in Table 3.2. Grey shaded areas represent the relationship between expected variance in others' other-regarding preferences and own other-regarding preferences; i.e., the boundaries of the grey areas are: $\mu(\widetilde{\alpha}) \pm 1.65\sigma(\widetilde{\alpha})$ and $\mu(\widetilde{\beta}) \pm 1.65\sigma(\widetilde{\beta})$. Note that $\pm 1.65\sigma(\widetilde{\beta})$ refer to 90% confidence intervals. The solid lines within these areas represent the relationship between average expected other-regarding preferences and own other-regarding preferences.



lowest and highest likely α_i values is only about 0.64. Thus, although the slope is high, the difference in predicted $\mu_{\widetilde{\alpha}_i}(\alpha_i)$ is not that high due to the small size and variance of α_i estimates. The U-shaped association between own other-regarding preferences and beliefs about the *variance* in others' other-regarding preferences for the α parameter is also apparent. The estimated polynomial function relating $\sigma^2(\widetilde{\alpha})$ and α is $\ln \sigma^2(\widetilde{\alpha}) = 1.02 - 1.79\alpha + 9.52\alpha^2$. The global minimum of this function is 0.09. In other words, those who have larger absolute values of α expect much more variation in the population than those with smaller absolute values of α. The relationship between $\sigma^2(\widetilde{\beta})$ and β, however, is statistically insignificant.

Figure 3.3 describes the nature of the relationship between other-regarding preferences and beliefs in a graphical form. Figure 3.3 is obtained using the posterior means displayed in Table 3.2 as point estimates of the parameters. This curvilinear relationship between own other-regarding preferences and expected variance in others' other-regarding preferences for the α parameter is depicted in Figure 3.3, with wider grey regions as absolute own other-regarding preferences increases. Although the variance in beliefs about β seem to decrease in own β, the relationship is highly insignificant. Note that the variance in beliefs is in general higher for the α parameter than for the β parameter.

To report, the estimate of $\rho_{\widetilde{\alpha}_i,\widetilde{\beta}_i}$, which is the belief regarding the correlation between α and β, is .65 with a posterior standard deviation of 0.11. τ_p is estimated as 1.19 with a posterior standard deviation of 0.02.

It is possible to improve the model in Table 3.2 by adding higher-order polynomial terms to equation 3.5. For example, a representation where $\mu(\widetilde{\alpha})$ is a third-order polynomial function of α yields a better fit. Yet, in this alternative representation, results still indicate a monotonic and increasing, though non-linear, relationship between $\mu(\widetilde{\alpha})$ and α. Similarly, a model where $\ln \sigma^2_{\widetilde{\alpha}_i}(\alpha_i)$ is a fourth-order polynomial function of α yields a better fit. But, again, the results of this model also indicate a U-shaped relationship between $\ln \sigma^2(\widetilde{\alpha})$ and α, with an almost identical local minimum. Thus, although it is possible to obtain better fitting, more complex polynomial representations, the model in Table (3.2) is sufficient to describe our main results. In the discussion below, we provide some insights on why this particular shape between other-

regarding preferences and beliefs about others' other-regarding preferences emerges.

3.4.3 Bayesian assessment of fit: posterior predictive checking

The question examined here is how well our model for other-regarding preferences and beliefs fits data. If the discrepancy between our model and data is too high, then the results discussed above might be misleading. Consider the following discrepancy statistics for the choice of subject i in game j:

$$D_{cij} = \frac{(y_{ij} - \Pr(U^r_{ijA} > U^r_{ijB}))^2}{(\Pr(U^r_{ijA} > U^r_{ijB}))(1 - \Pr(U^r_{ijA} > U^r_{ijB}))} \tag{3.6}$$

where $\Pr(U^r_{ijA} > U^r_{ijB})$ is the *predicted* probability of choosing option A in a game j for subject i (cited from equation 3.2), and y_{ij} is the *observed* choice for i in game j. Statistics (3.6) is chosen for its similarity to Pearson's χ^2 statistics, a widely used statistics to assess fit. We construct an overall fit statistic, $\bar{D}_{c++}$, by averaging over all i and j.

A similar discrepancy statistics can be constructed for beliefs as below:

$$D_{pij} = (\Phi^{-1}(p_{ij}) - \Phi^{-1}(\pi_{ij}))^2 \tag{3.7}$$

where $\Phi^{-1}(\pi_{ij})$ is the predicted inverse cumulative probability for subject i in game j that (3.2) holds and $\Phi^{-1}(p_{ij})$ is the *observed* inverse cumulative probability (see equations 3.3 and 3.4). As above, an overall fit statistic, $\bar{D}_{p++}$, is constructed by averaging over all i and j.

Had we known the theoretical distributions of $\bar{D}_{c++}$ and $\bar{D}_{p++}$ under the null hypothesis that our model fits data, we could compare the $\bar{D}_{c++}$ and $\bar{D}_{p++}$ scores in our sample to the theoretical distributions and calculate a p-value for each discrepancy score. These p-values then would show how likely it is to observe the calculated or higher $\bar{D}_{c++}$ and $\bar{D}_{p++}$ scores, given that our model fits. However, only for very few such discrepancy statistics the theoretical

distribution under the null is known and $\bar{D}_{c++}$ and $\bar{D}_{p++}$ are not among those few cases. Fortunately, the Bayesian framework offers an alternative, i.e., posterior predictive checking. Below we briefly describe this procedure and refer to Gelman et al. (1996) and Gelman & Hill (2007) for a fuller treatment of the topic.

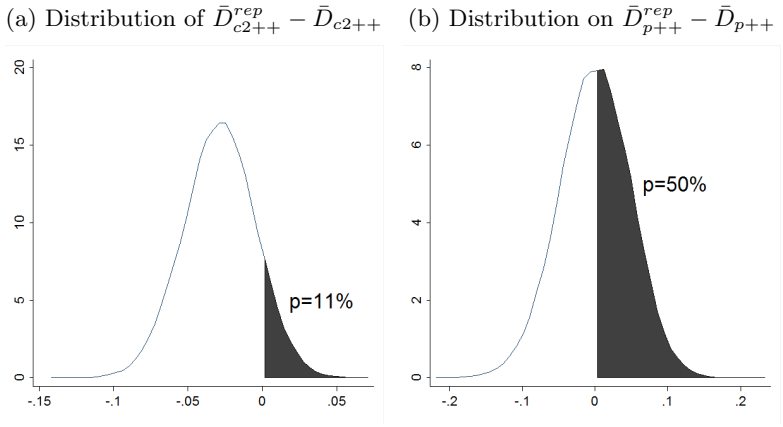
For each draw from the Markov Chain used to fit our model, we create a replicated dataset given our model. For each replicated dataset, we calculate the $\bar{D}_{c++}$ and $\bar{D}_{p++}$ scores, which we call $\bar{D}^{rep}_{c++}$ and $\bar{D}^{rep}_{p++}$, respectively. Subsequently, we obtain the discrepancy p-values for the parts of the model that explains the choices and beliefs of subjects by calculating $\Pr(\bar{D}^{rep}_{c++} > \bar{D}_{c++})$ and $\Pr(\bar{D}^{rep}_{p++} > \bar{D}_{p++})$.[6] This procedure is very general and could be used for any (discrepancy) statistics. In fact, we used a slightly different smoother discrepancy statistics for choices of subjects, to deal with the boundary cases using $D_{c2ij} = \frac{(y_{ij} - \Pr(U^r_{ijA} > U^r_{ijB}))^2}{(0.1+0.8\Pr(U^r_{ijA} > U^r_{ijB}))(0.9-0.8\Pr(U^r_{ijA} > U^r_{ijB}))}$. The exact form of the discrepancy statistics does not matter to calculate a p-value, since for any statistics posterior predictive samples could be created.

Figure 3.4 shows the distributions of $\bar{D}^{rep}_{c2++} - \bar{D}_{c2++}$ and $\bar{D}^{rep}_{c++} - \bar{D}_{c++}$, as well as $p_{c++} = \Pr(\bar{D}^{rep}_{c2++} > \bar{D}_{c2++})$ and $p_{p++} = \Pr(\bar{D}^{rep}_{c++} > \bar{D}_{c++})$. The discrepancy p-values are $p_{c++} = 11\%$ and $p_{p++} = 50\%$. Neither of these figures is low, which means that under the null hypothesis that the model fits, neither of the discrepancy statistics is extremely unlikely. In other words, we do not find much evidence against the hypothesis that the model fits, thus conclude that overall our model fits data relatively well.

It is also possible to obtain person fit statistics as $\bar{D}_{ci+}$, that is, for each subject averaging the discrepancy statistics over all games. Similarly, it is possible to obtain item fit statistics $\bar{D}_{c+j}$ by averaging the discrepancy statistics over all subjects for each Dictator Game (Fox, 2010). Subsequently, discrepancy p-values could be calculated for each subject and each Dictator Game, using posterior predictive sampling. This way, subjects whose choices and beliefs are not captured by the model or games in which choices and beliefs of subjects deviate from the predictions of the model can be detected. Since

[6] Note also that within the Bayesian framework the discrepancy statistics $\bar{D}_{c++}$ and $\bar{D}_{p++}$ are not single scores, but each has an entire posterior distribution.

Figure 3.4: Posterior predictive checking: Distributions of $\bar{D}^{rep}_{c2++} - \bar{D}_{c2++}$ and $\bar{D}^{rep}_{p++} - \bar{D}_{p++}$, as well as corresponding discrepancy p-values.



the overall fit of our model is rather satisfactory, we do not pursue assessing fit at lower levels. However, we would like to stress that this aspect of posterior predictive sampling on assessing individual level and item/game level discrepancies is potentially very useful to assess fit at different levels.

3.5 Discussion and conclusions

In this study, we investigate the relationship between an actor's beliefs about others' other-regarding preferences and her own other-regarding preferences using a hierarchical Bayesian method. We estimated the other-regarding preferences parameters, α and β, of actors using choice data from binary Dictator Games. Simultaneously, we estimated the distribution of actors' beliefs about others' α and β, conditional on own α and β, with incentivized belief elicitation. We demonstrated some advantages of the Bayesian method over its hierarchical frequentist counterparts including its flexibility in dealing with relatively complex models with many free parameters, as well as the possibility of a sound assessment of model fit even for complex models using posterior predictive sampling.

Besides describing the benefits of the hierarchical Bayesian method, the pa-

per presents interesting results. We found that there is a positive monotonic relationship between own other-regarding preferences and the belief about *average* other-regarding preferences. This is not an unprecedented finding (e.g., Blanco et al., 2009), although our study is perhaps novel in demonstrating this relationship in light of a specific preference model. What is probably more novel is that we also found a strong U-shaped association between own other-regarding preferences and *variance* in beliefs about others' other-regarding preferences for the α parameter. This result can also be interpreted in the following way. Selfish actors, i.e., those with approximately zero α, also expect most others to be selfish as well. Thus, these selfish actors fit the classical economic model in terms of both their preferences and their beliefs. As α deviate from selfishness, however, actors expect more variation in the population while still expecting others to be similar to themselves on average. The association between preferences and variance in beliefs for the β parameter however is insignificant. Another finding to be noted is that variance in beliefs is in general much higher for the α parameter than for the β parameter.

Why beliefs vary with own type and mechanisms yielding the particular relationships between own motives and beliefs that we describe in this paper is an open question. This question is tackled mainly within social psychology. We hope that our study will bring the social psychological literature on the relationship between types and beliefs about others' types to the attention of experimental economists. The positive relationship between own preferences and beliefs about preferences in the population is in line with what social psychologists call the "false" consensus effect (Ross et al., 1977).[7] The consensus effect literature typically does not investigate *variance* in beliefs (or uncertainty) about others' preferences. Consequently, our result on the relationship between own other-regarding preferences and expected variance of others' preferences is as novel in the social psychological literature as it is for economics. The triangle hypothesis (Kelley & Stahelski, 1970), structured as-

[7]Dawes (1989) and Engelmann & Strobel (2000) show that this effect is not necessarily *false*. That is, if available people use information on others' choices, and even assigning higher weights to others' choices than one's own choice. A truly *false* consensus effect would require ignoring information about others' choices. In our case subjects did not receive feedback about others' choices prior to belief elicitation.

sumed similarity bias (Kuhlman et al., 1992), and the cone model (Iedema, 1993) are three hypotheses proposed in the social psychological literature each of which indirectly proposes a certain relationship between own preferences and expected variance in others' preferences. We refer to Aksoy & Weesie (2012b) for a detailed description of those hypotheses. Among those three hypotheses, our results support partially the cone model. The cone effect, as discussed by Iedema (1993) is caused by several factors. We apply those factors to the context of this paper. First, in line with the consensus effect all types—actors with certain values of (α, β)—expect their own type to be more common in the population. Secondly, in addition to their own types, all types expects selfishness to be another common type in the population, because selfishness is a common stereotype about others. These two effects overlap when the expectations of selfish people are considered. As a result, expected variance is smaller for selfish people. We should note, however, that the cone effect is observed for the α parameter. For the β parameter variance in beliefs is stable, that is, it does not depend on own β. Without further research, we can only speculate why the cone pattern emerges for the α parameter but not for the β parameter. As the social psychological literature shows, biases in beliefs such as the cone or the consensus effects are stronger for situations where information is scarce and uncertainty is high (Ross et al., 1977). Our findings show that uncertainty about others' preferences is much higher for the α parameter than that for the β parameter as the variance in beliefs is smaller for β than for α. In line with this, also the relationship between own other-regarding preferences and the belief about *average* other-regarding preferences is much stronger α than for β. Thus, probably the cone and the consensus effects are stronger for the α parameter because subjects are more uncertain about others' α than others' β.

Irrespective of the exact causal mechanisms, these clear associations between own other-regarding preferences and beliefs about others' other- regarding preferences that we document in this paper call for more elaborate and accurate application of other-regarding utility models. In the experimental economics literature, beliefs are typically assumed to be rational. That is, actors are assumed to know the actual distribution of preferences, and this

distribution is independent of own preferences. We showed that these assumptions are problematic. If one disregards egocentric biases in beliefs by assuming rational expectations, one may obtain misleading results, such as incorrect predictions or inflated/incorrect estimates of social motives. Without modeling beliefs, modeling social preferences is not enough to derive accurate behavioral predictions for many interaction situations. We also hope that our findings on the distribution of own other-regarding preferences, beliefs about others' other-regarding preferences, and the relationship between the two will provide an empirical basis for future theoretical work that may incorporate these biases in beliefs into more complex game-theoretic models.

Before closing, we want to address some methodological caveats. As we explicitly discuss in the methods section, in our experimental design subjects decide in multiple binary Dictator Games. Additionally, a subject is paired with multiple other subjects as a recipient. We pay for all decisions, dictator and recipient. We analyze data assuming that subjects treat each Dictator Game as a one-shot game, rather than all games as one big game. This is a potentially problematic assumption. We believe that this feature of our design corresponds to a more general and important methodological issue in the literature. To estimate other-regarding preference parameters with good statistical precision, one needs several conditionally independent observations from each subject and a relatively large subject pool. Similarly, constraining subjects to play in a single decision role, e.g., the dictator or the recipient role reduces the sample size substantially. Paying for all decisions has the potential drawback that a subject may not treat a particular game independent from other games, e.g., maximize utility over all sets of games. It is possible to devise statistical models to analyze choices taking potential dependence between games. Yet, we think it is unlikely that subjects are maximizing utility over all games. Thus, even one could adapt the statistical model to include potential dependences between a subjects' choices over several games, such a model would be implausible, at least for most subjects. An alternative payment protocol, paying for a randomly selected game instead of all games may solve this potential dependence problem. In simple single person setups, paying for all or paying for a randomly selected game does not seem to matter (Laury, 2005). How-

ever, paying for a randomly selected game has its own problems in situations that involve more than 1-person, such as Dictator Game. For example, Laury (2005) shows that introducing a random payment scheme may reduce the influence of game outcomes on choice by introducing potential path-dependent utility. Thus, it is not clear if random payment protocol solves the problem at all. We think that this is an important methodological issue and leave the discussion to future research that will systematically compare the repercussions of using alternative designs and payment methods.

Chapter 4

Hierarchical Bayesian analyses of outcome- and process-based social preferences and beliefs in Dictator Games and sequential Prisoner's Dilemmas*

Abstract

In this paper, using a within-subjects design, we estimate the utility weights that subjects attach to the outcome of their interaction partners in four decision situations: (1) binary Dictator Games (DG), second player's role in the sequential Prisoner's Dilemma (PD) after the first player (2) cooperated and (3) defected, and (4) first player's role in the sequential Prisoner's Dilemma

*This chapter is written in collaboration with Jeroen Weesie and is currently under review.

game. We find that the average weights in these four decision situations have the following order: (1)>(2)>(4)>(3). Moreover, the average weight is positive in (1) but negative in (2), (3), and (4). Our findings indicate the existence of strong negative and small positive reciprocity for the average subject, but there is also high interpersonal variation in the weights in these four nodes. We conclude that the PD frame makes subjects more competitive than the DG frame. Using hierarchical Bayesian modeling, we simultaneously analyze beliefs of subjects about others' utility weights in the same four decision situations. We compare several alternative theoretical models on beliefs, e.g., rational beliefs (Bayesian-Nash equilibrium) and a consensus model. Our results on beliefs strongly support the consensus effect and refute rational beliefs: there is a strong relationship between own preferences and beliefs and this relationship is relatively stable across the four decision situations.

4.1 Introduction

Social dilemmas are an important research area in sociology (e.g., Dawes, 1980; Kollock, 1998). Standard rational choice models explain the emergence and persistence of cooperation in *embedded* settings with several factors such as network embeddedness, conditional cooperation, rewards, sanctions, termination of the relation, renegotiation of profits, and so on (e.g., Axelrod, 1984; Schuessler, 1989; Heckathorn, 1990; Weesie & Raub, 1996; Fudenberg & Maskin, 1986; Buskens & Raub, 2002). Yet, quite some social dilemma situations take place in *non-embedded settings* and among strangers where actors interact only once and will not see each other in the future. Such non-embedded settings lack the previously mentioned factors that could sustain cooperation. Thus, given classical models in the rational choice literature, one should not observe cooperation in non-embedded social dilemmas. However, we consistently observe otherwise (e.g., Sally, 1995; Camerer, 2003; Aksoy & Weesie, 2013b). Explaining cooperation in non-embedded settings, thus, has been a puzzle.

A rich body of literature, especially in social psychology and experimental economics but to a lesser extent in rational choice sociology, suggests that the

emergence and persistence of cooperation in non-embedded settings are due to *social preferences*. That is, in non-embedded settings cooperation is observed because (some) people do not try to maximize only own outcomes but are interested also in others' outcomes or hold other non-monetary motivations such as reciprocity. Many models of social preferences have been proposed to capture such non-selfish social preferences (for an overview see Fehr & Schmidt, 2006). One can distinguish roughly two types of social preferences: *outcome-based* and *process-based* (McCabe et al., 2003). Outcome-based social preferences are about how actors evaluate a certain outcome distribution between self and others. Social value orientations, e.g., individualism, cooperativeness, altruism, competitiveness (Schulz & May, 1989), and inequality aversion (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000) are examples of outcome-based preferences. In process-based social preferences, actors take the history of previous interactions into account. Responding kind intentions with more pro-social preferences (positive reciprocity) and unkind intentions with less pro-social preferences (negative reciprocity) are examples of process-based social preferences (Gautschi, 2000; Falk & Fischbacher, 2006; Vieth, 2009). In social dilemmas both outcome and process-based preferences could be at work. For example, in a sequential Trust Game when the trustor places trust, the trustee could be motivated by the objective outcomes that both actors would get in case she honors or abuses trust. But if trust is placed, the trustee may also want to reciprocate the kindness of the trustor irrespective of the monetary outcomes in the game. To understand cooperation in non-embedded settings, one should consider both outcome and process-based social preferences.

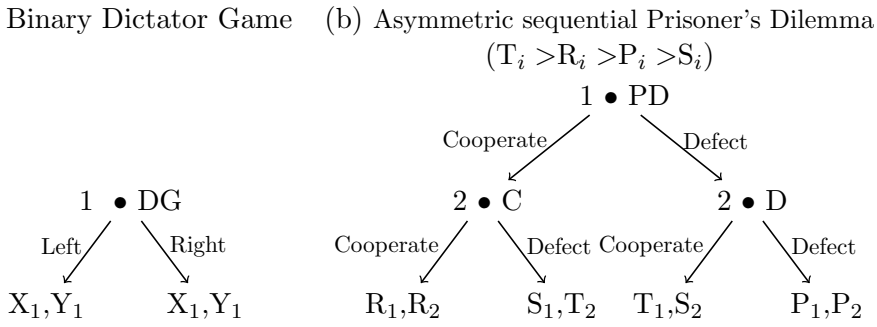
Social preferences are only part of the explanation. Social dilemmas are interdependent situations. In interdependent situations, behavior depends not only own (social) preferences but also on beliefs about others' choices. For example, one will not cooperate, however socially motivated, if one expects that others will free ride on one's cooperation. Thus, to predict the cooperative choice of people we should also deal with their *beliefs about the choices* of others. Although the economics and rational choice literatures on social preferences are vastly developed, the literature on beliefs is relatively scarce (see

also Blanco et al., 2009; Aksoy & Weesie, 2013a,b). In experimental economics and rational choice sociology, beliefs are often dealt with as an ingredient of the Bayesian-Nash equilibrium concept (Harsanyi, 1968). The Bayesian-Nash equilibrium is based on very strong assumptions about the beliefs that people have. For instance, people are assumed to know the distribution of social preferences in the population and that the interaction partner is a random draw from this distribution. Consequently, one's beliefs about others' social preferences are *independent* of one's own social preferences. These strong assumptions ensure that in the Bayesian-Nash equilibrium beliefs and choices are consistent. Throughout the paper we will use the term "rational beliefs" to denote beliefs that satisfy the aforementioned assumptions of Bayesian-Nash equilibrium (Bellemare et al., 2008). Although being mathematically elegant, in reality people's beliefs deviate from rational beliefs. There is a strong empirical relationship between preferences and beliefs which refute Bayesian-Nash beliefs (e.g., Blanco et al., 2009; Aksoy & Weesie, 2013a,b, 2012b). Still, it is yet to be studied how much harm assuming rational beliefs does in predicting social dilemma choices in non-embedded settings.

We should note that there are studies in the experimental economics literature that elicit beliefs experimentally rather than relying on Bayesian-Nash equilibrium (e.g., Bellemare et al., 2008; Blanco et al., 2009). These studies restrict the focus exclusively on the *beliefs about the choices* of others (see for a brief overview Aksoy & Weesie, 2013a). In our view, as one explains choices through a micro-model of social preferences, one should also explain beliefs about others' choices through the same micro-model of social preferences. That is, beliefs about the choices of others should be explained by *beliefs about social preferences* of others given by the model of social preferences. Extending the use of a model of social preferences to explain beliefs about the choices of others will, firstly, facilitate the empirical test of the social preference model (Aksoy & Weesie, 2013b). Secondly, explaining choices and beliefs about others' choices using the same social preference model provides a more parsimonious account than taking beliefs about others' choices as exogenous variables measured empirically.

In this paper, we employ a within-subjects experimental design with a set

Figure 4.1: Games used in the experiment. DG, PD, C, D are symbols that denote the decision nodes.



of binary Dictator Games and a set of non-embedded sequential Prisoner's Dilemma (PD) games, see Figure 4.1. Using a simple model for social preferences with a single social value orientation parameter, our analysis focuses on the following three questions. First, how does the social value orientation parameter differ between situations with and without relationship history (*process*). For example, is there positive or negative reciprocity? Second, how does the belief about the social value orientation parameters of others vary with own preferences, and does the relationship between own preferences and beliefs vary across histories. Third, if there is a relationship between one's own social value orientations and one's beliefs about others' social value orientations, and hence the Bayesian-Nash equilibrium does not hold, how much harm does assuming rational beliefs do in predicting choices in non-embedded social dilemmas? Answering these questions, we take advantage of hierarchical Bayesian statistical modeling.

4.2 Experimental design and procedure

4.2.1 Subjects

We recruited 155 subjects with the Online Recruitment System for Economic Experiments (ORSEE; Greiner (2004)). Subjects earned 16 Euros, on average, for taking part in the experiment.

4.2.2 Procedure

The experiment comprised eight sessions with 18 to 20 subjects and each session lasted about one hour. Subjects in each session were seated randomly in one of the cubicles in the Experimental Lab for Sociology and Economics (ELSE) at Utrecht University so that they could not see each other or the experimenter. After the general instructions stressing the important elements of the experiment such as incentive compatibility and anonymity, subjects played eight sequential Prisoner's Dilemma games (see Figure 4.1).[1] Because we want to analyze how the social value orientation of a subject varies across all decision situations that we include in the experiment, each subject played four of these eight PDs as the first player and the remaining four PDs as the second player. For the second player role, the so called *strategy method* was used to elicit decisions (Selten, 1967). These eight PDs differed with respect to the game outcomes T, R, P, S. Aksoy & Weesie (2013b) discusses the advantages of using asymmetric PDs. Following Aksoy & Weesie (2013b), one of these eight games was a symmetric game, that is T_1=T_2, R_1=R_2, P_1=P_2, S_1=S_2, and the remaining seven PDs were asymmetric. See Appendix C.1 for a description of these eight PDs. To explain the game to subjects in a comprehensible way, the game is described as an investment decision (Aksoy & Weesie, 2013b, 2009). The order in which a subject received these eight PDs was randomized. Because we are interested in non-embedded situations, in each of the PDs the interaction partner was a randomly selected other participant, i.e., we used a stranger matching protocol.

After subjects played these eight PDs (four as the first player, four as the second player), subjects made decisions in 18 binary Dictator Games (DG). See Appendix C.1 for a description of these 18 games. While subjects made decisions in these 18 DGs, simultaneously we elicited their beliefs about the decisions of other participants for these DGs. In particular, we asked what a subject thought was the percentage of subjects who chose option A in each DG. These beliefs were also incentivized as explained below. The order in which a subject received these 18 DGs was randomized. As for PDs, in each of the

[1] All instructions used in the experiment are available from the corresponding author.

DGs the recipient was a randomly selected other participant, i.e., stranger matching.

After these 18 DG decisions and beliefs, we elicited subjects' beliefs about the decisions of other participants in the eight PDs that they played before. To be precise, for each of the eight PDs we elicited beliefs of subjects about the percentage of others who they thought cooperated as the first player, as the second player after the first player's cooperation, and as the second player after the first player's defection. Note that in DGs, beliefs of subjects were elicited simulatenously with their own decisions whereas in PDs beliefs were elicited only *after* subjects played the PDs and DGs. Also remember that a subject played four of the eight PDs as the first player, and the other four as the second player. But we asked subjects' beliefs about the first players' as well as the second players' choices in all games. Hence, the number of responses per subject is higher for beliefs than for own decisions.

Also, following Aksoy & Weesie (2013a, 2012b), beliefs in all DGs and PDs were incentivized in the following way. For each correct guesses, i.e., the guessed percentage of A choices was the same as actual percentage of A choices, subjects received 500 points. For every percentage point deviation from the actual percentage, subjects received 20 points less. In case the guess was off by more than 25%, subjects received zero points.[2] Subjects received points for all games, DG and PD. Also, as in Aksoy & Weesie (2013a, 2012b), each subject passively earned points as the recipient of a randomly selected dictator in all 18 DGs. See Aksoy & Weesie (2013a) for a discussion of this payment scheme and incentivising all decisions and guesses.

4.3 Model of social preferences

In our setup, we are interested in four decision nodes that we refer to by the symbols DG, C, D, and PD: (DG) Dictator Game, (C) second player's choice in the sequential asymmetric PD after the first player cooperated; (D) the

[2]We assume that subjects report their average beliefs. See Aksoy & Weesie (2013a) for a discussion of eliciting and incentivizing beliefs as done here. There is also a statistical literature on elicitation of subjective probability distributions, see O'Hagan et al. (2006) and Lunn et al. (2013), pp 90–91.

second player's choice in the PD after the first player defected; (PD) the first player's choice in the PD (see Figure 4.1). These nodes correspond to different *histories* or futures: in nodes DG and PD there is no history, whereas in nodes C and D, the first player has cooperated or defected. Note that although there is no history in node PD, there is future. We define a subject i's utility in node $h \in \{\text{DG}, \text{C}, \text{D}, \text{PD}\}$ in game j for an outcome allocation x_{hj} for Ego and y_{hj} for Alter as:

$$\begin{aligned} U(x_{hj}, y_{hj}; \boldsymbol{\theta}_i, \boldsymbol{\epsilon}_{ij}, h) &= x_{hj} + \theta_{ih} y_{hj} + \epsilon_{ihj} \quad \text{and} \\ \boldsymbol{\theta}_i &= (\theta_{i\text{DG}}, \theta_{i\text{C}}, \theta_{i\text{D}}, \theta_{i\text{PD}}), \\ \boldsymbol{\epsilon}_{ij} &= (\epsilon_{i\text{DG}j}, \epsilon_{i\text{C}j}, \epsilon_{i\text{D}j}, \epsilon_{i\text{PD}j}). \end{aligned} \tag{4.1}$$

This model is a variant of the *social value orientation* model with single parameter (Schulz & May, 1989; Aksoy & Weesie, 2012b). We refer to Aksoy & Weesie (2012b) for the details of the *social orientation* model and only briefly discuss it here. In this model the vector $\boldsymbol{\theta}_i = (\theta_{i\text{DG}}, \theta_{i\text{C}}, \theta_{i\text{D}}, \theta_{i\text{PD}})$ denotes the θ weights for Alter's outcomes which differ across subjects i and nodes h. For example, $\theta_{i\text{DG}}$ corresponds to the weight that actor i attaches to the outcome of Alter in a Dictator Game. A subject with a higher θ in a given node is considered to be more cooperative than another subject with a lower θ in the same node. It is also possible that a subject attaches a negative weight $\theta < 0$ to the outcome of Alter, and thus is competitive.

The term $\boldsymbol{\epsilon}_{ij}$ in (4.1) is a random variable capturing evaluation noise. For statistical convenience we assume (multivariate) normality for $\boldsymbol{\theta}_i$ and $\boldsymbol{\epsilon}_{ij}$. Written formally,

$$\boldsymbol{\theta}_i = (\theta_{i\text{DG}}, \theta_{i\text{C}}, \theta_{i\text{D}}, \theta_{i\text{PD}}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{4.2}$$

where the 4×1 vector $\boldsymbol{\mu}$ and the 4×4 matrix $\boldsymbol{\Sigma}$ are the means and (co)variances of $\boldsymbol{\theta}_{ih}$, respectively. Subject i's weights in different nodes could be correlated, i.e., $\boldsymbol{\Sigma}$ need not be a diagonal matrix, because some people are more cooperative than others in all circumstances. Moreover, we make two simplifying assumptions on the correlations of $\boldsymbol{\theta}_i$ and $\boldsymbol{\epsilon}_{ij}$. First, we assume that $\boldsymbol{\theta}_i$ and

$\boldsymbol{\epsilon}_{ij}$ are uncorrelated. Second, we assume that ϵ_{ihj} are uncorrelated with each other. Thus,

$$\boldsymbol{\epsilon}_{ij} = (\epsilon_{i\mathrm{DG}j}, \epsilon_{i\mathrm{C}j}, \epsilon_{i\mathrm{D}j}, \epsilon_{i\mathrm{PD}j}) \sim N(\mathbf{0}_4, \boldsymbol{T}), \quad \boldsymbol{T} = \mathrm{diag}\left((\tau^2_{\mathrm{DG}}, \tau^2_{\mathrm{C}}, \tau^2_{\mathrm{D}}, \tau^2_{\mathrm{PD}})\right) \tag{4.3}$$

where $\boldsymbol{T}$ is a 4×4 diagonal matrix. We expect that the variance of the evaluation error increases in the cognitive complexity in the node: It will be the smallest in DG, then in C and D, and perhaps the highest in PD: $\tau^2_{\mathrm{DG}} < \tau^2_{\mathrm{C}}, \tau^2_{\mathrm{D}} \leq \tau^2_{\mathrm{PD}}$. For simplicity, and also because we see no obvious reason to consider otherwise, throughout our analyses, we assume that $\tau^2_{\mathrm{C}} = \tau^2_{\mathrm{D}}$.

Note that a subject makes multiple decisions in the nodes (to be precise 18, 4, 4, 4 decisions in DG, C, D, and PD, respectively). These games differ with respect to game outcomes (see Appendix C.1), both within and between the four nodes. We assume that the θ weight of a subject may vary across nodes, but not across different games within a node. In other words, the outcomes in the game do not influence θ. Similarly, we also assume that beliefs about other's θ in a node are not influenced by particular game outcomes in that node.

We need to discuss why we focus on this simple single-parameter social orientation model rather than any other, more complex model. Past research has shown that in addition to the outcomes of Alter, some respondents also consider the difference between the outcomes for Ego and Alter, e.g., equality or maximin preferences (e.g., Fehr & Schmidt, 1999; Schulz & May, 1989). An extension of the model in (4.1) by adding terms $\beta_{ih}|x_{hj} - y_{hj}|$ could capture such preferences. In principle, such extensions are possible within our framework. However, in this paper we have to refrain from complicating the utility function. First of all, as we discuss below the empirical analysis of preferences and beliefs is already very complex with the single social motive which is allowed to vary between nodes. Adding additional social motives to the utility function would complicate the implementation and presentation of analysis substantially. Fitting such higher dimensional models with sufficient precision also require much bigger datasets. Moreover, Aksoy & Weesie (2012b) explicitly compare the *social orientation* model in (4.1) with an extension for

inequality, and conclude that results on the social orientation θ parameter are robust with respect to the addition of a weight for inequality. Finally, although we discuss the fit of the simple utility model in Appendix C.2, the focus of this paper is not on finding the best model of social preferences. Rather, given this simple model, we would like to analyze how the θ parameter varies across nodes, the relationship between θ and beliefs about others' θ, and the predictive performances of alternative models for dealing with beliefs. We have to leave extending our analysis with additional social motives to future research.

4.4 Model of beliefs about others' social preferences

Following Aksoy & Weesie (2013a, 2012b), we model beliefs in the following way. Subject i has a belief about the distribution of θ. Let $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{i}} = (\widetilde{\theta}_{i\mathrm{DG}}, \widetilde{\theta}_{i\mathrm{C}}, \widetilde{\theta}_{i\mathrm{D}}, \widetilde{\theta}_{i\mathrm{PD}})$ denote the beliefs of actor i about $\boldsymbol{\theta}$ in the population. There could be differences between actors with respect to their beliefs. For example, compared with subject j, subject i may believe that people are in general more cooperative in node h. In model terms, $\mathrm{E}(\widetilde{\theta}_{ih}) > \mathrm{E}(\widetilde{\theta}_{jh})$. For mathematical convenience, we assume that the beliefs $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{i}}$ can be represented by a multivariate normal distribution. In preliminary statistical analyses we found that the correlations between the beliefs across the four nodes are all small and insignificant. For presentation purposes, throughout the paper we constrain all these correlations to 0. Hence, the beliefs $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{i}}$ could be represented by four independent normal distributions:

$$\widetilde{\theta}_{ih} \sim N(\widetilde{\mu}_{ih}, \widetilde{\sigma}^2_{ih}) \quad \text{for} \quad h \in \{\mathrm{DG}, \mathrm{C}, \mathrm{D}, \mathrm{PD}\}, \tag{4.4}$$

where $\widetilde{\mu}_{ih}$ and $\widetilde{\sigma}^2_{ih}$ are the mean and variance of i's beliefs about θ_h. An ego-centered bias in beliefs, e.g., the consensus effect, implies a positive relationship between one's own social orientation and (the mean of) one's beliefs about others' social orientation. Alternatively, in the Bayesian-Nash equilibrium concept beliefs and own preferences are assumed to be independent. These and other alternative hypotheses can be conveniently represented in terms of the regression of $\widetilde{\mu}_{ih}$ (and $\widetilde{\sigma}^2_{ih}$) on θ_{ih}, that is, regressing the mean (and

variance) of beliefs on own preferences. Therefore,

$$\widetilde{\mu}_{ih} = b_{0h} + b_{1h}\theta_{ih} + \eta_{ih} \quad \text{with} \quad \eta_{\mathrm{ih}} \sim \mathrm{N}(0, \varsigma_{\mathrm{h}}^2). \tag{4.5}$$

Note that equation (4.5) includes also error terms η_{ih}. This implies that two subjects with the same social orientation in node h may have different means for their beliefs about others' social orientation in that node. Thus, in our approach differences in beliefs across subjects are only partially explained by their own types θ_i if $\varsigma_h^2 > 0$.[3]

Given the model of social preferences and beliefs formulated so far, one can analyze possible relationships between preferences and beliefs via restrictions on the parameters in (4.5). For example, a strict consensus effect implies a full projection of one's own θ to others' social orientations, that is, $b_{0h} = 0$ and $b_{1h} = 1$, and maybe $\varsigma_h^2 = 0$. The Bayesian-Nash equilibrium assumption, on the other hand, implies that actors would *know* the actual distribution of $\boldsymbol{\theta}_i$. Thus, $b_{1h} = 0$ and $b_{0h} = \mu_h$, where μ_h is the true mean as in (4.2), and $\varsigma_h^2 = 0$.

We model the variances in beliefs, $\widetilde{\sigma}_{ih}^2$, to be the same for all subjects:

$$\widetilde{\sigma}_{ih}^2 = \widetilde{\sigma}_{0h}^2. \tag{4.6}$$

Actually, Aksoy & Weesie (2012b) have shown that the variance in beliefs about others' social preferences $\widetilde{\sigma}_{ih}^2$ varies (non-linearly) with one's own social preferences (roughly, the larger $|\theta_{ih}|$, the larger $\widetilde{\sigma}_{ih}^2$). In the current analysis there are four different social preference parameters, one per node. We considered model specifications in which all these four variances depended on the corresponding own social preference parameter. Yet, estimation was very time consuming and yielded convergence problems for many of our MCMC runs. Using a simpler set-up (only DG choices), we checked how this simpli-

[3] In principle, it is possible to model the mean of an actor's beliefs in node h, $\widetilde{\mu}_{ih}$, as a function of all elements of $\boldsymbol{\theta}_i$. For example, the mean of i's beliefs about others' θ weight in node DG, $\widetilde{\mu}_{i\mathrm{DG}}$, could be a function of i's own θ in node PD *as well as* of i's own θs in nodes $\mathrm{D}, \mathrm{C}, \mathrm{PD}$. However, such an extension would increase the number of parameters to be estimated dramatically. Moreover, we also think that it will be only of minor theoretical and empirical importance to make the mean of the belief about other's θ in node h depend on all elements of $\boldsymbol{\theta}_i$.

fying assumption of constant variance in beliefs influenced other parameter estimates. Fortunately, this simplifying assumption did not influence other estimates substantially. Thus, we stick to the simpler model in (4.6).

4.5 Statistical analysis of preferences and beliefs

We will present our analyses step-by-step. First, will start with analyzing the social preferences of subjects in the first three nodes, that is, DG, C, and D. Then, we will introduce the analysis of subjects' beliefs in nodes DG, C, and D. Finally, we will present the analysis of the social preferences of subjects in the fourth node, PD. The reason for presenting our results step-by-step rather than presenting immediately a simultaneous analysis of choices and beliefs in all four nodes is the following. As we will discuss below, in each step, we compare several alternative specifications of our model. We proceed to the subsequent step building on the best fitting specification and discarding worse fitting specifications in a current step. Presenting simultaneous analyses with all possible combinations of alternative specifications in all steps would require a vast amount of space. Also, the fourth PD node is the most complicated node. To analyze subjects' social preferences and choices in node PD, we also need to deal with subjects' beliefs about their interaction partners' choices in nodes C and D. This is because the choices in node PD depends theoretically on the beliefs about the second player's behavior. Because of this complexity, we present the analysis of the PD node after analyzing social preferences and beliefs in nodes DG, C, and D.

4.5.1 Social preferences in nodes DG, C, and D

Specifications

Here we compare four specifications of the distribution of $(\theta_{i\mathrm{DG}}, \theta_{i\mathrm{C}}, \theta_{i\mathrm{D}})$. In these four specifications we decrease complexity step by step. These specifications reflect different assumptions on how history influences social orientations (e.g., Gautschi, 2000).

Specification A1 (3-dimensional $\boldsymbol{\theta}$): This is the most general specification. Besides the general assumptions given in equations (4.1), (4.2), and (4.3), there is no further constraining assumption. Other specifications below are nested in this general specification. If there is both positive and negative reciprocity, with high probability, we should observe $\theta_{\mathrm{C}} > \theta_{\mathrm{DG}} > \theta_{\mathrm{D}}$: the weight given to the outcome of Alter will be the highest after Alter cooperated and the lowest after Alter defected. The "neutral" weight in a DG will be somewhere in the middle of the two. Note that this specification allows individuals to have different θs in different nodes. This means that the magnitude of reciprocity $(\theta_{\mathrm{C}} - \theta_{\mathrm{DG}}, \theta_{\mathrm{DG}} - \theta_{\mathrm{D}})$ can also vary between subjects.

Specification A2 (2-dimensional $\boldsymbol{\theta}$): This specification imposes the following constraint:

$$\theta_{\mathrm{C}} = \theta_{\mathrm{D}} + k. \tag{4.7}$$

This specification assumes that the social orientations in nodes C and D, θ_{C} and θ_{D}, are perfectly correlated and have the same variance. Compared to the node after the first player cooperated (node C), the social orientation after the first player defected (node D) only shifts down or up. This shift k is constrained to be the same across individuals. Reciprocity would mean $k > 0$. This specification does not make any additional assumption about the relationship between the social orientation θ_{DG} and the social orientations θ_{C}, θ_{D}. Yet, if there is positive reciprocity, one expects on average θ_{C} to be larger than θ_{DG}, or that $\Pr(\theta_{\mathrm{C}} > \theta_{\mathrm{DG}})$ is large.

Specification A3 (1-dimensional $\boldsymbol{\theta}$): This specification imposes the following two constraints:

$$\theta_{\mathrm{C}} = \theta_{\mathrm{DG}} + k_{\mathrm{C}}, \qquad \theta_{\mathrm{D}} = \theta_{\mathrm{DG}} + k_{\mathrm{D}}. \tag{4.8}$$

These two constraints imply that the social orientations in DG, C, and D, are perfectly correlated and have the same variance. The social orientation only shifts up or down depending on history. In addition, the magnitudes of the shifts, k_{C} and k_{D}, are the same for all individuals, i.e., homogeneous reci-

procity. A positive reciprocity would mean $k_{\mathrm{C}} > 0$, and negative reciprocity $k_{\mathrm{D}} < 0$.

Specification A4 (Single $\boldsymbol{\theta}$): This specification is the most parsimonious of the four. It assumes purely outcome-based social preferences that do not depend on history:

$$\theta_{\mathrm{DG}} = \theta_{\mathrm{C}} = \theta_{\mathrm{D}}. \tag{4.9}$$

Methods

Remember that each subject made 18, 4, and 4 decisions in nodes DG, C, and D, respectively. Also remember that the game outcomes in these 26 games differ (see Appendix C.1). We stack all these choices nested in subjects. All of these 26 decisions are binary. A subject chooses option 1 in node h in game j with outcomes (x_{1hj}, y_{1hj}) instead of option 2 with outcomes (x_{2hj}, y_{2hj}) if the utility U_{1ihj} for option 1 exceeds the utility U_{2ihj} for option 2. With equations (4.1), (4.2), and (4.3), the probability that subject i chooses the first option in node h in game j can be conveniently written as:

$$\begin{aligned}
\Pr(U_{1ihj} > U_{2ihj} | \boldsymbol{\theta}_i, h) &= \Pr\left((x_{1hj} + \theta_{ih} \cdot y_{1hj} + \epsilon_{1ihj}) > (x_{2hj} + \theta_{ih} \cdot y_{2ihj} + \epsilon_{2ihj})\right) \\
&= \Pr\left((x_{1hj} - x_{2hj}) + \theta_{ih} \cdot (y_{1hj} - y_{2hj}) + (\epsilon_{1ihj} - \epsilon_{2ihj}) > 0\right) \\
&= \Phi\left(\frac{(x_{1hj} - x_{2hj}) + \theta_{ih} \cdot (y_{1hj} - y_{2hj})}{\sqrt{2}\tau_h}\right),
\end{aligned} \tag{4.10}$$

where Φ is the cumulative standard normal distribution.

Statistically oriented readers will recognize that, assuming normality and independence of decisions conditional on $\boldsymbol{\theta}$, equation (4.10) implies a multilevel probit regression with heteroskedastic error with respect to history (see Aksoy & Weesie (2012b) for more details). The dependent variable is a subject's choice. The two independent variables are (1) the difference between the outcomes for Ego in the two options $x_{1hj} - x_{2hj}$ with a coefficient constrained to 1, and (2) the difference between the outcome for Alter in the two options $y_{1hj} - y_{2hj}$ with a random coefficient θ_h. Note that there is no (fixed or

random) intercept in the model. A non-zero intercept would represent an intrinsic motivation for choosing option 1 over option 2, controlling for game outcomes.[4] The parameters to estimate are the means, standard deviations, and correlations of $\boldsymbol{\theta}_i$, and the standard deviations τ_h of the evaluation error. Depending on the theoretical model (Specification A1 to A4), some of these parameters are constrained to be the same, or constrained to 1.

In our analyses we use a Bayesian approach simply because the more complicated statistical models that will be discussed below are too hard to fit in a frequentist framework. See Aksoy & Weesie (2013a) for a comparison of frequentist and Bayesian statistical frameworks in a similar substantial context. In all analyses in this paper, we use (weakly) uninformative priors for the parameters. Hence, the posterior means can be treated as point estimates that are approximately equal to the maximum likelihood estimates. In Appendix C.2 we provide the details of Bayesian estimation including the priors used. To compare the four specifications of social preferences, we use the Deviance Information Criteria (DIC) (Spiegelhalter et al., 2002), a Bayesian analogue to Akaike Information Criterion (AIC).[5]

For all models reported in this paper, we run two chains of 100,000 MCMC iterations using OpenBugs (Lunn et al., 2009). The OpenBugs codes are available upon request. Unless stated otherwise, we exclude the first 30,000 burn-in iterations, and record only every 10th iteration. Convergence is checked using the Gelman & Rubin (1992) $\hat{R}$ statistic as well as by visual inspection of the MCMC trajectories, and their autocorrelation.

[4] A non-zero intercept in nodes C and D would correspond to normative utility, utility derived from the act of cooperation, discussed by (Aksoy & Weesie, 2013b). We fitted models with intercepts in nodes C and D. In our case these intercepts were small and insignificant. For brevity we dropped these intercepts in all analyzes reported here.

[5] We also calculated posterior predictive p-values (PPP) (Gelman et al., 1996) to assess fit. Appendix C.2 includes a discussion of PPPs for our models. We found that the power of posterior predictive testing in our case is rather low, probably due to small number of choices per subject, especially in nodes C and D. Consequently, it was hardly possible to reject any of the models using PPPs.

Results

Table 4.1 presents our results for Specifications A1 to A4. Consider first the DIC scores. A DIC difference larger than 10 strongly favors the model with the lower DIC (Spiegelhalter et al., 2002). Thus, our results strongly favor Specification A1. This means that history matters and the constraints imposed by A2 to A4 are not consistent with the data. For example, Specification A2 assumes that the social orientations after the first player cooperated and defected in the PD are perfectly correlated and have the same variance, albeit allowed to shift up or down by a fixed amount. This assumption does not hold, implying that how much θ differs between nodes C and D varies between subjects. The most parsimonious specification, A4, assumes that history does not matter as θ does not vary across nodes. Rejection of this model clearly shows that history effects cannot be ruled out, thus social preferences are partly process-based.

We now interpret the results of the selected specification, A1. On average, subjects attach a positive weight to the outcome of Alter in a DG: $\text{mean}(\theta_{\text{DG}}) = 0.113$, 75% of subjects are expected to have positive social orientations ($\widehat{\Pr}(\theta_{\text{DG}} > 0) = \Phi(\frac{0.113}{0.169}) \approx .75$). Surprisingly, in the PD, on average, the second players attach a negative weight to the outcome of the first player, even if the first player cooperated: $\text{mean}(\theta_{\text{C}}) = -0.178$. Only about 40% of subjects are expected to have a positive social orientation in node C. In other words, if we compare nodes DG and C, we do not find positive reciprocity. It seems that being in a PD makes subjects more competitive than being in a DG. In the "discussion and conclusion" section we will turn back to this finding. We do find, however, strong negative reciprocity. The lowest average θ is in node D: $\text{mean}(\theta_{\text{D}}) = -0.578$. Only about 16% of subjects have a positive social orientation in node D.

Given standard properties of the multivariate normal distribution of $\boldsymbol{\theta}$ for Specification A1, one can obtain additional results. For example, $\widehat{\Pr}(\theta_{\text{C}} > \theta_{\text{DG}}) = \Phi\left(\frac{-0.178-0.113}{\sqrt{0.169^2+0.721^2-2\cdot 0.169\cdot 0.721\cdot 0.726}}\right) \approx 0.32$, i.e., for 32% of the subjects $\theta_{\text{C}} > \theta_{\text{DG}}$ and thus they show positive reciprocity. Similarly, $\widehat{\Pr}(\theta_{\text{C}} > \theta_{\text{D}}) = 0.79$, i.e., for 79% of the subjects $\theta_{\text{C}} > \theta_{\text{D}}$. Finally, $\widehat{\Pr}(\theta_{\text{C}} > \theta_{\text{DG}} > \theta_{\text{D}}) \approx 0.25$.

These findings further support the existence of mainly negative reciprocity when one compares the DG and second player PD choices.

It should be noted that the variation among subjects with respect to social orientations, as indicated by the standard deviations of θ_h, is higher in a PD than in a DG. Also, the social orientations correlate highly across histories (all above 0.5). As we expected, the evaluation error is smaller in the simpler node DG than in the more complex nodes C and D: the standard deviation of the evaluation error in node PD τ_{DG}=0.178, whereas $\tau_{\mathrm{C}} = \tau_{\mathrm{D}} = 1.235$ in the PD.

We will now proceed with analyzing beliefs using the selected specification, A1.

4.5.2 Beliefs in nodes DG, C, and D

Similar to social preferences, for beliefs in nodes DG, C, and D we compare five alternative specifications: an unrestricted one, two with a relaxed and a strict versions of the consensus effect, and two with a restricted and relaxed versions of rational beliefs.

Specifications

Specification B1 (Unconstrained beliefs): This is the general model for beliefs described above in equations (4.4) to (4.6). In specification B1, the mean of beliefs in node h is modeled as a linear regression on the social orientation in node h with residual variables that depend on history but not on subject. The only difference now is that we consider only the first three decision nodes whereas the general model above also included the fourth node. B1 encompasses B2-B5 discussed below in the sense that it does not impose constraints on the intercept and slope parameters for the regressions of the means of beliefs on own social orientations.

Specification B2 (Relaxed consensus): Specification B2 imposes directly a version of the consensus effect. In this model, a person fully projects her social orientation to her belief about the social orientations of others. That is,

Table 4.1: Hierarchical Bayesian models for 4030 choices (18 DG, 4 C, and 4 D) by N=155 subjects. Multivariate normal distribution of social orientation parameters and evaluation error across histories. We report posterior means and, in parentheses, posterior standard deviations of the parameters and DIC. Prior specifications are given in Appendix C.2.

	Spec. A1 (3-dim. $\boldsymbol{\theta}$)	Spec. A2 (2-dim. $\boldsymbol{\theta}$)	Spec. A3 (1-dim. $\boldsymbol{\theta}$)	Spec. A4 (Single $\boldsymbol{\theta}$)
Distribution of social orientation parameters				
Mean(θ_{DG})	0.113**	0.113**	0.123**	0.081^{c**}
	(0.016)	(0.015)	(0.018)	(0.015)
Mean(θ_{C})	-0.178*	-0.158*	0.070**	0.081^{c**}
	(0.098)	(0.094)	(0.031)	(0.015)
Mean(θ_{D})	-0.578**	-0.655**	-0.238**	0.081^{c**}
	(0.142)	(0.120)	(0.036)	(0.015)
SD(θ_{DG})	0.169**	0.167**	0.202^{c**}	0.164^{c**}
	(0.014)	(0.014)	(0.016)	(0.013)
SD(θ_{C})	0.721**	0.642^{c**}	0.202^{c**}	0.164^{c**}
	(0.104)	(0.098)	(0.016)	(0.013)
SD(θ_{D})	0.581**	0.642^{c**}	0.202^{c**}	0.164^{c**}
	(0.123)	(0.098)	(0.016)	(0.013)
Corr($\theta_{\mathrm{DG}}, \theta_{\mathrm{C}}$)	0.726**	0.715**	1.000^{f}	1.000^{f}
	(0.067)	(0.068)	–	–
Corr($\theta_{\mathrm{DG}}, \theta_{\mathrm{D}}$)	0.620**	0.715**	1.000^{f}	1.000^{f}
	(0.088)	(0.068)	–	–
Corr($\theta_{\mathrm{C}}, \theta_{\mathrm{D}}$)	0.725**	1.000^{f}	1.000^{f}	1.000^{f}
	(0.094)	–	–	–
Evaluation error				
SD($\epsilon_{\mathrm{DG}j}$) = τ_{DG}	0.178**	0.178**	0.181**	0.180**
	(0.006)	(0.006)	(0.006)	(0.006)
SD($\epsilon_{\mathrm{C}j}$)=SD($\epsilon_{\mathrm{D}j}$) = $\tau_{\mathrm{C}} = \tau_{\mathrm{D}}$	0.873**	1.018**	1.020**	0.713**
	(0.107)	(0.109)	(0.058)	0.039
DIC	2812	2856	3017	3274

(*)** (90%) 95% credibility interval excludes 0
^{c}parameter equality constrained; ^{f}parameter fixed

the expected value of the mean of beliefs about other's social orientation in a node equals one's own social orientations in the same node:

$$\begin{aligned} \widetilde{\mu}_{ih} &= \theta_{ih} + \eta_{ih} \quad \text{with} \quad \eta_{\text{ih}} \sim \text{N}(0, \varsigma_{\text{h}}^2), \\ \widetilde{\sigma}_{ih}^2 &= \widetilde{\sigma}_{0h}^2. \end{aligned} \tag{4.11}$$

However, Specification B2 still includes the error term η_{ih}. Thus, although people project their own social orientations to the mean of their beliefs, two subjects with the same social orientation may still have different beliefs due to the error term. Because of the existence of the error term, we call this specification "relaxed" consensus.

Specification B3 (Strict consensus): Specification B3 imposes a stricter version of the consensus effect. On top of Specification B2 above, B3 drops the error term η_{ih}, or equivalently, the variance of η_{ih}, ς_h^2, is constrained to be 0. Thus, Specification B3 assumes that subjects' mean beliefs are exactly the same as their own social orientations, and two subjects with the same social orientation have exactly the same beliefs:

$$\begin{aligned} \widetilde{\mu}_{ih} &= \theta_{ih}, \\ \widetilde{\sigma}_{ih}^2 &= \widetilde{\sigma}_{0h}^2. \end{aligned} \tag{4.12}$$

Specification B4 (Relaxed Bayesian-Nash): The Bayesian-Nash equilibrium approach, used extensively in (behavioral) game-theory, assumes that beliefs are rational. In our case, this implies that actors are assumed to *know* the actual distribution of social orientations $\boldsymbol{\theta}$ in the population. Since the actual distribution of social orientations is unique, this specification assumes that all subjects have the same belief, and hence there is no relationship between someone's own social preferences and someone's beliefs. Written formally:

$$\begin{aligned} \widetilde{\mu}_{ih} &= \text{mean}(\theta_h) + \eta_{ih} \quad \text{with} \quad \eta_{\text{ih}} \sim \text{N}(0, \varsigma_{\text{h}}^2), \\ \widetilde{\sigma}_{ih}^2 &= \text{var}(\theta_h). \end{aligned} \tag{4.13}$$

where mean(θ_h) and var(θ_h) are respectively the "true" means and variances of θ_h estimated from our sample. Note that despite assuming that actors know the true distribution of $\boldsymbol{\theta}$, Specification B4 still allows subjects to err about the mean by keeping the error term η_{ih} in the equation.

Specification B5 (Strict Bayesian-Nash): A stricter interpretation of Bayesian-Nash beliefs would require discarding the error term η_{ih}, or constraining ς_h^2 to be 0, because in Bayesian-Nash equilibrium people are assumed to know the true distribution without any error. This implies:

$$\begin{aligned} \widetilde{\mu}_{ih} &= \text{mean}(\theta_h), \\ \widetilde{\sigma}_{ih}^2 &= \text{var}(\theta_h). \end{aligned} \tag{4.14}$$

Methods

We first explain the data and the statistical procedure. Remember that all decisions in all nodes are binary but beliefs are elicited as percentages of subjects who are expected to choose option 1. Let $\widetilde{\pi}_{ihj}$ denote i's belief about the percentage of others choosing option 1 rather than option 2 in game j of node h. Under the social orientation model, subject i's belief about this percentage depends on the mean $\widetilde{\mu}_{ih}$ and the variance $\widetilde{\sigma}_{ih}^2$ of beliefs about $\boldsymbol{\theta}$ and on the game outcomes, $(x_{1hj}, y_{1hj}; x_{2hj}, y_{2hj})$. Formally, this percentage can be written as i's belief about the probability that a random Alter favors option 1 over option 2, that is

$$\widetilde{\pi}_{ihj} = Pr\left(U(x_{1hj}, y_{1hj}; \widetilde{\theta}_{ih}) > U(x_{2hj}, y_{2hj}; \widetilde{\theta}_{ih}) \mid \widetilde{\theta}_{ih} \sim N(\widetilde{\mu}_{ih}, \widetilde{\sigma}_{ih}^2), \widetilde{\tau}^2\right). \tag{4.15}$$

As $\widetilde{\theta}_{ih}$ is assumed to be normally distributed, $\widetilde{\pi}_{ihj}$ satisfies

$$\widetilde{\pi}_{ihj} = \Phi\left(\frac{(x_{1hj} - x_{2hj}) + \widetilde{\mu}_{ih} \cdot (y_{1hj} - y_{2hj})}{\sqrt{2\widetilde{\tau}^2 + \widetilde{\sigma}_{ih}^2 \cdot (y_{1hj} - y_{2hj})^2}}\right) \tag{4.16}$$

where Φ is the cumulative standard normal distribution. x_{khj} is the outcome for Ego in option $k = 1, 2$ and y_{khj} is the outcome for Alter in option k in

game j of node h.

In (4.15) and (4.16), $\widetilde{\tau}$ is the belief about the variance of the evaluation error which is, for simplicity, assumed to be the same for all subjects and all nodes. Alternatively, we could have assumed that it depends on node h, or even that beliefs about τ_h are rational, i.e., identical to the "true" standard deviation of evaluation noise. These alternative specifications for $\widetilde{\tau}$ yielded worse fit than in (4.16).

Furthermore, to be able to use (4.16) in statistical analysis of elicited beliefs p_{ihj} about $\widetilde{\pi}_{ihj}$, we introduce an error term such that a subject makes an unsystematic error in reporting $\widetilde{\pi}_{ihj}$:

$$p_{ihj} = \Phi\left(\frac{(x_{1hj} - x_{2hj}) + \widetilde{\mu}_{ih} \cdot (y_{1hj} - y_{2hj})}{\sqrt{2\widetilde{\tau}^2 + \widetilde{\sigma}^2_{ih} \cdot (y_{1hj} - y_{2hj})^2}} + \upsilon_{ihj}\right) \quad \text{with} \quad \upsilon_{\text{ihj}} \sim \text{N}(0, \zeta^2_{\text{h}}). \tag{4.17}$$

Here υ_{ihj} is added within the parentheses to ensure $0 < p_{ihj} < 1$. In our analyses, we found that the variance of the response error ζ^2_h differed across nodes. Accounting for this difference proved to be important for other parameter estimates. Consequently, in our analyses below, we allow ζ^2_h to vary with node, with the constraint $\zeta^2_{\text{C}} = \zeta^2_{\text{D}}$.

We also increased the number of burn-in iterations to 200,000 (it was 30,000 before) when we include beliefs. This is because the models with beliefs are more complex than models for only own preferences, and convergence takes longer.[6]

Results

Table 4.2 presents the DIC scores for the joint statistical models for social preferences and beliefs. Note that each statistical model in the table has two parts, the model for preferences (Specification A1) and the model for beliefs

[6] We also experienced a numerical difficulty in fitting B4 and B5. Ideally, in fitting B4 and B5, we should include the variances in beliefs $\widetilde{\sigma}^2_h$ (see equation (4.17)) and variances in social orientations σ^2_h as *latent variables* in the statistical estimation with the constraints $\widetilde{\sigma}^2_h = \sigma^2_h$ (beliefs correspond to "true" values). However, this yielded a numerical error in the OpenBugs routine. Instead, fitting B4 and B5, we plugged in directly the *estimated scores* for σ^2_h obtained from A1 as substitutes for $\widetilde{\sigma}^2_h$.

Table 4.2: DIC statistics for the statistical models that represent Specifications B1 to B5. Each model has two parts, a part for social orientations (Specification A1) and a part for beliefs. A DIC for each of these two parts and an overall DIC are provided.

Model	DIC (preferences)	DIC (beliefs)	DIC (overall)
B1 (Unconstrained)	2805	14020	16830
B2 (Relaxed consensus)	2810	14030	16840
B3 (Strict consensus)	2977	14160	17130
B4 (Relaxed Bayesian-Nash)	2812	14590	17400
B5 (Strict Bayesian-Nash)	2816	16610	19420

(Specifications B1-B5). Consequently, for each statistical model, three DIC scores are calculated: a DIC for the part on preferences, a DIC for beliefs, and an overall DIC. The difference between overall DICs in Specifications B1 and B2 is exactly 10. The DIC difference for the belief parts of B1 and B2 is also 10. A DIC difference of 10 is considered to be sizable, but not decisive. The DIC difference for the preference part of B1 and B2 is smaller, in fact only 5. This shows that imposing a relaxed version of the consensus effect, that is, constraining the expected mean of beliefs about others' social orientations to be the same as own social orientations in all decision nodes deteriorates the fit, but not a lot. All other specifications, viz. B3, B4, and B5, are flatly rejected based on their overall DIC or DIC for the belief parts, compared to those of Specifications B1 and B2. Both relaxed and strict versions of Bayesian-Nash beliefs and the stricter interpretation of the consensus effect are, thus, refuted. Note that the DIC scores for the social preference part of the specifications are quite similar. This is because in Specifications B1 to B5, the part for preferences (Specification A1) is the same, but the parameters for the preference part are somewhat different as they are also included as predictors in the belief part of models. The main difference between Specifications B1 to B5 is on how they model beliefs, thus DIC differences for the belief parts are much larger.[7]

[7]The DIC for the preference part of Specification B3 is, however, much higher than other specifications. This is because imposing strict consensus on beliefs also influences the parameters in the preference part, that is, the parameters for the distribution of social orientations and the evaluation error, which in turn deteriorates the fit for the preference

Table 4.3: Results for hierarchical Bayesian models for (the relationship between social orientations and) beliefs about others' social orientations: we report posterior means and, in parentheses, posterior standard deviations of the parameters for Specifications B1 and B2. N(belief)=5,270, N(decision) = 4,030, N(subject) = 155. Moreover, R^2s are provided for mean of beliefs. Note that η_{ih} are assumed independent across subjects and histories.

	Specification B1 : Unconstrained beliefs
	Means of beliefs :
$\widetilde{\mu}_{i\mathrm{DG}} =$	$-0.037(0.024) + 0.825(0.143)\theta_{i\mathrm{DG}j} + \eta_{i\mathrm{DG}},\ R^2 = 0.41,$
	$\eta_{i\mathrm{DG}} \sim N(0, 0.177^2(0.019^2))$
$\widetilde{\mu}_{i\mathrm{C}} =$	$-0.853(0.172) + 1.979(0.335)\theta_{i\mathrm{C}j} + \eta_{i\mathrm{C}},\ R^2 = 0.67,$
	$\eta_{i\mathrm{C}} \sim N(0, 0.933^2(0.123^2))$
$\widetilde{\mu}_{i\mathrm{D}} =$	$-0.241(0.177) + 0.999(0.404)\theta_{i\mathrm{D}j} + \eta_{i\mathrm{D}},\ R^2 = 0.64,$
	$\eta_{i\mathrm{D}} \sim N(0, 0.456^2(0.140^2))$
	Standard deviations of beliefs :
$\widetilde{\sigma}_{i\mathrm{DG}} =$	$0.334(0.019),\ \widetilde{\sigma}_{i\mathrm{C}} = 2.118(0.149),\ \widetilde{\sigma}_{i\mathrm{D}} = 0.915(0.255)$
	Evaluation and response errors :
$\widetilde{\tau} =$	$0.172(0.006),\ \zeta_{\mathrm{DG}} = 1.442(0.039),\ \zeta_{\mathrm{C}} = \zeta_{\mathrm{D}} = 0.390(0.012)$
	Specification B2 : Relaxed consensus
	Means of beliefs :
$\widetilde{\mu}_{i\mathrm{DG}} =$	$0^f + 1^f\theta_{i\mathrm{DG}} + \eta_{i\mathrm{DG}},\ R^2 = 0.39,$
	$\eta_{i\mathrm{DG}} \sim N(0, 0.196^2(0.019^2))$
$\widetilde{\mu}_{i\mathrm{C}} =$	$0^f + 1^f\theta_{i\mathrm{C}} + \eta_{i\mathrm{C}},\ R^2 = 0.65,$
	$\eta_{i\mathrm{C}} \sim N(0, 0.830^2(0.108^2))$
$\widetilde{\mu}_{i\mathrm{D}} =$	$0^f + 1^f\theta_{i\mathrm{D}} + \eta_{i\mathrm{D}},\ R^2 = 0.73,$
	$\eta_{i\mathrm{D}} \sim N(0, 0.370^2(0.064^2))$
	Standard deviations of beliefs :
$\widetilde{\sigma}_{i\mathrm{DG}} =$	$0.353(0.019),\ \widetilde{\sigma}_{i\mathrm{C}} = 1.689(0.107),\ \widetilde{\sigma}_{i\mathrm{D}} = 0.824(0.084)$
	Evaluation and response errors :
$\widetilde{\tau} =$	$0.172(0.006),\ \zeta_{\mathrm{DG}} = 1.199(0.016),\ \zeta_{\mathrm{C}} = \zeta_{\mathrm{D}} = 0.630(0.010)$

^{f}parameter is fixed.

Because we reject Specifications B3, B4, and B5, in Table 4.3 we present detailed results on the parameter estimates for only B1 and B2, for brevity. Detailed results for Specifications B3, B4, and B5 are available from the authors. To enhance readability and focus attention of readers to what is new, Table 4.3 omits the results for the preference part (Specification A1); the results for preferences are very similar to those reported in Table 4.1 and also very similar across Specifications B1 and B2. Table 4.3 shows that the mean of a subject's beliefs about the distribution of θ in the population is almost the same as one's own θ, maybe except for node C. In B1, it seems that the relationship between own preferences and beliefs is steeper and the intercept is smaller for node C than for nodes D and DG. Yet, as we discussed above, constraining all slopes to 1 and all intercepts to 0 does not deteriorate the fit dramatically. This yields a strong support for the consensus hypothesis. Also remember that the stricter version of the consensus effect which discards the error terms η_{ih} is rejected. Thus, while as a *general trend* the consensus effect holds, two subjects with the same social orientation do not necessarily have the same beliefs. Note also that Table 4.3 includes R^2 values for the regressions of mean beliefs on own social orientations.[8] These R^2 values show that own social orientations explain a great deal of variance in mean beliefs.

We now move on to the most complicated node, the first player's decision in the PD.

4.5.3 Social preferences in node PD

In this part, we extend the analyses to the fourth and final node, namely the first player's decision in the sequential PD. In node PD there is no history, but there is future. The analysis for node PD is more complicated than that for nodes DG, C, and D, because the consequences of player 1's decisions are

part.

[8] These R^2s are calculated using the formula $R^2 = \frac{b_{1h}^2 Var(\theta_h)}{b_{1h}^2 Var(\theta_h)+Var(\eta_h)}$. Note that the R^2 value for node D in B2 is slightly higher than that in B1, which is surprising since B1 encompasses B2. However, these R^2s are not exactly the same as conventional R^2s in linear regressions with observed variables. This is because in B1 and B2, the predictors, θ_{ih}, are unobserved variables that also change—so does $Var(\theta_h)$—between B1 and B2, which in turn influence the R^2 values.

strategically uncertain: The consequences of player 1's decisions depend on player 2's decisions. We assume that player 1's decisions can be explained in terms of *expected* consequences. Here by "expected" we mean averaging over possible decisions of player 2. That is, in node PD the outcomes that Ego expects for Ego and Alter are derived from Ego's beliefs about Alter's decisions.

Using the PD notation (see Figure 4.1) for outcomes T_{kij}, R_{kij}, P_{kij}, and S_{kij} for $k = 1, 2$, the expected outcomes for Ego ($x_{1\mathrm{PD}j}$) and Alter ($y_{1\mathrm{PD}j}$) when Ego chooses the first option (Ego cooperates) in game j in node PD can be written as

$$\begin{aligned} x_{1\mathrm{PD}j} &= \widetilde{\pi}_{i\mathrm{C}j} R_{1j} + (1 - \widetilde{\pi}_{i\mathrm{C}j}) S_{1j} \\ y_{1\mathrm{PD}j} &= \widetilde{\pi}_{i\mathrm{C}j} R_{2j} + (1 - \widetilde{\pi}_{i\mathrm{C}j}) T_{2j} \end{aligned} \tag{4.18}$$

where $\widetilde{\pi}_{i\mathrm{C}j}$ is Ego i's belief about the probability that Alter cooperates after Ego cooperated in game j. Similarly, the expected outcomes for Ego ($x_{2\mathrm{PD}j}$) and Alter ($y_{2\mathrm{PD}j}$) when Ego chooses the second option (Ego defects) in game j in node PD can be written as

$$\begin{aligned} x_{2\mathrm{PD}j} &= \widetilde{\pi}_{i\mathrm{D}j} T_{1j} + (1 - \widetilde{\pi}_{i\mathrm{D}j}) P_{1j} \\ y_{2\mathrm{PD}j} &= \widetilde{\pi}_{i\mathrm{D}j} S_{2j} + (1 - \widetilde{\pi}_{i\mathrm{D}j}) P_{2j}. \end{aligned} \tag{4.19}$$

where $\widetilde{\pi}_{i\mathrm{D}j}$ is Ego i's belief about the probability that Alter cooperates after Ego defected in game j in node PD.

Thus, once $\widetilde{\pi}_{i\mathrm{C}j}$ and $\widetilde{\pi}_{i\mathrm{D}j}$ are specified, the expected outcomes $x_{1\mathrm{PD}j}$, $y_{1\mathrm{PD}j}$, $x_{2\mathrm{PD}j}$, $y_{2\mathrm{PD}j}$ in node PD can be easily calculated. Furthermore, once these expected outcomes are calculated, social orientations in node PD θ_{PD} can be estimated using the same multilevel probit model that is used to estimate social orientations in nodes PD, C, and D (see equation 4.10). We conclude that the real issue is how to specify the beliefs $\widetilde{\pi}_{i\mathrm{C}}$ and $\widetilde{\pi}_{i\mathrm{D}}$. This is ultimately a substantive theoretical issue. In this paper, we compare four alternative assumptions about these beliefs. The first one is empirical beliefs, that is, substituting directly elicited beliefs for $\widetilde{\pi}_{i\mathrm{C}}$ and $\widetilde{\pi}_{i\mathrm{D}}$. In the second alternative

we use Specification B1, the best fitting model for beliefs in nodes DG, C, and D, to *predict* $\widetilde{\pi}_{i\mathrm{C}}$ and $\widetilde{\pi}_{i\mathrm{D}}$, i.e., beliefs are endogenized. In the third and fourth alternatives, we use Specifications B2 and B4, the relaxed interpretations of consensus and rational beliefs, respectively, to *predict* $\widetilde{\pi}_{i\mathrm{C}}$ and $\widetilde{\pi}_{i\mathrm{D}}$.

Specifications

Specification B1.C1 (Reported beliefs): Remember that in the experiment, we elicited subjects' beliefs about the decisions of others in nodes C and D, namely $p_{i\mathrm{C}j}$ and $p_{i\mathrm{D}j}$. To be precise, we asked subjects' guesses about the percentages of others who would cooperate given player 1 cooperated and defected. In Specification B1.C1, we use these self-reported beliefs of subjects, $p_{i\mathrm{C}j}$ and $p_{i\mathrm{D}j}$, as estimates of $\widetilde{\pi}_{i\mathrm{C}j}$ and $\widetilde{\pi}_{i\mathrm{D}j}$. Thus, we simply substitute $p_{i\mathrm{C}j}$ and $p_{i\mathrm{D}j}$ for $\widetilde{\pi}_{i\mathrm{C}}$ and $\widetilde{\pi}_{i\mathrm{D}}$ in (4.18) and (4.19) to calculate expected outcomes $x_{1\mathrm{PD}j}$, $y_{1\mathrm{PD}j}$, $x_{2\mathrm{PD}j}$, and $x_{2\mathrm{PD}j}$ for Ego and Alter.

Specification B1.C2 (Model-based beliefs): Equation (4.16) yields model-based predictions of these beliefs. Note that equation (4.16) expresses $\widetilde{\pi}_{i\mathrm{C}j}$ and $\widetilde{\pi}_{i\mathrm{D}j}$ in terms of the means and variances of a subject's beliefs about others' social orientations in nodes C and D. The means of a subject's beliefs, in turn, depend on subject's own social orientation. In Specification B1.C2, the means and variances of a subject's beliefs about other's social orientations in nodes C and D are specified using equations (4.5) and (4.6): we use Specification B1 to predict beliefs. As the results of Specification B1 show (Table 4.3), there is a strong positive association between the means of beliefs and own social orientations. Thus, although Specification B1.C2 does not directly impose the consensus effect, it can be seen as an application of the (empirically calibrated) consensus effect.

Specification B1.C2 endogenizes beliefs $\widetilde{\pi}_{i\mathrm{C}j}$ and $\widetilde{\pi}_{i\mathrm{D}j}$ ("latent variables") whereas Specification B1.C1 treats beliefs $\widetilde{\pi}_{i\mathrm{C}j}$ and $\widetilde{\pi}_{i\mathrm{D}j}$ as exogenous variables ("data"). Thus, Specification B1.C2 provides a more parsimonious account of the first player's decisions in the sequential PD than B1.C1. Besides this difference in the specification of $\widetilde{\pi}_{i\mathrm{C}j}$ and $\widetilde{\pi}_{i\mathrm{D}j}$, B1.C2 and B1.C1 are identical with respect to how social orientations and beliefs in nodes DG, C, and D are

specified, i.e., for both specifications we use Specification B1 in nodes DG, C, and D for social orientations and beliefs. That is why we use B1.C1 and B1.C2 as names instead of simply C1 and C2.

Specification B2.C3 (Relaxed consensus beliefs): Specification B2.C3 builds on Specification B2. In B2.C3, the means and variances of a subject's beliefs about other's social orientations in nodes DG, C, and D are specified using B2. Remember that B2 imposes directly (a relaxed version of) the consensus effect. The means and variances of a subject's beliefs about others' social orientations, specified via the imposed consensus effect, are used to predict $\widetilde{\pi}_{i\mathrm{C}j}$ and $\widetilde{\pi}_{i\mathrm{D}j}$ in (4.18) and (4.19). Shortly, B2.C3 imposes the consensus effect to beliefs in nodes DG, C, and D, and uses the imposed consensus effect in predicting the PD decisions and social orientations of subjects in node PD.

Specification B4.C4 (Relaxed Bayesian-Nash beliefs): In section 4.5.2 we showed that Bayesian-Nash beliefs, both relaxed and stricter specifications, are rejected. However, we do not know yet the consequences of assuming Bayesian-Nash beliefs for situations where beliefs directly influence decisions. Decisions in node PD are influenced directly by beliefs about the second player's decisions. Thus, predicting $\widetilde{\pi}_{i\mathrm{C}j}$ and $\widetilde{\pi}_{i\mathrm{D}j}$ using Bayesian-Nash beliefs, and in turn, predicting decisions in node PD using these beliefs will give the opportunity for analyzing behavioral consequences of assuming Bayesian-Nash beliefs. In Specification B4.C4, we predict $\widetilde{\pi}_{i\mathrm{C}j}$ and $\widetilde{\pi}_{i\mathrm{D}j}$ using the relaxed interpretation of Bayesian-Nash beliefs, thus using Specification B4. In addition to predicting $\widetilde{\pi}_{i\mathrm{C}j}$ and $\widetilde{\pi}_{i\mathrm{D}j}$, we also use B4 to model beliefs in nodes DG, C, and D. We could also use the stricter version of Bayesian-Nash beliefs, i.e., Specification B5. However, we know that B5 fits data poorly. Moreover, compared to a specification that would use B5, B4.C4 is structurally more comparable to B1.C1, B1.C2, and B2.C3, as these latter models all include error terms in predicting the mean of beliefs just like B4.C4 does.

Methods

The statistical estimation procedure is analogous to the procedure described above for the simultaneous analysis of decisions and beliefs in the first three nodes DG, C, and D. Only now the data are expanded with the four additional decisions of subjects in node PD. The outcomes for Ego and Alter in game j in node PD, $x_{1\text{PD}j}$, $y_{1\text{PD}j}$, $x_{2\text{PD}j}$, $y_{2\text{PD}j}$, are specified in four alternative ways (B1.C1, B1.C2, B2.C3, B4.C4). Each of these alternatives correspond to a different statistical model. In Specification B1.C1, two observed variables are substituted for $\widetilde{\pi}_{i\text{C}j}$ and $\widetilde{\pi}_{i\text{D}j}$, whereas in Specifications B1.C2, B2.C3, and B4.C4 $\widetilde{\pi}_{i\text{C}j}$ and $\widetilde{\pi}_{i\text{D}j}$ are unobserved variables specified as explained above.

Although we discuss B1.C1, B1.C2, B2.C3, and B4.C4 for mainly the PD node, in each of these four alternatives, we estimate the parameters of the model for social orientations in all four decision nodes as well as the parameters of the model for beliefs in the first three decision nodes (DG, C, and D) *simultaneously*. B1.C1, B1.C2, B2.C3, and B4.C4 are identical with respect to social orientations in all four nodes: social orientations in these four nodes are modeled as a four dimensional multivariate normal distribution without any further constraint, i.e., extending Specification A1 including node PD. Moreover, B1.C1 and B1.C2 are also identical with respect to beliefs about others' social orientations in the first three nodes, DG, C, and D: they both use Specification B1 to predict beliefs in the first three nodes. B1.C1 and B1.C2 differ *only* on how $\widetilde{\pi}_{i\text{C}}$ and $\widetilde{\pi}_{i\text{D}}$ are specified. B2.C3 and B4.C4 differ from B1.C1 and B1.C2 on how beliefs in the first three nodes are modeled, i.e., instead of B1, B2.C3 uses B2 and B4.C4 uses B4. Note that although some parts of these three specifications are identical, since estimation is done *simultaneously* for all nodes, preferences and beliefs, parameter estimates can differ between these four specifications, also for the identical parts.

Results

Table 4.4 shows the DIC scores for specifications B1.C1, B1.C2, B2.C3, and B4.C4. The overall DIC scores show that, surprisingly, model-based beliefs (B1.C2) outperform reported beliefs (B1.C1), although the DIC difference is

Table 4.4: DIC statistics for the statistical models that represent Specifications B1.C1, B1.C2, B2.C3, and B4.C4. Each statistical model has two parts, a part for social orientations and a part for beliefs. A DIC for each of these two parts and an overall DIC are provided.

Specification	DIC (preferences)	DIC (beliefs)	DIC (overall)
B1.C1 (Reported)	3253	14020	17270
B1.C2 (Model-based)	3239	14020	17260
B2.C3 (Relaxed consensus)	3230	14040	17270
B4.C4 (Relaxed Bayesian-Nash)	3237	14590	17830

not very high. In other words, the model in which beliefs of player 1 about the decisions of player 2 are predicted using modeled beliefs (B1.C2) fits the data slightly better than the model in which self-reported beliefs are directly used. That is, we can predict PD decisions better from model-based beliefs than elicited beliefs. This is most probably due to the fact that in model-based beliefs we account for various forms of noise and error in beliefs. Self-reported beliefs, however, are contaminated by unsystematic error. Moreover, model-based beliefs yield a more parsimonious model than reported beliefs because in reported beliefs two additional observed variables are included in the model. DIC statistics reflect both fit and parsimony. Note also that the DIC scores of B1.C1 and B1.C2 for the parts that deal with beliefs in the first three nodes DG, C, D are exactly the same. This is not very surprising because models B1.C1 and B1.C2 use the same specification B1 for beliefs in these nodes. The DIC scores of B1.C1 and B1.C2 differ relatively highly for the preference part (3253 versus 3239). This difference is mainly due to social orientations in node PD which again indicates that Specification B1.C2 predicts PD decisions somewhat better with a more parsimonious model compared to Specification B1.C1.

The overall DIC score of Specification B2.C3 is the same as that of B1.C1 and 10 higher than that of B1.C2. Thus, overall, B2.C3 fits equally well as B1.C1 and somewhat worse than B1.C2. The overall DIC difference between B2.C3 and B1.C2 is mainly due to the DIC scores for the part of the models that deal with beliefs (14040 vs. 14020). When the DIC scores for the parts of the models that deal with social orientations are considered, the DIC of

B2.C3 is in fact somewhat lower than the DIC of B1.C2 (3230 vs. 3239). This improvement in DIC is most likely due to the fact that B2.C3 predicts $\widetilde{\pi}_{iCj}$ and $\widetilde{\pi}_{iDj}$—which are used to estimate θ_{PD}—using a more parsimonious specification (B2) than B1.C2 (which uses B1). Shortly, imposing directly the consensus effect deteriorates the fit of the part that deals with beliefs but improves somewhat the fit of the model that deals with preferences.

When its overall DIC is considered, the specification that uses Bayesian-Nash beliefs B4.C4 is again flatly rejected. This clearly shows that subjects' beliefs deviate substantially from rational beliefs assumed within the Bayesian-Nash equilibrium concept. This is reflected with the huge DIC score (14590) of B4.C4 for its part on beliefs. However, to our surprise, when the DIC scores for the part that models social orientations and thus predicts subjects' decisions in nodes DG, C, D, and PD are considered, the fit of B4.C4 does not seem bad. The DIC score of B4.C4 for the preference part is higher than that of B2.C3 but it is very similar to, in fact slightly lower than that of B1.C2 (3237 vs. 3239). This is surprising because it seems that using wrong beliefs to predict decisions in node PD does not deteriorate fit much.[9] We will turn back this issue in the discussion and conclusions session.

Table 4.5 shows the results for the best fitting specification, namely Specification B1.C2, for social orientations, beliefs, and evaluation and response errors in all four nodes. First consider the parameters for the distribution of social orientations. The estimated mean of θ_{PD} is negative. Thus, on average,

[9] A technical reason may have contributed to the surprisingly satisfactory fit of B4.C4 for the preference part. Ideally, in fitting B4.C4 we should have included the variances in beliefs $\widetilde{\sigma}^2_{DG}$, $\widetilde{\sigma}^2_{C}$, $\widetilde{\sigma}^2_{D}$ and variances in social orientations σ^2_{DG}, σ^2_{C}, σ^2_{D}, σ^2_{PD} as *latent variables* in the statistical model with the constraints $\widetilde{\sigma}^2_h = \sigma^2_h$ (beliefs correspond to "true" values). However, we could not implement this in OpenBugs for numerical reasons as we also mentioned above when we discussed the same issue for Specifications B4 and B5. Instead, to fit B4.C4 we plugged in directly the *estimated values* for σ^2_{DG}, σ^2_{C}, σ^2_{D} obtained from Specification A1 as substitutes for $\widetilde{\sigma}^2_{DG}$, $\widetilde{\sigma}^2_{C}$, $\widetilde{\sigma}^2_{D}$. Plugging in pre-estimated values rather than treating variances in beliefs as parameters to be estimated does not deteriorate fit much, as those plugged values are quite similar to "would be" estimates. However, fixing parameters rather than estimating them reduces the complexity of part of the statistical model that deals with social preferences. The DIC statistic depends on both fit and complexity. Thus, had we fitted B4.C4 using the alternative estimation with latent variables, the DIC score of B4.C4 for the preference part would likely be higher. Unfortunately, we have to leave a deeper statistical inquiry for future research.

subjects attach a negative weight to the outcome of the other in node PD. This again supports the claim that being in a PD makes people more competitive than being in a DG. It is important to note that the mean of θ_{PD} is between the means of θ_{C} and θ_{D} for both models. Thus, although we do not find positive reciprocity when we compare nodes DG and C, there is *some* positive reciprocity when we make the theoretically more compelling comparison of nodes C and PD: on average, the theta weight slightly shifts up after the first player cooperated compared to the first player's decision in PD. To demonstrate this more precisely, $\widehat{\Pr}(\theta_{\mathrm{PD}} < 0) = .68$, and so 68% of subjects are expected to have negative social orientations in node PD. Moreover, $\widehat{\Pr}(\theta_{\mathrm{C}} > \theta_{\mathrm{PD}}) = .59$, so 59% of subjects show positive reciprocity. Finally, $\widehat{\Pr}(\theta_{\mathrm{D}} < \theta_{\mathrm{PD}}) = .75$, and so 75% of subjects show negative reciprocity, and $\widehat{\Pr}(\theta_{\mathrm{C}} > \theta_{\mathrm{PD}} > \theta_{\mathrm{D}}) = .36$, thus 36% of subjects show both positive and negative recpirocity. These results point to the existence of strong negative reciprocity and mild positive reciprocity.

It is also interesting to note that the correlation between θ_{PD} and θ_{DG} is smaller than the correlations between θ_{PD} and θ_{C}, and between θ_{PD} and θ_{D}. The variance of the evaluation error in node PD is the highest among all four nodes. We think that this reflects the highest cognitive complexity of the decision problem in node PD compared to nodes C, D, and DG.

The results of Specification B1.C2 that deal with beliefs of subjects about the social orientations of others in nodes DG, C, and D (Table 4.5(b)) are virtually identical to the results of Specification B1 reported in Table 4.3.

4.6 Discussion and conclusions

In this paper, using a within-subjects design, we estimate the utility weights ("social orientations") that the subjects attach to the outcome of their interaction partners in four decision situations: binary Dictator Games (DG), the second player's role in the sequential Prisoner's Dilemma after the first player cooperated (C) and defected (D), and the first player's role in the sequential Prisoner's Dilemma game (PD). In addition, we analyze the relationship between subjects' social orientations and their beliefs about the social orienta-

Table 4.5: Results for Specification B1.C2. Parameters for the distribution of social orientations, beliefs, and evaluation and response errors. Prior specifications are given in Appendix C.2. Posterior means (P.M.) and—in parentheses in (b)—posterior standard deviations (P.SD.) of the parameters. N(belief)= 5.270, N(decision) = 4.650, N(subject) = 155.

(a) social orientations

Distribution of social orientation parameters		
	P.M.	P.SD.
Mean(θ_{DG})	0.113**	0.015
Mean(θ_{C})	-0.160**	0.089
Mean(θ_{D})	-0.606**	0.136
Mean(θ_{PD})	-0.265**	0.091
SD(θ_{DG})	0.164**	0.014
SD(θ_{C})	0.651**	0.097
SD(θ_{D})	0.571**	0.110
SD(θ_{PD})	0.548**	0.111
Corr($\theta_{\mathrm{DG}}, \theta_{i\mathrm{C}}$)	0.592**	0.075
Corr($\theta_{\mathrm{DG}}, \theta_{i\mathrm{D}}$)	0.463**	0.094
Corr($\theta_{\mathrm{DG}}, \theta_{i\mathrm{PD}}$)	0.366**	0.106
Corr($\theta_{\mathrm{C}}, \theta_{i\mathrm{D}}$)	0.800**	0.067
Corr($\theta_{\mathrm{C}}, \theta_{i\mathrm{PD}}$)	0.705**	0.086
Corr($\theta_{\mathrm{D}}, \theta_{i\mathrm{PD}}$)	0.587**	0.104
Evaluation error		
SD($\epsilon_{\mathrm{DG}j}$) = τ_{PD}	0.178**	0.006
SD($\epsilon_{\mathrm{C}j}$)=SD($\epsilon_{i\mathrm{D}j}$) = $\tau_{\mathrm{D}} = \tau_{\mathrm{C}}$	0.953**	0.113
SD($\epsilon_{\mathrm{DG}j}$) = τ_{PD}	1.083**	0.192

(*)** (90%) 95% credibitility interval excludes 0.

(b) beliefs

Means of beliefs :

$$\widetilde{\mu}_{i\mathrm{DG}} = -0.036(0.025) + 0.826(0.147)\theta_{i\mathrm{DG}j} + \eta_{i\mathrm{DG}},\ \mathrm{R}^2 = 0.37,\quad \eta_{i\mathrm{DG}} \sim N(0, 0.177^2(0.019^2))$$

$$\widetilde{\mu}_{i\mathrm{C}} = -0.879(0.172) + 2.135(0.383)\theta_{i\mathrm{C}j} + \eta_{i\mathrm{C}},\ \mathrm{R}^2 = 0.69,\quad \eta_{i\mathrm{C}} \sim N(0, 0.901^2(0.137^2))$$

$$\widetilde{\mu}_{i\mathrm{D}} = -0.142(0.078) + 0.887(0.181)\theta_{i\mathrm{D}j} + \eta_{i\mathrm{D}},\ \mathrm{R}^2 = 0.63,\quad \eta_{i\mathrm{D}} \sim N(0, 0.376^2(0.059^2))$$

Standard deviations of beliefs :

$$\widetilde{\sigma}_{i\mathrm{DG}} = 0.335(0.019),\quad \widetilde{\sigma}_{i\mathrm{C}} = 2.074(0.191),\quad \widetilde{\sigma}_{i\mathrm{D}} = 0.752(0.076)$$

Evaluation and response errors :

$$\widetilde{\tau} = 0.172(0.006),\quad \zeta_{\mathrm{DG}} = 1.200(0.016),\quad \zeta_{\mathrm{C}} = \zeta_{\mathrm{D}} = 0.367(0.095)$$

tions of others in the first three of the decision situations. We first discuss the findings on social orientations, then discuss the findings on subjects' beliefs about others' social orientations.

In line with many studies, (e.g., Schulz & May, 1989; Simpson, 2004; Fehr & Schmidt, 2006) we find significant variation between subjects with respect to social orientations. In addition, social orientations differ significantly between the four decision nodes: the decision context and the relationship "history" between Ego and Alter influence social preferences (Gautschi, 2000; Falk & Fischbacher, 2006; Vieth, 2009). Yet, the social orientations of a subject across the four decision nodes are highly correlated. This shows that social orientations are partially dispositional *traits* and partially *states* influenced by the decision context (Steyer et al., 1999). Furthermore, comparing several specifications of the effects of context and history on social orientations, we find that the effects of context and history on social orientations differ between subjects. Thus, social orientations vary not only between subjects and contexts, but also how much social orientations vary between contexts varies between subjects (see also De Cremer & Van Vugt, 1999; Van Lange, 2000).

We now discuss the nature of the influence of history and decision context on social orientations. First consider the social orientations in the three decision situations in a sequential Prisoner's Dilemma (C, D, and PD). We find that the mean of social orientations in these three decision situations have the following order: $\mu_{\mathrm{C}} > \mu_{\mathrm{PD}} > \mu_{\mathrm{D}}$. This shows that there is both positive and negative reciprocity. The average social orientation is higher after Alter cooperated (positive history) than in the situation where there is no positive or negative history, i.e., when Ego decides as the first player, PD. The average social orientation is by far the lowest after Alter defected (negative history). These findings are in line with past research on history effects (Gautschi, 2000; Falk & Fischbacher, 2006; Vieth, 2009). We also find that negative reciprocity is stronger than positive reciprocity, that is, $\mu_{\mathrm{PD}} - \mu_{\mathrm{D}} > \mu_{\mathrm{C}} - \mu_{\mathrm{PD}}$. This is in line with a recent study which also reports that in one-shot situations negative reciprocity is stronger than positive reciprocity (Al-Ubaydli et al., 2010).

Additionally, we find that the Prisoner's Dilemma context makes subjects more competitive than the Dictator Game context. In fact, the average social

orientation in the Dictator Game is larger than the average social orientation in all three decision situations in the sequential Prisoner's Dilemma. Moreover, the average social orientation is negative in the sequential Prisoner's Dilemma, even after Alter cooperated. It has been shown that subtle features of the game, or how the game has been presented to the subjects, may influence social preferences (e.g., Lindenberg, 2008; Liberman et al., 2004; Burnham et al., 2000). The asymmetric investment game framework that we used in the experiment may have made subjects less pro-social, i.e, decreased their social orientations, compared to the more naturally presented Prisoner's Dilemma. Aksoy & Weesie (2013b) use the same asymmetric investment game framework in a simultaneous play Prisoner's Dilemma and report that the weight for Alter's "payoff" is also negative in that case. Another possible explanation for our subjects being more cooperative in a Dictator Game than in a Prisoner's Dilemma concerns the notion of responsibility as discussed by Camerer (2003) and Blanco et al. (2011). In a Dictator Game, the decision maker is fully responsible for the outcomes for Ego and Alter. Consequently, the decision maker in DG may try to make a fair decision by placing a high weight to the outcome of Alter. In the Prisoner's Dilemma, however, the outcomes for Ego and Alter are determined by the decisions of *both* Ego and Alter, thus the feeling of responsibility is likely lower. This may, in turn, have reduced the weight attached to the outcomes of Alter in the Prisoner's Dilemma. This alternative responsibility explanation can be studied with the following simple "partial Dictator" game. In a partial Dictator game, the player in the Dictator role makes a decision, say in a binary Dictator Game as the ones included in our design. Then, the experimenter tosses a coin. If it is heads, the decision of the Dictator is implemented and if it is tails, one of the binary decisions in the Dictator Game is randomly implemented. In this partial Dictator game, the Dictator is not fully responsible for Ego's and Alter's outcomes. If the responsibility explanation is correct, then social orientations would be lower in the partial Dictator game than in the conventional Dictator Game. We find it interesting to study further how social preferences vary across different types of games, not only Dictator Game and the Prisoner's Dilemma, but also other types of games and which features of games influence social preferences

the most.

In addition to social orientations, we simultaneously analyze beliefs of subjects about others' social orientations in three decision situations, DG, C, and D. We compare several alternative theoretical models on beliefs, e.g., some variants of a consensus model and rational Bayesian-Nash equilibrium beliefs. Our results on beliefs support the model that incorporates a form of consensus effect and reject rational beliefs: there is a strong relationship between own preferences and the mean of beliefs. In addition, the relationship between a subject's own social orientations and her beliefs about others' social orientations is relatively stable across the decision situations. This means that differences in beliefs about others' social orientations between the decision situations are mainly due to differences in own social orientations between these decision situations. However, although own social orientations explain the variance of beliefs to a large extent—in fact explains more than 50% of the interpersonal variance in most cases—, there is still significant unexplained variance. We do not analyze other factors than own social orientations that could potentially explain the variation of beliefs. No clear factors come to mind that can explain the variance of beliefs other than own social orientations. Consequently, we interpret this unexplained variance in beliefs as noise. Irrespective of its source, accounting for this noise proves to be important for model fit. Ignoring this noise in beliefs by, for instance, imposing a strict version of the false consensus effect which omits unexplained interpersonal variation in beliefs, deteriorates fit substantially.

Although the rational beliefs assumption employed in the Bayesian-Nash equilibrium concept is clearly refuted, the evidence for negative consequences of using rational beliefs for the quality of predictions of behavior in our case is not crystal clear. We find that using obviously wrong rational beliefs in predicting the first players' PD choices does not seem to deteriorate much the fit—relative to complexity—of the part of the statistical model for social orientations. (Yet, it does deteriorate a lot the fit of the part of the model for beliefs). This satisfactory fit—relative to complexity—may be due to the fact that the rational belief assumption yields substantial parsimony. The rational beliefs assumption implies that beliefs about the distribution of

others' social orientations correspond to the true distribution of social orientations. Consequently, parameters describing beliefs and parameters describing the distribution of social orientations are collapsed, yielding a huge reduction in model complexity. Yet, when subjects' beliefs are considered, the rational beliefs assumption is flatly rejected. Moreover, imposing directly the consensus effect yields a model as parsimonious as imposing rational beliefs, since under the consensus effect a subject's belief about others' social orientations is a projection of her own social orientation. Additionally, the model that directly imposes the consensus effect yields a better fit than the model that imposes rational beliefs, both in predicting beliefs and behaviors of subjects. Consequently, we recommend analyzing behavior taking ego-centered biases in beliefs into account rather than automatically assuming rational beliefs. At least, the researcher should carefully consider behavioral consequences of assuming rational beliefs, if in fact this assumption is violated.

We acknowledge that allowing for deviations from rational beliefs opens the door for adjusting beliefs *ex-post* to fit any theory to data. That is, one may explain many phenomena by changing the assumptions about subjects' beliefs ex-post. For an extensive discussion on the consequences of dropping the rational beliefs assumption see Morris (1995). However, we do not suggest abolishing the rational beliefs assumption and letting the researcher arbitrarily adjust the assumptions on subjective beliefs of subjects. We recommend replacing the rational beliefs assumption with a well defined model that reflects ego-centered biases in beliefs. Another issue that comes about when one deviates from rational beliefs is higher order beliefs. Higher order beliefs are crucial in modeling behavior in multistage and simultaneous action games. Under the rational beliefs assumption, first and higher order beliefs are all given by a common prior. When one abolishes the rational beliefs assumption, modeling higher order beliefs becomes difficult. Precisely because of this reason, we did not analyze beliefs about the social orientations of others in the first player's role in the sequential Prisoner's Dilemma (node PD). Remember that in order to assess the social orientation θ_{PD} in node PD we had to deal with the *first order* beliefs of subjects about the social orientations of the second players, θ_{C} and θ_{D}. If we want to deal with *beliefs* of subjects about

the θ_{PD} of others, we need to deal with *second order* beliefs: what subjects believe about other first players beliefs about θ_{C} and θ_{D}. We leave a theoretical and empirical treatment of higher order beliefs to future research, as the paper is already dense. Yet, the following *"egocentric"* model will be a good starting point to model higher order beliefs taking ego-centered biases into account (McKelvey & Palfrey, 1992). Remember that we model first order beliefs about others' social orientations as a normally distributed random variable, centered around a subject's own social orientation. In principle, the same ego-centered normal distribution can be used to derive all higher order beliefs. Using a subject's ego-centered normally distributed beliefs to derive all higher order beliefs, an equilibrium analogous to the Bayesian-Nash equilibrium can be calculated. In turn, using the distribution of social orientations and the theoretical model that specifies beliefs about others' social orientations, a *distribution* of Bayesian-Nash equilibria can be obtained, which then can be contrasted with experimental data.

In addition to higher order beliefs, there are other open issues. For instance, in our analyses we focused on a single parameter social orientation model, but had to ignore other types of social preferences, such as inequality aversion (Schulz & May, 1989; Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000; Aksoy & Weesie, 2012b). We expect that ignoring inequality aversion does not influence our results on the social orientations and beliefs about others' social orientations as Aksoy & Weesie (2012b) explicitly show. Yet, it would be interesting to study how history might influence inequality aversion as it is theoretically not clear what reciprocity implies for inequality aversion. For example, it is unclear if a negative history between Ego and Alter makes Ego less inequality averse or a positive history makes Ego more inequality averse. Perhaps, in this case, it makes more sense to distinguish between advantageous and disadvantageous inequality aversion (Fehr & Schmidt, 1999): probably a positive (negative) history will decrease (increase) disadvantageous inequality aversion and increase (decrease) advantageous inequality aversion. Another open issue, not only for this particular paper but for the social preference literature in general, is explaining the variance in social preferences, both as traits and states. We show that social preferences vary between subjects

and contexts. Moreover, the effect of context on social preferences varies between subjects. Yet, we leave providing an explanation for these inter-personal and inter-contextual variances in preferences for future research. A third open issue is the following. Although we showed that the rational beliefs assumption gives a poor description the beliefs of subjects, more work is needed to clearly document consequences of assuming rational beliefs for the predictions of behavior. Among the four decision situations we investigated in this study, DG, C, D, and PD, only in the last one beliefs of subjects directly influence their behavior. Extending our work to other decision situations where beliefs about others' preferences have potentially strong effects on behavior will help study further behavioral consequences of making wrong assumptions about beliefs. Examples of such situations include the proposer behavior in the Ultimatum Game (Güth et al., 1982), the trustor's decision in the Trust Game (e.g., Snijders, 1996), and contribution decisions in Public Goods Games with non-linear production functions (Erev & Rapoport, 1990).

This paper has, what we believe, a methodological strength. Here we present several examples of model testing using hierarchical Bayesian statistical analysis of social preferences and beliefs. Hierarchical Bayesian methods are quite flexible. The simultaneous analysis of own social orientations and beliefs about others' social orientations in several different decision situations is almost impossible within the frequentist paradigm (Aksoy & Weesie, 2013a). Yet, as we show in this paper, such complex analyses can be carried out within the Bayesian statistical framework. Because of their flexibility, in the future we expect to see more applications of Bayesian methods in social science research.

As we showed in this paper and others elsewhere (e.g., Falk & Fischbacher, 2006; Vieth, 2009), in social dilemmas both outcome and process-based social preferences are at work. Moreover, social preferences are not enough to explain behavior in social dilemmas as theoretically behavior in interdependent situations depends on also beliefs about others' decisions. In addition to outcome and process-based social preferences, to account fully for behavior in social dilemmas one needs to tackle beliefs of actors about others' decisions and preferences. We believe that rather than talking about a social preferences model, it makes much more sense to talk about a social preferences-belief

model, where preferences and beliefs are explicitly modeled. As we demonstrate here, accounting for various forms of social preferences and at the same time explicitly modeling beliefs is complex but possible.

Chapter 5

Social motives and expectations in one-shot asymmetric prisoner's dilemmas*

Abstract

We propose a formal-behavioral framework with three components: non-selfish motives, expectations about others' non-selfish motives, a game-theoretic component. For non-selfish motives, three nonstandard utility models representing altruism, inequality aversion, and norms are considered. Expectations are modeled as certain vs. uncertain expectations. The game-theoretic component predicts behavior of actors and actors' expectations about behaviors of others. This framework is applied to asymmetric one-shot prisoner's dilemmas, predictions are tested experimentally. Formal analyses show that asymmetry provides new predictions through which non-standard utility/expectation

*This chapter is written in collaboration with Jeroen Weesie and published in *Journal of Mathematical Sociology* (Aksoy & Weesie, 2013b).We thank Vincent Buskens, Rens van de Schoot, participants of the ICS Cooperative Relations Seminars, the June 2009 Maastrict Behavioral and Experimental Economics symposium, and the June 2008 International Institute of Sociology congress for helpful comments.

models can be distinguished. Empirical tests show that the inequality aversion model does considerably worse than altruistic and normative variants. Statistical tests for own motives, expected motives, and the association between the two are provided, while accounting for decision noise.

5.1 Introduction

Game theory is widely used in sociology to investigate problems of cooperation. The Prisoner's Dilemma Game (PD) is the most extensively used tool for this purpose. The PD is a 2-person game where each actor involved has a cooperative and a defective option. Each actor has an incentive to choose the defective option since no matter what the other actor chooses, the defective choice yields a higher outcome. As decisions are taken individually, actors are expected to end up defecting mutually and to expect that the other actor defects. In game-theoretic language, defection is the dominant strategy for both actors and mutual defection is the Nash equilibrium of the game. However, mutual defection is a Pareto-inferior outcome, i.e., both actors would be better off by joint cooperation which is, however, unstable due to individual incentives for defection. Hence, the PD represents a social dilemma capturing the conflict between individual and collective interests (Kollock, 1998). It should be stressed that this game-theoretic analysis applies to one-shot PDs. When the PD is repeated for an infinite or indefinite number of times, cooperation can be part of equilibrium behavior.

Experimental studies with the PDs, however, show that considerable levels of cooperation are observed even in one-shot situations. In other words, subjects are significantly more cooperative than what the standard application of game theory predicts (Camerer, 2003). These findings call for a revision in one of the core assumptions of the standard application of game theory: What people try to maximize in social interactions may not only be the self-interest as assumed in standard applications. People are also interested in the outcomes of others, e.g., people have altruistic motives, normative concerns etc. In other words, what the subjects are playing in those experiments is not the game that the experimenter assumes they are playing, but another game with

different payoffs which includes the 'subjective' payoffs or values that subjects associate with the outcomes of the 'objective' game. In this direction, non-standard utility models that include non-selfish (social) motives are employed very extensively in recent behavioral economics and rational choice literature (see Fehr & Schmidt, 2006; Fehr & Gintis, 2007). Already three decades ago, scholars of the social psychological research also incorporated social motives in a utility function called the social orientation model (McClintock, 1972; Griesinger & Livingston, 1977). By extending the standard game-theoretic model by assuming such social preferences, one may indeed explain cooperation in one-shot PDs.

In this paper, we carry on rigorous formal analyses and derive testable predictions on how different non-selfish (social) motives influence cooperation in one-shot PDs. One important aspect of this study is that we employ *asymmetric* PDs, whereas past research is typically restricted to symmetric situations (see also: Vogt, 2008). Asymmetric PDs are important in two respects. First, asymmetric PDs are important methodologically. Each observation can be explained by adjusting the assumptions on (non-selfish) preferences of actors *ex-post* (Buskens & Raub, 2013). To avoid this fallacy, one should not only give explanations for well-known empirical regularities, but derive new predictions for new settings and test these predictions (Buskens & Raub, 2013). In symmetric PDs, utility models that incorporate certain types of social motives are successful in explaining behavior. Yet, under asymmetry, new testable predictions are be obtained. Thus, testing models under asymmetry provides a more critical test. In addition, under asymmetry, different utility models yield more distinctive predictions than under symmetry, thus model differentiation is easier. Secondly, asymmetric PDs are also important substantially. Many real life social dilemmas are asymmetric in nature (Aksoy & Weesie, 2009). Understanding the factors influencing cooperation under asymmetric situations will advance our understanding of the cooperation problems in asymmetric social dilemma situations in real life.

Another important aspect of our study is that we pay specific attention to the expectations of actors. We differentiate and analyze two types of expectations. First, we deal with actors' expectations about *social motives* of others.

If an actor has a dominant strategy after taking social motives into account, behavior can be predicted straightforwardly. However, if the transformed subjective utility matrix has no dominant strategy, actors should predict the behavior the other. Predictions about what the opponent would do, in principle, depends on actors expectations of social motives of others. These expectations are particularly important for one-shot interactions since no learning of the motives of the other takes place. Thus, assumptions about actors' expectations influence game-theoretic predictions. Consequently, empirical tests of nonstandard utility models are inconclusive, since observed behavior can be an outcome of expectations about social motives of others rather than social motives *per se*. Fehr & Schmidt (2006) provide a detailed overview of the recent research on non-standard utility models, but fully ignore expectations about social motives of others.[1] Second, we also consider actors' expectations about *behaviors* of others. In the literature, game-theoretic models are usually tested only comparing formal predictions with the observed behavior of actors. Game-theoretic equilibrium predictions are, however, richer, i.e., they involve also actors' expectations about behavior of others (in fact all higher order expectations). Thus, successful theoretical models should account for not only actors' behavior, but also actors' expectations of others' behavior. We extend the application of our formal behavioral model, by including predictions with respect to actors' expectations about the behavior of their interaction partners.

After presenting our formal model and analyses, we provide the details of the experimental procedure through which we test the formal predictions. We then present the statistical analyses of the experimental data. With a Bayesian hypothesis testing procedure, we test our formal predictions on behavior and expected behavior at the aggregate level. With a classical frequentist method, we provide statistical tests for own non-selfish motives, expectations about

[1] This does not apply to intention based models that employ psychological game theory (e.g., Rabin (1993)), which include beliefs about others' intentions in the equilibrium solution. In this paper we don't include intention based models because of their high degree of complexity owing to the use of psychological game theory with multiple free parameters and multiple equilibrium predictions (see Fehr & Schmidt (2006)). Moreover, we investigate simultaneous play one-shot games with limited feedback where there is no opportunity to observe intentions.

others' non-selfish motives, and the association between own and expected non-selfish motives, while accounting for decision noise. A final discussion concludes the paper.

5.2 Theory

Every behavioral theory that aims at describing interdependent behavior in encounters (one-shot interactions) deals either explicitly or implicitly with three issues: preferences of actors, expectations of actors about preferences of others, and a decision-theoretic component. The formal behavioral model that this paper proposes thus has three components.

5.2.1 Preferences: nonstandard utility models

We use nonstandard utility models as micro-theories on preferences. We consider three alternative nonstandard utility models: the social orientation model (McClintock, 1972), an inequality aversion model (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000), and finally a normative model (Hegselmann, 1989; Crawford & Ostrom, 1995; Aksoy & Weesie, 2009). The social orientation model is quite popular in the social psychological literature, and widely applied to explain behavior in PDs (e.g., Simpson, 2004). Inequality aversion models are extremely popular in experimental economics, and extensively used to explain behavior in various interaction situations, but mainly in dictator and ultimatum games (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000). We also included a sociologically relevant normative model, which we think might have promising applications in rational choice theory, especially for social dilemma situations (see Harsanyi, 1980; White, 2004; Aksoy & Weesie, 2009).

Social orientation: Formally, the social orientation utility function $u(x, y; \theta)$ where (x, y) is an outcome allocation for the self (x) and the other (y) is defined as[2]

[2]This representation is mathematically equivalent to the representation $U(x, y; w_i) = w_{i1}x + w_{i2}y$, given that utility is defined up to increasing affine transformations and $w_{i1} > 0$,

$$u(x, y; \theta) = x + \theta y, \qquad \text{with } \theta \in [-1, 1].$$

The parameter θ is called a person's *social orientation* which represents the utility weight for the outcomes of the other. A person with $\theta > 0$ gains positive utility from the outcomes of the other and is called an 'altruist' or a 'cooperator'. If an actor has $\theta \approx 0$, then (s)he is an 'individualist'. Finally if $\theta < 0$, the actor derives disutility from the outcome of the other and is labeled as 'competitive'. The boundary values on θ imply that actors never value the outcomes of others more than their own outcomes neither positively nor negatively. Although this boundary assumption may not hold in certain conditions, e.g., from a biological point of one we may expect a $\theta > 1$ for a grandmother toward her grandchild, we deal with one-shot games (encounters among strangers) and this assumption is less problematic in our case.

Inequality aversion: The inequality aversion utility function that we consider in this paper, $u(x, y; \sigma)$ where (x, y) is an outcome allocation for the self (x) and the other (y) is:

$$u(x, y; \sigma) = x - \sigma|x - y|, \qquad \text{with } \sigma > 0.$$

The first term of the right side of the equation represents objective outcome for the self, the second term represents the utility loss from inequality. The assumption of $\sigma > 0$ implies that people dislike inequality. Although one may set an upper boundary for σ such as 1, which would assure that utility is always increasing in x when $x > y$, in our analyses we do not set such a boundary value. We acknowledge that when $x > y$, utility may decrease in own outcome, x. When $x < y$, the assumption of $\sigma > 0$ ensures that utility is always increasing in x.

Readers may note that this is a simplified version of Fehr & Schmidt utility function that has the restriction $\beta = \alpha = \sigma$. This implies that advantageous and disadvantageous inequality yield the same utility loss. This idea of single

i.e., people prefer more for self than less. The social orientation model is a special case of the utility model of Charness & Rabin (2002).

inequality aversion parameter originates in Bolton & Ockenfels (2000), but we express own outcomes and social inequality in the same unit. We have done theoretical and empirical analyses with the original model of Fehr & Schmidt (1999). But we present the simple form above rather than the original one because our specific formulation avoids the complexities of the Fehr & Schmidt model, yet in the asymmetric PD case yields very similar predictions. Also in terms of empirical success this simple form is not worse than the original one. Thus, we choose this simple form for presentation purposes. Appendix D.2 includes a discussion of the predictions of the original model of Fehr & Schmidt (1999) in our case.

Normative model: The normative utility function $u(x; c; \Delta)$ where x is the outcome and c is the strategy implemented by the actor in the PD is defined as

$$u(x; c; \Delta) = x + \Delta I_{(c=c^*)}.$$

where c^* is the cooperative strategy so that $I = 1$ if actor cooperates, and $I = 0$ if actor defects. This model is a special case of the normative model of Crawford & Ostrom (1995). Δ represents the extra utility gained from obeying the norm of cooperation. Equivalently, people derive negative utility from behaving in conflict with the norm. Or still equivalently, Δ represents the difference in the utilities of obeying and breaking the norm. For a discussion of this model and an empirical application, see Aksoy & Weesie (2009). This model is also in line with the Kantian imperative model (Hegselmann, 1989; White, 2004). The most widely cited version of Kant's categorical imperative is the *formula of universal law* which prescribes that people should act according a maxim whereby at the same time they want this maxim to become a universal law. According to this categorical imperative, cooperation in a social dilemma becomes rational because of the very nature of the social dilemma: a mutually cooperative outcome is better for everyone and thus serve as a categorical imperative, whereas the alternative one, defection is not desired to become a universal strategy.

The social orientation and inequality aversion models are outcome based, i.e., the outcome of the other person directly influences one's utility. In the

normative model, on the other hand, others' outcomes do not influence utility directly. Actors gain utility just by cooperating, i.e., choosing behavior that is part of a Pareto optimal outcome, irrespective of what the other person gets. Of course, the *ordinal* outcomes of the other and the self determine the PD situation, and thus indirectly influence whether actor gains a normative utility in case of cooperation. But besides this, the cardinal value of other's outcome does not enter in the utility function. This is the reason why the model is called *normative*.

Applications of utility models: Once one of these nonstandard utility models is assumed, then the given PD is transformed into an *effective decision matrix* (see Kelley & Thibaut, 1978). We assume that the 'objective' outcomes are common knowledge, both players use the same transformation process and this process is common knowledge as well. However, players may not know each other's social motive parameters, i.e., the strength of each other's social motives. In other words, after assuming a certain nonstandard utility model, we do not further assume that the sizes of the social motive parameters of the players are necessarily equal, i.e., we allow for heterogeneity of actors in terms of the size of their social motives. However, we exclude the possibility that, e.g., one actor is described by the social orientation model whereas her opponent is an inequality averse, and their different utility functions are common knowledge.

In our analyses, we assume that social motives are stable dispositions. This implies that these social motives are not influenced by the context. For example, the social orientation parameters of actors are not influenced by external factors such as the outcomes in the game or the process through which outcomes are allocated. However, the formal model that we propose can easily be extended by altering the social motive parameters in line with theories on the effects of social context on preferences, e.g., Aksoy & Weesie (2009).[3]

[3]In this paper, we also exclude interdependent utility, i.e., a player is interested in the 'utility' of the other player not just the outcomes (Becker, 1993).

5.2.2 Expectations about social motives of others

When an actor has a dominant strategy after matrix transformation, behavior can be predicted without questioning what the actor thinks about the payoffs of the other player. However, when there is no dominant strategy, expectations about the behavior of the other become crucial. An actor's expectations about the behavior of the other depend on actor's expectations about the preferences of the other. In other words, when there is no dominant strategy in the effective decision matrix, actors should consider the effective decision matrix of the other player, which requires expectations about the preferences of the other player. We deal with expectations about the preferences of others with two alternative approaches: certain symmetric expectations and uncertain symmetric expectations.

Certain symmetric expectations (CSE): This approach is in line with what social psychologists call the *false consensus* hypothesis which states that people often expect others to be similar to themselves, e.g., to have similar tastes, similar problem solving strategies etc. (Ross et al., 1977). This is especially true if they don't have sufficient information about other people, which is typically the case in laboratory experiments with anonymous subjects and stranger matching. We model this by assuming that actors expect others to have the same social motive parameters as the self. Thus, actors behave as if they are playing a game with complete information, where the given decision matrix is transformed into an effective decision matrix with a common social motive parameter. For example, within the social orientation model, an actor with θ behaves as if (s)he plays a complete information game with her opponent who transforms the given decision matrix with the same θ.

Uncertain symmetric expectations (USE): In this second approach, actors are uncertain about the preferences of their interaction partners. Consequently, expectations are modeled as in games with incomplete information. Social motive parameter of an actor is his/her private information, and nature randomly and independently selects the type of the other player. In this application, a type of an actor is simply his/her social motive parameter value,

and the population consists of different types of actors. The types in the population have a certain probability distribution F. This type distribution and the matching process are assumed to be common knowledge. That is, all people's social motive parameters are drawn independently from F, and actors only know their own motives, i.e., f. Thus, expectations of an actor about the social motive parameter of the interaction partner is described by F. In addition, player 1's expectations about the expectations of player 2 with respect to the social motives of player 1 as well as all higher order expectations are also given by the same F.

CSE and USE can be seen as the two extremes. The latter provides a full-rationalistic model and is the assumption in Harsanyi's (1968) solution to games with incomplete information which is the standard way of modeling expectations in behavioral game theory (e.g., Fehr & Schmidt (1999)). The former is a simple heuristic, based on social psychological findings. One may also use hybrid approaches. For example, the subjective probability distribution of social motives of others can be modeled as a function of the social motive of the self, where subjective probability distribution is peaked at the social motive parameter of the self. Such an extension will complicate analyses significantly, while in our case it would almost always yield very similar results as CSE.

5.2.3 Game theory as the decision model

As the decision-theoretic component, we retain the assumption of strategic rationality of game theory. In other words, we assume that given the social motives and expectations about the motives of others, actors maximize utility by manifesting equilibrium behavior. When CSE is assumed, the equilibrium is the classical Nash equilibrium. In the USE case, which results in games with incomplete information, Bayesian-Nash equilibrium concept is applied (Harsanyi, 1968).

Equilibrium of a game yields two kind of predictions. First, it predicts behavior of actors, second, it predicts expectations about the behavior of in-

teraction partners.[4] For example if the game-theoretic solution prescribes mutual cooperation to be the equilibrium of a given game under certain conditions, then this solution predicts cooperative behavior of the actor, as well as his/her expectation about the cooperative behavior of the interaction partner. In other words, the formal behavioral model that we propose which includes social motives, expectations about social motives, and the game-theoretic decision model constitutes the explanatory factors and *behavior* and *expected behavior* in the prisoner's dilemma are the phenomena to be explained.

Equilibrium selection: A technical issue that is worth mentioning in our game-theoretic application is the equilibrium selection criterion. As shown below, in some conditions, multiple equilibria exist. This hinders empirical testing. There are equilibrium selection criteria proposed in the literature, however, there is still much dispute among theorists (see Elberfeld, 1997). In our case, equilibrium selection problems occur only in the inequality aversion model.[5] In our study, in case of multiple equilibria, we select the equilibrium which maximizes the joint probability of cooperation. Readers can find a discussion in Appendix D.4 about the consequences of applying risk dominance and Pareto dominance criteria (Harsanyi & Selten, 1988).

Parallel Prisoner's Dilemma (PPD): In this paper, we restrict our analyses to the Parallel PD (PPD). PPD has an outcome structure where $T_i - R_i = P_i - S_i = X_i$. The difference between the outcomes for the self when (s)he chooses to cooperate and defect is the same, regardless of what his/her opponent does (see Table 5.1). This property has both theoretically and practically important outcomes, thus we do not generalize our formal and empirical findings that we obtain from PPDs to more general PDs, let alone to more general 2-by-2 games. Although providing solutions that account for more general non-parallel conditions is straightforward, it requires a lot of

[4] In fact it also predicts all higher order expectations about behavior.

[5] In the social orientation and normative models, multiple equilibria may arise, but only in cases where an actor is indifferent between cooperating and defecting, which yields an infinite number of equilibria. These cases can be ignored since they happen only for actors with very specific values of social motive parameters.

Table 5.1: (a) Prisoner's dilemma where $T_i > R_i > P_i > S_i$; (b) Parallel Prisoner's Dilemma (PPD), where $X_i = T_i - R_i = P_i - S_i > 0$, thus $Y_i = R_i - S_i = T_i - P_i > 0$.

(a)

	Cooperate	Defect
Cooperate	R_1, R_2	S_1, T_2
Defect	T_1, S_2	P_1, P_2

(b)

	Cooperate	Defect
Cooperate	R_1, R_2	S_1, $R_2 + X_2$
Defect	$R_1 + X_1$, S_2	$S_1 + X_1$, $S_2 + X_2$

space and is not relevant for the empirical tests that we provide in this paper. PPDs are very common in the literature and has wide application areas. For example, simultaneous one-shot favor exchange games (Van der Heijden et al., 2001), simultaneous social support games (Vogt, 2008), any public good game with a linear production function (Kollock, 1998), resource dilemma games with dichotomous choice (Dawes, 1980) and many more such applications are in fact parallel PDs. Moreover, besides those games which are structurally PPDs, in most studies, scholars use PPDs rather than non-parallel PDs. In this paper, we also show that using PPDs rather than nonparallel PDs may have important consequences.

5.3 Predictions

5.3.1 Social orientation

Certain symmetric expectations (CSE)

To define CSE formally, let θ_e be the social orientation parameter of the focal actor and θ_a be that of the interaction partner. For convenience, think that ego is the row player (player 1), and alter is the column player (player 2). Our application of CSE assumes that ego acts as if $\theta_a = \theta_e$ and this is common knowledge. In this case, C is dominant for ego if $\theta_e > \max(\frac{T_e - R_e}{R_a - S_a}, \frac{P_e - S_e}{T_a - P_a})$; and D is dominant if $\theta_e < \min(\frac{T_e - R_e}{R_a - S_a}, \frac{P_e - S_e}{T_a - P_a})$. Note that for the PPD, $\frac{T_e - R_e}{R_a - S_a} = \frac{P_e - S_e}{T_a - P_a}$, since $(T_i - R_i) = (P_i - S_i) = X_i$ and thus $Y_i = (R_i - S_i) = (T_i - P_i)$

for $i = e, a$. Let $\pi_e(\theta_e; CSE)$ be the probability that ego chooses to cooperate, and let $\pi_{a|e}(\theta_e; CSE)$ be ego's expectation about the probability that alter cooperates under CSE. Then, the Nash-equilibrium solution for the social orientation model with CSE prescribes that

$$\begin{aligned}
\pi_e(\theta_e; CSE) &= \begin{cases} 1 & \text{if } \theta_e \geq \frac{X_e}{Y_a} \\ 0 & \text{if } \theta_e < \frac{X_e}{Y_a} \end{cases} \quad \text{(Ego}'\text{s behavior)} \\
\pi_{a|e}(\theta_e; CSE) &= \begin{cases} 1 & \text{if } \theta_e \geq \frac{X_a}{Y_e} \\ 0 & \text{if } \theta_e < \frac{X_a}{Y_e} \end{cases} \quad \text{(Ego}'\text{s expect. about alter}'\text{s behavior)}
\end{aligned} \tag{5.1}$$

According to (5.1), an actor with $\theta_e \geq \max(\frac{X_e}{Y_a}, \frac{X_a}{Y_e})$ cooperates and expects alter to cooperate as well. An actor with θ_e, $\frac{X_e}{Y_a} \leq \theta_e < \frac{X_a}{Y_e}$ cooperates but expects alter to defect, and if $\frac{X_a}{Y_e} < \theta_e \leq \frac{X_e}{Y_a}$ defects but expects alter to cooperate. Note that these latter situations arise only under asymmetry, since in symmetric PDs, $\frac{X_e}{Y_a} = \frac{X_a}{Y_e}$. Finally an actor with θ_e where $\theta_e < \min(\frac{X_e}{Y_a}, \frac{X_a}{Y_e})$, defects and expects alter to defect as well.

(5.1) involves predictions at the micro level. To predict the behavior of a single individual, information about his/her θ parameter is needed. However, at the aggregate level, comparative statics can easily be derived from (5.1): cooperation rates decrease in $\frac{X_e}{Y_a}$, and the proportion of people who expect their interaction partners to cooperate decreases in $\frac{X_a}{Y_e}$. Another interesting comparative statics prediction of (5.1) is about the association between own behavior and expected behavior. Under symmetry, where $\frac{X_e}{Y_a} = \frac{X_a}{Y_e}$, (5.1) predicts a perfect positive correlation between own behavior and expected behavior at the population level, since if ego cooperates (defects), expects alter to cooperate (defect) as well. However, as the difference between $\frac{X_e}{Y_a}$ and $\frac{X_a}{Y_e}$ increases, that is the more the game becomes asymmetric, the correlation coefficient decreases. The reason is, those subjects with θ values between $\frac{X_e}{Y_a}$ and $\frac{X_a}{Y_e}$ expect alter to choose the other option than they themselves do. Note that according to (5.1), there is no direct association between $\pi_e(\theta_e; CSE)$ and $\pi_{a|e}(\theta_e; CSE)$, since either cooperation or defection is dominant depending on θ_a. The correlation is due to the expectation that $\theta_a = \theta_e$. This finding is

particularly important and we will turn back to this later on.

Uncertain symmetric expectations (USE)

Now, let's examine the second model of expectations. Under USE, ego knows his/her own θ, θ_e, but is uncertain about the θ of alter, θ_a. The θ of alter is an independent draw from the population with a certain distribution. This distribution and the matching process are assumed to be common knowledge. Assume that $\theta \sim F$. To avoid randomized strategies, we assume that F is atomless. Let θ_e^* be the θ value which would make ego indifferent between cooperating and defecting in terms of expected utility. Then, ego cooperates (plays C) if $\theta_e \geq \theta_e^*$, and defects otherwise, given that $\theta_e^* > 0$. Similarly, alter cooperates if $\theta_a \geq \theta_a^*$, given that $\theta_a^* > 0$. θ_e^* depends on the probability of alter's behavior, which can be derived from F. More precisely, the expectation of ego about the probability that alter defects is $Pr(\theta_a < \theta_a^* | F) = F(\theta_a^*)$. Then, straightforward calculation shows that:

$$\theta_e^* = \frac{T_e - R_e - F(\theta_a^*)[(T_e - R_e) - (P_e - S_e)]}{R_a - S_a - F(\theta_a^*)[(P_a - S_a) - (T_a - R_a)]} \tag{5.2}$$

Similarly, θ_a^* and, thus, $F(\theta_a^*)$ depends on θ_e^* as:

$$\theta_a^* = \frac{T_a - R_a - F(\theta_e^*)[(T_a - R_a) - (P_a - S_a)]}{R_e - S_e - F(\theta_e^*)[(P_e - S_e) - (T_e - R_e)]} \tag{5.3}$$

Here we are dealing with an implicit solution. (5.2) and (5.3) define a relationship between θ_e^* and θ_a^* via $F(\theta_a^*)$ and $F(\theta_a^*)$. (5.2) and (5.3) can easily be written as a single fixed-point equation for θ_a and the distribution function F. Thus, solving for θ_e^* and θ_a^* requires two scalar equations that depend on the outcomes and F. However, these equations simplify in case of the PPD as:

$$\theta_e^* = \frac{X_e}{Y_a}, \qquad \theta_a^* = \frac{X_a}{Y_e}. \tag{5.4}$$

We conclude that the Bayesian-Nash equilibrium solution for the social orientation model with USE for $\pi_e(\theta_e; USE; F)$, the probability that an actor

cooperates, and $\pi_{a|e}(\theta_a; USE; F)$, ego's expectation about the probability that alter defects can be written as:

$$\begin{aligned} \pi_e(\theta_e; USE; F) &= \begin{cases} 1 & \text{if } \theta_e \geq \frac{X_e}{Y_a} \\ 0 & \text{if } \theta_e < \frac{X_e}{Y_a} \end{cases} && \text{(Behavior)} \\ \pi_{a|e}(\theta_a; USE; F) &= 1 - F(\tfrac{X_a}{Y_e}) && \text{(Expected behavior)} \end{aligned} \tag{5.5}$$

Equation (5.5) yields important predictions. First, the behavior of actors derived by assuming USE does not depend on F for the parallel PD. In fact in this case, the two models of expectations, CSE and USE yield the same predictions; cooperation decreases in $\frac{X_e}{Y_a}$. Different from CSE, however, ego's expectations about the behavior of alter do depend on F. Yet, at the population level, comparative statics predictions can be derived without assuming a specific form of F: egos' expectations about alters' cooperation decreases in $\frac{X_a}{Y_e}$. Thus, also for expected behavior, CSE and USE yield the same comparative statics predictions. The predictions of CSE and USE differ, though, with respect to the association between own behavior and expected behavior. CSE predicts a positive correlation between own behavior and expected behavior, yet USE predicts zero correlation between own behavior and expected behavior, since given the outcomes in the game ($\frac{X_e}{Y_a}$ and $\frac{X_a}{Y_e}$), $\pi_{a|e}(\theta_a; USE; F)$ will be the same for each actor in the population.

To sum up, the predictions of social orientation model for the PPD are mostly robust with respect to assumptions on expectations about the social motives of others. More precisely, comparative statics predictions on behavior and expected behavior are the same for CSE and USE, yet they differ with respect to the predicted association between own behavior and expected behavior.

5.3.2 Inequality aversion

In this section, we report our analyses for a special case where $T_i - S_j > |P_i - P_j|$ for $i \neq j; i, j = e, a$, which typically holds in experimental literature as well as

in our experiments. For the general case, depending on the outcome structure of the game, the effective decision matrix may have many different equilibria. Accounting for all those cases requires space and not of a primary concern, since they are not relevant for our empirical tests. Moreover, the condition $T_i - S_j > |P_i - P_j|$ is satisfied in almost all PDs used in the literature and our experiments. Under this condition, predictions are more manageable; DD is always a Nash-equilibrium of the effective decision matrix. However, under certain conditions, playing C can also be part of equilibrium behavior. These conditions are given below.

Certain symmetric expectations (CSE)

Under CSE, ego acts as if $\sigma_a = \sigma_e$ and this is common knowledge. First let's define ϕ_e as

$$\phi_e = \frac{T_e - R_e}{T_e - S_a - |R_e - R_a|}$$

This ratio can be seen as the cooperation threshold for ego, if ego thinks alter cooperates. That is, if ego knows that alter cooperates, ego chooses to cooperate if $\sigma_e \geq \phi_e$, given that $\phi_e > 0$. If $\phi_e < 0$, then cooperation would be attractive for ego if $\sigma_e < \phi_e$, which is impossible due to the assumption that $\sigma > 0$. Thus if $\phi_e < 0$, ego defects for sure whatever the alter does. Similarly, ϕ_a is defined as:

$$\phi_a = \frac{T_a - R_a}{T_a - S_e - |R_e - R_a|}$$

Under CSE, $\sigma_a = \sigma_e$, and CC is equilibrium if $\sigma_e \geq \max(\phi_e, \phi_a)$, given that $\min(\phi_e, \phi_a) > 0$. If $\min(\phi_e, \phi_a) < 0$, then the single equilibrium is DD. The solution of the inequality aversion model with CSE for $\pi_e(\sigma; CSE)$, the probability that ego cooperates, and $\pi_{a|e}(\sigma_e; CSE)$, ego's expectation for the probability that alter cooperates can be written as:

$$
\begin{aligned}
&\text{if } \min(\phi_e, \phi_a) > 0 \\
&\qquad \pi_e(\sigma_e; CSE) \quad = \begin{cases} 1 & \text{if } \sigma_e \geq \max(\phi_e, \phi_a) \\ 0 & \text{otherwise} \end{cases} \qquad \text{(Beh.)} \\
&\qquad \pi_{a|e}(\sigma_e; CSE) \quad = \begin{cases} 1 & \text{if } \sigma_e \geq \max(\phi_e, \phi_a) \\ 0 & \text{otherwise} \end{cases} \qquad \text{(Exp.)} \\
&\text{if } \min(\phi_e, \phi_a) < 0 \\
&\qquad \pi_e(\sigma_e; CSE) \quad = 0 \qquad \text{(Beh.)} \\
&\qquad \pi_{a|e}(\sigma_e; CSE) \quad = 0 \qquad \text{(Exp.)}
\end{aligned}
\tag{5.6}
$$

Comparative statics can easily be derived from (5.6): cooperation rates decrease in $|R_e - R_a|$, and the proportion of people who expect their interaction partner to cooperate also decreases in $|R_e - R_a|$. Moreover, cooperation rates decrease in $X_e = T_e - R_e$ and increase in $T_e - S_a$ if $\phi_e > \phi_a$, otherwise cooperation rates decrease in $X_a = T_a - R_a$ and increase in $T_a - S_e$. The same is true for the proportion of people who expect alters to cooperate. Note that if $\min(\phi_e, \phi_a) < 0$, or equivalently if $|R_e - R_a| > \min(|T_e - S_a|, |T_a - S_e|)$, then ego defects for sure and expects alter to defect as well. Finally, (5.6) predicts always a perfect positive correlation between the behavior of ego, and ego's expectation about the behavior of alter. That is, if ego cooperates, she also expect alter to cooperate as well, and if ego defects, she expect alter to defect as well, whatever the outcomes in the game are.

Uncertain symmetric expectations (USE)

Assume that $\sigma_i \sim F$ for $i = e, a$, and F is atomless. Ego cooperates if $\sigma_e > \sigma_e^*$ and defects otherwise, given that $\sigma_e^* > 0$. Similarly, alter cooperates if $\sigma_a > \sigma_a^*$ and defects otherwise, given that $\sigma_a^* > 0$, where

$$
\sigma_e^* = \frac{T_e - R_e}{T_e - S_a - |R_e - R_a| - F(\theta_a^*)(T_e - S_a - |R_e - R_a| + T_a - S_e - |P_e - P_a|)} \tag{5.7}
$$

$$\sigma_a^* = \frac{T_a - R_a}{T_a - S_e - |R_e - R_a| - F(\theta_e^*)(T_a - S_e - |R_e - R_a| + T_e - S_a - |P_e - P_a|)} \tag{5.8}$$

(5.7) and (5.8) do not simplify further. Consequently, the Bayesian-Nash equilibrium solution for $\pi_e(\sigma; USE; F)$ and $\pi_{a|e}(\sigma_e; USE; F)$ is:

$$\begin{aligned}
&\text{if } \min(\sigma_e^*, \sigma_a^*) > 0 \\
&\qquad \pi_e(\sigma_e; USE; F) &&= \begin{cases} 1 & \text{if } \sigma_e \geq \sigma_e^* \\ 0 & \text{otherwise} \end{cases} && \text{(Beh.)} \\
&\qquad \pi_{a|e}(\sigma_e; USE; F) &&= 1 - F(\sigma_a^*) && \text{(Exp.)} \\
&\text{if } \min(\sigma_e^*, \sigma_a^*) < 0 \\
&\qquad \pi_e(\sigma_e) &&= 0 && \text{(Beh.)} \\
&\qquad \pi_{a|e}(\sigma_e) &&= 0 && \text{(Exp.)}
\end{aligned} \tag{5.9}$$

Now, the solution stated by (5.9) is significantly more complex than the solution for the inequality aversion model with CSE. Deriving comparative statics from (5.9) is not straightforward since in equations (5.7) and (5.8), we are dealing with an implicit solution of σ^*. Comparative statics can be derived either using implicit function theorem or writing (5.7) and (5.8) as a single fixed point equation for σ^* and F. However, those comparative statics are non-monotonic, i.e., in addition to the outcomes in the game, they depend on F, as well as the values of σ_e^* and σ_a^*. To be more precise, $\pi_e(\sigma_e; USE; F)$ decreases in X_e and $|R_e - R_a|$, increases in $T_e - S_a$ and $P_e - P_a$ if $\sigma_e^* \sigma_a^* f(\sigma_e^*) f(\sigma_a^*) > \frac{X_e X_a}{a^2}$, given that $\min(\sigma_a, \sigma_a) > 0$, where $a = T_a - S_e - |R_e - R_a| + T_e - S_a - |P_e - P_a|$. If $\sigma_e^* \sigma_a^* f(\sigma_e^*) f(\sigma_a^*) < \frac{X_e X_a}{a^2}$, the comparative statics predictions are reversed. Later in the paper, when we test these formal predictions experimentally, we provide solutions for (5.9) for a number of different games and different type-F distributions, thus the predictions of the inequality aversion model with USE will be clearer. The most important theoretical finding is, the predictions of USE and CSE differ. Most simply, the solutions of USE depend on F, as well as $|P_e - P_a|$, and comparative statics predictions are non-monotonic, which are

not the case for CSE solutions. Thus, the predictions of the inequality aversion model are not robust with respect to the assumptions on actors' expectations about social motives of others. Finally, in contrast to CSE,(5.9) predicts a zero correlation between own behavior and expected behavior, given the outcomes of the game.

5.3.3 Normative model

By now, the applications of CSE and USE should be clear, thus we directly provide the solutions.

Certain symmetric expectations (CSE)

Assume that ego acts as if $\Delta_a = \Delta_e$ and this is common knowledge. Then, the probability that ego cooperates,$\pi_e(\Delta_e; CSE)$, and the probability that ego expect alter to cooperate, $\pi_{a|e}(\Delta_e; CSE)$, in the PPD are:

$$\begin{aligned} \pi_e(\Delta_e; CSE) &= \begin{cases} 1 & \text{if } \Delta_e \geq X_e \\ 0 & \text{if } \Delta_e < X_e \end{cases} \quad \text{(Behavior)} \\ \pi_{a|e}(\theta_e; CSE) &= \begin{cases} 1 & \text{if } \Delta_e \geq X_a \\ 0 & \text{if } \Delta_e < X_a \end{cases} \quad \text{(Expected behavior)} \end{aligned} \tag{5.10}$$

(5.10) implies that cooperation rates decrease in X_e, and the proportion of people that expect alter to cooperate decreases in X_a. Under symmetry where $X_e = X_a$, (5.10) predicts a perfect positive correlation between own behavior and expected behavior, and as the difference between X_e and X_a increases, this correlation decreases, but remains positive.

Uncertain expectations

Assume that $\Delta_i \sim F$ for $i = e, a$, and F is atomless. We define a Bayesian-Nash equilibrium Δ_i^* in which player i cooperates (plays C_i) if $\Delta_i \geq \Delta_i^*$, and i defects otherwise. Denote $F(\Delta_i^*)$ for the probability that i defects. The

strategy-vector $\tilde{\pi}(\Delta_e^*, \Delta_a^*)$ is a *Bayesian-Nash equilibrium* in the PPD, if and only if

$$\Delta_i^* = X_i \tag{5.11}$$

Thus, the solution for $\pi_e(\Delta_e; USE; F)$, and $\pi_{a|e}(\Delta_e; USE; F)$ is

$$\begin{aligned} \pi_e(\Delta_e; USE; F) &= \begin{cases} 1 & \text{if } \Delta_e \geq X_e \\ 0 & \text{if } \Delta_e < X_e \end{cases} && \text{(Behavior)} \\ \pi_{a|e}(\Delta_a; USE; F) &= 1 - F(X_a) && \text{(Expected behavior)} \end{aligned} \tag{5.12}$$

Similar to the social orientation model, for the normative model, ego's behavior does not depend on F for the PPD. Cooperation rates decrease in X_e monotonically. At the individual level, expectations about the behavior of alter do depend on F, although at the population level CSE and USE yield the same comparative statics: the proportion of people who expect alter to defect decreases in X_a. However, similar to the social orientation model, CSE and USE solutions for the normative model differ with respect to the association between ego's behavior and ego'e expectation about alter's behavior. CSE predicts a positive correlation, yet USE predicts that given X_a and X_e, there is no correlation.

5.4 Experimental design and procedure

Subjects were recruited using the Online Recruitment System for Economic Experiments (ORSEE; Greiner (2004)) to participate in a study of "decision making in different situations". 134 subjects participated in the experiment in one of the 8 sessions that were held with 14 to 20 respondents in December 2007 in the ELSE lab at Utrecht University.

We followed standard procedures of experimental economics (e.g., provision of monetary payoffs, real but anonymous partners, no deception of subjects etc., see Friedman & Cassar (2004)). To express the decision situation PPD to the subjects in a comprehensible way, a variant of the *asymmetric investment*

Table 5.2: Prisoner's Dilemma and Asymmetric Investment Game.

(a) Prisoner's Dilemma where $T_i > R_i > P_i > S_i$

	Cooperate	Defect
Cooperate	R_1, R_2	S_1, T_2
Defect	T_1, S_2	P_1, P_2

(b) Asymmetric Investment Game where $T_i - R_i = P_i - S_i$

	Invest	Not Invest
Invest	$\lambda_1\kappa(\mu_1+\mu_2)$, $\lambda_2\kappa(\mu_1+\mu_2)$	$\lambda_1\kappa\mu_1$, $\mu_2+\lambda_2\kappa\mu_1$
Not Invest	$\mu_1+\lambda_1\kappa\mu_2$, $\lambda_2\kappa\mu_2$	μ_1, μ_2

game framework proposed in Aksoy & Weesie (2009) is used (Table 5.2). In the two person asymmetric investment game, the actors have budgets of μ_i, $i = 1, 2$. Each actor can invest or decide not to invest his/her budget. The total investment grows at a rate $\kappa > 1$ and is redistributed to the actors. The first (second) actor gets a share of λ_1 (λ_2) from the common budget where $\lambda_1 + \lambda_2 = 1$, regardless of their investment decisions. Under certain restrictions on the parameters μ, λ, and κ, the asymmetric investment game is a Prisoner's Dilemma, where mutual investment is Pareto-optimal, while each actor has an incentive not to invest. Moreover, this variant of the asymmetric investment game is a PPD. There is almost infinite number of ways of creating asymmetry in the PD, e.g., introducing asymmetry in one cell, some cells, or in all cells in varying degrees. The asymmetric investment game systematizes the asymmetry with only three parameters. Moreover, with this conceptual framework, asymmetry can be expressed to subjects in an easy and sensible way using the parameters μ, λ, and κ.

The experiment employs a full factorial $\mu(3) \times \lambda(3)$ within subjects design with nine asymmetric investment games shown in Table 5.3. In those nine games, total budget $\mu_1 + \mu_2$ is fixed at 1000, and κ is fixed at 1.5. Thus, those nine games can be represented by only two parameters, μ_1 and λ_1. In

Table 5.3: 9 asymmetric investment games (Γ), $\mu_2 = 1000 - \mu_1$, $\lambda_2 = 1 - \lambda_1$, κ is fixed at 1.5.

		λ_1		
		0.4	0.5	0.6
μ_1	400	Γ_1	Γ_2	Γ_3
	500	Γ_4	Γ_5	Γ_6
	600	Γ_7	Γ_8	Γ_9

the instructions, the one-shot nature of the decision situation was emphasized. Subjects went through two examples and were checked that they understood the task before real decision making. Throughout the experiment, each subject made ten game play decisions with stranger matching. Feedback about the decision of others was provided only after the 5th period and after the 10th period where a period comprises a game-play with a stranger. The full symmetric game (Γ_5) was played twice: one before the first feedback and one after the first feedback. After the first five periods, but before the feedback, subjects were asked to state their expectations about the behavior of their interaction partners in the first five games. Expectations about the behavior of others in the first five games were asked only after subjects selected their own behavior in these games, to avoid the possibility that asking about expectations influences own behavior (Croson, 2000; Gaechter & Renner, 2010). Moreover, the order of asking about expectations and own behavior was found to have no effect on expectations (Messe & Sivacek, 1979; Iedema & Poppe, 1994a). As in Bellemare et al. (2008), in contrast to own behavior, we chose not to incentivize expectations about the behavior of others based on several studies which show that incentivizing expectations do not yield significantly different responses (e.g., Friedman & Massaro, 1998; Gaechter & Renner, 2010). The five games played before and after the first feedback and the order of these games were varied in four factors. At the end of the experiment, subjects were paid their earnings in cash where 400 points were worth 1 Euro. On average, subjects earned 12 Euros. The experiment took place on computers with the software z-tree (Fischbacher, 2007) and the whole procedure took 1 hour.

5.5 Methods and results

We test the predictions of nonstandard utility-expectation models in two steps. In these steps, we employ a combination of Bayesian statistics as well as more conventional frequentist methods. Employing both Bayesian and frequentist methods facilitates hypothesis testing as we demonstrate below. At the first step, we compare the comparative statics predictions of the utility/expectation models with the aggregate experimental data. Applying the formal analyses presented in section 3, we obtain for each utility-expectation model a ranking of the nine asymmetric investment games used in the experiment, with respect to the predicted frequency of cooperative behavior. In addition, each of these utility/expectation models predicts a ranking of these nine games in terms of subjects' expectations about the cooperative behaviors of their interaction partners. Thus in the first step, using a Bayesian model selection method, we assess the overall fit of the utility-expectation models by comparing their ranking predictions with aggregate data. After eliminating poorly fitting models at this step, in a second step we study individual level behavior, using a special probit regression analysis similar to Palfrey & Prisbrey (1997). This probit analysis enables us to combine the two well fitting models from the Bayesian analysis of aggregate data into one encompassing model. Consequently, we perform subsequent statistical tests involving the means and standard deviations of social motive parameters of both own motives and expectations about motives of others, while accounting for some sort of decision noise. Finally, we analyze the association between own social motives and expected social motives, providing a statistical test for the two alternative models of expectations, CSE and USE.

5.5.1 Preliminary analyses

Table 5.4 provides the threshold values of the social motive parameter for each utility-expectation model, so that an actor with a social motive parameter exceeding this threshold is expected to cooperate. For example, consider the social orientation model. In Γ_1, the social orientation model predicts that actors with θs larger than 0.44 will cooperate, otherwise defect. Similarly,

the threshold for Γ_2 is 0.33. Thus, everyone with $\theta \geq 0.44$ will cooperate both in Γ_1 and Γ_2, those with $0.33 \leq \theta < 0.44$ will cooperate only in Γ_2, and those with $\theta < 0.33$ in neither game. At the aggregate level, the social orientation model predicts that the number of subjects who cooperate in Γ_2 will be higher than that in Γ_1. In the same manner, all these nine games can be ranked per utility-expectation model. An ∞ sign indicates that the solution is defection for sure, and zero cooperation is predicted. The predicted rankings of those games are also shown in the table, with 1=the game with the highest cooperation rate, and ties marked with an '='. For some models, some games are tied, e.g., for the social orientation model Γ_1 and Γ_4. Below we explain how we handle ties. As we showed above, for the social orientation and the normative models, behavioral predictions of CSE and USE are the same. However, for the inequality aversion model, predictions differ for CSE and USE solutions (see the predicted rankings for CSE and the two USE solutions in Table 5.4).

Table 5.4 also includes observed cooperation rates and the Spearman's rank correlations between predicted and observed rankings. For the inequality aversion model, we obtained predictions for a variety of additional type distributions than those presented in Table 5.4. See Appendix D.1 for some examples and a discussion. The general pattern for the inequality aversion is as follows. First, for any possible distribution, the inequality aversion model predicts that the single equilibrium is DD in Γ_1 and Γ_9. For reasonably inequality aversive populations, CC can be an equilibrium only in the three games where equal outcomes are obtained in case of CC, that is, in games where $\lambda_1 = \lambda_2 = 0.5$. For all other games, DD is the single equilibrium, since cooperation yields inequality, even in case of mutual cooperation. Cooperation can be part of the equilibrium behavior in games where $\lambda_1 \neq \lambda_2$, only if one assumes a population with extreme inequality aversion, as shown in the last two columns of Table 5.4. But in such a highly inequality aversive population, the solution requires also very high levels of cooperation, which is not observed in our data. Thus, for any type distribution, the predictions do not approximate the observed pattern in our data. Nevertheless, for presentation purposes, in subsequent analyses we choose to report results for the two dis-

tributions included in Table 5.4, one with moderate inequality aversion, and one with extreme inequality aversion.

The predictions of the normative and the social orientation models, however, seem more promising. The likelihood of cooperation increases in λ, which is in line with the predictions of the social orientation model and the normative model. There is a slight tendency in the data that cooperation decreases in μ, which is in line with the normative model, while the social orientation model predicts that the likelihood of cooperation is independent of μ. Yet, this tendency is small and not consistent across all games.

Table 5.5 includes analogous information as Table 5.4, but this time for actors' expectations on alter's behavior. Expectations about the behavior of the other are asked only for the first five games, thus the total number of decisions is half of that for own behavior. In contrast to Table 5.4, data with respect to expected behavior do not follow a very clear pattern. There is a slight tendency that expected cooperation increases in λ_2, but this trend is not strong and consistent.

5.5.2 Bayesian model selection

As explained above, each utility-expectation model predicts a ranking of the nine games both for own and expected behavior, with respect to cooperation rates as presented in Tables 5.4 and 5.5. In other words, if we map the frequency of cooperative choices in these nine games in a 9×2 contingency table (9 games and 2 behavioral options of cooperation and defection), each utility-expectation model imposes a certain set of (in)equality constraints in the cell frequencies of this contingency table. We refer to each set of these (in)equality constraints proposed by each utility-expectation model as a hypothesis.

We assess the overall fit of these hypotheses by using a Bayesian model selection method for equality and inequality constrained hypotheses for contingency tables. For a detailed explanation of this method and some applications, see Laudy & Hoijtink (2007); Hoijtink et al. (2008); Klugkist et al. (2010). The analysis starts with specifying a non-informative prior distribution for the cell probabilities in the contingency table, representing *ex ante* expectations about the distribution of data in the table. An uninformative prior is chosen in or-

Table 5.4: Cooperation thresholds and rankings of 9 asymmetric investment games by the predicted levels of ego's cooperation. Observed cooperation rates and the Spearman's rank correlations (ρ) between observed and predicted rates are also included. N=268 for Γ_5, N=134 for all other games.

Game			Experimental Data		Social Orientation		Normative Model		Inequality Aversion					
									CSE		USE, $F = U[0,2]$		USE, $F = U[0,10^2]$	
	μ_1	λ_1	% C-choices	rank	$\frac{X_e}{Y_a}$	rank	X_e	rank	$\max(\phi_e, \phi_a)$	rank	σ_e^*	rank	σ_e^*	rank
Γ_1	400	.4	15	7	.44	6=	160	7	∞	8=	∞	4=	∞	8=
Γ_2	400	.5	26	3	.33	4=	100	4	0.25	1=	0.375	2	0.251	3=
Γ_3	400	.6	33	1	.17	1=	40	1	1.333	4=	∞	4=	0.147	1=
Γ_4	500	.4	10	9	.44	6=	200	8	4	6=	∞	4=	4.089	7=
Γ_5	500	.5	27	4	.33	4=	125	5	0.25	1=	0.5	3	0.251	3=
Γ_6	500	.6	32	2	.17	1=	50	2	4	6=	∞	4=	0.156	2=
Γ_7	600	.4	10	8	.44	6=	240	9	1.333	4=	∞	4=	1.339	6=
Γ_8	600	.5	17	6	.33	4=	150	6	0.25	1=	0.33	1	0.251	3=
Γ_9	600	.6	23	5	.17	1=	60	3	∞	8=	∞	4=	∞	8=
	ρ			1		.84		.93		.26		.14		.76

Table 5.5: Cooperation thresholds and rankings of 9 asymmetric investment games by the predicted levels of expectations of ego about cooperative behavior of alter. Observed percentages of ego's expectations about alter's cooperative behavior and the Spearman's rank correlations (ρ) between observed and predicted rates are also included. N=134 for Γ_5, N=67 for all other games.

Game			Experimental Data		Social Orientation		Normative Model		Inequality Aversion					
									CSE		USE, $F = U[0,2]$		USE, $F = U[0,10^2]$	
	μ_1	λ_1	% C-expectations	rank	$\frac{X_a}{Y_e}$	rank	X_a	rank	$\max(\phi_e, \phi_a)$	rank	σ_a^*	rank	σ_a^*	rank
Γ_1	400	.4	39	1	.17	1=	60	3	∞	8=	∞	4=	∞	8=
Γ_2	400	.5	28	5	.33	4=	150	6	0.25	1=	0.33	1	0.251	3=
Γ_3	400	.6	24	7=	.44	6=	240	9	1.333	4=	∞	4=	1.339	6=
Γ_4	500	.4	31	3	.17	1=	50	2	4	6=	∞	4=	0.156	2=
Γ_5	500	.5	34	2	.33	4=	125	5	0.25	1=	0.5	3	0.251	3=
Γ_6	500	.6	25	6	.44	6=	200	8	4	6=	∞	4=	4.089	7=
Γ_7	600	.4	30	4	.17	1=	40	1	1.333	4=	∞	4=	0.147	1=
Γ_8	600	.5	22	9	.33	4=	100	4	0.25	1=	0.375	2	0.251	3=
Γ_9	600	.6	24	7=	.44	6=	160	7	∞	8=	∞	4=	∞	8=
ρ				1		.44		.55		-.18		-.15		.27

der not to bias the conclusions in any particular direction. Then, this method updates the probability of each hypothesis composed of a set of (in)equality constraints proposed by a specific utility-expectation model, by taking the likelihood of data derived from the observed cell frequencies into account. Finally, this analysis yields a *posterior model probability* (PMP) for each model. This PMP represents the probability that a particular model is "true", given the alternative models and given that one of the proposed models is true. If the observed data is in line with the set of (in)equality constraints that are predicted by a utility-expectation model, the PMP of that model is higher. The PMPs within a given set of alternative models add up to 1. Thus, each PMP is interpreted in the context of all other sets of hypotheses.

A final issue is how we handle ties. In come cases a model predicts the same cooperation rates in certain games. For example, the cooperation thresholds for the social orientation model in the three games Γ_1, Γ_4, and Γ_7 are all 0.44. If we apply the model very strictly, we would formulate the hypothesis that the cooperation rates in these three games are exactly the same. However, we strongly feel that such a use of formal models is very restrictive and unrealistic. Minor misspecification of the model such as a nonlinear relationship between outcomes and utilities, or factors that are external to the model, e.g., a small random factor such as decision noise or additional social motives, or yet other factors that are not captured by the model would likely differentiate the cooperation rates of tied games. A similar situation arises for the cases where a model predicts zero or full cooperation. The hypothesis of no cooperation in such games is rather unrealistic, and almost surely rejected, for the same reasons as discussed for the tied games. Thus, we choose to formulate the hypotheses without imposing equality constraint among games with tied predictions and without specific value constraints. Thus, our hypotheses involve only inequality constraints, no equality constraints. In addition to these 'philosophical' reasons, there is also a technical reason for reporting the results without imposing equality constraints. See Appendix D.3 for a detailed discussion on the issue and the details of the tested hypotheses. We have also conducted analyses with a stricter interpretation of equality constraints, but this variation of the analysis did not substantially change our conclusions.

Table 5.6 presents the PMPs for each utility-expectation model, both for own behavior, and for expectations about alter's behavior. There are two kinds of PMPs reported in the table. First consider PMPs that are obtained by comparing each model to the unconstrained model, i.e., $Pr(H_i|H_0)$, where H_0 denotes the unconstrained model that does not impose any restriction on the cell probabilities. All other (in)equality constraints are nested in this unconstrained model. The PMP of a model i compared to the unconstrained model, $Pr(H_i|H_0)$, is conventionally interpreted as follows (Kass & Raftery, 1995). If $Pr(H_i|H_0)$ is smaller than 0.75, the evidence is not worth more than "*a bare mentioning*" and if it is smaller than 0.5, then the fit is *extremely poor*. If $0.75 < Pr(H_i|H_0) < 0.99$, the evidence is *strong* and if $Pr(H_i|H_0) > 0.99$, the evidence is *decisive*. The PMPs in Table 5.6 show a clear pattern. When $Pr(H_i|H_0)$s are considered, the inequality aversion model is flatly rejected for both own behavior of subjects and their expectations about the behavior of their interaction partners. When own behavior is considered, both the social orientation model and the normative model have very high PMPs compared to the unconstrained model, which means that these models fit data well. When expectations are considered, the social orientation model is again strongly supported, but support for the normative model is less convincing.

Table 5.6 also includes the PMP of a model given all models in the table, i.e., $Pr(H_i|All)$. $Pr(H_i|All)$ represents the probability that model i is true given all other models in Table 5.6 including the unconstrained model. Thus, $Pr(H_i|All)$s can be used to compare the models with each other. If a decision has to be made on which of the social orientation and the normative model is better, this analysis favors the normative model for own behavior but the social orientation model for expected behavior. Again, compared to all other models, the inequality aversion model receives no support.

To sum, there is no support for the inequality aversion model, neither compared to the unconstrained model, nor compared to other alternative models, thus this model is clearly rejected. Both the social orientation model and the normative model receive support from this analysis. When these two models are compared, support for the normative model is higher for own behavior of subjects. When, however, expectations are considered, the social orientation

model is better than the normative model.

One of the main issues in Bayesian analyses is the specification of prior distributions for the model parameters. Laudy & Hoijtink (2007) present evidence from extensive simulation that the choice of the prior distribution hardly effects the PMPs of models with only inequality constraints, but can influence the PMPs of models with equality constraints. This issue is not problematic in our case. First, and most importantly, since our hypotheses do not include equality constraints, this is not a problem at all in our case. Moreover, even if hypotheses are formulated with equality constraints, the PMPs are so decisive, i.e., either very big or very small, that the specification of the prior distribution can hardly change our qualitative conclusions. Finally, even if equality constraints are included the specification of the prior may influence only the choice between the social orientation and the normative models, since the fit of these two models are relatively close. Yet, as we explain below, we do not select one of these two models using the Bayesian methodology, since we compare these two utility models using more conventional statistical methods. [6]

5.5.3 The probit model

In the previous section, we rejected the inequality aversion model, however, we did not conclude whether the normative model or the social orientation model fits the data best. In this section, we combine these two utility models in an encompassing model, so that we can perform subsequent statistical tests of nested models. Moreover, predicting own behavior and expectations about the behavior of alter simultaneously, we investigate the association between own and expected social motives. Here we build on the method proposed by

[6] There is a final caveat. This Bayesian model selection method assumes that individual observations are independent (see Mulder (2010) for the generalization of the Bayesian model selection method to the multilevel linear regression case). Because each person decided in several games, this assumption is violated, that is, decisions are nested in subjects. However, since the evidence against the inequality aversion model is so decisive, we conjecture that this violation of the independence assumption does not influence our substantial conclusions. Moreover, we do not use this Bayesian model selection method as our final method, but perform more conventional statistical methods below which *do* take into account that decisions are nested in individuals.

Table 5.6: Posterior Model Probabilities (PMPs) for the utility - expectation models. $Pr(H_i|H_0)$ is the PMP of the model, given the unconstrained model and $Pr(H_i|All)$ is the PMP of the model, given all of the models in the table plus the unconstrained model.

Model	Own behavior		Expected behavior	
	$Pr(H_i\|H_0)$	$Pr(H_i\|All)$	$Pr(H_i\|H_0)$	$Pr(H_i\|All)$
Social orientation with CSE/USE	0.992	0.085	0.922	0.786
Normative model with CSE/USE	0.999	0.915	0.677	0.139
Inequality aversion with CSE	0.000	0.000	0.015	0.001
Inequality aversion with USE, $F = U[0, 2]$	0.001	0.000	0.049	0.003
Inequality aversion with USE, $F = U[0, 10^2]$	0.119	0.000	0.057	0.004

Palfrey & Prisbrey (1997). When the social orientation and the normative models are combined, the utility of a subject i in period t for an outcome allocation for the self (x) and the other (y) becomes:

$$U_i(x, y, ; \theta_i, \Delta_i) = x + \theta_i y + \Delta_i I_{c=c*} \tag{5.13}$$

Following Palfrey & Prisbrey (1997), we extend this model into a statistically estimable model by including a random utility component into (5.13). There is a variety of options for modeling this random utility component. One can add the random component to θ and Δ. Moreover, one may not add the term to these two utility parameters but to each strategy option, namely the cooperative and the defective option. There is not a strong rationale for choosing any of these options. Palfrey & Prisbrey (1997) include the random component to what one may analogously call the Δ parameter. Due to the restrictions of our design, we add the random term to θ; if we add the random utility term to Δ or to the strategy options, the resulting model suffers from multicollinearity problems, and regression results become unreliable. Consequently, we assume that there is a random utility component ϵ_{it} added to θ_i for each i's decision at each period t. This error term represents an additional small random propensity added to the outcome of the other actor. One may interpret this error term as subjects tremble about how they value the outcomes of others. We further assume that ϵ_{it}'s are independent, identical, and normally distributed with zero mean and non-zero $sd(\epsilon_{it})$. The parallelness

property of the asymmetric investment game allows us to model the investment probability of an actor i in period t in a conventional way. It is easy to verify that subject i in period t chooses to invest in an asymmetric investment game if and only if

$$\epsilon_{it} \geq \frac{X_e}{Y_a} - \theta_i - \frac{\Delta_i}{Y_a}. \tag{5.14}$$

The simplicity of (5.14) is due to the fact that there is always a dominant strategy for both the social orientation model and the normative model in the PPD. This is also the case when the additional error term is included. Had we not eliminated the inequality aversion model, the decision situation would not have dominated strategies for certain parameter values and we could not model the probability of investing as straightforward as in (5.14), consequently cannot use the probit regression. In addition to (5.14), we also assume that $sd(\epsilon_{it})$ decreases with period t; we hypothesize the error variance decreases as subjects gain experience by making more decisions. We model this as:

$$sd(\epsilon_{it}) = e^{\gamma_1 + \gamma_2(t-1)} \tag{5.15}$$

e^{γ_1} represents the standard deviation at the first period, and γ_2 represents the decrease in $\ln(sd(\epsilon_{it}))$ per period.

The probability of expecting alter to invest, and the error term involving the expected behavior is modeled similarly as in (5.14) and (5.15). Ego expects alter to invest if and only if

$$\epsilon_{at|i} \geq \frac{X_a}{Y_e} - \theta_{a|i} - \frac{\Delta_{a|i}}{Y_e} \tag{5.16}$$

Where $\theta_{a|i}$ is i's expectation about the θ of alter and $\Delta_{a|i}$ is i's expectation about the Δ of alter. We model the error involved in (5.16) as:

$$sd(\epsilon_{at|i}) = e^{\gamma_1 + \gamma_2(t-1)} \tag{5.17}$$

In other words, we assume that the model for the error term is the same when an actor is making his/her own decision and an actor is forming an expectation about the behavior of the other. This assumption is not necessary,

thus, we can easily relax it. The rationale behind this assumption is convenience; with a relatively small dataset as ours not too many parameters can be estimated reliably. Moreover, theoretically, we are not directly interested in the error structure underlying own and expected behavior. Finally, and more convincingly, we have tested this assumption with our data as we report below and found no evidence against this assumption.

We treat θ_i, Δ_i, $\theta_{a|i}$, and $\Delta_{a|i}$ as random parameters that are normally distributed in the set of subjects. We estimate the means and variances of social motive parameters, θ_i, Δ_i, $\theta_{a|i}$, $\Delta_{a|i}$, as well as the two error parameter γ_1, and γ_2 in (5.14), (5.16), (5.15), and (5.17) simultaneously with a probit regression model with random intercepts (for θ) and random slopes (for Δ) and level-1 multiplicative heteroskedasticity for period specific decision noise, using the Stata program GLLAMM (Rabe-Hesketh et al., 2002). The details of the statistical model is as follows. Readers who are not interested in these technical details can skip directly to the next paragraph where we interpret the results. The dependent variable in the model is invest (1=invest, 0=not invest), where the five decisions for expected behavior are appended to ten decisions for own behavior per subject. When the dependent variable is about own behavior, there are three predictor variables that take nonzero values: $\frac{X_e}{Y_a}$, a constant (call it "own"), and $\frac{1}{Y_a}$; and all other predictor variables are recoded to take a value of zero. Similarly, when the dependent variable is about expected behavior, the three predictor variables that take nonzero values are: $\frac{X_a}{Y_e}$, a constant (call it "expected"), and $\frac{1}{Y_e}$; and all other predictor variables are recoded as zero. In other words, we included interaction terms involving the predictors and a variable that codes whether the dependent variable is about own behavior or expected behavior. In line with (5.14) and (5.16), the coefficients of $\frac{X_e}{Y_a}$ and $\frac{X_a}{Y_e}$ are constrained to be 1 . Since decisions are nested in individuals, random slopes for variables "own", "expected", $\frac{1}{Y_a}$, and $\frac{1}{Y_e}$ are estimated. The estimated coefficients of "own" and "expected" are the estimates of mean(θ_i) and mean($\theta_{a|i}$), respectively; whereas the variances of the random slopes of those coefficients gives us estimates of the variances of those social orientation parameters. Similarly, the coefficients of $\frac{1}{Y_e}$ and $\frac{1}{Y_a}$ are the estimates of mean(Δ_i) and mean($\Delta_{i|e}$), respectively; and random slopes

are the estimates for the variances of the normative parameters. Within this framework, covariances of those random slopes are easy to estimate. Finally the heteroskedastic probit specification allows us to estimate γ_1 and γ_2 in (5.15) and (5.17) directly. Note that this model is mathematically equivalent to the model where own behavior and expected behavior is predicted with the variables in (5.14) and (5.16) separately, but the error structure is constrained to be the same.

Table 5.7 displays the estimation results. First of all, both mean(θ_i) and mean(Δ_i) are significantly different from zero, which assures the existence of both motives. Mean(θ_i) is estimated as -.58 and mean(Δ_i) is as 122. Note that θ is in utils/points while Δ is in utils. Thus, on average, subjects share the norm of cooperation but at the same time they are competitive, i.e., they value the outcomes of their opponents negatively. When expectations are considered, both mean($\theta_{a|i}$) and mean($\Delta_{a|i}$) are insignificant, where mean($\theta_{a|i}$) is estimated as almost zero, and mean($\Delta_{a|i}$) as negative. Thus we find no evidence that the means of expected motives are different from zero. Since Wald tests for hypotheses of the variances of random intercepts and slopes reported in Table 5.7 would not be appropriate, we provide correct likelihood-ratio boundary tests using the mixture distribution $\overline{\chi}^2(01)$ of $\chi^2(1)$ and $\chi^2(0)$ (see Self & Liang, 1987). The variance of θ_i is statistically significant (likelihood-ratio $\overline{\chi}^2(01) = 4.641$, p(2-sided)<0.05), which means that subjects vary significantly with respect to how they value the outcomes of their opponent. Moreover, this significant variation also shows that, although on average subjects are competitive, some are actually altruist, i.e., do have positive θ. To be more precise, the probability of having a positive θ_i, $\Pr(\theta_i > 0)$, computed from the estimates of the mean and variance of θ_i, is around 8%. These θ estimates seem to imply that our subjects are less altruist/more competitive than those of some other studies, e.g., Simpson (2004). But, it should also be noted that our θ estimates are about *net* altruism, i.e., controlling for both the cooperative norms and the random noise component of the altruism parameter.

The variance of $\theta_{a|i}$ is also significant (likelihood-ratio $\overline{\chi}^2(01) = 4.029$, p(2-sided)<0.05). Thus, subjects vary significantly with respect to their expectations about the θ of their interaction partners. When the random slopes

are included in the model, the variances of Δ_i and $\Delta_{i|e}$ are found to be highly insignificant (likelihood-ratio $\overline{\chi}^2(01) = 0.000$, p(2-sided)=1, for each variance). Thus we find no evidence that subjects vary with respect to how they value cooperation and what they expect about how others value cooperation.

We also obtained interesting results for the differences between own and expected social motives. The difference between mean($\theta_{a|i}$)=0.04 and mean(θ_i) =-0.58 is .62 and statistically significant (Wald $\chi^2(1) = 5.98$, p(2-sided)<0.05). The conditional distribution of $\theta_{a|i}$ given θ_i, namely $\theta_{a|i}|\theta_i$, is easily attained from the parameter estimates in Table 5.7 (Miller & Miller, 2004, see: theorem 6.9). Mean($\theta_{a|i}|\theta_i$)=$0.93 + 1.5\theta_i$ and var($\theta_{a|i}|\theta_i$)= 0.04. This small conditional variance is due to the high correlation between θ_i and $\theta_{a|i}$. These results show that subjects expect others to be less competitive, or equivalently, more altruist than themselves. One of the most striking finding is, however, the correlation between θ_i and $\theta_{a|i}$ which is estimated as 0.92. Thus, although subjects expect others to be less competitive, own motives and expected motives highly and significantly co-vary (Wald $\chi^2(1) = 4.49$, p(2-sided)<0.05). This finding strongly supports CSE. The difference between $\Delta_{a|i}$ and Δ_i is roughly -213 and statistically significant (Wald $\chi^2(1) = 5.81$, p(2-sided)<0.05). Since the variances of both Δ_i and $\Delta_{a|i}$ are highly insignificant, $\Delta_{a|i}|\Delta_i$ is estimated to have a mean of $\Delta_i - 213$ with zero variance. Thus, subjects expect others to give less value to cooperation.

Finally, consider the parameters estimated for the error model. γ_1 is estimated as -0.277. Thus, $sd(\epsilon)$ is about 0.76 in the first period. γ_2 is estimated as -0.08, and highly significant (Wald $\chi^2(1) = 21.11$, p(2-sided)<0.01). Thus, in each period, $sd(\epsilon)$ decreases roughly as $100 \times (e^{-0.08} - 1) = 8\%$. Moreover, we have tested whether the model for the error is different when subjects are forming expectations compared to deciding their own behavior. Neither the standard deviation of error in the first period (Wald $\chi^2(1) = 0.865$, p(2-sided)=0.353) nor the effect of period on the standard deviation of error (Wald $\chi^2(1) = 0.723$, p(2-sided)=0.395) are significantly different for expected behavior compared to own behavior.

Our results are not directly comparable to the results of Palfrey & Prisbrey (1997) for several reasons. First, they study n-person public good games, thus

the game structure is different. Moreover, although the utility model they use is similar to ours, it is a different one. In addition, there are methodological problems with their approach. Our probit model is methodologically more advanced compared to the model of Palfrey & Prisbrey (1997) in several respects. Palfrey & Prisbrey (1997) model dependency in decisions that are made by the same subject by fixed effects for subjects, estimated by maximum likelihood, an estimation known to produce inconsistent estimates of the parameters (Andersen, 1980, chapter 9). Moreover, they do not estimate a random slope for what one may analogously call θ_i, and conclude that θ is not a prominent motive in their subject pool, since the estimated coefficient was not significant. If the mean of θ is not significantly different from zero, it is not sufficient to conclude that such motive is not existent. Had they estimated the variance of θ_i, they might find a significant variance. Dissimilar to our approach, they estimated γ_1 indirectly, e.g., estimating the coefficient of $\frac{1}{Y_a}$ freely and then with algebraic manipulation obtain γ_1. This approach is mathematically equivalent to ours and there is no problem. However, they estimate γ_2 incorrectly, by including an interaction of $\frac{1}{Y_a}$ with period and interpret it as γ_2. Yet, a decrease in error variance yields a proportional decrease in all coefficients in the model, thus they should have included the interaction of period with all other coefficients in the model, which they have not. Thus, their model is misspecified. Estimating γ_1 and γ_2 directly as we did waives the problems that Palfrey & Prisbrey (1997) had. Finally, and most importantly, Palfrey & Prisbrey (1997) model only own behavior, thus are not able to compare own motives with expectations about the motives of others.

5.6 Discussion and conclusions

In this paper, we study a formal behavioral model with three components: social motives, expectations about social motives of others, and a decision theoretic component. We consider three non-standard utility models as alternative micro theories on social motives: the social orientation model, an inequality aversion model, and the normative model. Expectations are modeled as certain symmetric expectations as well as uncertain symmetric expec-

Table 5.7: Heteroskedastic random-effects probit regression estimates for the decision error parameters and for the means and variances of social motive parameters. Differences between own and expected social motives, covariance and correlation between own and expected θ, and their standard errors are also given.

	Own motives			Expected motives	
Parameter	Estimate	Std. Err.	Parameter	Estimate	Std. Err.
mean(θ_i)	-.576**	.289	mean($\theta_{a\|i}$)	.041	.247
mean(Δ_i)	121.774**	60.256	mean($\Delta_{a\|i}$)	-91.520	67.464
var(θ_i)	.173**	.085	var($\theta_{a\|i}$)	.489**	.237
var(Δ_i)	.000	.000	var($\Delta_{a\|i}$)	.000	.000

Own vs. expected motives

Parameter	Estimate	Std. Err.
mean(θ_i) - mean($\theta_{a\|i}$)	-.617**	.252
mean(Δ_i) - mean($\Delta_{a\|i}$)	-213.294**	88.453
Cov($\theta_i, \theta_{a\|i}$)	.267**	.126
Corr($\theta_i, \theta_{a\|i}$)	.917	

Error structure ($sd(\epsilon_{it}) = e^{\gamma_1+\gamma_2(t-1)}$)

Parameter	Estimate	Std. Err.
γ_1	-.271	.237
γ_2	-.081***	.018

Model summary

Log-likelihood	-936.455
N(Subject)	134
N(Own Behavior)	1340
N(Expected Behavior)	670

**p(2-sided)<0.05; ***p(2-sided)<0.01 for Wald tests.

tations. The (Bayesian-)Nash-equilibrium concept is applied as the decision theoretic component. This formal behavioral model with three components predicts two phenomena. First, it predicts behavior of actors. Second, it also predicts expectations of actors about the behavior of others. We applied this formal behavioral model to an asymmetric prisoner's dilemma case, and tested predictions with an experiment.

The results of our analyzes can be grouped into two: theoretical results obtained from formal analyses and empirical results obtained from experimental data. We first discuss results obtained from formal analyses. First, when asymmetry is introduced in the decision situation, nonstandard utility models can be told apart. Predictions of different utility models differ under asymmetry, and thus model differentiation is easier. There is another advantage of introducing asymmetry. Each empirical observation can be explained by assuming preferences *ex post*. Thus, to test nonstandard utility models in a more compelling way, one should not only account for known empirical regularities, such as behavior in symmetric PDs, but also derive new predictions for new phenomena or in new contexts and test these new predictions. Asymmetry provides such a new context. We obtained a variety of new predictions under asymmetry, which is not the case if one considers only symmetric games.

The second important theoretical finding concerns expectations about others' preferences. Assuming non-selfish preferences is the most popular trend in explaining observations that are at odds with the predictions of the standard application of game theory. However, when non-selfish motives are introduced, a further complication arises. How to model actors' expectations about non-selfish motives of others properly? This issue is largely neglected in the literature. We showed that, how actors' expectations about the social motives of others are modeled may influence the predictions substantially, but only in some cases. In other cases, robust predictions can be obtained. We found that the predictions of the inequality aversion model vary substantially, depending on how these expectations are modeled. However, behavioral predictions of the social orientation model and the normative model are more robust in this sense. Thus, empirical tests of nonstandard utility models are incomplete unless one shows that predictions are robust with respect to the

assumptions about actors' expectations about social motives of others, and if not, these expectations are modeled properly.

Thirdly, we extended the application of the formal behavioral model, by including predictions with respect to actors' expectations about the behavior of their interaction partners. We also consider the association between own behavior and expected behavior. We showed that the association between own behavior and expected behavior need not be a direct association. Such an association occurs if actors expect others to have similar social motives as themselves. Croson (2007) tests different nonstandard utility models by considering the association between actors' own behavior and their expectations about the behavior of others. For example she hypothesizes that the normative model[7] predicts zero correlation between one's contribution to the public good and expectations about contributions of others. Yet, her analyses ignore possible association between one's social motives and one's expectation about social motives of others. Thus, modeling own behavior and expected behavior simultaneously helps us understand the underlying mechanisms better than modeling own behavior, expected behavior, or the association between own behavior and expected behavior independently from each other.

We obtained important empirical results as well. We analyzed experimental data rigorously, with both Bayesian and frequentist methods. These two approaches nicely complement each other in our case. No matter how expectations are modeled, the inequality aversion model is in flat contrast with our data. Although we report the results of a simplified version of the original inequality aversion models of Fehr & Schmidt (1999), this result can also be generalized to the original model. Appendix D.2 presents analyses with the Fehr & Schmidt model showing that the original version also fails to explain the pattern in our data. Thus, we conclude that inequality aversion is not a prominent motive in (asymmetric) prisoner's dilemma situations. This conclusion is in line with previous studies which also shows that subjects are less concerned with reducing inequality in outcomes than other motivations such as increasing social welfare and efficiency (Charness & Rabin, 2002; Engelmann

[7]She uses the terms *commitment theories* as a generic term for Kantian or normative theories.

& Strobel, 2004). Note that the rejection of the inequality aversion model is largely due to its unsuccessful predictions in asymmetric PDs. Restricting attention to only symmetric PDs could have easily yielded a false support for inequality aversion. One final issue, however, is that the inequality aversion model predicts multiple equilibria. One can argue that our conclusion on the empirical success of the inequality aversion model might differ if other equilibrium selection criteria are considered. Appendix D.4 includes a discussion on possible equilibrium selection criteria, and it seems that none of such criteria saves the inequality aversion model from being rejected.

The social orientation model and the normative model both receive certain support from the Bayesian model selection procedure. Thus, we combine these two models in an encompassing model and fitted an innovative random coefficients probit model which allowed us to perform subsequent statistical tests, while accounting for decision error. These tests show that when own motives are considered, (1) both social motives proposed by the social orientation model and the normative model exist simultaneously. Subjects share a norm of cooperation, but at the same time on average they are competitive, i.e., they value the outcomes of others negatively. (2) There is significant variation among subjects in terms of the value they give to the outcomes of others. Thus, although on average subjects value the outcomes of others negatively, some subjects derive positive utility from the outcomes of others. Yet, there is no significant variation among subject in terms of how they value cooperation. When expectations about social motives of others are considered, (3) on average, social motives proposed by the social orientation and the normative models are not significantly different from zero. But (4) subjects differ significantly in terms of their expectations about how others value the outcomes of the self.

Moreover, analyzing own behavior and expected behavior simultaneously, we also performed statistical tests regarding the relationship between own social motives and expected social motives. (5) Subjects expect others to be significantly less competitive/more altruist than the self, but to give less value to the norm of cooperation. A number of social psychological studies found evidence for a *uniqueness* bias, i.e., the tendency for people to see themselves

as better than others by underestimating the proportion of people that perform socially desirable actions (Goethals et al., 1991). This uniqueness bias is partially observed in our 5th empirical finding. Expecting others to value cooperation normatively less than the self is in line with the uniqueness bias. But we also find that people also expect others to be less competitive (more altruist) than the self. This second finding seems to contradict the uniqueness bias. We do not have a clear explanation why the uniqueness bias is only partially supported. Past research typically demonstrates the uniqueness bias on single motives (Iedema & Poppe, 1999), but our findings involve simultaneous investigation of multiple social motives. More research is needed to figure out the nature of the uniqueness bias in different types of motives. It has been shown that competitive subjects associate high cooperativeness with less intelligence (Van Lange & Liebrand, 1991). Perhaps, expecting others to be less competitive (more altruist) is related to this phenomenon, that is, expecting others to be less intelligent given that most of our subjects are competitive, thus in fact in line with the uniqueness bias. With our data, we cannot test this explanation.

Finally, (6) there is a strong association between subjects' own social motives and their expectations about others' social motives. This finding is in line with the false consensus effect and supports strongly certain symmetric expectations approach rather than uncertain symmetric expectations. In the behavioral economics literature, expectations about others' social motives are mostly dealt with the latter, where it is assumed that actors know the actual distribution of social motive parameters of others, and this distribution is independent of one's own social motive parameter. Thus, refuting uncertain symmetric expectations approach, our 6th empirical finding addresses an important drawback of the behavioral economics literature.

Rational choice sociology has shown that cooperation in social dilemmas can be sustained under certain conditions, assuming that actors are *selfish*. For example, when the Prisoner's Dilemma is embedded in a broader social context such as a social network, or when there are external formal or informal institutions of social control, cooperation can be sustained (Heckathorn, 1990; Macy & Skvoretz, 1998; Buskens & Raub, 2002). In addition, if the PD is

played repeatedly, because of the strategic concerns due to future interactions, cooperation in social dilemmas is also feasible (Axelrod, 1984). These are *external* mechanisms. In our study, we focused on *internal* factors, by studying social dilemmas in "encounters", i.e., in one-shot situations where there is no social embeddedness, institutions or the possibility of future interactions. One-shot encounters is an important characteristics of modern large open societies. Understanding the behavioral and motivational mechanisms in one-shot situations will not only advance our understanding of cooperation among strangers in social encounters, but also provide a micro-behavioral framework that will complement cooperation in socially embedded and repeated settings with the existence of external institutions. In certain situations, internal mechanisms such as internalized norms, social motives as well as expectations of others' norms and motives will accentuate the effects of these external factors. Yet in other cases, internal factors may diminish the effects of external factors on cooperation. In addition, external factors may work through internalization, i.e., influencing motivation and expectations and cognition. Thus, accounting for internal factors and the interplay between internal and external factors, one will draw a more accurate and complete picture of social dilemmas.

Chapter 6

Inequality and procedural justice in social dilemmas*

Abstract

This study investigates the influence of resource inequality and the fairness of the allocation procedure of unequal resources on cooperative behavior in social dilemmas. We propose a simple formal behavioral model that incorporates conflicting selfish and social motivations. This model allows us to predict how inequality influences cooperative behavior. Allocation of resources is manipulated by three treatments that vary in terms of procedural justice: allocating resources randomly, based on merit, and based on ascription. As predicted, procedural justice influences cooperation significantly. Moreover, gender is found to be an important factor interacting with the association between procedural justice and cooperative behavior.

6.1 Introduction

Social dilemmas are situations where individual interests and collective interests are in conflict. One example is the Prisoner's Dilemma (PD). In the

*This chapter is written in collaboration with Jeroen Weesie and published in *Journal of Mathematical Sociology* (Aksoy & Weesie, 2009). We thank Vincent Buskens, Manuela Vieth, and Werner Raub for their comments on earlier drafts and their help in conducting the experiment.

PD, every party involved would be better off by joint cooperation. But as decisions are taken individually, each party has an incentive to defect, with a Pareto-inferior outcome at the group level as the unintended consequence due to joint defection (Dawes, 1980). Real world situations that can be represented as PDs are numerous (Kollock, 1998). For example, a wide array of public goods such as well-functioning labor unions or a clean environment can be studied within the PD framework. In these phenomena, each individual has an incentive not to contribute to the public good while enjoying the benefits without costs. But if everyone behaves in this manner, the collective outcome is socially undesirable.

The PD received great interest from researchers, but most attention is devoted to situations where actors are equal in terms of their resources. However, many real life interactions that yield cooperation problems are asymmetric (De Jasay et al., 2004). Thus, in employer-employee relations or in interethnic encounters where one party can mobilize more resources than the other, our understanding of cooperation problems is incomplete unless we take this asymmetry into account.

The few existing studies that investigate the effects of inequality on behavior in social dilemmas are equivocal. Some find that participants with larger endowments cooperate more than those with lower endowments (Rapoport, 1988; Van Dijk & Wilke, 1995); others find the exact opposite (Rapoport et al., 1989). These mixed results constitute the motivation of this paper.

As Smith et al. (2003) put it, one reason of these conflicting findings could be the differences in the perceptions of participants about the reasons of inequality. Justice research shows that the procedure through which resources are allocated may influence people's reactions. When the inequality is perceived as legitimate or when the allocation procedure is fair, people perceive the situation as fair even though the outcomes are personally adverse (Thibaut & Walker, 1975). Smith et al. (2003) compared merit based and random allocations, and showed that the reasons of inequality affected cooperative behavior. In this paper, we introduce one further manipulation, allocating resources via a more illegitimate way than random allocation, i.e., through ascription. In real life, randomly occurring inequalities are rather rare, and a more com-

pelling way to create illegitimate allocation is allocation by ascription. Would similar mechanisms be at work when affluence is obtained by ascription and by pure luck?

In addition to the legitimacy of inequality, we believe that the conflicting results in the literature can be unraveled theoretically by defining and modeling conflicting interests of individuals more precisely. In a social dilemma, individual motives such as greed and selfishness are active simultaneously with social motives that result from feelings of efficiency, fairness, and group well-being (Wilke, 1991; Eek et al., 1998). Using models to represent these conflicting interests and analyzing interaction situations using game-theoretic tools allow us to come up with detailed predictions on the influence of inequality on cooperative behavior. In this paper, we use a simple formal model and show that this model is fruitful in understanding behavior in asymmetric social dilemmas.

6.2 Theory and hypotheses

6.2.1 Decision situation

To introduce inequality in the PD and simultaneously express this inequality to participants in a substantial and comprehensive way, we propose an *asymmetric investment game* (Figure 6.1). In an asymmetric investment game, two actors face an opportunity to form a partnership. Actors have certain budgets (μ_1 and μ_2) that need not be equal. Each actor can invest or decide not to invest the budget. Total investment is multiplied by a rate of returns ($\kappa > 1$) and redistributed to actors. The first actor gets a share of λ_1 from the common budget after investment, whereas the other actor gets λ_1, $\lambda_1 + \lambda_2 = 1$, regardless of their own investments. In an asymmetric investment game, inequality is systematized by setting $\mu_1 \neq \mu_2$ or $\lambda_1 \neq \lambda_2$. Under certain restrictions on κ, μ, and λ, an asymmetric investment game is a PD, since individual interests suggest mutual disinvestment, but both actors would be better off by joint investment.

Asymmetric investment game represents "objective" outcomes that actors face in a social dilemma situation. In other words, an asymmetric investment

Table 6.1: (a) Asymmetric investment game, symbols in cells represent outcomes for actors 1 and 2, respectively; (b) actor 1's payoffs after transforming (a) into subjective utility, where γ_1 represents the value for the cooperative option.

(a) Outcomes

	Invest	Not Invest
Invest	$\lambda_1\kappa(\mu_1+\mu_2)$, $\lambda_2\kappa(\mu_1+\mu_2)$	$\lambda_1\kappa\mu_1$, $\mu_2+\lambda_2\kappa\mu_1$
Not Invest	$\mu_1+\lambda_1\kappa\mu_2$, $\lambda_2\kappa\mu_2$	μ_1, μ_2

(b) Utilities

	Invest	Not Invest
Invest	$\lambda_1\kappa(\mu_1+\mu_2)+\gamma_1$	$\lambda_1\kappa\mu_1+\gamma_1$
Not Invest	$\mu_1+\lambda_1\kappa\mu_2$	μ_1

game constitutes the given decision matrix which is known in advance by actors that involve in the dilemma situation (Kelley & Thibaut, 1978). However, in addition to these objective outcomes, our game-theoretic model also includes a subjective utility component which is represented by the utility model that we introduce now.

6.2.2 Conflicting interests

To formally define the conflicting selfish and social motivations, we use the following utility model:

$$U(x, c, \gamma_i) = x + I_c\gamma_i \tag{6.1}$$

where x is the material or objective outcome that actor i gets, c is the strategy implemented by the actor and γ_i represents utility gained by choosing the cooperative option for actor i. $I_c = 1$ if $c =$*Invest*, $I_c = 0$ if $c =$*Not invest*. Thus, the utility that actors try to maximize includes both selfish (x) and social (γ) motives.

This utility model is consistent with two interrelated micro-behavioral theories. First, it is in line with the Kantian imperative formulation (Hegselmann, 1989; White, 2004). In a PD, the cooperative strategy maximizes the collective outcome when implemented universally, thus cooperation in a PD can serve as the Kantian categorical imperative which suggests "act according to a maxim whereby at the same time you want this maxim to become a universal law". In this sense, γ_i can be seen as the individual Kantian parameter, which represents utility gained by complying with the Kantian categorical imperative (Elster, 1989). This Kantian interpretation of the utility model can be generalized also to non-PD cases (see Harsanyi, 1980).

Second, in a PD, γ can also be defined as $\gamma = \gamma^c - \gamma^d$, where γ^c represents the extra utility that results from obeying the norm of cooperation, and γ^d is the utility loss (—or depending on the sign of the term—gain) that results from breaking the norm. Given this representation, this utility model is a special case of the normative model of Crawford & Ostrom (1995).[1] The Kantian imperative and the normative interpretations of (6.1) are equivalent in the PD case (see Crawford & Ostrom, 1995, pp. 590).

Finally, γ can also be seen as utility gained from warm glow giving. In this sense, this model is also similar to the social orientation model (McClintock, 1972; Van Lange, 1991).

In our application, people may vary with respect to the value they give to cooperation, i.e., γ is an individual parameter. We do not set boundary values for γ, i.e., $\gamma \in (-\infty, +\infty)$. Those with $\gamma > 0$ are cooperators, and the extreme case where $\gamma \to +\infty$ represents an ideal situation where the actor is a perfect Kantian (Harsanyi, 1980). Those with $\gamma = 0$ are interested only in the outcomes for the self and thus are selfish. It is also possible that some actors have negative γ, which implies spitefulness, i.e., the desire to harm the opponent or a motive to break the cooperation norm. As one can see below, our predictions are robust against the assumptions on the specific distribution of γ.

[1] Crawford & Ostrom (1995) further differentiate between internal and external sources of utility that result from obeying or breaking the norm of cooperation. In our design, there is no external source which imposes the norm of cooperation, thus this differentiation is not necessary. We thank an anonymous JMS reviewer who reminded us of this equivalence.

Readers who are familiar with the experimental economics literature may wonder why we use this specific social utility model, but not another more popular one, e.g., an inequality aversion model (e.g., Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000).[2] The first reason is parsimony; the utility model that we use is simpler than the inequality aversion models and most of the other social utility models proposed in the literature. In contrast to others, it involves only one unknown parameter and imposes a simpler shape of the utility function. In addition, as we show below, the predictions derived from this utility model are very robust against the assumptions made on the specific distribution of the γ parameter, as well as actors' information about the γ of their interaction partners, which is not the case for the inequality aversion models and for other more complex models. Second, as we will elaborate more on this in the discussion, inequality aversion models would do empirically rather poor in our case, since inequality aversion motives, at least the ones proposed by Fehr & Schmidt (1999) and Bolton & Ockenfels (2000) do not seem to be prominent social motives in the asymmetric PD, and are unable to explain the pattern in our data.

6.2.3 Formal behavioral model

We assume that actors transform a given (asymmetric) investment game into an effective decision matrix by replacing the objective outcomes with the subjective utility represented by (6.1), see Figure 6.1 (Kelley & Thibaut, 1978). In our model, we assume that actors know their own motivations but may not necessarily be certain about the γ of their interaction partners. After this matrix transformation, we then apply the classical Nash-equilibrium concept to the effective decision matrix, i.e., we assume that people manifest equilibrium behavior. The equilibrium of the effective decision matrix depends both on the objective outcomes as well as the subjective utility parameters of actors, i.e., γ of actors. The following theorem states the behavioral predictions derived from the equilibrium solution of this effective decision matrix.

[2] One can extend the list of possible alternative social utility models even further by including many other utility models proposed in the literature. An overview of these social utility models can be found in Fehr & Schmidt (2006).

Theorem 6.2.1. *The Nash equilibrium solution for an asymmetric investment game with utility transformation stated in (6.1) prescribes that actor i with social motive γ_i invests with probability*

$$\pi_i^e(\gamma_i) = \begin{cases} 1 & \text{if } \gamma_i \geq \mu_i(1-\lambda_i\kappa) \\ 0 & \text{otherwise} \end{cases} \tag{6.2}$$

$\pi_i^e(\gamma_i)$ is independent of γ_j, i.e., the γ of the interaction partner.

Proof. Let $\widetilde{\pi}^e_{2|1}(\gamma_2)$ denote the subjective belief of actor 1 about the probability that actor 2 chooses to invest, which depends on the subjective utility matrix of actor 2, thus objective outcomes of actor 2 as well as γ_2. In this case, investing has a higher expected value for actor 1 if and only if:

$$\gamma_1 > (\mu_1 + \lambda_1\kappa\mu_2 - \lambda_1\kappa(\mu_1+\mu_2))\,\widetilde{\pi}^e_{2|1}(\gamma_2) + \mu_1(1-\lambda_1\kappa)(1-\widetilde{\pi}^e_{2|1}(\gamma_2)) \tag{6.3}$$

The right hand side of (6.3) simplifies to $\mu_1(1-\lambda_1\kappa)$, which is independent of $\widetilde{\pi}^e_{2|1}(\gamma_2)$. Thus this theorem is valid for any subjective probability assessment and any expectation about γ_j. □

Theorem 6.2.1 states that given the utility model in (6.1) and the asymmetric investment game, actors always have a dominant strategy in the effective decision matrix, which is either defection or cooperation. Thus, in our case, behavioral predictions of the Nash equilibrium, Bayesian-Nash equilibrium (Harsanyi, 1968), and (expected) utility maximization concepts are the same. Note that Theorem 6.2.1 does not provide the equilibrium solution, but the predicted behavior derived from the equilibrium solution. Equilibrium is a broader concept, which includes the behavior of actors as well as expectations of actors with respect to the behavior of each other, and all higher order expectations. The comparative statics of Theorem 6.2.1 yield clear predictions on the effects of inequality on cooperation. The likelihood of cooperation, $\pi_i^e(\gamma_i)$ monotonically,

H1: increases in share (λ_i), and

H2: decreases in budget (μ_i).

Thus, Theorem 6.2.1 unravels the differential effects of unequal endowments on the likelihood of cooperation: the budget of actors that they can

invest in has a negative effect, whereas the share of actors from the collective good has a positive effect on the likelihood of cooperation.

In addition to those comparative statics, Theorem 6.2.1 also predicts that:

H3: The higher the γ_i, the higher the likelihood of cooperation, $\pi_i^e(\gamma_i)$.

Testing H3 directly requires estimates of γ, the subjective social utility parameters of subjects, which is beyond the scope of this paper. However, using this prediction, the influence of procedural justice on cooperation can be incorporated in the formal behavioral model.

6.2.4 Procedural justice

Once the effects of structural parameters on the likelihood of cooperation are established with Theorem 6.2.1, the influence of the legitimacy of inequality on behavior can be incorporated in this formal representation. Combining procedural justice theories with framing theories on the effects of social context on individual motives (Hertel & Fiedler, 1998), we claim that the legitimacy of inequality influences the strength of the social motive, γ, of actors.

Procedural justice refers to the perceptions of actors about the fairness of the procedure through which resources are allocated (Thibaut & Walker, 1975; Leventhal, 1976; Greenberg, 1987). When the allocation procedure is fair, people perceive the situation as fair even though the outcomes are personally adverse (Thibaut & Walker, 1975). One important procedural mechanism that affects perceptions about procedural justice is "setting ground rules" (Leventhal, 1980). This principle requires explaining to potential receivers the nature of available rewards and what must be done to attain them. According to this rule, legitimacy occurs if clear and relevant evaluation criteria are defined before the allocation takes place and these criteria are communicated to the receivers. However, if this principle is not satisfied, the allocation procedure is perceived as unjust and thus the inequality becomes illegitimate. In turn, this elicits dissatisfaction with the status quo and actors' motives shift in the direction of moving to a fairer allocation (Leventhal, 1980). As distributive justice scholars also argue, reactions of subjects to unfair distributions can be incorporated in the utility model represented in (6.1) (Greenberg, 1987). To model the case where actors are clearly distinguished as advantaged and

disadvantaged in terms of their resources, we argue that the legitimacy of this inequality influences the γ of those who benefit from this inequality differently than those who suffer from it. More precisely, as the distribution process becomes less legitimate, the γ of advantaged parties increases whereas the γ of disadvantage parties decreases, and combining with *H3* we expect that:

H4a: The more illegitimate an allocation is, the more likely an actor cooperates given that she is in the advantageous position.

H4b: The more illegitimate an allocation is, the less likely an actor cooperates given that she is in the disadvantageous position.

H4 claims that those who benefit from illegitimate inequality would be more cooperative whereas those who suffer from illegitimate inequality would be less cooperative compared to the situation where procedural justice is satisfied. Industrialized western societies define themselves as meritocratic societies, and the prevalent norm is that social status should be, at least in principle, attained through merit rather than ascription (Leventhal, 1970). Thus, in our application we claim, and verify empirically by a manipulation check, that if unequal resources are allocated via merit, then the resulting inequality will be perceived as legitimate. However, if resources are allocated by pure luck, and even worse, affluence is obtained by ascription, then inequality is perceived as unfair and illegitimate.

There is already some evidence in the literature for *H4a*. Hoffman et al. (1994) show that in the Ultimatum and Dictator games, when the proposer "earns" his/her position to be a proposer by scoring high on a knowledge quiz, the proposer behaves more selfishly compared to a situation where the proposer role is assigned randomly. Unfortunately, Hoffman et al. (1994) do not assign roles via ascription, thus we do not know how the proposers would act if they obtained their position illegitimately.

Finally, Smith et al. (2003) show that gender is an important factor that interacts with the effects of inequality and procedural justice on cooperative behavior. Various, though speculative explanations of this gender effect can be offered. For example, due to different socialization practices, a provider or helper schema is activated more strongly among men (Eagly & Crowley, 1986; Smith et al., 2003). In addition, men might be more responsive to in-

equality and procedural justice due to sensitivity to social status and hierarchy (Sidanius & Pratto, 1999). Moreover, in line with these theoretical arguments, brain research also has shown that men are more responsive in terms of neural activity to perceived fairness compared to women (Singer et al., 2006). Combining these, we also expect gender effects:

H5: The effect of legitimacy of inequality on behavior is larger for men than women.

6.3 Methods

6.3.1 Participants

134 (82 women, 52 men) participants were recruited using the Online Recruitment System for Economic Experiments (ORSEE Greiner (2004)) at Utrecht University. They were offered on average 12 Euros but were also informed that this amount might change depending on their decisions during the experiment.

6.3.2 Design

The basic design was three reasons of inequality (allocation by merit, random assignment, allocation by ascription) by three inequality conditions (three different asymmetric investment games). Also a symmetric investment game was included. A combination of within and between subjects design was used as explained below.

6.3.3 Task and procedure

Sessions were held in March 2007 in the ELSE lab at Utrecht University on computer using the software z-Tree (Fischbacher, 2007). Participants were assigned randomly to a cubicle where they could not communicate with each other and identify whom they were dealing with. The participants' task was to decide whether to invest their budgets or not.

Participants ran through two examples and they were quizzed to check their understanding. The experiment began after each participant verified answers and it was ensured that everyone understood the task.

In all conditions, first a symmetric investment game was played with $\mu_1 = \mu_2 = 500$, and $\lambda_1 = \lambda_2 = \frac{1}{2}$. Throughout the experiment, κ was fixed at 1.5. After this game, each participant received two of the three treatments where the allocation procedure was manipulated. One of these was always the *random assignment* treatment and the other was either *allocation by merit* or *allocation by ascription*. The order of these treatments was varied in two factors.

In the *allocation by merit* treatment a feeling that participants deserved their advantageous or disadvantageous positions was induced by the following description:

> Parties who are more experienced in the stock exchange market, who are more knowledgeable and experienced in economic and financial transactions, get more resources and obtain higher shares of profit. Knowledge about the economy is a key indicator of success in risky businesses. Thus, in this section of the experiment, the budgets and shares of profit will be allocated according to your skills and knowledge which are important in these kinds of investment and relevant for this decision situation.

Then they received four questions asking for an estimate of the inflation rate, growth rate, unemployment rate and the increase rate of the share prices of AEX (Amsterdam stock exchange) companies in 2006. According to their Knowledge Grades, participants were split into two halves: top half with higher grades, and bottom half with lower grades. Participants in the top half had higher budgets and/or shares and always randomly matched with participants from the bottom half who received lower budgets and shares.

In the *allocation by ascription* treatment, participants were again split into two halves but this time according to the education levels of their fathers. Every participant received questions about their demographics after the symmetric game and before the allocation procedure manipulations as a distractive task. One of those questions was their fathers' level and field of education. In the allocation by ascription condition, a feeling that participants

did not deserve their disadvantageous or advantageous positions was induced by the following text:

> We will give you budgets and shares not according to your own knowledge and skills but according to your father's level and field of education. The level and fields of studies are sorted reflecting social prestige. For example if your father is lower educated or working in the social science field, you will receive lower resources than the other person whose father is higher educated or working in a more prestigious field such as engineering.

Then participants were again split into two halves: ones with higher educated fathers and ones with lower educated fathers. If some participants are tied in terms of education level, the prestige of the field of education is used to differentiate them. Participants in the top half with higher educated fathers received higher budgets and/or shares and were always matched with another person from the bottom half.

In the *random assignment* treatment, participants were randomly divided into two; one half received higher budgets and/or shares and the other half received lower budgets and/or shares.

After these manipulations, participants played three asymmetric investment games per condition, with stranger matching, where parameters were set to put one player in an advantageous position in terms of at least one parameter. These games are (1) $\mu_1 = 600$, $\mu_2 = 400$, $\lambda_1 = 0.6$, $\lambda_2 = 0.4$; (2) $\mu_1 = 500$, $\mu_2 = 500$, $\lambda_1 = 0.6$, $\lambda_2 = 0.4$; (3) $\mu_1 = 600$, $\mu_2 = 400$, $\lambda_1 = 0.5$, $\lambda_2 = 0.5$.

After each of the three legitimacy treatments, participants were asked to evaluate the fairness of the distribution of resources. In order not to reveal the purpose of the experiment, participants were given 5 adjectives: exciting, fair, confusing, comprehensible, unjust and boring and asked to indicate on 5 scale answer categories (1= not at all, 5= very) to what extent these adjectives described the games played after the last legitimacy treatment.

Before ending the experiment, a third condition was applied for ethical reasons and to balance the earnings of the participants at the end. The roles in

the allocation by merit or ascription treatments were simply reversed. For example, participants that were previously in the bottom half and received lower resources because of their fathers' low education received higher resources in this last treatment. These treatments are not included in the analyses. The experiment ended by debriefing participants truthfully. On average, the whole procedure took one hour.

6.4 Results

6.4.1 Manipulation check

The adjectives used for manipulation check were factor analyzed with oblique oblimin rotation and "fair" and "unjust" loaded on the same factor. The correlations between these two variables vary from -0.52 to -0.58 depending on the legitimacy treatment type. One perceived fairness measure was constructed by adding up answers given to these two adjectives after reverse coding "unjust".

Perceived fairness scores are 5.97, 5.91, and 4.78 for allocation by merit, random allocation, and allocation by ascription conditions, respectively. Note that each participant rated the fairness after each legitimacy treatment. A test of these differences was conducted using a multilevel regression model. The differences of Ascription/Random and Ascription/Merit are both significant (p(two-sided)<0.01), however the Merit/Random difference is not significant. There is no significant gender effect, position effect, position×gender, position×allocation procedure and position×gender×allocation procedure interaction effects on perceived fairness. We conclude that the legitimacy manipulations were successful.

6.4.2 Behavioral data

Since each participant made seven investment decisions and received two legitimacy treatments, the data is analyzed using fixed effect logistic regression models (Fischer & Molenaar, 1995). The dependent variable is cooperation/defection with a fixed effect for the participant (1=cooperation, 0=defection). First a model is fitted for the entire sample. In addition, since we are

interested in gender differences in the influence of legitimacy manipulations on cooperation, separate models are fitted for men and women. Table 6.2 presents regression results as well as differences in coefficients between men and women. Note that besides separate effects of budget and share parameters on the likelihood of cooperation, we do not have a hypothesis about whether advantaged actors cooperate more or less than disadvantaged actors. Consequently, we included the interaction of advantageous position with the legitimacy of inequality in a less conventional way in the regression models reported in Table 6.2. Rather than including the direct effect of position and two interaction variables that involve the interaction of position with two of the three legitimacy dummies, we specify another model that is mathematically equivalent. Models in Table 6.2 do not include the direct effect of position, but the interaction of all three treatment dummies with position.

Overall, *H1* and *H2* are supported. The likelihood of cooperation has a negative association with budget and a positive association with share, although the influence of budget is more prominent for women and the influence of share is more for men.

Figure 6.1 provides a general overview of the influence of the allocation procedure of inequality on cooperation: the significance tests are based on Wald tests for the models in Table 6.2, also in the cases where the hypotheses involve linear combinations of the reported coefficients. Overall, in line with *H4*, disadvantaged participants cooperated significantly more when they received their disadvantageous position by merit compared to both random allocation ($z = -2.20$, p(two-sided)<0.05) and allocation by ascription ($z = -1.73$, p(two-sided)$=0.083$). There is no significant difference between the random assignment and allocation by ascription conditions for disadvantaged parties. For advantaged participants, no significant influence of the legitimacy of inequality on the likelihood of cooperation is found.

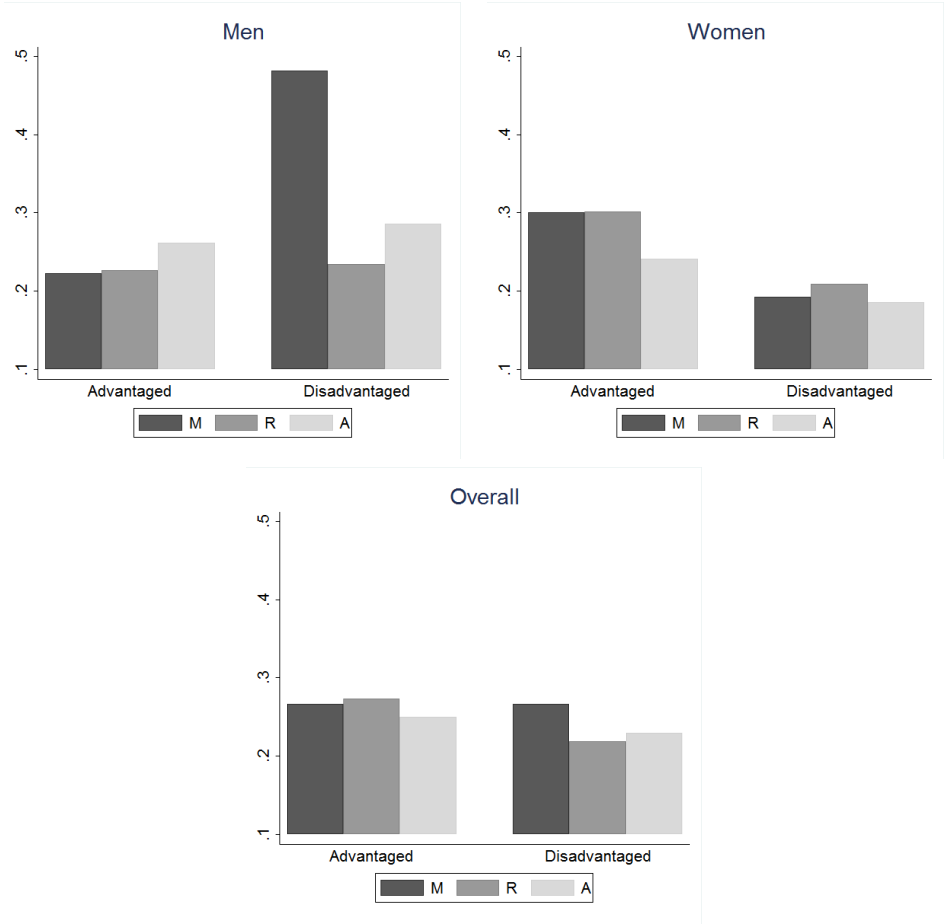
As expected, there are important gender effects. The pattern for advantaged men is in line with *H4a*, although the differences are insignificant. Advantaged women cooperated slightly more in the random assignment condition compared to allocation by merit, yet the difference is insignificant. In allocation by ascription condition, however, advantaged women cooperated less

Table 6.2: Fixed effect logistic regression models predicting cooperative behavior (1=cooperation, 0=defection), conditional maximum likelihood estimates. Equality condition is a dummy for the symmetric game; period is the period in the experiment when the decision is made. The last two columns include differences in coefficients between men and women and their standard errors.

Model	Overall		Men		Women		Men - Women	
Variable	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
Budget (μ), reference: $\mu = 400$								
$\mu = 500$	-1.263***	0.414	-1.350*	0.705	-1.214**	0.522	-0.136	0.877
$\mu = 600$	-1.292***	0.493	-0.935	0.833	-1.546**	0.626	0.611	1.042
Share (λ), reference: $\lambda = 0.4$								
$\lambda = 0.5$	0.264	0.312	0.218	0.527	0.318	0.393	-0.100	0.657
$\lambda = 0.6$	0.945**	0.438	2.472***	0.872	0.365	0.540	2.107**	1.025
Treatment (ref=Merit)								
Random	-0.898**	0.408	-2.346***	0.812	-0.345	0.505	-2.001**	0.956
Ascription	-0.897*	0.518	-2.328**	1.008	-0.390	0.628	-1.938	1.187
Treatment×Position								
Adv×Random	0.955	0.652	-0.734	1.237	1.529*	0.813	-2.263	1.480
Adv×Merit	-0.130	0.720	-3.191**	1.372	1.033	0.897	-4.224***	1.639
Adv×Ascription	0.300	0.756	-0.664	1.351	0.552	0.947	-1.216	1.649
Equality condition	0.820	0.649	-0.486	1.190	1.157	0.800	-1.643	1.433
Period	-0.078	0.063	-0.085	0.111	-0.097	0.080	0.012	0.136
Log Likelihood	-258.039		-84.099		-165.546			
N(participants)	134		52		82			
N(decisions)	938		364		574			

*p<0.1; **p<0.05; ***p<0.01 for two sided tests

Figure 6.1: Mean cooperative behavior (1=cooperation, 0=defection). M = allocation by merit, R = random allocation, A = allocation by ascription.



than the two other conditions where only the difference between random assignment is statistically significant (z = 2.02, p(two-sided)<0.05). Another interesting finding is obtained for disadvantaged men; they cooperated significantly and dramatically more in allocation by merit compared to both random assignment (z = -2.89, p(two-sided)<0.05) and allocation by ascription (z = -2.31, p(two-sided)<0.05). Although disadvantaged men cooperated more in allocation by ascription than random assignment, this difference is highly insignificant. Hardly any influence of the legitimacy of inequality on cooperation is observed for disadvantaged women. When the differences in the effects of the legitimacy of inequality on cooperative behavior between men and women are tested, H5 is supported for mainly disadvantaged participants. For advantaged participants, there is no significant difference between men and women in terms of the effect of the legitimacy of inequality on cooperation. Note that the significance tests reported are obtained controlling for all other variables in the models in Table 6.2, including budget and share. For all insignificant test results, p(two-sided)>0.1.

6.4.3 Additional analyses

One may suspect that Knowledge Grades and participants fathers' education levels may have direct effects on the likelihood of cooperation. As a check, although not reported here, random effect logistic regression models were fitted, with random effects at the participant level, and Knowledge Grade and father's education as participant level variables. These variables were found to have no significant effects on the likelihood of cooperation.

Finally, the order that the participants received the legitimacy manipulations was found to have no effect on the likelihood of cooperation.

6.5 Discussion and conclusions

Many naturally occurring social dilemmas, e.g., cooperation problems in employer -employee relations or interethnic encounters where one party can mobilize more resources than the other are asymmetric. We showed that modeling conflicting interests of actors explicitly and analyzing interaction situations

formally provides a better understanding of the influence of inequality on cooperation and may unravel the reasons of conflicting results obtained in the literature. A simple utility model that includes conflicting selfish and social motives proved fruitful to predict cooperative behavior among unequal actors: as the budget of actors that they can invest in the collective good increases, the likelihood of cooperation decreases. In contrast, as the share of actors that they get from the collective good increases, the likelihood of cooperation also increases.

Besides this specific social utility model we used, there is a bewildering list of other social utility models that we could have considered as well, e.g., the inequality aversion models (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000). See Fehr & Schmidt (2006) for a detailed overview of different utility models proposed in the literature. The normative/Kantian model that we use is relatively simple; the predictions derived from this model are robust against assumptions about the distribution of the unknown social motive parameter and actors' information about the social motives of others. At the same time it received certain empirical support from our data. Moreover, being a normative approach, this particular model is sociologically relevant. Different comparative statics predictions would be obtained for the asymmetric investment game had different utility models been used. For example, both inequality aversion models predict that cooperation would be maximal under symmetry where actors obtain equal outcomes in case of mutual cooperation, i.e., actors have equal shares in the asymmetric investment game. Moreover, the Fehr & Schmidt (1999) model predicts no cooperation at all in the games that we used in our experiments, unless actors obtain equal outcomes in case of mutual cooperation.[3] Clearly, these predictions of the inequality aversion

[3]Fehr & Schmidt (1999) utility function for an actor with an outcome allocation (x, y) for the self (x) and the other (y) is defined as $U = x - \beta \max(0, x - y) - \alpha \max(0, y - x)$, where β and α represent sensitivities to advantaged and disadvantaged inequality, respectively. Fehr & Schmidt (1999) further assume that $\beta \in [0, 1]$ and and $\alpha > \beta$. With the assumption of $\alpha > \beta$, defection is a dominant strategy for at least one actor in the asymmetric games that we use, and the single Nash equilibrium is mutual defection unless the share parameters of actors are equal, i.e., $\lambda_1 = \lambda_2 = 0.5$. Cooperation can be an equilibrium only if the share parameters of actors are equal, and even in this case, mutual defection still remains an equilibrium.

models are not supported by our data, since cooperation is maximal when budget is minimal, and share is maximal, i.e., when the game is highly asymmetric. Inequality aversion does not seem to be a prominent motive in the asymmetric investment game. Thus, we think that the normative/Kantian utility model that we use serves as a good theoretical base to understand the nature of cooperation in asymmetric social dilemmas.

In line with procedural justice theories and previous findings, the legitimacy of inequality is found to be an important factor influencing behavior in social dilemmas. We introduced an important manipulation by allocating resources based on ascription, besides random allocation, and allocation by merit. Our study shows that allocation by ascription may elicit different type of responses than random allocation. Naturally occurring illegitimate inequalities are more often due to ascription, rather than by pure luck and behavioral outcomes of this ascription based allocation is socially and scientifically more relevant (Ganzeboom et al., 1991).

We also observed very interesting gender effects. Compared to women, male behavior is in general more in line with our hypotheses, although men seem to be more sensitive to the allocation process when they are disadvantaged instead of advantaged. As we predicted, the legitimacy of inequality influences male behavior more strongly compared to women, yet only for disadvantaged actors. For advantaged actors, gender differences are less prominent. The association between the legitimacy of inequality and cooperative behavior, and the interaction of gender with the legitimacy of inequality, however, are not as straightforward as we hypothesized. For example, at odds with our predictions, advantaged women cooperated significantly less when resources are allocated by ascription compared to the situation where resources are allocated randomly.[4] Also, unexpectedly high cooperation rates are observed

[4]One reviewer offered the fear-greed hypothesis (Simpson, 2003; Kuwabara, 2005) as a possible explanation. This hypothesis claims that men defect in the PD in response to greed, i.e., to earn more; whereas women defect in response to fear, i.e., to avoid being exploited by cheaters. It is possible that under the ascription condition the advantaged women fear that their partners will be defective because of the illegitimacy, and due to this fear do not cooperate themselves. Although this explanation seems plausible, Theorem 6.2.1 shows that the probability of cooperation does not depend on the expectation about the behavior of the opponent in the asymmetric investment game. Moreover, in our case, the greed component,

among disadvantaged men, when inequality is due to merit. We predict that disadvantaged actors would cooperate more when inequality is due to merit compared to random allocation or allocation by ascription, and this effect to be stronger for men. Thus, this finding is not contradictory to our predictions. Yet, it is remarkable that the cooperation rates for disadvantaged men in the merit condition exceed by far the cooperation rates in any other condition. Similar gender effects are also reported by Smith et al. (2003) and particularly interesting since they point to different decision making mechanisms among men and women. We think that more research is needed before speculating about the reasons of these gender differences, since hitherto explanations are unsatisfactory. It could be that the legitimacy of inequality influences the motives of men and women differently, or man and women may have different social motives and utility functions. Alternatively, the differences are not due to social motives but due to other factors such as differences in the norms about appropriate behavior under various allocation procedures, divergent risk preferences, or a combination of those. We leave investigating these mechanisms for future research.

(T-R), is always the same as the fear component, (P-S).

Chapter 7

Going interethnic: altruism and inequality aversion in intra- and inter-group interactions*

Abstract

This study reports on an international inter-ethnic binary Dictator Game experiment with Turkish, Dutch, and Turkish-Dutch (subjects of Turkish origins from the Netherlands) subjects. With this design we investigate the influence of Ego's and Alter's social group on the weight that Ego attaches to (1) the outcome of Alter—"altruism"—and (2) the absolute difference between the outcomes for Ego and Alter—"inequality aversion"—. We find clear differences between Turkish, Turkish-Dutch, and Dutch subjects with respect to altruism and inequality aversion. However, we find no in-group favoritism.

*This chapter is written in collaboration with Jeroen Weesie and a Dutch language translation of the chapter is published as Aksoy & Weesie (2012c). We thank I. Ercan Alp and the Psychology Department of Bogazici University for permission to conduct the experiment at BU and their help and hospitality during the experiment. We also thank Werner Raub, Tobias Stark, and Andreas Flache for their comments on earlier drafts; Edwin Poppe for his input in designing the experiment; and Ali Aslan Yildiz for his help in subject recruitment.

7.1 Introduction

Recent experimental studies on social dilemmas from various disciplines have identified differences in levels of cooperation and trust across societies and social groups (e.g., Habyarimana et al., 2007; Kuwabara et al., 2007; Hermann et al., 2008; Buchan et al., 2009; Gheorghiu et al., 2009; Bornhorst et al., 2010; Gaechter et al., 2010). In addition to the differences between social groups on levels of cooperation and trust, *intergroup* processes are an important research area. As a result of increasing globalization and mass immigration, interactions increasingly take place between members of different social and ethnic groups (Chuah et al., 2007). This trend, however, seems to be accompanied by a rising anti-immigration climate and an increase in the disbelief in multiculturalism in the host countries (Sniderman & Hagendoorn, 2007). In the last decade, political parties that actively oppose non-EU immigration increased their votes dramatically in many EU countries. It is thus important to investigate empirically whether inter-personal decisions are influenced by the social groups of actors and if hostility between social groups is reflected in the interactions of individuals from these social groups (Bouckaert & Dhaene, 2004).

In this study we report on an international inter-ethnic laboratory experiment conducted with *Turkish* subjects from Turkey, native *Dutch* subjects from the Netherlands, and *Turkish-Dutch* subjects with Turkish ethnic origins who spent most of their lives in the Netherlands. Most experimental studies use minimal group treatments where groups are formed based on arbitrary distinctions, such as preferences for Klee versus Kandinsky paintings (see, e.g., Tajfel & Billig, 1974; Simpson, 2006). In our experiment we use real social groups. Turks are the second largest immigrant population in the Netherlands. Using this design, we investigate whether and how (1) the decision maker's and (2) her interaction partner's social/ethnic group influence how much a subject, relative to her own outcome, weights the outcomes of her interaction partner and the outcome inequality between self and partner (Aksoy & Weesie, 2012b). We call these two social motives *altruism* and *inequality aversion*. We use the choices of Turkish, Dutch, and Turkish-Dutch subjects

in binary Dictator Games to measure measure altruism and inequality aversion. A binary Dictator Game is a two-person game with two options, each option specifies a certain outcome allocation between self and another person. Only one player faces the choice between the two options and the other person is simply a recipient. The subject's choices directly reflect his/her her social preferences. A binary Dictator Game is a variant of Decomposed Games (see: Schulz & May, 1989) which are quite familiar to social psychologists.

The distinctive features of our study compared with previous experimental inter-group social dilemma research and our motivations for conducting the experiment are the following. First, as opposed to *behavior* of Ego, we investigate the influence of the social group of Ego on Ego's *altruism* and *inequality aversion*. Previous studies show that the level of trust and cooperation among strangers is lower in more traditional societies such as southern European countries compared with less traditional, e.g., northern European countries (Bornhorst et al., 2010; Gaechter et al., 2010; Gheorghiu et al., 2009). Osterberk et al. (2004) report similar cultural differences in bargaining behavior in Ultimatum Game experiments. Those studies employ more complex games, though still highly stylized, such as the Trust Game, Prisoner's Dilemma or Public Goods Game (Bornhorst et al., 2010; Bouckaert & Dhaene, 2004; Güth et al., 2008; Kuwabara et al., 2007). In those games, behavioral outcomes like levels of cooperation and trust are influenced by several factors including social preferences of actors such as altruism and inequality aversion, beliefs of actors about social preferences of opponents, mechanisms for solving coordination problems etc.(see Aksoy & Weesie, 2013b; Habyarimana et al., 2007). One exception is Anderson et al. (2011) who consider explicitly differences with respect to social motives. Buchan et al. (2006) also investigate other regarding preferences in an inter-group context, but use behavior as a proxy for social motives. Thus, the differences between cultures in collective action problems could be attributed to multiple causes.

As mentioned above, we analyze the choices of subjects in binary Dictator Games. In binary Dictator Games, choices of actors just reflect their social preferences. We single out one factor that influence cooperation and trust, social preferences, and focus on two social motives, altruism and inequality

aversion. It has been shown, both theoretically and empirically that usually cooperation increases in the weight that actors assign to the outcome of the interaction partner (Aksoy & Weesie, 2012b, 2013b). In this study we investigate whether the levels of altruism and inequality aversion vary across social groups in line with reported inter-cultural differences in levels of cooperation and trust.

Second, we investigate how the social group of Alter influences altruism and inequality aversion of Ego. Classical social psychological theories such as social identity theory predict in-group favoritism with respect to social preferences and cooperative behavior. There is ample support for in-group favoritism in various contexts (Tajfel, 1970). However, studies that test in-group favoritism and at the same time follow the protocols of experimental economics, e.g., incentive compatibility and anonymity (Friedman & Cassar, 2004) report quite mixed results. Some studies indeed report in-group favoritism (e.g., Whitt & Wilson, 2007; Goette et al., 2006), yet other studies report either no significant influence of the social identity of the interaction partner on cooperative behavior (Bouckaert & Dhaene, 2004; Güth et al., 2008), or even some level of out-group favoritism (Gil-White, 2003; Heap & Patrick, 2010). These mixed results call for additional inter-group experiments and motivate the current study. To reiterate the same issue, all these findings involve *behavior* of subjects in games where beliefs about the behavior of the interaction partner in addition to social preferences play a role. Thus, these findings on (lack of) Alter effects could involve beliefs of subjects confounding subjects' social preferences. In our design, altruism and inequality aversion could be measured independent of beliefs.

Other distinctive features of our experimental design are the following. With our specific composition of the subject pool, we are able to control for several factors which would be difficult to control for with a different design: Turks from Turkey and Turkish immigrants in the Netherlands share the same ethnic background and religion whereas Turkish immigrants in the Netherlands and the Dutch share the country of residence. Including Dutch-Turks, i.e., Dutch immigrants in Turkey as a fourth group would make the design more complete. However, there are very few Dutch people living in Turkey, and

their reason of immigration and social backgrounds are very different from the rest of our subjects. Finally, in addition to game-play behavior, our design includes social psychological measures of perceived social distance between the three social groups (Turks, Dutch, and Turkish-Dutch). This enables us to investigate the effect of perceived social distance on social preferences.

7.2 Theory and hypotheses

7.2.1 Social preferences: altruism and inequality aversion

We consider two types of social preferences. The first one is the weight that Ego attaches to the outcome of Alter. We call this motive "altruism". We also allow for negative altruism, that is, an actor may attach a negative weight to the outcome of Alter representing spite. The second type of social preference concerns the difference between the outcomes for Ego and Alter, inequality aversion. This is called inequality *aversion*, because typically people dislike inequality (Fehr & Schmidt, 1999) but we do not rule out that some people actually like inequality. These two types of social preferences can be represented by the following utility specification for an outcome allocation for Ego (x) and Alter (y):

$$U(x, y; \theta, \beta) \quad = \quad x + \theta y - \beta|x - y| + \epsilon \qquad \epsilon \sim N(0, \tau^2). \tag{7.1}$$

Where θ is the altruism parameter and β is the inequality aversion parameter. We do not impose any boundary values for (θ, β), so $\theta < 0$ or $\beta < 0$ are allowed. Subjects may differ with respect to these parameters. Also (θ, β) may be influenced by the social identities of Ego and Alter. An important issue is the following. The theoretical arguments and the hypotheses that we formulate below involve mainly the altruism parameter θ. Only in one case we have a specific hypothesis on the inequality aversion parameter β. We nevertheless keep β in the model because dropping it reduces the predictive power and fit of the model. Moreover, descriptive results on β are interesting in their own right.

Model (7.1) also includes a random component ϵ (McFadden, 1974). This random component can be interpreted as an evaluation error capturing the fact that not all choices are fully described by the altruism and inequality aversion motivations. Mood swings or different social preferences than altruism and inequality aversion all manifest themselves in this random component. This error also makes model (7.1) suitable for statistical analyses. We assume that ϵ is normally distributed with zero mean, independent across subjects, evaluations, and alternatives in the games. For statistical convenience, we assume homogeneity with respect to the variance of the evaluation error, ϵ, within an experimental condition. Experimental conditions comprise the combinations of social group of Ego and the social group of Alter. However, the variance of ϵ may vary between experimental conditions.

Model (7.1) is a fairly general and flexible utility specification employed in various disciplines. Ignoring the random component, ϵ, model (7.1) is an extension of the inequity aversion model of Fehr & Schmidt (1999) and a special case of the model of Charness & Rabin (2002) and of Tutic & Liebe (2009). Additionally, the two parameter version of the social value orientation model (Schulz & May, 1989) familiar to social psychologists is, ignoring the random component, mathematically equivalent to (7.1). For a detailed discussion and application of this utility model see Aksoy & Weesie (2012b).

To be clear, model (7.1) describes choices (behavior) of subjects in various decision situations, but there is not a clear theoretical background why social preferences exist. It is an active area of discussion, if not speculation, why people have such social preferences, especially toward unrelated strangers. The discussion spans through various disciplines including evolutionary psychology, behavioral game theory, sociology, and social psychology. A through discussion on the origins of these social preferences is beyond the scope of the paper. For a detailed discussion, we refer to Fehr & Gintis (2007) and Gintis et al. (2008). We say only a few words here. We discuss below several hypotheses on how certain characteristics of a society, such as cultural values, type of interactions etc. may shape social preferences. In addition, in line with some of our hypotheses, our findings show that there are substantial inter-cultural differences in intensity of these social preferences. Based on this, we like to

treat social preferences as values shaped through experiences from previous interactions, rather than only by long term evolutionary processes.

7.2.2 Ego effects on social preferences

Several experimental studies report the following pattern. Compared with respondents from northern Europe, (e.g., Denmark, Germany, and the Netherlands) subjects from southern Europe, (e.g., Greece and Turkey) show much less cooperative behavior in finitely repeated public good games (Gaechter et al., 2010; Bornhorst et al., 2010; Hermann et al., 2008). Cross-country surveys also support the results of those experiments: generalized trust and cooperation is much higher in northern European countries than in southern European countries (Buchan et al., 2009; Delhey & Newton, 2005; Gheorghiu et al., 2009; Hamamura, 2012). Several factors are found to contribute to this cross-country difference in generalized cooperation and trust. For instance, cultural values seem to have an effect. Collectivism, both at the individual and country level, has a negative influence on generalized trust and cooperation (Gheorghiu et al., 2009; Kuwabara et al., 2007). This cultural effect is often explained by the emancipation theory of Yamagishi & Yamagishi (1994). In collectivistic societies, actors interact more frequently within densely embedded long-term relations. This high level of embeddedness reduces the likelihood of being cheated, due to actual and anticipated direct and third party informal control (Axelrod, 1984; Buskens & Raub, 2002). This provides security against free riding but only between non-strangers. Interactions between strangers are still vulnerable to free riding. In individualistic societies, however, lower levels of social embeddedness, fewer relational obligations and less informal social control makes interaction and cooperation between strangers more frequent which in turn facilitate generalized cooperation and trust between strangers. Banfield (1958) and Putnam (1993) provide similar arguments to explain why levels of trust between strangers are very low in southern Italy.

In addition to cultural values, other structural factors such as national wealth, quality of government, income inequality are also found to influence generalized cooperation and trust (Delhey & Newton, 2005; Hamamura, 2012; Buchan et al., 2009). Several theoretical explanations have been proposed to

explain the effects of those structural factors on generalized cooperation and trust. For example, trust and cooperation can be seen as luxury goods and their consumption increases in wealth (Bornhorst et al., 2010). Income inequality reduces perceived similarity between people, which in turn reduces trust between strangers (Delhey & Newton, 2005). Stressing the equality of rights and duties of citizens, democratic institutions and good governments encourage trust and cooperation between individuals (Delhey & Newton, 2005).

Although these cultural and structural factors are used to explain differences between societies with respect to primarily the levels of generalized trust and cooperation, there is no reason why these factors influence *only* cooperation and trust. We argue that these explanations can be transferred to values governing inter-dependent relations among strangers in general (see: Delhey & Newton, 2005). In other words, we claim that these cultural and structural factors not only influence cooperation and trust among strangers, but also lead to positive sentiments toward strangers in general. In our experimental design with Turkish, Turkish-Dutch and Dutch subjects, this yields the expectation that subjects will have more "pro-social" preferences in the Netherlands than in Turkey (Delhey & Newton, 2005). Reformulated:

On average, altruism (θ) will be higher for Dutch subjects compared with Turkish subjects (H1a).

It is not obvious, however, how one can apply the arguments above to the inequality motive, β. Thus, we present results on β as descriptive.

We do not know of a study on game-play behavior of Turkish immigrants in the Netherlands, let alone on their social preferences. Thus, results on Turkish-Dutch subjects should be treated as exploratory. One might expect that, belonging to a more collectivistic ethnic background, but living in a "high-trust" country, Turkish-Dutch will be in between the Turkish and Dutch subjects with respect to their social preferences.

However, there is an additional twist. Hong & Bohnet (2007) show that minority groups in the US are much more averse to inequality than higher status groups. According to Hong & Bohnet (2007) minority groups are exposed to inequality more frequently which makes them more sensitive to inequality, especially to disadvantageous inequality where self is worse off than the other.

A structural analysis (e.g., Blau & Schwartz, 1984) also demonstrates that given that a minority group is in general worse off than the dominant group in a country, more or less random interactions between individuals would imply that the members of the minority groups will be exposed to inequality more often than the dominant group. Note that this argument mainly applies to disadvantageous inequality. In our model of social preferences, we do not differentiate between advantageous and disadvantageous inequality (see: Fehr & Schmidt, 1999). It is not possible to include both the θ term as well as two different terms representing advantageous and disadvantageous inequality in the utility function in two-person binary Dictator Games, due to identification reasons. Both forms of inequality aversion are captured by the β parameter in model (7.1). However, additional analyses—not reported here—excluding θ and including different weights to the two forms of inequality show that Turkish-Dutch subjects, on average, dislike both disadvantageous and advantageous inequality more than Dutch and Turkish subjects.

Thus, in our design we expect:

On average, inequality aversion (β) will be the highest for the Turkish-Dutch compared to Turkish and Dutch subjects. (H1b).

7.2.3 Alter effects on social preferences

Classical social psychological theories, such as social identity theory or realistic group conflict theory predict in-group favoritism. According to social identity theory, actors reduce cognitive complexity by placing objects including self and others in a few distinct categories. As actors also have a need for self enhancement, they attribute positive qualities to the in-group and negative attributes to the out-group (Tajfel, 1970). The tendency to treat in-group members more favorably than out-group members has been demonstrated with minimal group treatments (Tajfel, 1970). In addition, the realistic group conflict theory explains the hostility between groups as an outcome of the competition between groups for tangible material resources (Sherif et al., 1961). This hostility, in turn, yields positive discrimination towards the in-group and negative discrimination towards the out-group. Based on this social psychological literature on in-group favoritism, one expects that actors will have more pro-social

preferences towards the in-group than towards the out-group. Thus, in-group favoritism in our case would imply that:

Altruism (θ) will be higher towards the in-group than towards the out-group (H2a).

Again, what in-group favoritism would imply for inequality aversion is not clear. Thus, we refrain from formulating a hypothesis on the β parameter.

These social psychological theories are widely supported by numerous tests. However, studies that test in-group favoritism and at the same time follow the protocols of experimental economics, e.g., incentive compatibility and anonymity (Friedman & Cassar, 2004), report mixed results. Anonymity in the inter-group context implies that a subject knows the social identity of Alter, but not who the Alter is (Goette et al., 2012). Falk & Zhender (2007), Goette et al. (2012), Goette et al. (2006), and Simpson (2006) show that their subjects indeed display more cooperation and trust towards in-group members than towards out-group members. However, in these studies higher levels of cooperation and trust towards in-group is strategically justified, i.e., actors earn more by cooperating with and trusting an in-group member due to reciprocative behavior and self-fulfilling expectations. Moreover, Goette et al. (2006) show that while in-group cooperation is more frequent, an in-group member is punished more severely in case of free riding. This means that in-group favoritism observed in those studies does not necessarily mean that subjects have more pro-social motives towards in-group. Güth et al. (2008) and Bouckaert & Dhaene (2004), on the other hand, report that the social identity of Alter does not significantly affect trust and trustworthiness in the Trust Game. Bahry et al. (2005) also find no substantial difference between the levels of in-group and out-group trust. Whitt & Wilson (2007) conduct an experiment in postwar Bosnia-Herzegovina using subjects from ethnic groups that have been in conflict with each other for years. Even in this context, they observed very weak in-group favoritism. Finally, some studies even report *out-group favoritism*. Gil-White (2003) reports an inter-ethnic Ultimatum Game experiment with Mongols and Kazakhs in Western Monglia, and finds that subjects offer more to an out-group member than to an in-group member, and subjects are more reluctant to reject the offers of out-group members compared

to offers of in-group members. Heap & Patrick (2010) also report out-group favoritism in a Trust Game experiment with the Gisu people of Uganda. Out-group favoritism doesn't seem to be restricted to such "exotic" groups. Using representative data on native German subjects playing the Prisoner's Dilemma game with subjects of immigrant backgrounds, Veit & Koopmans (2010) also report out-group favoritism by native German subjects. Out-group favoritism is a very recent and surprising observation, given the dominance of the classical social psychological theories in the literature.

Usually, the first reaction to the absence of in-group favoritism or the existence of out-group favoritism is attributing it to social desirability. Subjects or respondents may have suppressed negative sentiments about the out-group during experiments or surveys. However, social desirability does not seem to be the major driving force behind the experiments that report no in-group favoritism. Bahry et al. (2005) discuss this issue at some length. Bahry et al. (2005) use interviewers from the same ethnic group as the respondents, which would reduce possible social desirability effects. Yet, they report substantial levels of out-group trust. Moreover, their subjects do not 'simply give rote positive responses about inter-ethnic relations', as their out-group trust varies across different out-groups. In addition, if social desirability is solely responsible for the lack of in-group favoritism, than that should also influence the responses of subjects on other psychological measures such as perceived social distance, stereotype traits etc. In other words, we should not observe, for instance, that subjects feel more distant towards the out-group than towards the in-group or that they report different stereotypical traits about the in-group and the out-group. However, as we will report below, subjects do discriminate out-group and in-group when, for example, perceived social distance is measured. Being a recently observed phenomenon, more research is needed to flesh out the exact mechanisms that yield "true" out-group favoritism. Some explanations have been proposed. Heap & Patrick (2010) argue that out-group favoritism functions as an "unconditional gift" which helps prevent potential conflicts between groups from being actualized. Gil-White (2003) interviewed the subjects after his experiment on why his subjects offered more to the out-group. One of the two most common answers he got was indeed in line

with this "unconditional gift" explanation: "I don't want to foster any misunderstanding between the groups". The function of out-group favoritism on preventing potential conflicts resonates also with a common feature of contexts where out-group favoritism is observed. Out-group favoritism seems to occur where "real" ethnic groups rather than artificial groups, e.g., minimal groups interact. This definitely applies to our case. Out-group favoritism implies the exact opposite of *H2a*, thus:

Altruism (θ) will be higher towards the out-group than towards the in-group (H2b).

Another mechanism may explain a special form of out-group favoritism. Research has shown that high status group members behave more altruistically towards low status group members than towards other high status group members (Liebe & Tutic, 2010; Hong & Bohnet, 2007). Liebe & Tutic (2010) and Hong & Bohnet (2007) explain this status effect using social exchange theory (Blau, 1964). Low status groups transfer privilege and status to high status groups, and in turn high status groups take more responsibility and behave more altruistically towards low status groups. Thus, in cases where interacting groups are clearly distinguished as low and high status groups, out-group favoritism displayed by the high status group can be explained by this status effect. This status effect can explain possible out-group favoritism displayed by Turkish and Dutch subjects towards Turkish-Dutch subjects. Turkish-Dutch are among the least accepted and most discriminated social groups in the Netherlands (Verkuyten & Zaremba, 2005; Verkuyten & Yildiz, 2006). Things are not much better for Turkish-Dutch when the attitudes of non-emigrant Turks toward Turkish-Dutch are considered. Turkish immigrants in Europe in general and Turkish-Dutch in particular are also subject to rejection and discrimination by non-emigrant Turks (Mandel, 1990). In other words, Turkish-Dutch could be perceived as a relatively lower status group by both Turkish and Dutch subjects. Consequently, this yields the following hypothesis:

On average, altruism (θ) of Turkish (Dutch) subjects will be higher towards Turkish-Dutch subjects than towards Turkish (Dutch) subjects (H2c).

It is unclear, however, how Turkish and Dutch subjects will perceive each

other with respect to status. Thus, we restrict *H2c* to (i) the differences between Turkish vs. Turkish-Dutch (T.TD) and Turkish vs. Turkish (T.T) conditions and (ii) the difference between Dutch vs. Turkish-Dutch (D.TD) and Dutch vs. Dutch (D.D) conditions.

Note that *H2a* and *H2b* are inconsistent, and a nil result would be inconclusive. However, these two hypotheses refer to different mechanisms and more direct tests of those mechanism are available. For example, *H2a* predicts not only in-group favoritism but also a positive association between perceived social distance toward Alter and Altruism. In the results section below, we investigate the relationship between perceived social distance and social preferences as a more elaborate test of *H2a*. In addition, the mechanism underlying *H2c* also provide a somewhat more elaborate prediction that the extent of out-group favoritism will be higher when D.D. versus D.TD. conditions are considered. This is because the Dutch and Turkish-Dutch are the real interacting groups and the potential for conflict is higher between these groups. Shortly, there are ways to test these two opposing hypotheses in a more elaborate way.

7.3 Methods

7.3.1 Subjects

Following Hermann et al. (2008), all of our subjects are undergraduate Turkish, Dutch, and Turkish-Dutch students. This way, except for the social group of Ego and Alter, the subjects would be more similar with regard to age, education, and probably also to socio-economic status, than a non-student sample. Thus, the observed effects of social identity on experimental responses are unlikely to be due to differences in the composition of subject pools.

Turkish Sample: We recruited 113 Turkish students from Bogazici University, Istanbul. These students were recruited by posting a list on the Psychology Department board. Students from any study field could subscribe in one of the six sessions by writing their names in the list. Bogazici University is a public university with a western oriented education system. With mainly middle class students, it is comparable to its western counterparts.

Turkish data used in Hermann et al. (2008) and Gaechter et al. (2010) were also collected at Bogazici University.

Dutch Sample: The Dutch sample includes 103 Dutch students from Utrecht University recruited using ORSEE (Greiner, 2004). They were invited to one of the 8 sessions conducted in the Experimental Lab for Sociology and Economics (ELSE). Recruiting these Dutch students, we paid specific attention to make the Dutch sample as similar as possible to the Turkish sample, with respect to gender, age, and study field. Especially students studying economics tend to behave slightly differently from other students, thus it was one of the factors we tried to keep balanced in subject pools. Table 7.1a summarizes the composition of the subject pools with respect to these characteristics.

Turkish-Dutch Sample: This sample was gathered differently. Because it was significantly more difficult to reach Turkish-Dutch students, we used several methods for recruitment. We e-mailed people with Turkish names in the Utrecht University e-mail database, and other student e-mail groups, posted an advertisement in a university newspaper, and used social contacts and a Turkish *döner* shop to reach students of Turkish origins who spent most of their lives in the Netherlands. These Turkish-Dutch students were invited to the preplanned sessions at the ELSE lab of Utrecht University. In case students were unable to attend these sessions, they filled in the experiment booklet at home. This resulted in participation of 40 Turkish-Dutch subjects in the experiment.

7.3.2 Design and procedure

In the experiment, subjects from the three social groups were matched with each other. We employed a 3 (Ego's social group) by 3 (Alter's social group) minus 1 condition, between subjects design. Originally, we planned a 3×3 full factorial design, but due to the relatively low number of subjects in the Turkish-Dutch sample, we had to exclude one condition, Turkish-Dutch vs. Turkish-Dutch. The social group of Alter with whom Ego was matched was randomized. This resulted in the distribution of the number of subjects per condition as shown in Table 7.1b. We opted for a between subjects design

Table 7.1: (a) Subject pool composition. (b) number of cases per condition.

(a) Composition of subject pools

Pool	% male	% economics	average age
Turkish	27	13	21
Turkish Dutch	54	3	22
Dutch	27	16	22

(b) N(Subjects) per condition

	Alter's identity		
Ego's identity	Turkish	Turkish Dutch	Dutch
Turkish	38	37	38
Turkish Dutch	23		17
Dutch	34	35	34

in order to disguise the overall design of the experiment and thus reduce the possible impact of social desirability. A subject knew only the social identity of the group that she is matched with, but didn't know that the experiment involved three social groups.

Subjects were given the experiment booklet, which started with general instructions stressing important elements of the experiment such as incentive comparability, anonymity, and interacting with real partners (Friedman & Cassar, 2004). It is nearly impossible to ensure that all subjects trusted the procedure fully, such as the existence of real other participants. To reduce this possibility, however, we underlined that the experiment was approved by the ethic committee and included a brief assurance letter from the supervisor of the project with the contact details of the researchers for any further inquiry. Turkish subjects completed the experiment in the Turkish language, Dutch subjects in the Dutch language. Turkish-Dutch subjects were given the option to do the experiment either in the Turkish or Dutch language. Besides the authors, translations to those languages were double checked by two other experts.

Game-play. After these general instructions, subject played 18 binary Dictator Games without feedback with stranger matching from the target social group. Each of these Dictator Games involved two options where each option comprised an outcome allocation for the self and another person from the target social group. In each DG, a different recipient is randomly selected from the target group. See the Appendix E.1 for details of the DGs used. Subjects were instructed to choose one of the options in the DGs, according to their preference. Subjects were explicitly informed about the social identity of Alter. We explained them that we recruited a number of (depending on the condition, Turkish, Turkish-Dutch or Dutch) students (from Bogazici University or Utrecht University) and the recipient in the Dictator Game would be a randomly chosen subject from the group that the subject was matched with. We also made it clear that in each DG a new recipient from the target group would be randomly selected, i.e., games were one-shot. Next to their decisions as "senders" in the DGs, as a subsequent step, subjects also passively earned points as "recipients" of other randomly selected participants from the target group.

The 18 DGs were constructed based on Aksoy & Weesie (2013a). The outcomes in these 18 DGs were chosen to facilitate the statistical estimation of θ and β parameters (Aksoy & Weesie, 2013a). The order of the 18 DGs was randomized per subject. Subjects were paid for all games.

Perceived social distance of Alter. In the experiment, we manipulated the target group that Ego was matched with. But it could be possible that subjects did not discriminate the out-group, perceiving the target group as socially non-distant. Thus, in each of the three groups we administrated two classical measures of attitudes towards Turks, Turkish-Dutch, and Dutch. These were the Bogardus social distance measure and the feeling thermometer (Hagendoorn et al., 1998). The Bogardus social distance measure comprised three items asking how a subject would feel if a Turkish/Turkish-Dutch/Dutch person became a relative through marriage/classmate/neighbor (1= very negative, 9=very positive). The scale reliability coefficient α for these three items is larger than 0.8 in all of our experimental conditions. The Bogardus social distance measure is obtained by averaging these three items. The thermome-

ter measure asked subjects to express their feelings on a scale from 0 (very negative) to 100 (very positive) towards Turks/Turkish-Dutch/Dutch.

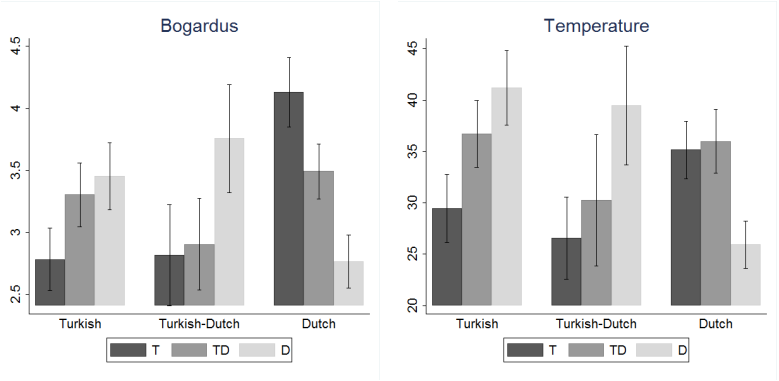
Payment. Before leaving, each subject was paid a show-up fee of 5 Euros (or 10 Liras in Turkey) directly in cash. After subjects completed the experiment, they also received a final payment instruction for the earnings they collected throughout the experiment. As the matching of subjects with other subjects and the calculation of their final earnings could only be done after all subjects completed the experiment, payment took place 4 weeks later. Subjects were given the choice to receive the payment on their bank accounts or to collect the payment in person using a code slip. Only three subjects chose to collect the payment in person. To ensure anonymity, subjects gave their bank account numbers and names in a closed and sealed envelope, submitted separately from the experiment booklet. After matching subjects' decisions with other subjects using unique subject identifier numbers printed on experiment booklets, overall earnings of the subjects were calculated and bank-transferred or given in person. Subjects were fully informed on this payment procedure. We received no expression of confusion on the payment procedure, of disbelief in whether the subjects played those games with real people, or of distrust on whether the payments were actually going to happen.

7.4 Results

7.4.1 Manipulation check, perceived social distance and feeling thermometer

Figure 7.1 shows the average perceived social distance towards Turkish, Turkish -Dutch, and Dutch. Clearly, subjects do feel differently towards the three social groups. All differences discussed here are statistically significant at the $\alpha = 0.05$ level for two-sided paired t-tests, unless stated otherwise. Turks perceive other Turks as least distant and the Dutch as most distant, the Turkish-Dutch being in the middle. Yet, the difference between the perceived social distance of Turks towards Turkish-Dutch and Dutch is statistically insignificant. Dutch subjects perceive Turks as most distant and the Dutch as least distant, again with the Turkish-Dutch being in the middle. Turkish-Dutch

Figure 7.1: Average perceived social distance of Turkish, Turkish-Dutch, and Dutch subjects towards Turkish (T), Turkish-Dutch (TD), and Dutch (D) people, and associated 95% confidence intervals. The left panel is the Bogardus social distance measure, the right panel is the feeling thermometer measure. Scores are coded such that a higher score means a higher perceived social distance.



subjects perceive Turks as least distant, and the Dutch as most distant. Although Turkish-Dutch subjects feel closer to Turks than to the Turkish-Dutch, the difference between the two is statistically insignificant (t(36)=-0.24, p(2-sided)=.8).

Almost the same pattern is observed for the feeling thermometer. The only different finding is that now Dutch subjects do not differentiate Turks and Turkish-Dutch (t(101)=-0.69, p(2-sided)=.5). Again, Turkish-Dutch subjects seem to feel warmer towards Turks than to other Turkish-Dutch, but the difference is statistically insignificant (t(34)=-0.73, p(2-sided)=.47). The Pearson correlation between the thermometer measure and the Bogardus measure across all 8 conditions is 0.52 (p(2-sided)<0.01).

To summarize, subjects do discriminate the three social groups, perceiving their own social group to be least distant. The only exception is Turkish-Dutch subjects who perceive both Turks and the other Turkish-Dutch as in-group, and the Dutch as out-group.

Table 7.2: Means and in parentheses standard errors of the means of the altruism (θ) and inequality aversion (β) parameters, per condition. T=Turkish, TD=Turkish-Dutch, D=Dutch. See Appendix E.2 for additional results.

	Alter					
Ego	T		TD		D	
	μ_θ	μ_β	μ_θ	μ_β	μ_θ	μ_β
T	-0.066$^+$	0.030	0.070	0.094$^+$	-0.069	0.112*
	(0.036)	(0.034)	(0.070)	(0.056)	(0.053)	(0.046)
TD	-0.245	0.589**			-0.092	0.167*
	(0.215)	(0.250)			(0.066)	(0.080)
D	0.104*	0.081**	0.123**	0.091**	0.049	0.104**
	(0.041)	(0.024)	(0.039)	(0.031)	(0.030)	(0.033)

**p<0.01;*p<0.05;$^+$p<0.1 for two sided tests.

One may claim that these two measures capture social distance towards the target group as a whole, but not towards students from the target group. One item in the Bogardus social distance was explicitly about students, asking what would a subject feel if a Turkish/Turkish-Dutch/Dutch becomes a classmate. A separate analysis of this single item yields the same results as of the combined scale.

7.4.2 Social preferences: altruism and inequality aversion

In a binary Dictator Game a subject prefers option A over option B if the utility from option A is larger than option B, that is $U_A > U_B$. Due to the random component in the utility specification, whether $U_A > U_B$ depends not only the outcomes in the game and the social preference parameters, but also on the evaluation error. Thus, we can only say something about the probability that a subject prefers option A over B. Under our specification,

$$\begin{aligned}\Pr(U_A > U_B) &= (x_A + \theta \cdot y_A - \beta \cdot |x_A - y_A| + \epsilon_A) > \\ &\quad (x_B + \theta \cdot y_B - \beta \cdot |x_B - y_B| + \epsilon_B) \\ &= \Pr\left(\Delta_x + \theta \cdot \Delta_y - \beta \cdot \Delta_{xy} + (\Delta_\epsilon) > 0\right) \\ &= \Phi\left(\frac{\Delta_x + \theta \cdot \Delta_y - \beta \cdot \Delta_{xy}}{\sqrt{2}\tau}\right). \qquad (7.2)\end{aligned}$$

where $\Delta_x = x_A - x_B$ is the outcome difference between option A and option B for Ego, $\Delta_y = y_A - y_B$ is the outcome difference between option A and B for Alter, $\Delta_{xy} = |y_A - x_A| - |y_B - x_B|$, Φ is the cumulative standard normal distribution, and $\Delta_\epsilon \sim N(0, 2\tau^2)$ (Aksoy & Weesie, 2012b). We assume that (θ, β) follow a bivariate normal distribution, conditional on Ego's and Alter's social group. Given these assumptions, equation (7.2) is just a probit model. The means and (co)variances of (θ, β) per experimental condition can be estimated with a multilevel probit model with random coefficients. In this multilevel probit model, the dependent variable is a subject's choice in a Dictator Game. The independent variables are the three terms capturing the utility differences between the two options of a Dictator Game, namely, Δ_x, Δ_y and Δ_{xy}. Note that Δ_x is an offset, an independent variable with coefficient fixed to 1.

This random coefficients model implies that, given the experimental condition, there is a population of subjects described by the means and variances of θ and β parameters, and our subjects are sampled from those populations. We simply estimate those means and variances in the population. Using fixed effects, it is possible to estimate a θ and a β per subject without making the additional assumption of bivariate normality. Using fixed effects, a subject's θ and β are estimated independent of her peers in the experimental condition. We opted for a random effects approach because fixed effects are known to produce inconsistent estimates of the parameters, especially if the number of cases per subject is, as in our design, low (Andersen, 1980). For details of this estimation method and a discussion of the assumptions of the model, see Aksoy & Weesie (2012b).

We fitted a multilevel probit model explained above for each of the eight Ego-Alter combinations, using the Stata program GLLAMM (Rabe-Hesketh et al., 2002). In all of the eight experimental conditions, Wald tests show that the correlation between θ and β were statistically insignificant. Consequently, we report results for models in which this correlation is fixed to zero. We performed the same analyses without constraining this correlation to zero, and the results were effectively identical. Table 7.2 shows a selection of the results relevant for our hypotheses: the means of θ and β per condition, and the standard errors of those means. Appendix E.2 includes a detailed table with estimates of all parameters including variances of θ and β, log-likelihoods, number of cases etc. per experimental condition. Using the means and their standard errors in Table 7.2, it is possible to compute a test for any (linear) combination of those means: as we employed a between subjects design, the results across the eight groups are independent. Below, we report a selection of statistical tests of our hypotheses.

Ego effects

First compare the Turkish vs. Turkish (T.T) and Dutch vs. Dutch (D.D) conditions. *H1a* predicts that, on average, altruism (θ) will be lower in the T.T condition than in the D.D condition. Indeed, the mean of θ is lower in the T.T condition compared to the D.D condition (z=-2.43, p(1-sided)<0.01). According to *H1b*, Turkish-Dutch subjects would be the most inequality averse group. Indeed on average Turkish-Dutch subjects seem to have the highest inequality aversion (β) parameter values. Since we lack a Turkish-Dutch vs. Turkish-Dutch (TD.TD) condition, however, it is not obvious how to compare the Turkish-Dutch sample with the Turkish and Dutch sample. As reported in Figure 7.1, Turkish-Dutch subjects perceive Turks as in-group and they do not differentiate Turkish and other Turkish-Dutch with respect to social distance. Thus, we use the Turkish-Dutch versus Turkish (TD.T) condition to compare Turkish-Dutch subjects with Turkish and Dutch subjects. As *H1b* predicts, the mean of β in the TD.T condition is higher than that in the T.T (z=2.21, p(1-sided)<0.05) and in the D.D (z=1.92, p(1-sided)<0.05) conditions.

For exploratory purposes, we report that the mean of the inequality aver-

sion parameter β is not significantly different between the T.T and the D.D conditions (z=-1.55, p(2-sided)=0.12). We did not formulate hypotheses on the altruism parameter θ for the Turkish-Dutch sample. Still, the overall mean of θ in the TD.T condition is not significantly different from that in the T.T (z=0.82, p(2-sided)=0.41) and that in the D.D condition (z=-1.36, p(2-sided)<0.18).

Alter effects

We do not observe any in-group bias in the altruism parameter θ (*H2a*). On the contrary, in line with *H2b* the mean of θ is higher towards the out-group than the in-group for the Turkish and Dutch samples. However, the cases where the effect of Alter approaches statistical significance are the ones predicted by *H2c*: The mean of θ is higher in the T.TD condition than in the T.T condition (z=1.72, p(1-sided)<0.05). Similarly, the mean of θ is higher in the D.TD condition than in the D.D condition (z=1.52, p(1-sided)=0.06). For the inequality aversion parameter β, the effect of Alter is quite minor, and the only sizable but statistically insignificant differences are between the T.TD and T.D conditions (z=1.44, p(2-sided)=0.15) and between the TD.T and TD.D conditions (z=1.61, p(2-sided)=0.11).

Since the results on Alter effects are not very conclusive, and we have two opposing hypotheses *H2a* and *H2b*, we perform further analyses. Using the Mplus software (Muthén & Muthén, 1998–2010), we fitted three separate models for the Turkish, Turkish-Dutch, and the Dutch sample. In those models, the means of θ and β are estimated conditional on Alter's social group, whereas for statistical convenience the variance of θ and β as well as the decision error τ are constrained to be invariant over Alter's social group. Aksoy & Weesie (2012c) includes a detailed description of the Mplus syntax to fit such models. Simultaneously, θ is regressed on a latent measure of perceived social distance toward alter which is constructed from the items of the Bogardus and the thermometer measures. Perceived social distance has no significant effect on θ for the Turkish and Turkish-Dutch sample but has a statistically significant *positive* effect on θ for the Dutch sample (b=0.042, z=2.67, p-2sided<0.01). In addition, when the social distance is controlled for, the means of θ in the D.T

(b=.106, z=2.14, p-2sided= 0.03) and D.TD (b=0.081, z=.734, p-2sided=0.08) conditions are both significantly higher than that in the D.D condition. Also in the Turkish sample, when social distance is controlled for, the mean of θ in the T.TD condition is significantly higher than that in the T.TD condition (b=0.147, z=2.22, p-2sided=0.03). These results do not support *H2a* and partially corroborate *H2b* and *H2c*.

Finally, in our subject pool, gender has no significant effect on (θ, β). This is important because there are relatively more male subjects in the Turkish-Dutch group.

7.5 Discussion and conclusions

In this study we report on an international inter-ethnic laboratory experiment conducted with Turkish, Dutch, and Turkish-Dutch subjects (subjects of Turkish origin from the Netherlands). With this design we investigate the influence of Ego's and Alter's social group on Ego's social preferences. We consider two types of social preferences: (1) altruism reflecting a concern for the outcome of Alter and (2) inequality aversion reflecting a concern for the absolute difference between the outcomes for Ego and Alter. In our analysis, we assume that these two types of social preferences depend on the social groups of Ego and Alter.

We find that, on average, Turkish subjects attach a lower weight to the outcome of Alter and to the inequality between Ego and Alter compared to Dutch subjects. In other words, on average, Turkish subjects have more pro-self preferences compared to the Dutch subjects. This result is in line with the consistent finding of relatively lower levels of cooperative and trusting behavior observed in Turkey compared with Western European countries. As far as we know, our study is the first to report on social preferences of Turkish-Dutch subjects. As we predicted, we find that Turkish-Dutch subjects dislike inequality much more than both Turkish and Dutch subjects: being in the minority seems to increase inequality averseness.

We also analyze the effects of Alter's social group on Ego's social preferences. In contrast to what we expect from classical social psychological

theories, we find no in-group favoritism. On the contrary, we observe some "out-group favoritism": subjects tend to attach a higher value to the outcome of Alter when Alter is from an out-group than from an in-group. However, this out-group favoritism is not a strong and a general trend, but restricted mainly to the cases where Turkish and Dutch subjects are matched with Turkish-Dutch subjects. Also we find, for the Dutch respondents, that the weight that a subject attaches to the outcome of Alter *increases* in perceived social distance toward Alter. Thus, absence of in-group favoritism and some mild out-group favoritism seem to be present both at the individual and at the aggregate level.

Our results on the non-existence of in-group bias and presence of some out-group bias adds up to the inconsistent results reported in the literature, and calls for an alternative approach than naively applying the classical social psychological theories on behavioral-game theory experiments. It remains a puzzle to discover under which conditions in-group favoritism is observed, under which conditions it disappears or even out-group favoritism is observed. This is definitely an important direction for future research. The issue seems to have several dimensions. First, the nature of groups is an obvious factor that potentially determines whether in-group bias is observed: e.g., whether the groups are minimal or real groups, whether in-group members have personal contacts with each other, whether interactions are anonymous or not, and so on. Additionally, status differences between groups, as we discussed above, is a factor that influences in-group vs. out-group sentiments. Secondly, the nature of the interaction could be an important factor. When Ego interacts with an out-group Alter, Ego's social preferences toward Alter, Ego's choices, Ego's expectations about the preferences and behavior of Alter as well as Alter's actual choices could potentially influence Ego's outcomes. In our design, the outcomes in a given binary Dictator Game are fixed and what could Ego get from this single game is solely determined by Ego. Alter is a passive recipient. In many other types of interactions, however, Alter can influence Ego's outcomes, as Ego can influence Alter's. In such situations, out-group discrimination and in-group favoritism could emerge as a result of negative expectations not negative preferences. Imagine an alternative design. In this

alternative design Ego would play a Trust Game with an Alter, and Ego is given the option to play the game with either an in-group Alter or an out-group Alter. In such a design, the expected outcomes of the interaction situation could be influenced by Alter. Then, one would expect homophily from Ego if Ego expects out-group discrimination from Alter.

To conclude, our results suggest that, despite the rising anti-immigration climate and an increase in the disbelief in multiculturalism and the hostility between social groups, interactions between individuals from different ethnic groups do not necessarily yield less pro-social preferences. On the contrary, reiterating the famous contact hypothesis of Allport (1954), inter-group interactions have the potential to promote pro-social preferences.

Chapter 8

Discussion and conclusions

The core formal-theoretical framework of the thesis is developed in Chapters 2 to 5. These four chapters provide a game-theoretic framework that aims to explain cooperation in non-embedded settings. Each of these chapters studies a part of the overarching theoretical framework and builds on the previous chapter with increasing statistical and theoretical sophistication. The game-theoretic cogs and wheels laid out in these chapters are calibrated with experimental data. According to the theoretical and empirical analyses presented, cooperation in non-embedded settings can be explained by three factors: (1) heterogeneous social preferences, (2) ego-centered—consensus type—biases in beliefs, and (3) evaluation error. These three factors correspond to two core assumptions of the standard rational choice model. The first of the three factors revises the selfishness assumption, and the other two "stretch" the rationality assumption of the standard model.

The two final empirical chapters, Chapters 6 and 7, apply this framework to two sociologically important contexts. Chapter 6 experimentally studies the influence of procedural justice on cooperation. It reports that procedural justice indeed influences cooperation significantly. Moreover, gender is found to be an important factor interacting with the association between procedural justice and cooperative behavior. Chapter 7 reports an international inter-ethnic experiment with Turkish, Dutch, and Turkish-Dutch respondents. It shows clear inter-group differences in social preferences resembling the "North

vs. South" effect (Delhey & Newton, 2005), but the social distance between Ego and Alter does not seem to matter, i.e., no ingroup favoritism is found.

In this concluding chapter, we provide a "bird's-eye view" of general conclusions based on the main results. To avoid repetition, we refrain from providing an extensive summary of the findings presented in the chapters. Each chapter includes an abstract summarizing the main findings of that chapter.

8.1 Main results and conclusions

8.1.1 Social preferences as states and traits

The findings presented in this thesis show that social preferences are complex. They are influenced by several factors, such as culture (Chapter 7), procedural justice (Chapter 6), and history (Chapter 4). Even the game structure itself has an influence on social preferences. For example, as we show in Chapter 4, the average weight subjects attach to the material outcomes of others is significantly smaller in sequential social dilemmas compared with decision situations where people simply allocate outcomes between self and others—Dictator Games. Thus, social preferences are endogenous to the decision situation, i.e., social preferences are *states*, at least partially (Steyer et al., 1999). This means that it is not possible to fully measure social preferences "out-of context" because context has an effect on social preferences. These findings resonate with the literature on "framing" (e.g., Hertel & Fiedler, 1994, 1998; Liberman et al., 2004; Lindenberg & Steg, 2007). This literature shows that relatively subtle contextual cues, such as cooperative adjectives presented before game play or labeling the social dilemma game as a "Community Game" instead of a "Wall Street Game", may substantially influence social preferences and cooperative behavior (Hertel & Fiedler, 1994, 1998; Liberman et al., 2004). Goal framing theory may help understand the nature of the influence of context on social preferences (Lindenberg, 2001, 2008). This theory proposes that different motivational goals coexist. In our case, some of these different goals could be maximizing only one's own utility, minimizing inequality, or maximizing others' outcomes, i.e., different types of social preferences. These different goals compete with each other for the limited cognitive resources of the individual.

The goal that dominates this competition determines the frame through which a person evaluates the situation. Game structure, history, and other contextual factors, such as procedural justice, may work via the situational selection of goals, making certain types of goals or social preferences more salient.

We also find a high level of individual variation with respect to social preferences. Moreover, although contextual factors such as game structure and history influence social preferences, the social preferences of a person are highly correlated across different contexts. Such correlations of the social preferences of a person across time and contexts is also documented in other research (e.g., Van Lange, 1999; Van Lange et al., 2007; Benz & Meier, 2008). For instance, Benz & Meier (2008) report that their subjects' pro-social behavior, such as voluntary donations during the experiment, correlates moderately with the field behavior of those subjects up to two years before and after the experiment was conducted. This finding indicates that social preferences are also dispositional *traits*, at least partially.

To complicate the matter further, our results also suggest that the effect of context on social preferences is not stable across people. That is, people's social preferences are (partially) influenced by context, but some people's social preferences are influenced more than others. For example, in Chapter 4 we show that the weight one attaches to the material outcomes of others varies with the game structure and history. But how much this weight varies, also varies between persons. This indicates an *interaction* between dispositional and contextual determinants of social preferences. Similar interactions between disposition and context within the social dilemma setup are also reported in previous research (e.g., Van Lange et al., 1997; De Cremer & Van Vugt, 1999; Van Lange, 2000; Bogaert et al., 2008). For instance, De Cremer & Van Vugt (1999) show that inducing group identity experimentally makes dispositionally pro-self individuals behave more cooperatively compared with a treatment where group identity is not salient. The same manipulation, however, has no significant effect on dispositionally pro-social people who already assign high weights to others' outcomes, irrespective of group identity.

To summarize these findings under a single framework, let vector $\boldsymbol{\theta}_{ij}$ in-

formally denote person i's social preferences, such as how much i cares about others' material outcomes and dislikes inequality, in context j, which may include, for example, history, game structure, or the social identity of interaction partners. Our results, as well as those of the literature discussed here, suggest that $\boldsymbol{\theta}_{ij}$ can be broken down into three components:[1]

$$\boldsymbol{\theta}_{ij} = \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \boldsymbol{\gamma}_{ij}. \tag{8.1}$$

where $\boldsymbol{\alpha}_i$ represents the trait component, $\boldsymbol{\beta}_j$ represents the state component, and $\boldsymbol{\gamma}_{ij}$ represents the interaction between the two.[2]

As we discussed in the introductory chapter, explaining the origins of social preferences as traits and providing a thorough explanation of why and how context influences social preferences is beyond the scope of the thesis. Our findings and the related literature discussed thus far call for a general "meta-theory" on social preferences. Such a meta-theory would describe where social preferences as traits come from and would predict how situational factors influence social preferences. In other words, it identifies the factors that influence the terms in equation (8.1), $\boldsymbol{\alpha}_i$, $\boldsymbol{\beta}_j$, and $\boldsymbol{\gamma}_{ij}$. In the section "future research" below, we will briefly discuss how one could systematically study the contextual and dispositional determinants of social preferences.

Irrespective of their origins, the complex nature of social preferences is most likely not what a rational choice theorist welcomes. Reality is more complex than the standard model assumes. Of course, in formal modeling some complexities have to be ignored. In fact, simplicity is the strong suit of formal modeling, provided that ignoring the complexity does not result in a poor empirical fit and that including the complexities in formal models is not feasible for analytical tractability.[3] Neither of these conditions is met in our case. First, ignoring the complexity yields a very poor fit. As we

[1] For social psychologists, equation (8.1) would be quite natural, as it is very similar to the well-known formula by Lewin (1936) (see also Van Lange, 2000). In Lewin's formula, behavior is a function of the environment and the person. The most important difference between equation (8.1) and Lewin's formula is that equation (8.1) is not about behavior, but about social preferences.

[2] $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ here should not be confused with α and β in Chapter 2 or $\boldsymbol{\gamma}$ with γ in Chapter 5, which denoted particular types of social preferences.

[3] For an informal discussion on the aims of formal modeling, see Carvalho (2007).

show in Chapter 4, models in which contextual influences on social preferences are ignored are flatly rejected by experimental data. Second, as we show in several chapters, there are statistical and theoretical tools available for accommodating these complexities. Thus, there seems to be no compelling reason to shy away from these complexities.

We should admit that there are other forms of complexities that we ignored in this thesis. The non-exhaustive list of factors we ignored within the social dilemma context includes risk preferences (Raub & Snijders, 1997), other types of social preferences than those considered in this thesis, such as regret or guilt aversion (Loomes & Sugden, 1982; Ellingsen et al., 2010), the role of pre-play communication (Balliet, 2010), and signals and signs that help reveal social preferences (Frank, 1988; Bacharach & Gambetta, 2001). We believe that, at least in principle, many of those factors can be incorporated into the framework of the thesis.

8.1.2 Beliefs about others' preferences: game theory as decision theory

> No man really knows about other human beings. The best he can do is to suppose that they are like himself (John Steinbeck, *The Winter of Our Discontent*).

Although social preferences are complex, our results suggest that beliefs about others' social preferences are not that complex, at least considering the cases we study in this thesis. Beliefs about others' social preferences are highly correlated with one's own preferences in *all* conditions and manipulations we analyzed: the simultaneous Prisoner's Dilemma (PD), sequential PD, binary Dictator Games (DGs), beliefs measured after one's own choices, beliefs measured simultaneously with one's own choices, and incentivized and non-incentivized elicitation of beliefs.[4] This feature of beliefs helps the researcher because incorporating ego-centered beliefs in formal models is easy, as we show in Chapter 5. At the same time, however, ego-centered beliefs have

[4]Strictly speaking, we refer, in particular, to correlations between social preferences and the *means* of beliefs about others' social preferences.

implications for the rational beliefs component of the rationality assumption in the standard model, as we will discuss below briefly.

Our findings on the correlation between social preferences and beliefs are in line with past research in social psychology (e.g., Kelley & Stahelski, 1970; Ross et al., 1977; Kuhlman et al., 1992; Iedema, 1993) and, more recently, in experimental economics (e.g., Blanco et al., 2009, 2011), which demonstrate similar ego-centered—consensus-type—biases in beliefs. Our study documents these biases ameliorating certain methodological shortcomings in the social psychological literature and in light of clearly defined social preference models. Another contribution of our study is that we explicitly analyze the *variance* of beliefs about others' social preferences as a function of one's own social preferences. We find, as a general trend, that the variance of beliefs is the smallest for selfish people, and thus, selfish people are more certain about their beliefs. This variance increases as social preferences deviate from selfishness, corroborating the cone effect (Iedema, 1993).[5] Although Iedema (1993) predicted this cone pattern two decades ago, it has not received much attention in the literature thus far.

In this thesis, we do not extensively discuss the causal mechanisms underlying the relationship between social preferences and beliefs about others' social preferences. In Chapter 2, we test the *implications* of three social psychological hypotheses, but do not delve much into the causal mechanisms. The consensus effect, expecting others to be similar to the self, underlies all three hypotheses. There are several mutually non-exclusive mechanisms that could in principle cause the consensus effect. Iedema (1993) discusses these mechanisms in detail. For example, the consensus effect occurs if beliefs are reflections of one's own preferences: when people do not have enough information about the tastes of others, they use the available observation, their own tastes, to guess the tastes of others. This is called the cognitive availability bias (Marks & Miller, 1987). Even if there are other observations available, e.g., people's encounters with others throughout their life, they tend to be similar to the self. This is because people tend to associate with others who are similar to themselves with respect to, for example, background, interests,

[5] See Chapters 2 and 3 for details.

tastes, which further reinforces the "false" consensus effect. This mechanism is called selective exposure (Ross et al., 1977). Selective exposure and the availability bias point to the causal direction of preferences causing beliefs. Alternatively, the consensus effect occurs if one adjusts one's preferences to one's beliefs, due to, e.g., reciprocity or conformity. For example, one may expect another person to be highly pro-socially motivated toward the self, and because of this belief one may also feel pro-socially towards the other due to reciprocity. Similarly, if one expects a specific type of social preference to be the common norm across others, one may adjust one's own preferences to conform to this social norm (Asch, 1951). In these latter two examples, beliefs cause preferences. We do not concern ourselves over which causal mechanism is at work, whether more than one is at work, or, if so, which one is stronger as long as the consensus effect is present. Irrespective of the exact causal mechanisms, the strong correlation between preferences and beliefs is an important factor in the analysis of non-embedded social dilemmas.

The rational beliefs assumption, which underlies the Bayesian-Nash equilibrium framework of games with incomplete information, implies that people share a common prior belief about the distribution of others' (social) preferences. Moreover, this common prior belief corresponds to the true distribution of social preferences in the population (Harsanyi, 1968). However, the high correlation between beliefs and preferences, together with the heterogeneity in social preferences we document in this thesis, shows that people are also heterogeneous with respect to their beliefs. In other words, people do not share a common prior belief about the distribution of social preferences. In fact, we explicitly test and refute the rational beliefs assumption in several chapters. This means that interpreting the Bayesian-Nash equilibrium as an accurate representation of the macro state, at least for the non-embedded case, would be too naive. Nevertheless, this does not mean that we should abolish game theory for studying non-embedded interactions. The uncommon prior beliefs we document in this paper correspond to what McKelvey & Palfrey (1992) call the *egocentric model*. In the egocentric model, everybody plays the game in his or her own world, believing that his or her own world is the "true" world. This belief of actors is obviously not true, at least for some actors,

as there is most likely only one true world. However, as long as people play the game *as if* their personal world is the true world, game-theoretic models can be used to predict the choices of actors and their beliefs about others' choices. For example, imagine a person who attaches a very high weight to the outcomes of others; this person is thus quite pro-social. In addition, as the consensus effect implies, this person believes that most other people are also pro-socially motivated. Assume that this person will be playing a game, say a Prisoner's Dilemma game, in an experiment with an anonymous opponent. We can still solve for the Bayesian-Nash equilibrium of this Prisoner's Dilemma game where the "type" distribution corresponds to the (overly optimistic) beliefs of this particular person.[6] Because the person plays the game as if her own (overly optimistic) beliefs are the true type distribution, her choices and her beliefs about the other's choices in the PD can be predicted using the Bayesian-Nash equilibrium framework. Moreover, this prediction can be tested empirically.

There is a caveat, however, for dropping the common prior beliefs assumption. It is the same caveat as for introducing heterogeneous social preferences: one can explain many phenomena by adjusting the beliefs *ex post* arbitrarily (for an overview, see Morris, 1995). Thus, explaining phenomena by adjusting beliefs *ex post* is not scientific. To demonstrate this point, continuing with the above example, assume that we predicted our highly pro-social respondent would cooperate in the PD, but she chose not to cooperate. We can explain this non-cooperative behavior of our respondent by adjusting her beliefs *ex post* arbitrarily, e.g., this person did not cooperate because she *believed* that most others were pro-selfish and thus would defect.[7] To avoid this fallacy, one should be careful in introducing heterogeneous prior beliefs. We do not advocate adjusting beliefs *ex post*, but rather modeling the relationship between one's own preferences and beliefs *ex ante* and carefully testing the part of the model regarding beliefs.

[6] Calculating the Bayesian-Nash equilibrium, all higher order beliefs of this particular person can also be derived from this person's subjective prior distribution.

[7] In some cases, equilibrium behavior does not depend on beliefs, as we show in Chapter 5. For example, in certain PDs, such as parallel PDs, a pro-social respondent's cooperative behavior may not depend on her beliefs about others' social preferences.

As we discuss at length in Chapters 4 and 5, in some situations beliefs about others' social preferences are as important as social preferences for the explanation of cooperative behavior. In the behavioral game theory literature, the focus so far has been on the latter. There are a vast number of models proposed in the literature that include some form of social preferences (see Fehr & Schmidt, 2006). Despite their importance, however, beliefs are not studied to the same extent. We suggest reporting explicitly which model for beliefs is used next to the model for social preferences. We prefer using the terminology "social preferences-belief model" as an explanation of behavior rather than mentioning only the social preferences model. For example, we prefer "... in this study, we explain the behavior of our respondents in a series of Trust Games with an inequality aversion-rational beliefs model" over "... in this study, we explain the behavior of our respondents in a series of Trust Games with an inequality aversion model".

8.2 Future research

8.2.1 Explaining variation in social preferences

Research on the explanation of the existence and variation of social preferences is an obvious next step. We treat social preferences partially as individual *traits*, namely, as an aspect of an individual's psychological makeup, and partially as *states* influenced by the context. There is also heterogeneity in social preferences as traits and states. How can one explain the existence of and the variance in social preferences as traits and states?

There is a debate over whether pro-social preferences towards strangers can evolve because pro-social preferences incur costs to the self while benefiting genetically unrelated others. There are several evolutionary models showing that such preferences can indeed evolve and that heterogeneity in preferences can be sustained under some conditions. These models involve various mechanisms such as group selection (e.g., Bowles & Gintis, 2004) or the presence of certain observable features that reliably signal the social preferences of others (e.g., Frank, 1988), gene-culture interaction (e.g., Henrich & Henrich, 2007) and social preferences as byproducts of skills for social interaction (De Waal,

2010). We believe that further research on the evolutionary roots of social preferences may provide a theoretical basis for the assumptions of the game-theoretic framework developed in this thesis. In addition to the evolutionary roots of social preferences, further sociological research on the proximate causes of social preferences is important. Already in this thesis we investigated one factor, namely, one's culture, in Chapter 7, and showed that social preferences indeed vary between social groups. This indicates that social preferences are similar to internalized norms and are influenced by macro-sociological conditions. A further study with large- scale data that regresses social preferences on macro variables, such as indicators of macro-economic conditions and several social institutions (e.g., Delhey & Newton, 2005; Buchan et al., 2009), micro variables such as age, education, religion, and gender (e.g., Gheorghiu et al., 2009; Bekkers & Wiepking, 2011), and socialization variables (Gattig, 2002), would help develop a sociological explanation of heterogeneous social preferences.

We also showed that contextual factors influence social preferences ($\boldsymbol{\beta}_j$ in (8.1)). Moreover, contextual and dispositional factors interact ($\boldsymbol{\gamma}_{ij}$ in (8.1)). However, we did not systematically analyze which feature of the context influences exactly what type of social preferences and the nature of the interaction between context and disposition. A systematic study of social preferences in an array of paradigmatic games, including various 2×2 games, Trust Games, Dictator Games, Ultimatum Games, using a within-subject design (e.g., extending Guyer & Rapoport, 1972; Blanco et al., 2009, 2011) may help understand the exact nature of the effect of the game structure and history on social preferences as well as the nature of the interaction between the contextual and dispositional determinants of social preferences.

8.2.2 Other games

In this thesis, we analyzed only non-embedded interactions. Moreover, in our experiments the subjects received no feedback about the choices of their interaction partners or of other actors.[8] Thus, our subjects did not have much

[8]Except Chapter 4 where we analyze the sequential PD. In the sequential PD, the second player makes a decision conditional on the behavior of the first player's decision. Besides

information to learn and to update their beliefs. There are several ways to extend our work to more dynamic settings.

One extension will be analyzing repeated games or sequential games with several decision nodes. In such games, at least theoretically, actors would update their beliefs about others' social preferences based on others' previous choices and take into account that the actors' own behaviors also affect others' beliefs. Analyzing such cases, one can also combine outcome-based social preferences and process-based social preferences in a single model. For example, we are currently working on the following approach to model reciprocity. Assume that people have "effective" social preferences. These are preferences actors use when they make a decision. Effective preferences are a function of a person's "true" social preferences and her beliefs about others' social preferences. For example, a "true" cooperative actor, i.e., an actor who assigns a large weight to the outcome of the other, may lower her effective weight on the other's outcome if she believes that the other is selfish, i.e., has a zero "true" weight for others' outcomes. Note that this model can also be applied to non-dynamic settings, such as Dictator Games or the PD. However, in dynamic games, depending on others' observed behavior, people's beliefs about others' social preferences change. Hence, their "effective" social preferences also change.

Another extension of our work would involve the dynamics of learning others' social preferences. The literature on learning others' preferences includes Dawes (1989); Iedema & Poppe (1994b); Engelmann & Strobel (2000); Hyndman et al. (2012) and Danz et al. (2012). In these studies, subjects are given feedback about others' choices, and thus, the subjects could update their beliefs based on this feedback. In most of the experiments reported in this thesis, the subjects do not receive feedback about others' choices. Thus, what we measure in our experiments is subjects' *prior* beliefs about others' preferences. Although we measure prior beliefs, our study yields clear predictions about how actors with different social preferences may update their beliefs differently. For example, we show in Chapters 2 and 3 that selfish actors' prior

eliciting these conditional decisions, in Chapter 4 we did not provide feedback on the actual choices of others before all subjects completed their choices.

beliefs are more informative—the variance in their beliefs about others' preferences is less—than prior beliefs of cooperative or competitive actors. Based on this finding, one can derive the following hypothesis: When repeated feedback is provided, selfish actors will update their beliefs more slowly than cooperative or competitive actors because they have a more informative prior belief. However, if people update their beliefs rationally, i.e., in line with Bayes' theory, the effect of this difference in prior beliefs will disappear as the amount of new information—feedback on others' choices—increases. Ultimately, beliefs of different types will converge to a common belief, which will correspond to the true distribution of preferences, and the Bayesian-Nash equilibrium will prevail.[9] However, people may not update their beliefs rationally, e.g., they may systematically ignore information (Iedema & Poppe, 1994b). In that case, even with repeated feedback people's beliefs will not converge to the correct common belief.

In this thesis, we show that the rational beliefs assumption featured in the Bayesian-Nash equilibrium framework is clearly refuted. However, more research is needed to clearly document the behavioral consequences of making wrong assumptions about actors' beliefs about others' preferences. That is, under what conditions do inaccurate assumptions about beliefs lead to inaccurate predictions of behavior? In Chapters 2 and 3, we directly test the rational beliefs assumption without analyzing the consequences of assuming rational beliefs for actors' decisions. In Chapter 4, we extend the analysis to the first player's decision in the sequential Prisoner's Dilemma, which is directly influenced by the first player's beliefs about the second player's preferences. Chapter 4 shows that although rational beliefs are refuted and that a model that imposes ego-centered biases in beliefs yields a better fit, the assumption of rational beliefs does not seem to dramatically reduce the quality of the predictions of the first player's decision in the sequential PD. Chapter 5 analyzes simultaneously played Prisoner's Dilemma and again refutes the rational beliefs assumption. However, Chapter 5 focuses on a special type of Prisoner's Dilemma, namely, the parallel Prisoner's Dilemma, where the influ-

[9] Provided that prior beliefs of actors are well behaved, i.e., do not assign zero probability to certain types of preferences that do exist in the population.

ence of beliefs on behavior is limited. Experimentally analyzing cases where beliefs are expected to influence behavior more strongly will help document the consequences of making incorrect assumptions about the actors' beliefs. Such cases include the proposer's decision in the Ultimatum Game (Güth et al., 1982), the trustor's decision in the Trust Game (e.g., Berg et al., 1995; Snijders, 1996), the contribution behavior in Public Goods Games with nonlinear production functions (Erev & Rapoport, 1990), and the non-parallel version of the Prisoner's Dilemma (see Aksoy & Weesie, 2013b).

8.2.3 *N*-person games

In this thesis, we study only 2-person interactions. As we discuss in Chapters 2 and 3, in the 2-person case many alternative ways of modeling outcome-based social preferences are nearly mathematically equivalent, provided the same number of unknown parameters is included. For example, Fehr & Schmidt (1999) inequity aversion model with weights for advantageous and disadvantageous inequalities is mathematically nearly equivalent to the social orientation model with weights for the outcome for the other and for the absolute difference between the outcomes for the self and other.[10] This condition has an upside: the choice of which particular utility function is used does not substantially influence the fit. However, if one aims to study several types of social preferences simultaneously, one needs to consider games with more than 2 players. In addition, N-person cases will also help understand how much people care about the distribution of outcomes between others, not only relative to the self, but relative to each other. N-person Dictator Games will be a good starting point for studying social preferences in games with more than two players. Engelmann & Strobel (2007) provide an overview of experiments on N-person Dictator Games and conclude that there are not enough dictator experiments with more than two players (Engelmann & Strobel, 2007, pp. 290). Studies that analyze N-person Dictator Games experimentally include Stahl & Haruvy (2006); Engelmann & Strobel (2004); Charness & Rabin (2002), and there is definitely room for more research on this topic.

[10] As a technical side, they are not fully equivalent due to the random utility terms in the function.

Appendix A

Appendix Chapter 2

A.1 Decomposed games

See Table A.1.

A.2 Beliefs about others' social orientations

We define subject i's belief p_{ij} about the fraction π_{ij} of others who choose option A in Decomposed Game j as i's beliefs about the probability that the utility difference $(A.2)$ is positive. We assume that actor i with her social orientation (θ_i, β_i) has beliefs $(\widetilde{\theta}_i, \widetilde{\beta}_i)$ that can be represented by a bivariate normal distribution $\begin{pmatrix} \widetilde{\theta} \\ \widetilde{\beta} \end{pmatrix} \sim N(\widetilde{\mu}, \widetilde{\Sigma})$, where $\widetilde{\mu} = \begin{pmatrix} \mu(\widetilde{\theta}_i) \\ \mu(\widetilde{\beta}_i) \end{pmatrix}$ and $\widetilde{\Sigma} = \begin{pmatrix} \sigma(\widetilde{\theta}_i) & \rho(\widetilde{\theta}_i, \widetilde{\beta}_i) \\ \rho(\widetilde{\theta}_i, \widetilde{\beta}_i) & \sigma(\widetilde{\beta}_i) \end{pmatrix}$. $\mu(\widetilde{\theta}_i)$ and $\mu(\widetilde{\beta}_i)$ are i's beliefs about the means of θ and β; $\sigma(\widetilde{\theta}_i)$ and $\sigma(\widetilde{\beta}_i)$ are i's beliefs about the standard deviations of θ and β; $\rho(\widetilde{\theta}_i, \widetilde{\beta}_i)$ is i's belief about the correlation between θ and β. Then, π_{ij} satisfies

Table A.1: 18 Decomposed Games used to measure social orientations and beliefs about others' social orientations. The last three columns include some descriptives, (N(subjects)=155).

	Option A		Option B		% of subjects	Belief about the % of others choosing A	
Game	You	Other	You	Other	choosing A	Mean	St. Dev.
1	700	650	650	650	96	90.64	15.71
2	700	715	650	650	89	85.09	21.14
3	640	640	680	695	8	33.91	36.22
4	420	420	440	455	7	37.05	36.26
5	320	320	300	280	96	91.57	15.34
6	500	540	550	550	5	30.14	37.76
7	660	750	630	630	88	81.62	21.60
8	410	410	400	370	97	90.00	16.21
9	640	640	680	920	12	40.52	36.08
10	650	600	650	685	26	46.70	30.89
11	540	500	540	555	20	44.06	30.22
12	480	40	440	440	48	56.54	29.02
13	540	540	580	340	40	43.48	28.99
14	630	630	600	735	88	81.34	18.25
15	350	350	300	547	93	81.67	18.47
16	310	310	320	290	28	44.80	31.06
17	450	450	400	525	95	81.93	20.04
18	310	310	320	305	17	40.02	31.66

$$\begin{aligned}\pi_{ij} &= \Pr\left(U_{AB}(\widetilde{\theta},\widetilde{\beta})>0 \,\middle|\, \begin{pmatrix}\widetilde{\theta}\\ \widetilde{\beta}\end{pmatrix} \sim N(\widetilde{\mu},\widetilde{\Sigma})\right)\\ &= \Phi\left(\frac{\Delta_{jx}-\mu(\widetilde{\theta}_i)\cdot\Delta_{jy}-\mu(\widetilde{\beta}_i)\cdot\Delta_{jxy}}{\sqrt{2\widetilde{\tau}^2+\sigma^2(\widetilde{\theta}_i)\cdot\Delta_{jy}^2+\sigma^2(\widetilde{\beta}_i)\cdot\Delta_{jxy}^2+2\rho(\widetilde{\theta}_i,\widetilde{\beta}_i)\sigma(\widetilde{\theta}_i)\sigma(\widetilde{\beta}_i)\cdot\Delta_{jy}\Delta_{jxy}}}\right)\end{aligned} \tag{A.1}$$

where x_{jA} is the outcome for the self in option A of game j, y_{jB} is the outcome of the other in option B of game j, $\Delta_{jx} = x_{jA} - x_{jB}$, $\Delta_{jy} = y_{jA} - y_{jB}$, $\Delta_{jxy} = |x_{jA} - y_{jA}| - |x_{jB} - y_{jB}|$, and $2\widetilde{\tau}^2$ is the variance of $\widetilde{\epsilon_A} - \widetilde{\epsilon_B}$, the variance of the difference of random disturbances. Φ is the cumulative normal

distribution.

We model $p_{ij} = \pi_{ij} + \varepsilon_{ij}$, ignoring the boundary conditions on fractions and assuming that ε_{ij} has a zero expected value. We do not make any further assumption about the shape of the distribution of ε_{ij}. We estimate the parameters by the (nonlinear) least squares method using the cluster adjustment of the sandwich estimator of variance to deal with nesting within subjects. π_{ij} for the one-parameter social orientation model can be easily obtained by substituting zero for $\mu(\widetilde{\beta}_i)$ and $\sigma(\widetilde{\beta}_i)$ in (A.1). To ensure that $-1 \leq \rho(\widetilde{\theta}_i, \widetilde{\beta}_i) \leq 1$, we use the transformation $\rho(\widetilde{\theta}_i, \widetilde{\beta}_i) = \frac{e^{2t}-1}{e^{2t}+1}$ and estimate t in the non-linear regression.

A.3 Simultaneous analysis of own and expected social orientations using Mplus

Abstract

This note is a follow up to Aksoy & Weesie (2012b). Here we demonstrate how disadvantages of the two step estimation method used to analyze the relationship between one's own social orientations and one's beliefs about others' social orientations in Aksoy & Weesie (2012b) can be overcome by formulating a technical variation that can be fitted in Mplus 6. Results obtained with the method introduced here are qualitatively the same as the ones reported in Aksoy & Weesie (2012b), thus support the conclusions of Aksoy & Weesie (2012b).

A.3.1 Introduction

In Aksoy & Weesie (2012b) we propose a two step estimation method of a random utility model for one's own social orientations and one's beliefs about the social orientations of others. The first step involves the maximum likelihood estimation of a multilevel probit regression model for binary choices in a series of dictator games (DG), and the subsequent estimation of the social

orientation parameters of subjects by posterior means. The second step involves a nonlinear regression that explains the beliefs about others' choices in these DGs by the subjective beliefs of subjects about the distribution of the social orientations of others. In this nonlinear regression, the means and standard deviations of the subjective belief distribution are modeled as functions of subjects' own social orientations.

We discussed two main disadvantages of this method (Aksoy & Weesie, 2012b, pg. 8). First, the nonlinear regression fitted in the second step does not account for the fact that own social orientations are estimated scores rather than true scores. Generally, this would lead to a biased estimate of the nonlinear regression parameters and to an underestimation of the uncertainty in those parameters. Second, our two step procedure uses the assumption that one's beliefs about others' social orientations are a function of *only* one's own social orientations. That is, two subjects with the same social orientations are assumed to have the same beliefs. This is obviously an unrealistic assumption.

In this short note, we propose a one step method for estimating one's own social orientation *and* one's beliefs about others' social orientations *simultaneously*, using structural equation modeling with Mplus (Muthén & Muthén, 1998–2010). This one step method overcomes the two disadvantages of the two step estimation method discussed above.

A.3.2 Reformulating the problem in the Mplus framework

Consider the single parameter social orientation model where an actor assigns a weight θ to the outcome of the other actor in a two-actor binary DG. Equation (5) in Aksoy & Weesie (2012b) specifies a multilevel probit model with a random subject coefficient for θ, so that the probability that subject i ($i = 1, \ldots, 155$) in DG j ($j = 1, \ldots, 18$) prefers option A over option B is specified as

$$\Pr(U_{Aij} > U_{Bij}|\theta_{ij}) = \Phi\left(\frac{\Delta_{xj} - \theta_i \Delta_{yj}}{\sqrt{2}\tau}\right) \qquad \theta_i \sim N(\mu, \sigma). \tag{A.2}$$

Δ_{xj} and Δ_{yj} are outcome differences between the two options in DG j for self

and other, respectively. *Model* (*A*.2) contains three parameters: the mean μ and standard deviation σ of the random subject effect θ, and the standard deviation τ of the random additive component in the utility function. In Aksoy & Weesie (2012b), we fitted this model using GLLAMM in Stata 11.

It is possible to fit (A.2), for now ignoring the beliefs, in Mplus with a trick to make Mplus fit τ while constraining the coefficient of Δ_{xj} to 1. However, this specification makes Mplus fairly slow and numerically unstable. So we proceed with a reparametrization of (A.2) as a standard multilevel probit model with a fixed effect for Δ_{xj}, a random subject level slope for Δ_{yj}, and level 1 residual variance, τ fixed to 1, and so

$$\Pr(U_{Aij} > U_{Bij}|\theta_{ij}) = \Phi\left(\alpha\Delta_{xj} - \vartheta_i\Delta_{yj}\right). \tag{A.3}$$

where $\tau = \frac{1}{\sqrt{2}\alpha}$ and $\theta_i = \frac{\vartheta_i}{\alpha}$, and so $\mu(\theta) = \frac{\mu(\vartheta)}{\alpha}$ and $\sigma(\theta) = \frac{\sigma(\vartheta)}{\alpha}$. With this specification, Mplus gives reliable results identical to GLLAMM/Stata. Thus, we gain confidence to incorporate beliefs into the model.

Now, consider our model for a subject's belief about others' behavior, i.e., i's subjective probability π_{ij} that a random other subject prefers option A over option B in DG j. For ease of presentation and formulation in Mplus, we reordered options so that $\Delta_{yj} > 0$. In our model, beliefs are assumed to be described by a normal distribution. A statistically convenient formulation of the model for beliefs is defined in terms of the inverse cumulative normal of π_{ij},

$$\Phi^{-1}(\pi_{ij}) = \frac{\Delta_{xj} - \mu(\widetilde{\vartheta}_i)\Delta_{yj}}{\sigma(\widetilde{\vartheta}_i)\Delta_{yj}} + e_{ij} \qquad e_{ij} \text{ iid } N(0, \sigma_e^2). \tag{A.4}$$

Equation (A.4) includes a normally distributed error term e_{ij}. Note also that in (A.4) we omit the (common) belief about the standard deviation $\widetilde{\tau}$ of random utility. In analyses with GLLAMM we found that the parameter estimates of interest to us were hardly affected by whether $\widetilde{\tau}$ was fixed to 0, or estimated as a free parameter. Since there is no straightforward way to estimate $\widetilde{\tau}$ in the Mplus setup, we simply fix it to 0 here. Now (A.4) can be rewritten as

$$\begin{aligned}\Phi^{-1}(\pi_{ij}) &= \frac{\Delta_{xj}}{\Delta_{yj}} \cdot \frac{1}{\sigma(\widetilde{\vartheta}_i)} - \mu(\widetilde{\vartheta}_i) \cdot \frac{1}{\sigma(\widetilde{\vartheta}_i)} + e_{ij} \\ &= t_j \phi(\widetilde{\vartheta}_i) + \mu(\widetilde{\vartheta}_i)\phi(\widetilde{\vartheta}_i) + e_{ij} \end{aligned} \tag{A.5}$$

where

$$\begin{aligned} t_j &= \tfrac{\Delta_{xj}}{\Delta_{yj}} \\ \phi(\widetilde{\vartheta}_i) &= 1/\sigma(\widetilde{\vartheta}_i) \end{aligned} .$$

In our data, for some cases $\pi_{ij} = 0$ or $\pi_{ij} = 1$, thus $\Phi^{-1}(\pi_{ij}) \to \mp\infty$. For those boundary cases, we recode π_{ij} such that $\delta < \pi_{ij} < 1 - \delta$ where δ is a small number such as 0.001 or 0.005. We performed a sensitivity analysis by varying δ and found that the exact value of δ hardly changed the results. In the results presented here, δ is set to 0.005.

The final part of the structural equation model specifies linear and quadratic regressions of the mean $\mu(\widetilde{\vartheta})$ and precision $\phi(\widetilde{\vartheta})$ respectively on ϑ,

$$\begin{aligned} \mu(\widetilde{\vartheta}_i) &= b_{a0} + b_{a1}\vartheta_i &&+ d_{i1} \\ \phi(\widetilde{\vartheta}_i) &= b_{sa0} + b_{sa1}\vartheta_i + b_{sa2}\vartheta_i^2 &&+ d_{i2} \end{aligned} \tag{A.6}$$

where

$$\begin{pmatrix} d_{i1} \\ d_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{12} & \psi_{22} \end{pmatrix} \right).$$

Note that in (A.6), the relationships between $\sigma(\widetilde{\vartheta})$ and ϑ are modeled differently than in Aksoy & Weesie (2012b) in two respects. The first difference is in parametric form. In Aksoy & Weesie (2012b), $\ln \sigma(\widetilde{\vartheta})$ is modeled as a quadratic function of ϑ, here $\phi(\widetilde{\vartheta}) = \frac{1}{\sigma(\widetilde{\vartheta})}$ is modeled as a quadratic function of ϑ. The reason for this adjustment is simply that there is no obvious computationally feasible way to model a log-quadratic relation in latent variables in Mplus. The parametric form in (A.6) still adequately captures a potential u-shaped relationship between ϑ and $\sigma(\widetilde{\vartheta})$. However, the current specification does not ensure that the predicted values for $\phi(\widetilde{\vartheta})$ are necessarily positive; negative values would resist even the most stubborn attempts of interpreta-

tion. Second, (A.6) includes normally distributed error terms d_{i1} and d_{i2} at the level of subjects, representing interpersonal differences in the mean and precision of beliefs that are not captured by own value orientations. In Aksoy & Weesie (2012b) these residuals are omitted.[1]

The structural equation model that we fit here comprises the equations (A.3), (A.5), and (A.6) simultaneously. Besides the three residual terms, these three equations include three latent variables, namely ϑ (`theta`), $\mu(\widetilde{\vartheta})$ (`mu`), and $\phi(\widetilde{\vartheta})$ (`phi`). These three latent variables are assumed to be (conditionally) normally distributed. Equation (A.5) includes a latent interaction term of $\mu(\widetilde{\vartheta}) \cdot \phi(\widetilde{\vartheta})$ (`muphi`). Equation (A.6) includes a quadratic term in a latent variable ϑ^2 (`theta2`). Moreover, the dependent variable (`dg`) in (A.3) is a categorical (binary) variable. Mplus allows categorical variables as well as latent interaction and quadratic terms. A DG with $\Delta_{yj} = 0$ is excluded from the analyses. In the fitted model ψ_{12} in (A.6) is constrained to zero, that is, ψ_{11} and ψ_{22} are assumed to be independent, to make Mplus produce admissible results.

A.3.3 Results

Below, we include the Mplus syntax with detailed explanations. We experience that convergence could not be achieved unless we specify reasonable starting values. These starting values are obtained by the two-step method with GLLAMM and nonlinear regression in Stata. Table A.2a displays the results given in the Mplus output. Note that in the Mplus output, THETA is in fact ϑ (see Table A.2a). Thus, the mean of θ can be obtained by dividing $\mu(\vartheta)$ by $\hat{\alpha} = 4.047$ and the variance of θ can be obtained by dividing $\sigma(\vartheta)$ by $\hat{\alpha}^2$ (see (A.3)). Note that $2\tau^2$, the variance of the decision error for own choices, can be obtained by $\frac{1}{\hat{\alpha}^2}$. Similarly, the coefficient of ϑ should be multiplied by $\hat{\alpha}$ and the coefficient of ϑ^2 should be multiplied by $\hat{\alpha}^2$. Table A.2b shows the results after this reparametrization. Standard errors of those nonlinear terms are computed with the Delta method. $\mu(\theta)$, $\sigma^2(\theta)$, and $2\tau^2$ are effectively identical to those reported in Aksoy & Weesie (2012b).

[1] We thank Michał Bojanowski for commenting on this limitation in our original approach.

Figure A.1 displays the results on the relationship between own social orientation and beliefs about others' social orientations in a graphical form. Note that Figure A.1 is obtained after converting ϑ to θ for both own social orientations and beliefs. Results support the conclusions of Aksoy & Weesie (2012b): $\mu(\widetilde{\vartheta})$ increases in ϑ, thus $\mu(\widetilde{\theta})$ increases in θ. Moreover, $\phi(\widetilde{\vartheta})$ and ϑ have an inverse u-shaped relation, and thus $\sigma(\widetilde{\theta})$ and θ have a u-shaped association. These results show that those conclusions of Aksoy & Weesie (2012b) are robust with respect to using one or two-step estimation method.

In principle, the same method can be used for any outcome based social utility model, thus also for the two parameter social orientation model where the equality dimension is added. The resulting structural equation model for this two parameter social orientation model, however, includes many additional parameters. We tried to fit a model for the two parameter social orientation case, but could not obtain convergence, even with good starting values.

Table A.2: (a) parameters of the fitted model on actors' beliefs about the mean/variance of social orientations, actors own social orientations and the relationship between the two. (b) reparametrization of (a).

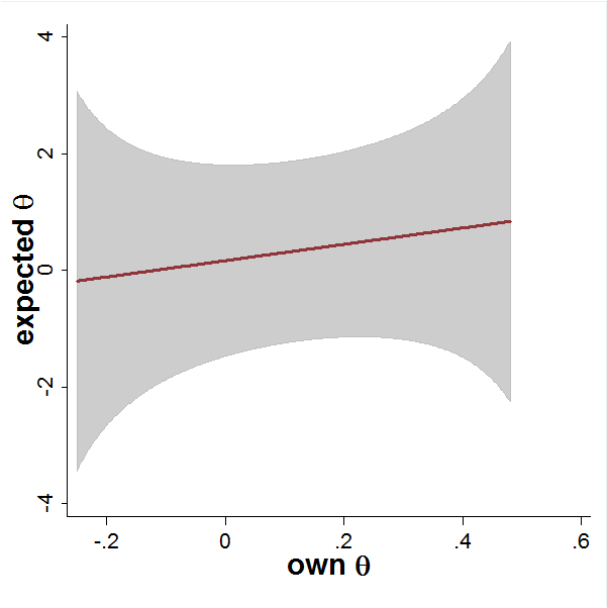
(a)

Var	$\mu(\widetilde{\vartheta})$	$\phi(\widetilde{\vartheta})$
	Coef. (S.E.)	Coef. (S.E.)
Cons	.161**	1.104**
	(.053)	(.129)
ϑ	.316**	.233
	(.046)	(.114)
ϑ^2		-.234**
		(.103)
α	4.047**	
	(.351)	
$\mu(\vartheta)$	.446**	
	(.069)	
$\sigma^2(\vartheta)$	.415**	
	(.132)	
σ_e^2	2.004**	
	(.144)	
ψ_{11}	.023	
	(.036)	
ψ_{22}	.725**	
	(0.082)	
ψ_{22}	0	

(b)

Var	$\mu(\widetilde{\theta})$	$\phi(\widetilde{\theta})$
	Coef. (S.E.)	Coef. (S.E.)
Cons	.161**	1.104**
	(.053)	(.129)
$\theta = \vartheta \times \alpha$	1.279**	.942**
	(.240)	(.329)
$\theta^2 = \vartheta \times \alpha^2$		-3.980**
		(1.839)
$2\tau^2 = \frac{1}{\alpha^2}$	.061**	
	(.011)	
$\mu(\theta) = \frac{\mu(\vartheta)}{\alpha}$	.110**	
	(.017)	
$\sigma^2(\theta) = \frac{\sigma^2(\vartheta)}{\alpha^2}$	.026**	
	(.005)	
σ_e^2	2.004**	
	(.144)	
ψ_{11}	.023	
	(.036)	
ψ_{22}	.725**	
	(0.082)	
ψ_{22}	0	

Figure A.1: Relationship between actors' social orientations and beliefs about other's social orientations in a graphical form, based on the results in Table A.2b. The gray-shaded region represents the relationship between beliefs about the variance of others' social orientations and one's own social orientation: the boundary of the area is $\mu(\widetilde{\theta}) \pm 1.65\sigma(\widetilde{\theta})$. The thick line in the gray region shows the relationship between beliefs about the mean of social orientations and actors' social orientation.



Mplus syntax

Below, we explain the Mplus syntax in detail. For reader-friendliness, we present the syntax in three steps. The first syntax fits the part of the structural equation model that deals with the social orientations of subjects. The second syntax is the final model that includes beliefs about the social orientations of others in addition to own orientations. The third syntax specifies starting values and parameter constraints that ensures convergence and admissibility.

Below, comments are included after `!`. Recall that in Mplus `@value` constrains a parameter to `value`, while `*value` specifies a starting value.

The data `dglong.dat` contains 5 variables and 2634 cases for 155 subjects. 155 (N subjects) $\times$ 17 (N games) = 2635. A subject did not respond to a game, thus the number of cases is 2634.

```
subject   1..155                    subject identifier
dg         0/1                     1=A>B  0=A<B
pij   -2.58..2.58                  phi-inverse(belief-proportion
                                                    !of others: A>B)
dx                                 xA-xB (outcome diff. for self)
dy                                 yA-yB (outcome diff. for other)
```

Syntax 1: Subjects' own social orientations

```
TTLE:
  Social orientations the single parameter model
DATA:
  FILE        = 'dglong.dat';         !specifies the data file & path
ANALYSIS:
  LINK        = PROBIT;               !specifies a probit link
  TYPE        = RANDOM TWOLEVEL;      !specifies that data are bi-level
  INTEGRATION = MONTE (385);          !user defined #integration points
  ALGORITHM   = INTEGRATION;          !due to complexity M.Carlo is used
VARIABLE:
  NAMES       = subject dg pij dx dy; !names of the variables in the dataset
  USEV        = subject dg dx dy;     !names of the variables in the model
  Missing     = ALL (9999);           !missings coded with 9999
  CATEGORICAL = dg;                   !lists the categorical variable(s)
  WITHIN      = dx dy dg pij t;       !lowest level variables
  CLUSTER     = subject;              !highest level indicator
MODEL:
  %WITHIN%
    theta | dg ON dy;                 !theta: random slope of dy on dg.
    dg ON dx;                                          ! theta = vartheta
    [dg$1@0];                         !threshold (intercept) fixed to 0
  %BETWEEN%                           !no between level equation yet
```

Syntax 2: Own orientations and beliefs

```
TTLE:
  Joint analysis of social orientations and beliefs
  for the single parameter model
DATA:
  FILE  = 'dglong.dat';
ANALYSIS:
  LINK        = PROBIT;
  TYPE        = RANDOM TWOLEVEL;
  INTEGRATION = MONTE (385);
  ALGORITHM   = INTEGRATION;
VARIABLE:
  NAMES       = subject dg pij dx dy;
  USEV        = subject dg pij dx dy;
  Missing     = ALL (9999);
  CATEGORICAL = dg;
  WITHIN      = dx dy dg pij t;
  CLUSTER     = subject;
DEFINE
  t = Dx/Dy;                          !variable construction
MODEL:
  %WITHIN%
    theta | dg ON dy;
    dg ON dx;
    [dg$1@0];
    phi | pij ON t;                   !phi=1/sd(belief|theta)
                                      !t is the only within var.
  %BETWEEN%                                    !for beliefs (eq:4)
    mu BY;                            !mu: belief mean theta, latent var.
                                      !latent vars defined by using BY;
    muphi | mu XWITH phi;             !muphi: latent interaction mu*phi
    theta2 | theta XWITH theta;       !theta2: theta*theta, latent var.
    mu ON theta;                      !mu depends on theta linearly
    phi ON theta theta2;              !phi depends on theta curvilinearly
    [pij@0];                          !zero intercept for pij
    pij@0;                            !no subj.level error varian. for pij
    pij ON muphi@1;                   !coefficient constrained to 1 (eq:4)
```

Syntax 3: Final model with starting values and constraints

```
TTLE:
  Joint analysis of social orientations and beliefs for the
  single parameter model constraints and starting values added
  to produce admissible solutions and convergence
DATA:
  FILE  = 'dglong.dat';
ANALYSIS:
  LINK        = PROBIT;
  TYPE        = RANDOM TWOLEVEL;
  INTEGRATION = MONTE (385);
  ALGORITHM   = INTEGRATION;
VARIABLE:
  NAMES       = subject dg pij dx dy;
  USEV        = subject dg pij dx dy;
  Missing     = ALL (9999);
  CATEGORICAL = dg;
  WITHIN      = dx dy dg pij t;
  CLUSTER     = subject;
DEFINE
  t = Dx/Dy;
MODEL:
  %WITHIN%
    theta | dg ON dy;
    dg ON dx*3.936;                !starting value (sv) for the coef.
    [dg$1@0];
    phi | pij ON t;
    pij*2;                         !sv for var(e_ij)
  %BETWEEN%
    mu BY;
    muphi | mu XWITH phi;
    theta2 | theta XWITH theta;
    mu ON theta*.363;
    phi ON theta*0.204 theta2*-0.222;
    [pij@0];
    [mu*0.099 phi*1.1 theta*0.329]; !sv for intercept mu/phi and mean theta
    pij@0;
    mu*0.021 phi*0.7 theta*0.272;   !sv for variances of residual errors
                                                          !and of theta
    pij ON muphi@1;
```

```
    pij WITH mu@0;                  !constraints to produce admissible results
    pij WITH phi@0;                 !residual errors assumed to be independent
    mu WITH phi@0;
OUTPUT:                             !invoked to obtain additional output
 TECH1                              !shows how Mplus calls parameters
 TECH3                              !prints the covariance matrix of parms
SAVEDATA:
 TECH3 = MyTech3.dta             !saves tech3 without rounding for future use
                                 !fed to Stata for nonlinear tests
```

Mplus output for Syntax 3

Constrained parameters are omitted in the output to conserve space

```
THE MODEL ESTIMATION TERMINATED NORMALLY
TESTS OF MODEL FIT
Loglikelihood
          H0 Value                        -5856.285
          H0 Scaling Correction Factor        2.590
            for MLR
Information Criteria
          Number of Free Parameters             11
          Akaike (AIC)                    11734.570
          Bayesian (BIC)                  11799.209
          Sample-Size Adjusted BIC        11764.259
            (n* = (n + 2) / 24)
MODEL RESULTS                                             Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value
Within Level
 DG_        ON
    DX                4.047      0.351     11.542      0.000
 Residual Variances
    PIJ               2.004      0.144     13.889      0.000
Between Level
 MU         ON
    THETA             0.316      0.046      6.926      0.000
 TAU        ON
    THETA             0.233      0.114      2.048      0.041
    THETA2           -0.234      0.103     -2.274      0.023
 Means
    THETA             0.446      0.069      6.484      0.000
 Intercepts
```

PIJ	0.000	0.000	999.000	999.000
MU	0.161	0.053	3.027	0.002
TAU	1.104	0.129	8.586	0.000
Variances				
THETA	0.415	0.132	3.137	0.002
Residual Variances				
MU	0.023	0.036	0.630	0.529
TAU	0.725	0.082	8.817	0.000

Appendix B

Appendix Chapter 3

B.1 Dictator Games used in the study

In the experiment, 18 binary Dictator Games, shown in Table B.1, were used to measure other-regarding preferences and beliefs about others' other-regarding preferences. We choose the outcomes in these 18 Dictator Games to facilitate the statistical estimation of the parameters α and β. In 16 out of 18 games, one option includes an equal distribution, whereas the other option includes an unequal distribution. Because of this characteristic of the games, in each of these games there is a critical value of either α or β such that a subject with α or β exceeding that threshold would choose the equal distribution option. These threshold values are shown in the last two columns of Table B.1. We chose these particular critical α and β values to capture more variation by considering the empirical distributions of α and β parameters reported in the literature, e.g., Fehr & Schmidt (1999); Bellemare et al. (2008); Morishima et al. (2012). When choosing these thresholds, we also included negative values because Bellemare et al. (2008) and Morishima et al. (2012) also report negative α and β. For the two games in which both options include unequal distributions (Game 17 and 18), there is no single critical α or β value but a critical linear combination of the two. Table B.1 includes those linear combinations such that a subject who satisfies the condition chooses option A.

Table B.1: 18 Dictator Games used to measure other-regarding preferences and beliefs about others' other-regarding preferences and some descriptive statistics. Columns 6 and 7 include the associated critical α or β values such that a subject with α or β exceeding that threshold would choose the equal distribution option. The last four columns include average A-choices, average belief about %A choices, standard deviation of beliefs about %A-choices, and the Pearson's correlation between A-choice and belief about % A-choices (N-subject=187). All correlations given in the last column are statistically significant—p(2-sided)<0.05.

	Design				Prediction		Descriptives			
Game	Option A		Option B		Threshold	Threshold	Mean	Mean belief	S.D. belief	Corr
	You get	Other gets	You get	Other gets	α	β	A-choice	%A-choice	%A-choice	(choice,belief)
1	320	320	300	280	.	-1	.963	91.241	15.073	0.522
2	410	410	400	370	.	-.333	.973	89.797	16.189	0.450
3	500	400	550	550	.	-.5	.043	30.326	37.692	0.174
4	450	450	400	525	-.4	.	.957	83.000	18.887	0.413
5	350	350	300	475	-.286	.	.936	81.759	19.086	0.370
6	630	630	600	735	-.222	.	.893	80.984	19.021	0.421
7	640	640	680	920	.167	.	.134	42.471	36.078	0.317
8	660	750	630	630	.333	.	.866	80.257	23.313	0.600
9	420	420	440	455	1.333	.	.086	37.091	35.587	0.298
10	640	640	680	695	2.667	.	.096	35.888	36.482	0.352
11	700	715	650	650	3.333	.	.904	84.139	21.971	0.626
12	480	40	440	440	.	.091	.476	56.754	28.022	0.586
13	540	540	580	340	.	.167	.412	44.572	28.692	0.354
14	310	310	320	290	.	.333	. 278	44.738	30.982	0.403
15	310	310	320	305	.	.667	.187	40.332	31.020	0.330
16	700	650	650	650	.	1	.952	89.679	16.097	0.430
17	540	500	540	555	$3\alpha - 8\beta > 0$		.198	44.733	30.379	0.422
18	650	600	650	685	$7\alpha - 10\beta > 0$		.273	46.380	31.023	0.509

B.2 Instructions

Please consider Example 1 below. In this table, you see 2 pairs of points to be allocated between yourself and another person. In this part of the experiment, you will receive a number of such pairs. For each pair you need to choose one of the options that you prefer. According to your choice, the points attached to that choice will be given to you and another person who will be randomly selected from all participants in this experiment. For example, if you choose option A in the example below, then you will receive 100 points and a randomly selected other participant will receive 200 points. Similarly, if you chose option B, then you will receive 100 points, and a randomly selected other participant will receive 1 points.

Moreover, for each of the pairs, we ask you to guess the choices of other participants. Again, you will earn points depending on the accuracy of your guess. We will compute the actual percentages from the decisions of all other participants. The closer your guess is to the actual percentages, the more points you will earn. In particular, if your guess is exactly equal to the actual percentage, than you will earn 500 points. For each percentage point deviation from the actual percentage, you will earn 20 points less. If your guess in a given situation is off more than 25 percentage points, then you earn 0 points from that situation.

For each of the pairs that you will see next, please indicate your choice by marking one of the check boxes given below the options, and state your guess about the percentage of participants who you think would chose option A by writing a number between 0 and 100 in the space given.

Note also that for each pair, other participants in this experiment will also state their choices; where you will be the other person for a randomly selected participant. You will also earn points depending on this randomly selected person's choice.

Example 1	Option A	Option B
You get	100	100
Other participant gets	200	1
Your choice	[ ]	[ ]
% participants who choose option A: [____]		

B.3 OpenBugs code

```
# ------------------------------------------------------------------------
# OpenBugs code for analyzing inequity aversion and beliefs
# Electronic supplement to biased beliefs and inequity aversion
# data: cdata.txt  dg[i,j]   = choices {0,1}; dg=1 => U(a)>U(b),
#                  ipdg[i,j] = beliefs [-2,6, 2.6]; inverse.cumulative(p)
#                  Dx[j]     = xa[j]-xb[j]
#                  Dxy[j]    = max(xa[j]-ya[j],0) - max(xb[j]-yb[j],0)
#                  Dyx[j]    = max(ya[j]-xa[j],0) - max(yb[j]-xb[j],0)
# See the OpenBugs manual by Spiegelhalter,Thomas,Best and Lunn (2011)
# "http://www.openbugs.info/Manuals/Manual.html"
# ------------------------------------------------------------------------
model {
# models for subjects characteristics
     for (i in 1:N) {
      theta[i,1:2]  ~ dmnorm(mu[], invS[,])
      alpha[i]      <-theta[i,1]
      beta[i]       <-theta[i,2]
      # beliefs about (alpha,beta)
      mu.alpha[i]   <-    ba.0 + ba.1 *alpha[i]+ e.alpha[i]
      mu.beta[i]    <-    bb.0 + bb.1 *beta[i] + e.beta[i]
      sd2.alpha[i]  <-exp(bsa.0+ bsa.1*alpha[i]+ bsa.2*pow(alpha[i],2))
      sd2.beta[i]   <-exp(bsb.0+ bsb.1* beta[i]+ bsb.2*pow( beta[i],2))
      ab.cov[i]     <-rho*sqrt(sd2.alpha[i]*sd2.beta[i])
      e.alpha[i]    ~ dnorm(0, e.ba.mu.tau)
      e.beta[i]     ~ dnorm(0, e.bb.mu.tau)
  }
# models for choices and beliefs
   for (i in 1:N) {
      for (j in 1:M) {
        dg[i,j]           ~ dbern(dg.p[i,j])
```

```
        du[i,j]           <-Dx[j]-alpha[i]*Dyx[j]-beta[i]*Dxy[j]
        probit(dg.p[i,j]) <-max(min(sqrt(e.tau)*du[i,j], 5),-5)
        ipdg[i,j]         ~ dnorm(ipdg.mu[i,j], ipdg.tau)
        ipdg.mu[i,j]      <-(Dx[j]-mu.alpha[i]*Dyx[j]-mu.beta[i]*Dxy[j])/
                        sqrt(sd2.alpha[i]*pow(Dyx[j],2)+
                             sd2.beta[i] *pow(Dxy[j],2)+
                             2*ab.cov[i]*Dyx[j]*Dxy[j] )
      }
    }

# PRIORS:
 # priors for the means of (alpha,beta)
    mu[1:2]      ~ dmnorm( mu0[], invS0[,] )
    mu0[1]       <- 0
    mu0[2]       <- 0
    invS0[1,1]   <- 0.1
    invS0[1,2]   <- 0
    invS0[2,1]   <- 0
    invS0[2,2]   <- 0.1

 # prior for the (co)variances of (alpha,beta)
    invS[1:2,1:2] ~ dwish(Omega0[,], 2)
    Omega0[1,1]  <- 0.01
    Omega0[1,2]  <- 0
    Omega0[2,1]  <- 0
    Omega0[2,2]  <- 0.01

 # priors for correlation and variance parameters
    rho          ~ dunif(-1,1)
    e.tau        ~ dgamma(0.01,0.01)
    ipdg.tau     ~ dgamma(0.01,0.01)
    e.ba.mu.tau  ~ dgamma(0.01,0.01)
    e.bb.mu.tau  ~ dgamma(0.01,0.01)

 # priors for regression coefficients
    ba.0  ~ dnorm(0,0.01)
    ba.1  ~ dnorm(0,0.01)
    bb.0  ~ dnorm(0,0.01)
    bb.1  ~ dnorm(0,0.01)
    bsa.0 ~ dnorm(0,0.01)
    bsa.1 ~ dnorm(0,0.01)
```

```
   bsa.2 ~ dnorm(0,0.01)
   bsb.0 ~ dnorm(0,0.01)
   bsb.1 ~ dnorm(0,0.01)
   bsb.2 ~ dnorm(0,0.01)

# Posterior predictive sampling: overall fit
   for (i in 1:N) {
      for (j in 1:M) {
         Dc[i,j] <- pow(dg[i,j]-dg.p[i,j],2) /
                    ((0.1+0.8*dg.p[i,j])*(0.9-0.8*dg.p[i,j]))
         Db[i,j] <- pow(ipdg[i,j] - ipdg.mu[i,j], 2)
      }
   }

   Dcbar <- mean(Dc[,])  #Average discrepancy in choices
   Dbbar <- mean(Db[,])  #Average discrepancy in beliefs

  # predictive sampling

   for (i in 1:N) {
      for (j in 1:M) {
         dg.rep[i,j]    ~ dbern(dg.p[i,j])
         Dc.rep[i,j]    <- pow(dg.rep[i,j] - dg.p[i,j],2) /
                          ((0.1+0.8*dg.p[i,j])*(0.9-0.8*dg.p[i,j]))
         ipdg.rep[i,j] ~ dnorm(ipdg.mu[i,j], ipdg.tau)
         Db.rep[i,j]    <- pow(ipdg.rep[i,j] - ipdg.mu[i,j], 2)
      }
   }
   Dcbar.rep <- mean(Dc.rep[,])
   Dbbar.rep <- mean(Db.rep[,])
   pc        <- step(Dcbar.rep-Dcbar)
   pb        <- step(Dbbar.rep-Dbbar)


# extract parameters of interest as vectors for monitoring

   bc[ 1] <- ba.0              #regression coefficients
   bc[ 2] <- ba.1              #that relates moments of
   bc[ 3] <- bsa.0             #beliefs to own (alpha,beta)
   bc[ 4] <- bsa.1
   bc[ 5] <- bsa.2
```

```
   bc[ 6] <- bb.0
   bc[ 7] <- bb.1
   bc[ 8] <- bsb.0
   bc[ 9] <- bsb.1
   bc[10] <- bsb.2
   d[1] <- mu[1]                #mu(alpha)
   d[2] <- mu[2]                #mu(beta)
   S[1:2,1:2] <- inverse(invS[,])
   d[3]   <- sqrt(S[1,1])         #sd(salpha)
   d[4]   <- sqrt(S[2,2])         #sd(beta)
   d[5]   <- S[1,2]/(a.sd*b.sd)   #rho(alpha,beta)
   d[6]   <- rho                  #belief rho(alpha,beta)
   d[7]   <- pc                   #Discrepancy p-value choices
   d[8]   <- pc                   #Discrepancy p-value beliefs
   d[9]   <- 1/sqrt(ipdg.tau)     #var. residual error for ipdg
   d[10]  <- 1/sqrt(e.ba.mu.tau)  #var. of res. error for belief on mean beta
   d[11]  <- 1/sqrt(e.bb.mu.tau)  #var. of res. error for belief on mean alpha
   d[12]  <- 1/sqrt(e.tau)        #Var. of evaluation error
}
```

Appendix C

Appendix Chapter 4

C.1 Games

See Table C.1 and Table C.2.

C.2 Some notes on Bayesian estimation

C.2.1 Priors

In the hierarchical Bayesian analyses, we used the following (weakly) uninformative prior distributions. We report several relatively complex models each of which takes considerable time (days rather than minutes) to converge. Due to time constraints, we performed less sensitivity analyses with respect to priors and model specification than we would want to. However, since we use (weakly) uninformative priors, we expect that the influence of priors on the posterior distribution of the parameters will be minimal and the posterior means will be approximately equal to maximum likelihood estimates. In fact, using very similar priors as we use here, Aksoy & Weesie (2013a) compare the Bayesian and frequentist results in a similar setup and show that they are indeed approximately equal.

For the full analyses of all four nodes, the prior for the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{\theta}$ is an inverse Wishart with 4 degrees of freedom with a scale matrix of $10 \cdot I_4$ where I_4 is the 4×4 identity matrix (Kunreuther et al., 2009; Aksoy &

Table C.1: 18 Decomposed Games used in the study. The last three columns include some descriptives, (N(subjects)=155).

	Option A		Option B		%subjects	Belief about the % of others choosing A	
Game	You	Other	You	Other	choosing A	Mean	St. Dev.
1	700	650	650	650	96	90.64	15.71
2	700	715	650	650	89	85.09	21.14
3	640	640	680	695	8	33.91	36.22
4	420	420	440	455	7	37.05	36.26
5	320	320	300	280	96	91.57	15.34
6	500	540	550	550	5	30.14	37.76
7	660	750	630	630	88	81.62	21.60
8	410	410	400	370	97	90.00	16.21
9	640	640	680	920	12	40.52	36.08
10	650	600	650	685	26	46.70	30.89
11	540	500	540	555	20	44.06	30.22
12	480	40	440	440	48	56.54	29.02
13	540	540	580	340	40	43.48	28.99
14	630	630	600	735	88	81.34	18.25
15	350	350	300	547	93	81.67	18.47
16	310	310	320	290	28	44.80	31.06
17	450	450	400	525	95	81.93	20.04
18	310	310	320	305	17	40.02	31.66

Weesie, 2013a). We adjusted the degrees of freedom down when we reduced the dimension of $\boldsymbol{\theta}$, thus also the dimension of the I matrix. For example, when we analyzed only the first three nodes, the inverse Wishard had 3 degrees of freedom. For the mean vector of $\boldsymbol{\theta}$, $\boldsymbol{\mu}$, we used a multivariate normal prior with 0_4 means and $10 \cdot I_4$ variance matrix. For the intercepts and slopes in equations (4.5) and (4.6)—the $b_\bullet$ and $c_\bullet$ parameters—we used univariate normals with mean 0 and variance 100. For the natural logarithm of variance of the evaluation error τ_h^2, we used a Normal(0,10) prior. For natural logarithms of the $\widetilde{\tau}^2$, ζ_h^2, and $\widetilde{\sigma}_{0h}^2$ parameters we used Normal(0,100) priors. Finally for the ς_h^2 parameter we used a Gamma(100,100) prior.

C.2.2 Posterior predictive checking

The Bayesian toolkit provides the useful method of Posterior Predictive Checking, a method to assess the fit of fairly arbitrary statistical models (see: Gelman et al., 1996; Gelman & Hill, 2007; Fox, 2010). Consider, for example, the Pearson discrepancy statistics for the choice of subject i in node h in game j:

$$D^c_{ihj} = \frac{(y_{ihj} - \pi_{ihj})^2}{\pi_{ihj}(1 - \pi_{ihj})}, \tag{C.1}$$

where $\pi_{ihj} = \Pr(U_{ihj1} > U_{ihj2})$ is the *predicted* probability that i chooses option 1 in game j in node h, and y_{ihj} is the *observed* choice for i in game j in node h. One can construct an overall fit statistic, $D^c_{\bullet\bullet\bullet}$ by averaging over all i, h, and j. Alternatively, one can also construct a node-level fit statistic by averaging over only the games in a particular node, i.e., $D^c_{\bullet h\bullet}$, a subject-level fit statistics (i.e., person-fit statistics in the parlor of item response theory) $D^c_{i\bullet\bullet}$, or game-level fit statistics (i.e., item-fit statistics) $D^c_{\bullet\bullet j}$.

Similarly, a discrepancy statistics can be defined for beliefs:

$$D^b_{ihj} = (\Phi^{-1}(p_{ihj}) - \Phi^{-1}(\pi_{ihj}))^2 \tag{C.2}$$

where $\Phi^{-1}(\pi_{ihj})$ and $\Phi^{-1}(p_{ihj})$ are the *predicted* and *observed* inverse cumulatives of subject i's beliefs about other's choices in game j in node h. The inverse cumulative transformation is used to transform the (0,1) range of beliefs to $(-\infty, +\infty)$ to detect discrepancies better. As above, overall ($D^b_{p\bullet\bullet\bullet}$), node-level ($D^b_{\bullet h\bullet}$), person-level ($D^b_{i\bullet\bullet}$), or game-level ($D^b_{\bullet\bullet j}$) fit statistics can be constructed.

Subsequently, one can simulate a dataset for each MCMC draw in the Bayesian estimation, and calculate the discrepancy scores. One then ends up with a a simulated sample from the distribution of discrepancy scores D for replicated datasets. This sample, in turn, can be used to calculate a posterior predictive p-value (PPP=$\Pr(D_{obs} < D_{replicated})$). These PPPs show how likely it is to obtain a discrepancy score more extreme than the observed discrepancy statistics under the null hypothesis that the model fits (for details see Gelman et al., 1996; Gelman & Hill, 2007; Fox, 2010). Obtaining the

replicated discrepancy distributions overall, node-level, person-level, or game-level, one can break down the assessment of fit into components.

We followed this procedure to asses fit for the models we test in this paper. We discovered that the PPPs were rather stable across many alternative models, even though we calculated a separate PPP for each node, and an overall PPP for both choices and beliefs. We found "significant" PPPs, i.e., values smaller than 0.05, only for the very poor fitting models, e.g., Model A4 and only for some nodes and only for choices, not beliefs. We suspect that in our case the power of the PPPs is rather low, perhaps due to relatively low number of subjects and low number of data points per subject, especially for nodes C, D, and PD. We leave a serious investigation of this issue to a future, more statistical, study.

C.3 List of symbols

See Table C.3.

C.4 OpenBugs code for Specification B1.C2

```
# ------------------------------------------------------------------------
# OpenBugs code for analyzing social orientations and beliefs in
# nodes DG,C,D,PD. Model C1.B2 in Aksoy & Weesie 2013c
# data: pdgb5.txt    D[i]= choices {0,1}; 1 => U(a)>U(b),
#                   iB[i]= beliefs [-2.48, 2.48]; inverse.cumulative(p)
#                   xa[i]= outcome for self in option a (1)
#                   xb[i]= outcome for self in option b (1)
#                   ya[i]= outcome for self in option a (1)
#                   yb[i]= outcome for self in option b (1)
#         isDG[i],hc[i],hd[i],hp[i],isPD[i], = {0,1} node==DG,C,D,PD,(C|D|PD)
#          T_k[i],R_k[i],P_k[i],S_k[i] for k=1,2 = outcomes in PD game
# See the OpenBugs manual by Spiegelhalter,Thomas,Best and Lunn (2011)
# "http://www.openbugs.info/Manuals/Manual.html"
# !!!Revised version, may contain typos!!!
# ------------------------------------------------------------------------
model {
     for (i in 1:Nsubj) {
## own parameters
```

```
      theta[i,1:4] ~ dmnorm(mu[], invS[,])
       a[i] <- theta[i,1]
      ca[i] <- theta[i,2]
      da[i] <- theta[i,3]
      pa[i] <- theta[i,4]
## belief parameters
       # mean of beliefs about (a,ca,da)
        ba.mu[i] <- a.b0 + a.b1* a[i] +   e.a.mu[i]
       bca.mu[i] <-ca.b0 +ca.b1*ca[i] +  e.ca.mu[i]
       bda.mu[i] <-da.b0 +da.b1*da[i] +  e.da.mu[i]
 # variance of beliefs about (a,ca,da)
       ba.sd2[i] <- exp(a.c0 )
      bca.sd2[i] <- exp(ca.c0)
      bda.sd2[i] <- exp(da.c0)
  # errors are independent
       e.a.mu[i] ~ dnorm(0, e.a.mu.tau)
       e.ca.mu[i] ~ dnorm(0, e.ca.mu.tau)
       e.da.mu[i] ~ dnorm(0, e.da.mu.tau)
}
for (i in 1:N) {
  junk[i] <-  0*pBC[i] + 0*pBD[i] + 0*epc[i] + 0*epd[i]

# mbc[i]= Belief prob. that player 2 cooperates in node C

  invmbc[i]<-max(min((R2_[i]-T2_[i]+(R1_[i]-S1_[i])*bca.mu[id[i]])/
            sqrt(bca.sd2[id[i]]*pow(R1_[i]-S1_[i],2)+pow(exp(eb0),2)),5),-5)

  probit(mbc[i])  <- invmbc[i]

# mbd[i]= Belief prob. that player 2 cooperates in node D

  invmbd[i]<-max(min((S2_[i]-P2_[i]+(T1_[i]-P1_[i])*bda.mu[id[i]])/
            sqrt(bda.sd2[id[i]]*pow(T1_[i]-P1_[i],2)+pow(exp(eb0),2)),5),-5)

  probit(mbd[i])  <- invmbd[i]

#outcome differences in nodes

   Dx[i] <- (xa[i] - xb[i])*(1-hp[i])
  pDx[i] <- ((mbc[i]*R1_[i]+(1-mbc[i])*S1_[i])
           - (mbd[i]*T1_[i]+(1-mbd[i])*P1_[i]))*hp[i]
```

```
   Dy[i] <- (ya[i] - yb[i])*isDG[i]
  cDy[i] <- (ya[i] - yb[i])*hc[i]
  dDy[i] <- (ya[i] - yb[i])*hd[i]
  pDy[i] <- ((mbc[i]*R2_[i]+(1-mbc[i])*T2_[i])
           - (mbd[i]*S2_[i]+(1-mbd[i])*P2_[i]))*hp[i]

          D[i] ~ dbern(D.p[i])
         du[i] <- Dx[i] + pDx[i] +
                          a[id[i]] *   Dy[i] +
                         ca[id[i]] *  cDy[i] +
                         da[id[i]] *  dDy[i] +
                         pa[id[i]] *  pDy[i]

probit(D.p[i]) <- max(min(du[i]/exp(e0+e1*isPD[i]+e2*hp[i]),5),-5)

        iB[i]  ~ dnorm(iB.mu[i], iB.tau[i])
     iB.mu[i] <- (Dx[i] + pDx[i]
                         + ba.mu[id[i]]* Dy[i]+ bca.mu[id[i]]*cDy[i]
                         + bda.mu[id[i]]*dDy[i]+ bpa.mu[id[i]]*pDy[i])
                         /
                  sqrt( exp(eb0)        +
                         ba.sd2[id[i]] *pow( Dy[i],2)  +
                         bca.sd2[id[i]]*pow(cDy[i],2)  +
                         bda.sd2[id[i]]*pow(dDy[i],2)  )
    log(iB.tau[i]) <- err0 + err1*isPD[i]
    }
# priors
   mu[1:4] ~ dmnorm( mu0[], invS0[,] )
   mu0[1] <- 0
   mu0[2] <- 0
   mu0[3] <- 0
   mu0[4] <- 0

   invS0[1,1] <- 0.1
   invS0[1,2] <- 0
   invS0[1,3] <- 0
   invS0[1,4] <- 0
   invS0[2,1] <- 0
   invS0[2,2] <- 0.1
   invS0[2,3] <- 0
   invS0[2,4] <- 0
```

```
  invS0[3,1] <- 0
  invS0[3,2] <- 0
  invS0[3,3] <- 0.1
  invS0[3,4] <- 0
  invS0[4,1] <- 0
  invS0[4,2] <- 0
  invS0[4,3] <- 0
  invS0[4,4] <- 0.1

  invS[1:4,1:4] ~ dwish(Omega0[,], 4)
  Omega0[1,1] <- 0.1
  Omega0[1,2] <- 0
  Omega0[1,3] <- 0
  Omega0[1,4] <- 0
  Omega0[2,1] <- 0
  Omega0[2,2] <- 0.1
  Omega0[2,3] <- 0
  Omega0[2,4] <- 0
Omega0[3,1] <- 0
  Omega0[3,2] <- 0
  Omega0[3,3] <- 0.1
  Omega0[3,4] <- 0
  Omega0[4,1] <- 0
  Omega0[4,2] <- 0
  Omega0[4,3] <- 0
  Omega0[4,4] <- 0.1

     e0 ~ dnorm(0, 0.1)
    e1 ~ dnorm(0, 0.1)
    e2 ~ dnorm(0, 0.1)
   eb0 ~ dnorm(0, 0.01)
  err0 ~ dnorm(0, 0.01)
  err1 ~ dnorm(0, 0.01)
  a.b0 ~ dnorm(0,0.01)
  a.b1 ~ dnorm(0,0.01)
 ca.b0 ~ dnorm(0,0.01)
 ca.b1 ~ dnorm(0,0.01)
 da.b0 ~ dnorm(0,0.01)
 da.b1 ~ dnorm(0,0.01)
  a.c0 ~ dnorm(0,0.01)
 ca.c0 ~ dnorm(0,0.01)
```

```
  da.c0 ~ dnorm(0,0.01)
  e.a.mu.tau ~ dgamma(0.01, 0.01)
 e.ca.mu.tau ~ dgamma(0.01, 0.01)
 e.da.mu.tau ~ dgamma(0.01, 0.01)

#extract relevant parameters
   S[1:4,1:4] <- inverse(invS[,])
   p[1]   <- mu[1]
   p[2]   <- mu[2]
   p[3]   <- mu[3]
   p[4]   <- mu[4]
   p[5]   <- sqrt(S[1,1])
   p[6]   <- sqrt(S[2,2])
   p[7]   <- sqrt(S[3,3])
   p[8]   <- sqrt(S[4,4])
   p[9]   <- S[1,2]/(sqrt(S[1,1])*sqrt(S[2,2]))
   p[10]  <- S[1,3]/(sqrt(S[1,1])*sqrt(S[3,3]))
   p[11]  <- S[1,4]/(sqrt(S[1,1])*sqrt(S[4,4]))
   p[12]  <- S[2,3]/(sqrt(S[2,2])*sqrt(S[3,3]))
   p[13]  <- S[2,4]/(sqrt(S[2,2])*sqrt(S[4,4]))
   p[14]  <- S[3,4]/(sqrt(S[3,3])*sqrt(S[4,4]))
   p[15]  <- exp(e0)
   p[16]  <- exp(e0+e1)
   p[17]  <- exp(e0+e1+e2)

   pb[1] <-  a.b0
   pb[2] <-  a.b1
   pb[3] <- ca.b0
   pb[4] <- ca.b1
   pb[5] <- da.b0
   pb[6] <- da.b1
   pb[7] <-  a.c0
   pb[8] <- ca.c0
   pb[9] <- da.c0
   pb[10]<- 1/sqrt(e.a.mu.tau)
   pb[11]<- 1/sqrt(e.ca.mu.tau)
   pb[12]<- 1/sqrt(e.da.mu.tau)
   pb[13]<- exp(eb0)
   pb[14]<- err0
   pb[15]<- err1
   pb[16]<- err2
```

```
#R-sqs in mean beliefs
   pb[17]<- (S[1,1]* pow(a.b1,2))/(S[1,1]* pow(a.b1,2)+1/ (e.a.mu.tau))
   pb[18]<- (S[2,2]*pow(ca.b1,2))/(S[2,2]*pow(ca.b1,2)+1/(e.ca.mu.tau))
   pb[19]<- (S[3,3]*pow(da.b1,2))/(S[3,3]*pow(da.b1,2)+1/(e.da.mu.tau))
 }
```

Table C.2: 8 Asymmetric Prisoner's Dilemma Games used in the study. Last 9 columns include some descriptive statistics, N(subj.) = 155.

									Node PD			Node C			Node D		
	Game Outcomes									Belief %coop.			Belief %coop.			Belief %coop.	
Game	T_1	T_2	R_1	R_2	P_1	P_2	S_1	S_2	%coop.	Mean	SD	%coop.	Mean	SD	%coop.	Mean	SD
1	965	740	750	500	650	350	585	210	24	25.80	22.87	12	18.54	21.61	09	13.70	18.44
2	959	978	687	687	350	650	328	609	16	28.12	26.56	16	22.01	22.48	22	25.77	26.98
3	978	959	937	937	650	350	609	328	33	41.81	29.98	51	39.02	29.21	23	23.87	25.12
4	995	720	975	650	800	200	780	130	25	37.88	28.63	20	24.56	23.32	13	13.47	18.78
5	800	950	600	900	500	500	300	450	18	22.36	21.00	46	35.91	29.62	22	22.25	24.27
6	991	772	975	650	650	350	633	227	39	42.97	30.15	15	26.66	25.28	06	16.70	20.21
7	912	837	750	750	650	350	487	262	12	24.92	22.48	25	28.55	25.56	06	17.25	21.70
8	906	906	812	812	500	500	406	406	39	32.86	27.44	32	31.28	27.92	09	20.54	23.59

Table C.3: Symbols and their descriptions

Symbol	...	Description
$h \in \{\text{DG}, \text{C}, \text{D}, \text{PD}\}$	...	decision node
DG	...	decision node in a Dictator Game
C	...	decision node in a sequential Prisoner's Dilemma after the first player cooperated
D	...	decision node in a sequential Prisoner's Dilemma after the first player defected
PD	...	decision node in a sequential Prisoner's Dilemma in the first player's role
x_{khj}	...	outcome for Ego in option $k = 1, 2$ in a decision node
y_{khj}	...	outcome for Alter in option $k = 1, 2$ in a decision node
T,R,P,S	...	outcomes in a sequential Prisoner's Dilemma
θ_{ih}	...	weight i attaches to the outcomes of Alter in node h
$\boldsymbol{\mu}$	...	the vector of means of $\boldsymbol{\theta} = (\theta_{\text{DG}}, \theta_{\text{C}}, \theta_{\text{D}}, \theta_{\text{PD}})$
$\boldsymbol{\Sigma}$	...	the variance-covariance matrix of $\boldsymbol{\theta}$
ϵ_{ihj}	...	evaluation error in node h
τ_h	...	standard deviation of ϵ_{ihj}
$\boldsymbol{T}$	...	diagonal containing $(\tau^2_{\text{DG}}, \tau^2_{\text{C}}, \tau^2_{\text{D}}, \tau^2_{\text{PD}})$
$\widetilde{\theta}_{ih}$	...	a random variable representing i's belief about θ in node h
$\widetilde{\mu}_{ih}$	...	the mean of $\widetilde{\theta}_{ih}$
$\widetilde{\sigma}^2_{ih}$	...	the variance of $\widetilde{\theta}_{ih}$
b_{0h}	...	intercept for the regression of $\widetilde{\mu}_{ih}$ on θ_{ih}
b_{1h}	...	slope for the regression of $\widetilde{\mu}_{ih}$ on θ_{ih}
η_{ih}	...	error term for the regression of $\widetilde{\mu}_{ih}$ on θ_{ih}
ς^2_h	...	variance of η_{ih}
$\widetilde{\sigma}^2_{0h}$	...	the $\widetilde{\sigma}^2_{ih}$, assumed constant across i
$\widetilde{\tau}^2$	...	belief about τ_h, assumed invariant across h
$\widetilde{\pi}_{ihj}$	...	i's beliefs about fraction of others chosing option 1 in game j in node h
p_{ihj}	...	i's elicited beliefs (response) about $\widetilde{\pi}_{ihj}$
υ_{ihj}	...	error in elicited beliefs
ζ^2_h	...	variance of υ_{ihj}
$\widetilde{\pi}_{i\text{C}j}$	...	i's belief about the probability that a random other cooperates in game j in node C
$\widetilde{\pi}_{i\text{D}j}$	...	i's belief about the probability that a random other cooperates in game j in node D

Appendix D

Appendix Chapter 5

D.1 Equilibria for alternative distributions for the inequality aversion parameter

Table D.1 presents the solutions of the equilibrium threshold σ_e^* for a variety of alternative type distributions. Compare the solutions for the uniform distribution to the solutions for Chi^2 distributions. The threshold values and the predicted orderings of the games are very close. We also computed equilibria also for many other distributions than those in Table D.1 (not reported). It seems that the exact shape of the distribution seem qualitatively of little importance. Thus, considering various uniform distributions is sufficient to come up with accurate predictions for the 9 asymmetric investment games. In the results part of the paper we present results for (1) with a moderately inequality aversive population, i.e, U[0,2], and (2) with highly inequality aversive population, i.e, U[0,100]. For less inequality aversive populations, than U[0,2], the equilibrium implies defection for sure, and thus is not different than the classical game-theoretic prediction with selfish actors. This prediction is obviously not supported by our data. On the other hand, a higher inequality aversive population is not plausible, since this requires actors to value inequality a lot more than the outcomes for the self. Nevertheless, we also report solutions for a highly inequality aversive population, U[0,100]. This distribution yields real (not infinite) solutions for all games except game 1 and game 9. In games 1

Table D.1: Solutions of σ_e^* for alternative type distributions

		Uniform distribution $U[0,a]$ for σ			
Game	μ_1-α_1	$U[0, a < 1.65]$	$U[0,2]$	$U[0,7]$	$U[0, a > 100]$
1	400-0.4	∞	∞	∞	∞
2	400-0.5	∞	0.375	0.270	0.251
3	400-0.6	∞	∞	0.25	0.147
4	500-0.4	∞	∞	∞	4.089
5	500-0.5	∞	0.5	0.271	0.251
6	500-0.6	∞	∞	∞	0.156
7	600-0.4	∞	∞	1.5	1.339
8	600-0.5	∞	0.33	0.264	0.251
9	600-0.6	∞	∞	∞	∞
		$\text{Chi}^2(a)$ and $F(1;a)$ distributions for σ			
Game	μ_1-α_1	$\text{Chi}^2(a < 1.5)$	$\text{Chi}^2(2)$	$F(1; a = 1, 3, 5, 50, 100)$	$\text{Chi}^2(10)$
1	400-0.4	∞	∞	∞	∞
2	400-0.5	∞	0.355	∞	0.25
3	400-0.6	∞	∞	∞	0.143
4	500-0.4	∞	∞	∞	4
5	500-0.5	∞	0.384	∞	0.25
6	500-0.6	∞	∞	∞	0.108
7	600-0.4	∞	∞	∞	1.333
8	600-0.5	∞	0.319	∞	0.25
9	600-0.6	∞	∞	∞	∞

and 9, defection is always dominant at least for one player, and thus the only equilibrium is mutual defection. We conjecture that our conclusions would not differ had we considered other distributions than the two that we report, since none of the solutions in table D.1 approximate our data. In fact, other distributions such as U[0,7] fit even poorer than the solutions for U[0,100].

D.2 Predictions for the Fehr and Schmidt (1999) model in asymmetric PDs

In the paper, we report results for a simplified version of the inequality aversion model of (Fehr & Schmidt, 1999). The inequality aversion model of (Fehr & Schmidt, 1999) for an outcome allocation for the self (x) and the other (y) is

defined as:

$$u(x, y, \alpha_i, \beta_i) = x - \beta_i max(x - y, 0) - \alpha_i max(y - x, 0) \qquad \beta \in [0, 1], \alpha > \beta.$$

The first term of the right side of the equation represents objective outcome for the self, the second term represents the utility loss from disadvantageous inequality, and the third term represents the utility loss from advantageous inequality. The β and α parameters reflect how much value is given to the difference between own and other's outcomes. The assumptions about the range of these two parameters ensure that for the same level of inequality, utility loss due to disadvantage inequality is always greater than the utility loss due to advantageous utility, utility is always increasing in x, and increasing in y only if $x > y$.

In all of the 9 asymmetric investment games that we employed in our experimental design, $P_e > S_e$ and $|P_e - P_a| < T_a - S_e$. This means that for the inequality aversion model of Fehr & Schmidt (1999), mutual defection (DD) is always an equilibrium, since DD always yield a higher outcome for ego, and a smaller utility loss due to inequality than the case where ego cooperates and alter defects (CD), for both players. When can cooperation be part of the equilibrium behavior? This is possible in only 3 out of 9 asymmetric investment games where $\lambda_e = \lambda_a = 0.5$. When $\lambda_e \neq 0.5$, which holds for 6 of 9 of asymmetric investment games used in the experimental design, the only equilibrium is DD.

Consider the three games where $\lambda_e = 0.4$. In these three games, the utility for ego in case of mutual cooperation, $\mathrm{U}_e(CC)$, is $R_e - \alpha_e(R_a - R_e)$, and $\mathrm{U}_e(CD)$ is $T_e - \beta_e(T_e - S_a)$. In this case $\mathrm{U}_e(CC) \geq \mathrm{U}_e(CD)$ if:

$$\beta_e \geq \frac{T_e - R_e}{T_e - S_a} + \alpha_e \frac{R_a - R_e}{T_e - S_a}. \tag{D.1}$$

In our design, for all of the three games where $\lambda_e = 0.4$, $\frac{Te - R_e}{T_e - S_a} + \frac{R_a - R_e}{T_e - S_a} > 1$. Fehr & Schmidt (1999) assume $\alpha_e > \beta_e$ and $\beta_e \leq 1$. This implies that in these three games, the condition (D.1) does not hold and $\mathrm{U}_e(CD) > \mathrm{U}_e(CC)$, thus the single equilibrium is DD. Analogously, in the remaining three games

where $\lambda_e = 0.6$, $\lambda_a = 0.4$. Starting from the alter's perspective, the same analysis concludes that the single equilibrium is DD. To sum up, the inequality aversion model of Fehr & Schmidt (1999) predicts that cooperation can be observed only in 3 of 9 games that we use in our experimental design, where $R_e = R_a$. In the remaining 6 games, the model of Fehr & Schmidt (1999) predicts no cooperation at all. This prediction is clearly at odds with our data.

D.3 Hypotheses tested with Bayesian model selection

Here we elaborate on the details of the formulation of hypotheses tested in the Bayesian model selection procedure. These hypotheses involve (in)equality constraints on the probability of cooperation in the nine games used in the experiment. The cooperation thresholds computed from the formal analyses are the basis of these inequality constraints. An important issue is the handling of ties between the predicted cooperation rates. For illustration purposes, let's consider the predictions of the inequality aversion model with CSE for the games Γ_1, Γ_2, Γ_4, and Γ_6. For these four games, cooperation thresholds of the inequality aversion model with CSE points to the following set of (in)equality constraints on the probability π of cooperation: $\pi(\Gamma_2) > \pi(\Gamma_6) = \pi(\Gamma_4) > \pi(\Gamma_1) = 0$. This particular set of constraints involve two equality constraints: Γ_6 and Γ_4 are tied, and zero cooperation is predicted in Γ_1. For the reasons discussed in section 5.2, equality constraints could be problematic. There are various ways of dealing with these equality constraints. The first way is simply not imposing any constraint when there is an equality sign. To be more precise, this first way of dealing with equality constraints would include only the following constraints in our example:

$$\pi(\Gamma_2) > \pi(\Gamma_6) > \pi(\Gamma_1);\ \pi(\Gamma_2) > \pi(\Gamma_4) > \pi(\Gamma_1). \tag{D.2}$$

Thus, (D.2) does not impose any direct constraint among the tied games Γ_6 and Γ_4, neither does it include a constraint of zero probability of cooperation

in Γ_1. We call this first option *strict inequality*.

Another, more restrictive way of dealing with equality cases is setting a boundary value δ on the differences in cooperation probabilities of tied games and of games with zero predicted cooperation. Again turning to our example, this option will involve the following constraints:

$$\pi(\Gamma_2) > \pi(\Gamma_6) > \pi(\Gamma_1); \quad \pi(\Gamma_2) > \pi(\Gamma_4) > \pi(\Gamma_1); \\ |\pi(\Gamma_6) - \pi(\Gamma_4)| \leq \delta; \quad \pi(\Gamma_1) \leq \delta. \quad \text{(D.3)}$$

where δ is a small number, such as 0.05 or 0.01. Thus (D.3) does not impose exact equality constraints, but instead uses a small boundary value. We call this option *weak equality*.

Finally, one may impose equality constraints on games with tied predictions and games with zero predicted cooperation. We call this *strict equality*. Strict equality condition is a special case of the weak equality condition with $\delta = 0$. Note that strict inequality condition is also a special case of weak inequality where $\delta = 1$.

We have computed the PMPs with assuming strict inequality, weak equality, and strict equality conditions, both for own behavior and expected behavior. We also computed PMPs with weak equality conditions for $\delta = 0.05$ and $\delta = 0.01$. For all of these variations, the same pattern emerges: the inequality aversion is flatly rejected, and both the social orientation and normative models receive strong support. We observe that only $\Pr(H_i|All)$s for the social orientation and normative models change, depending on how ties are handled. Yet, since we do not choose one of these two as the best fitting model based in this Bayesian model selection, this issue is not problematic. In the paper, we report PMPs computed with *strict inequality* condition for two reasons. First, with *strict inequality*, PMPs are most robust with respect to the choice on the prior distribution. Second, the inequality aversion model receives the highest PMPs with *strict inequality*, and even in this condition, the support for the model is too low.

D.4 Some notes on equilibrium selection

Multiple equilibria may arise in the games that we use in our experiments for the inequality aversion model with CSE: whenever CC is equilibrium, DD is also equilibrium. In the paper, we base predictions on the equilibrium which maximizes the joint probability of cooperation, namely CC, in case of multiple equilibria. Below we investigate the consequences of applying alternative equilibrium selection methods: risk dominance and Pareto dominance criteria (Harsanyi & Selten, 1988).

D.4.1 Risk dominance

DD risk dominates CC if:

$$[P_e - S_e + \sigma_e(|T_a - S_e| - |P_e - P_a|)][P_a - S_a + \sigma_a(|T_e - S_a| - |P_e - P_a)]$$
$$>$$
$$[R_e - T_e + \sigma_e(|T_e - S_a| - |R_e - R_a|)][R_a - T_a + \sigma_a(|T_a - S_e| - |R_e - R_a|)].$$

In the parallel PD (PPD), the inequality simplifies to:

$$\sigma_e X_a + \sigma_a X_e > \sigma_e \sigma_a (|P_e - P_a| - |R_e - R_a|). \tag{D.4}$$

The left hand side of (D.4) is always positive, as $\sigma > 0$ and $X > 0$ in the PD. The right hand side of the equation is negative for the 7 investment games except the two games, Γ_2 and Γ_8. Thus only in the two games where $R_e = R_a$, and $P_e \neq P_a$, CC may risk dominate DD. Even in the full symmetric game DD still risk dominates CC. Note that under CSE, (D.4) simplifies further as:

$$X_e + X_a > \sigma \left(|P_e - P_a| - |R_e - R_a|\right). \tag{D.5}$$

For Γ_2 and Γ_8, DD is always in equilibrium; CC is also in equilibrium iff $\sigma > 0.25$, and CC risk dominates DD iff $\sigma > 1.25$. In all of the remaining 7 games DD risk dominates CC. Thus, if we apply the risk dominance criterion,

cooperation becomes part of the equilibrium behavior only in two out of nine games that we use. Clearly, this prediction contradicts our data.

D.4.2 Pareto dominance

CC Pareto dominates DD if:

$$|R_1 - R_2| - |P_1 - P_2| < \min(\frac{R_1 - P_1}{\sigma_1}, \frac{R_2 - P_2}{\sigma_2}) \tag{D.6}$$

CC can be the Pareto dominant equilibrium only in three games out of the seven games that we use in our experiments, where C can be part of the equilibrium behavior: Γ_2, Γ_8, and the full symmetric game, Γ_5. Otherwise, either neither equilibria Pareto dominates the other, or DD Pareto dominates CC. Had we applied a stepwise selection, i.e., first apply the Pareto dominance and if there is no Pareto dominant equilibrium apply risk dominance, then the inequality aversion model would predict cooperation only in three out of the nine games that we use in our experiment, i.e., Γ_2, Γ_5, and Γ_8. This prediction is also clearly at odds with our experimental data.

Appendix E

Appendix Chapter 7

E.1 Binary Dictator Games games used in the experiment

See table E.1.

E.2 Detailed results

See Table E.2.

Table E.1: Binary Dictator Games games used in the experiment.

	Option A		Option B	
Game	You get	Other gets	You get	Other gets
1	410	410	400	370
2	500	400	550	550
3	320	320	300	280
4	450	450	400	525
5	350	350	300	475
6	630	630	600	735
7	640	640	680	920
8	660	750	630	630
9	420	420	440	455
10	640	640	680	695
11	700	715	650	650
12	480	40	440	440
13	540	540	580	340
14	310	310	320	290
15	310	310	320	305
16	700	650	650	650
17	540	500	540	555
18	650	600	650	685

Table E.2: Detailed results of the multilevel probit models: Means (μ), variances (σ^2) of θ and β parameters, variance of the evaluation error (τ^2), log-likelihoods (LL), and number of subjects (N-subj.) and decisions (N-dec.) per experimental condition.

Ego.Alter	μ_θ	μ_β	σ^2_θ	σ^2_{beta}	$2\tau^2$	LL	N-subj.	N-Dec.
T.T	-.066	.030	.035	.029	.103	-337.463	38	684
	(.037)	(.034)	(.014)	(.013)	(.016)			
T.TD	.070	.094	.150	.079	.175	-356.130	37	666
	(.070)	(.056)	(.060)	(.035)	(.033)			
T.D	-.069	.112	.073	.044	.336	-392.862	37	666
	(.053)	(.046)	(.033)	(.023)	(.071)			
TD.T	-.245	.589	.816	.835	.933	-246.201	23	414
	(.215)	(.250)	(.513)	(.528)	(.457)			
TD.D	-.092	.167	.024	.057	.711	-193.741	17	306
	(.066)	(.080)	(.028)	(.046)	(.286)			
D.T	.104	.081	.044	.008	.040	-206.048	34	612
	(.041)	(.024)	(.016)	(.005)	(.006)			
D.TD	.123	.091	.038	.021	.058	-250.788	35	629
	(.039)	(.031)	(.015)	(.010)	(.009)			
D.D	.049	.104	.020	.026	.047	-238.087	34	612
	(.030)	(.033)	(.010)	(.011)	(.008)			

*Standard errors in parentheses

Samenvatting: Discussie en conclusies

De kern van het formeel-theoretische framework van dit proefschrift wordt ontwikkeld in de hoofdstukken 2 tot en met 5. Deze vier hoofdstukken leveren een speltheoretisch framework dat een verklaring moet bieden voor samenwerking in niet-ingebedde situaties. Elk van deze hoofdstukken bestudeert een deel van het overkoepelende theoretische framework en bouwt voort op het voorgaande hoofdstuk met toenemende statistische en theoretische verfijning. Het speltheoretische model dat in deze hoofdstukken wordt beschreven wordt geschat met experimentele data. Volgens de gepresenteerde theoretische en empirische analyses is samenwerking in een niet-ingebedde setting te verklaren door drie factoren: (1) heterogene sociale preferenties, (2) ego-centrische consensus-achtige systematische fouten ('biases') in verwachtingen, en (3) toevallige inschattingsfouten (evaluation errors). Deze drie factoren corresponderen met twee kernaannames van het standaard rationele keuze model. De eerste factor herziet de zelfzuchtigheidsaanname en de andere twee betreffen de rationaliteitsaanname van het standaardmodel.

De twee laatste empirische hoofdstukken, de hoofdstukken 6 en 7, passen dit framework toe in twee sociologisch belangrijke contexten. Hoofdstuk 6 beschrijft experimenteel onderzoek naar de invloed van procedurele rechtvaardigheid op samenwerking. Procedurele rechtvaardigheid heeft inderdaad significante invloed op samenwerking. Bovendien blijkt sekse een belangrijke factor te zijn die het verband tussen procedurele rechtvaardigheid en samenwerking modereert. Hoofdstuk 7 doet verslag van een internationaal interetnisch ex-

periment met Turkse, Nederlandse, en Turks-Nederlandse respondenten. Het toont duidelijke 'inter-group'-verschillen in sociale preferenties die lijken op het 'Noord versus Zuid'-effect (Delhey & Newton, 2005), maar de sociale afstand tussen Ego en Alter lijkt niet van belang, i.e., we vinden geen 'ingroup favoritism'.

In dit slothoofdstuk behandelen we in vogelvlucht enkele algemene conclusies gebaseerd op de belangrijkste resultaten. Om herhaling te vermijden, zien we af van een uitgebreide samenvatting van de bevindingen die we reeds in de hoofdstukken presenteren. Ieder hoofdstuk omvat immers reeds een resumé van de belangrijkste bevindingen van dat hoofdstuk.

Belangrijkste resultaten en conclusies

Sociale preferenties als toestanden ('states') en eigenschappen ('traits')

De uitkomsten die we presenteren in dit proefschrift tonen aan dat sociale preferenties complex zijn. Ze worden beïnvloed door verscheidene factoren zoals cultuur (hoofdstuk 7), procedurele rechtvaardigheid (hoofdstuk 6), en relatiegeschiedenis (hoofdstuk 4). Ook de spelstructuur zelf heeft invloed op sociale preferenties. Zoals we bijvoorbeeld laten zien in hoofdstuk 4 is het gemiddelde gewicht dat proefpersonen toekennen aan de materiële uitkomsten van anderen significant kleiner in sequentiële sociale dilemma's dan in beslissituaties waar mensen eenvoudigweg moeten kiezen tussen verschillende uitkomsten voor zichzelf en een ander ('Dictator Games'). Sociale preferenties zijn dus afhankelijk van de beslissituatie, i.e., sociale preferenties zijn *toestanden*, in elk geval gedeeltelijk (Steyer et al., 1999). Dit betekent dat het niet mogelijk is om sociale preferenties buiten de context te meten, aangezien context invloed heeft op sociale preferenties. Deze uitkomsten komen overeen met de literatuur over 'framing' (e.g., Hertel & Fiedler, 1994, 1998; Liberman et al., 2004; Lindenberg & Steg, 2007). Deze literatuur laat zien dat relatief subtiele context-cues, zoals het tonen van coöperatieve adjectieven voorafgaand aan het spel of het labelen van het sociale dilemma spel als 'Community Game' in plaats van 'Wallstreet Game', sociale preferenties en coöperatief gedrag sub-

stantieel kunnen beïnvloeden (Hertel & Fiedler, 1994, 1998; Liberman et al., 2004). De 'goal framing'-theorie kan inzichtelijk maken hoe context sociale preferenties beïnvloedt (Lindenberg, 2001, 2008). Deze theorie veronderstelt dat verschillende motiverende doelen naast elkaar kunnen bestaan. In ons geval zouden zulke verschillende doelen kunnen zijn: maximaliseren van het eigen voordeel, minimaliseren van ongelijkheid, of maximaliseren van de uitkomsten van anderen, dus verschillende soorten sociale preferenties. Deze verschillende doelen concurreren met elkaar om de beperkte cognitieve vermogens van het individu. Het dominante doel in deze competitie bepaalt het frame waarmee iemand de situatie evalueert. De spelstructuur, relatiegeschiedenis en andere contextuele factoren, zoals procedurele rechtvaardigheid, kunnen werken via situationele selectie van doelen waarbij bepaalde soorten doelen of sociale preferenties saillanter zijn.

Verder vonden we een hoge mate van individuele variatie in sociale preferenties. Bovendien, ook al beïnvloeden contextuele factoren als spelstructuur en relatiegeschiedenis de sociale preferenties, toch is er een hoge correlatie tussen iemands sociale preferenties in verschillende contexten. Die correlatie van iemands sociale preferenties over tijd en contexten is ook gevonden in ander onderzoek (e.g., Van Lange, 1999; Van Lange et al., 2007; Benz & Meier, 2008). Benz & Meier (2008) rapporteren bijvoorbeeld dat pro-sociaal gedrag van hun proefpersonen, zoals vrijwillige donaties gedurende het experiment, correleert met gedrag van die proefpersonen van twee jaar voor tot twee jaar na het experiment. Dit geeft aan dat sociale preferenties ook *eigenschappen* (traits) zijn, op zijn minst gedeeltelijk.

Het wordt nog gecompliceerder doordat onze resultaten ook suggereren dat het effect van context op sociale preferenties van mensen niet constant is. Dat wil zeggen: sociale preferenties van mensen worden (gedeeltelijk) beïnvloed door context, maar bij sommige mensen is die invloed sterker dan bij anderen. In hoofdstuk 4 laten we bijvoorbeeld zien dat het gewicht dat iemand toekent aan de materiële uitkomsten van anderen varieert afhankelijk van spelstructuur en relatiegeschiedenis. Maar de mate waarin dit gewicht varieert, varieert tussen personen. Dit wijst verder op een interactie tussen dispositionele en contextuele determinanten van sociale preferenties. Vergelijkbare interac-

ties tussen dispositie en context binnen sociale dilemma's zijn ook in eerder onderzoek gevonden (e.g., Van Lange et al., 1997; De Cremer & Van Vugt, 1999; Van Lange, 2000; Bogaert et al., 2008). De Cremer & Van Vugt (1999) laten bijvoorbeeld zien dat het experimenteel opwekken van een groepsidentiteit zorgt dat dispositioneel zelfzuchtige individuen zich coöperatiever gedragen vergeleken met de conditie waar groepsidentiteit niet saillant is. Dezelfde manipulatie heeft echter geen significant effect op dispositioneel pro-sociale mensen die al veel gewicht toekennen aan de uitkomsten van anderen, los van groepsidentiteit.

Om deze onderzoeksbevindingen in één framework te incorporeren, beschouwen we de vector θ_{ij} van sociale preferenties van persoon i, zoals hoeveel i geeft om de materiële uitkomsten van anderen of hoezeer i een hekel heeft aan ongelijkheid, in context j die kan omvatten: relatiegeschiedenis, spelstructuur, sociale identiteit van interactiepartners etc. Zowel onze resultaten als de hier besproken literatuur suggereren dat θ_{ij} kan worden geschreven als de som van drie componenten:[1]

$$\theta_{ij} = \alpha_i + \beta_j + \gamma_{ij} \tag{1}$$

waarbij α_i de eigenschapscomponent weergeeft, β_j de toestandscomponent en γ_{ij} de interactie tussen die twee.[2]

Zoals besproken in het inleidende hoofdstuk, valt de verklaring van de herkomst van sociale preferenties en een diepgaande verklaring van waarom en hoe context sociale preferenties beïnvloedt, buiten het bestek van dit proefschrift. Onze uitkomsten en de tot dusver besproken literatuur vragen om een algemene metatheorie over sociale preferenties. Zo'n metatheorie zou beschrijven waar sociale preferenties als disposities vandaan komen en voorspellen hoe situationele factoren sociale preferenties beïnvloeden. Met andere woorden:

[1] Voor sociaal psychologen zal vergelijking (1) volkomen normaal zijn, aangezien deze sterk lijkt op de bekende formule van Lewin (1936) (zie ook Van Lange, 2000). In Lewins formule is gedrag een functie van de omgeving en de persoon. Het belangrijkste verschil tussen vergelijking (1) en Lewins formule is dat vergelijking (1) niet over gedrag gaat, maar over sociale preferenties.

[2] De componenten α en β in (1) moeten niet worden verward met de α en β in hoofdstuk 2 en γ in (1) niet met de γ in hoofdstuk 5 die verwijzen naar speciale typen sociale preferenties.

het identificeert de factoren die de termen α_i, β_j en γ_{ij} in vergelijking (1) beïnvloeden. In de paragraaf over toekomstig onderzoek hieronder zullen we kort bespreken hoe systematisch de contextuele en dispositionele determinanten van sociale preferenties kunnen worden bestudeerd.

Afgezien van vragen naar hun herkomst, is het complexe karakter van sociale preferenties waarschijnlijk niet datgene wat een rationele-keuze-theoreticus graag ziet. De werkelijkheid is complexer dan het standaardmodel veronderstelt. Natuurlijk moet in formele modellen enige complexiteit genegeerd worden. Eenvoudige modellen zijn juist erg geschikt, vooropgesteld dat het buiten beschouwing laten van complexiteit niet resulteert in slechte empirische 'fit' en dat complexere formele modellen niet goed geanalyseerd kunnen worden.[3] In ons geval gaan deze twee condities niet op. Ten eerste zorgt het negeren van de complexiteit voor een slechte beschrijving van onze data. Zoals we laten zien in hoofdstuk 4, worden modellen waarin contextuele invloed op sociale preferenties wordt genegeerd simpelweg verworpen door experimentele data. Ten tweede, zoals we laten zien in verschillende hoofdstukken, zijn er statistische modellen en theoretische benaderingen beschikbaar om deze complexiteit in te passen. Dus lijkt er geen dwingende reden te zijn om complicerende factoren te schuwen.

We geven toe dat er ook vormen van complexiteit zijn die we wel negeren in dit proefschrift. De niet uitputtende lijst van factoren die we negeren omvat risicopreferenties (Raub & Snijders, 1997); andere soorten sociale preferenties dan die we beschouwen in dit proefschrift, zoals de spijt- of schuldaversie (Loomes & Sugden, 1982; Ellingsen et al., 2010); de rol van 'pre-play'-communicatie (Balliet, 2010); en signalen die sociale preferenties zichtbaar maken voor anderen (Frank, 1988; Bacharach & Gambetta, 2001). We zijn overtuigd dat in principe veel van deze factoren in het framework van dit proefschrift kunnen worden geïncorporeerd.

[3] Voor een informele discussie over de doelstellingen van formele modellen, zie Carvalho (2007).

Verwachtingen over preferenties van de Ander: speltheorie als beslistheorie

> Geen mens kent andere mensen werkelijk. Het beste wat hij kan doen is veronderstellen dat ze zijn zoals hijzelf. (John Steinbeck, The Winter of Our Discontent).

Hoewel sociale preferenties complex zijn, suggereren onze resultaten dat verwachtingen over sociale preferenties van de ander niet zo complex zijn, in elk geval in de gevallen die we in dit proefschrift bestudeerd hebben. Verwachtingen over sociale preferenties van de ander zijn sterk gecorreleerd met eigen preferenties in alle condities en manipulaties die we geanalyseerd hebben: simultaan 'Prisoner's Dilemma' (PD), sequentieel PD, binaire 'Dictator Games' (DG), verwachtingen gemeten na eigen keuzes, verwachtingen gemeten tegelijk met eigen keuzes, en ook als er al dan niet beloningen gegeven worden voor de nauwkeurigheid van geëxpliciteerde verwachtingen.[4] Dit kenmerk van verwachtingen helpt de onderzoeker omdat het opnemen van ego-centrische verwachtingen in formele modellen relatief eenvoudig is, zoals we laten zien in hoofdstuk 5. Tegelijkertijd hebben ego-centrische verwachtingen wel enige implicaties voor de rationele-verwachtingencomponent van de rationaliteitsaanname van het standaardmodel dat we hieronder kort bespreken.

Onze bevindingen over de correlatie tussen sociale preferenties en verwachtingen komen overeen met eerder onderzoek in de sociale psychologie (e.g., Kelley & Stahelski, 1970; Ross et al., 1977; Kuhlman et al., 1992; Iedema, 1993) en recenter in experimentele economie (e.g., Blanco et al., 2009, 2011) die vergelijkbare ego-centrische, consensus-achtige 'biases' laten zien bij verwachtingen. Onze studie beschrijft deze biases, een aantal methodologische tekortkomingen in de sociaalpsychologische literatuur vermijdend, met behulp van formeel gedefinieerde sociale-preferentiemodellen. Een andere bijdrage van ons onderzoek is dat we expliciet de variantie onderzoeken van verwachtingen over sociale preferenties van de ander als functie van iemands eigen sociale preferenties. De algemene trend die we vinden is dat de variantie van

[4] Strikt genomen verwijzen we naar correlaties tussen sociale preferenties en de gemiddelden van verwachtingen over sociale preferenties van de Ander.

verwachtingen het kleinst is voor zelfzuchtige mensen; dus zelfzuchtige mensen zijn zekerder van hun verwachtingen. Deze onzekerheid neemt toe naarmate sociale preferenties afwijken van zelfzuchtigheid; dit bevestigt het kegel-effect ('cone effect'; zie Iedema, 1993).[5] Hoewel Iedema (1993) dit kegel-patroon twee decennia geleden al voorspelde, heeft het tot dusver niet veel aandacht gekregen in de literatuur.

In dit proefschrift bespreken we niet uitgebreid de causale mechanismen die ten grondslag liggen aan de relatie tussen sociale preferenties en verwachtingen over sociale preferenties van anderen. In hoofdstuk 2 onderzoeken we de implicaties van drie sociaalpsychologische hypotheses, maar we graven niet diep in de causale mechanismen. Het consensuseffect, de verwachting dat anderen vergelijkbaar zijn met jezelf, ligt ten grondslag aan al deze drie hypotheses. Er zijn verscheidene elkaar niet-uitsluitende mechanismen die het consensuseffect kunnen veroorzaken. Iedema (1993) bespreekt deze mechanismen in detail. Het consensuseffect treedt bijvoorbeeld op wanneer verwachtingen reflecties zijn van eigen preferenties: wanneer mensen niet genoeg informatie hebben over de voorkeuren van anderen, gebruiken ze de beschikbare informatie, namelijk hun eigen voorkeur, om de voorkeuren van anderen in te schatten. Dit wordt de cognitieve-beschikbaarheidsbias genoemd (Marks & Miller, 1987). Zelfs wanneer er andere informatie beschikbaar is—mensen ontmoeten anderen tijdens hun leven—vertoont die andere informatie dezelfde tendens als de eigen ervaringen. Dit komt doordat mensen banden aangaan met anderen die op hen lijken in achtergrond, interesses, voorkeuren, etc. Dit versterkt het 'false consensus'effect nog verder. Dit mechanisme noemt men 'selective exposure' (Ross et al., 1977). 'Selective exposure' en de beschikbaarheidsbias wijzen op de causale richting dat preferenties tot verwachtingen leiden. Anderzijds treedt het consensuseffect op als iemand zijn preferenties aanpast aan zijn verwachtingen, ten gevolge van bijvoorbeeld reciprociteit of conformisme. Bijvoorbeeld: iemand kan verwachten dat een ander sterk prosociaal is tegenover hem- of haarzelf, en door deze verwachting voelt men zich ook pro-sociaal tegenover de ander, als gevolg van reciprociteit. En zo ook: als iemand verwacht dat een bepaalde sociale preferentie de heersende norm is, kan iemand

[5] Zie de hoofdstukken 2 en 3 voor de details.

zijn eigen preferenties aanpassen om zich te conformeren aan die sociale norm (Asch, 1951). In deze laatste twee voorbeelden veroorzaken verwachtingen de preferenties. We maken ons er niet druk om welk causaal mechanisme aan het werk is, of, wanneer het er meerdere zijn, welke sterker is, zolang het consensuseffect aanwezig is. Ongeacht de exacte causale mechanismen is de sterke correlatie tussen preferenties en verwachtingen een belangrijke factor in de analyse van niet-ingebedde sociale dilemma's.

De rationele-verwachtingen-aanname die ten grondslag ligt aan het Bayesiaans-Nash evenwicht voor spellen met incomplete informatie impliceert dat mensen een gemeenschappelijke inschatting hebben van de verdeling van sociale preferenties bij anderen. Bovendien correspondeert deze gemeenschappelijke verwachting met het werkelijke voorkomen van sociale preferenties in de populatie (Harsanyi, 1968). De hoge correlatie tussen verwachtingen en preferenties tezamen met de heterogeniteit in sociale preferenties die we in dit proefschrift documenteren, laten echter zien dat mensen ook heterogeen zijn in hun verwachtingen. Met andere woorden, mensen hebben geen gemeenschappelijke inschatting van de verdeling van sociale preferenties. In feite testen en verwerpen we daarmee de rationele-verwachtingen-aanname in verschillende hoofdstukken. Dit betekent dat het te naïef zou zijn om het Bayesiaans-Nash evenwicht te interpreteren als een accurate representatie op macroschaal in elk geval voor niet-ingebedde interacties. Toch betekent dit niet dat we speltheorie moeten afschaffen bij het onderzoek naar niet-ingebedde interacties. Individuele 'priors' corresponderen met wat McKelvey & Palfrey (1992) het *ego-centrische model* noemen. In het ego-centrische model speelt iedereen het spel in zijn of haar eigen wereld, in de overtuiging dat zijn of haar eigen wereld de 'echte' wereld is. Deze overtuiging van de actoren klopt uiteraard niet, in elk geval voor sommige actoren, aangezien er zeer waarschijnlijk maar één echte wereld is. Echter, zolang mensen het spel spelen alsof hun persoonlijke wereld de echte wereld is, kunnen met speltheoretische modellen de keuzes van actoren en hun verwachtingen over keuzes van de ander worden voorspeld. Stel je bijvoorbeeld een persoon voor die zeer veel gewicht toekent aan de uitkomsten van anderen; een zeer pro-sociaal persoon dus. En zoals het consensuseffect impliceert, verwacht deze persoon bovendien dat de meeste an-

dere mensen ook pro-sociaal gemotiveerd zijn. Stel dat deze persoon een spel speelt, laten we zeggen het 'Prisoner's Dilemma'-spel, in een experiment met een anonieme opponent. We kunnen nog steeds het Bayesiaans-Nash evenwicht van dit 'Prisoner's Dilemma'-spel uitrekenen waarbij de 'type verdeling' correspondeert met de (veel te optimistische) verwachtingen van deze specifieke persoon.[6] Aangezien deze persoon het spel speelt alsof haar eigen (veel te optimistische) verwachtingen de werkelijke verdeling zijn, kunnen haar keuze en haar verwachtingen over keuzes van de ander in het 'Prisoner's Dilemma' worden voorspeld met het Bayesiaans-Nash evenwicht. Bovendien kan deze voorspelling empirisch worden getoetst.

Er valt echter een voorbehoud te maken bij het laten vallen van de aanname over gemeenschappelijke verwachtingen ('prior'). Het is hetzelfde voorbehoud als bij het introduceren van heterogene sociale preferenties: men kan veel verschijnselen verklaren door het achteraf arbitrair aanpassen van aannamen over de verwachtingen van actoren (voor een overzicht, zie Morris, 1995). Het verklaren van verschijnselen door het achteraf willekeurig aanpassen van verwachtingen is niet wetenschappelijk. Om dit aan de hand van bovenstaand voorbeeld te illustreren: stel dat we voorspelden dat onze zeer prosociale respondent zou samenwerken in het 'Prisoner's Dilemma'. Ze besloot echter niet mee te werken. We kunnen dit niet-coöperatieve gedrag van onze respondent verklaren door haar verwachtingen achteraf willekeurig aan te passen; bijvoorbeeld: deze persoon werkte niet mee omdat ze verwachtte dat de meeste anderen zelfzuchtig waren en niet zouden coöpereren.[7] Om deze misvatting te vermijden zou men voorzichtig moeten zijn met de introductie van heterogene 'priors'. We bevelen niet aan om verwachtingen achteraf aan te passen, maar om de relatie tussen eigen preferenties en verwachtingen ex ante in het model op te nemen en het model zorgvuldig te toetsen waar het de verwachtingen betreft.

[6] Wanneer we het Bayesiaans-Nash evenwicht berekenen, kunnen we alle hogere-orde verwachtingen van deze specifieke persoon afleiden uit de subjectieve 'prior'-verdeling van deze persoon.

[7] In sommige gevallen hangt evenwichtsgedrag ('equilibrium behavior') niet af van verwachtingen, zoals we laten zien in hoofdstuk 5. In sommige PDs, zoals parallelle PD's, kan het zijn dat het coöperatieve gedrag van een pro-sociale respondent niet afhangt van haar verwachtingen over sociale preferenties van de ander.

Zoals we uitvoerig bespreken in de hoofdstukken 4 en 5 zijn in sommige situaties verwachtingen over sociale preferenties van de ander even belangrijk als sociale preferenties ter verklaring van coöperatief gedrag. In de literatuur over speltheorie ('behavioral game theory') lag de focus totdusver op het laatste. Er is in de literatuur een groot aantal modellen voor sociale preferenties voorgesteld (zie Fehr & Schmidt, 2006). Maar verwachtingen zijn ondanks het belang ervan niet in dezelfde mate onderzocht. Wij adviseren om expliciet te benoemen welk model voor verwachtingen gebruikt wordt naast het model voor sociale preferenties. We gebruiken liever de term 'sociale preferenties-verwachtingen-model' als verklaring van gedrag, dan alleen 'sociale-preferentiemodel'. We prefereren bijvoorbeeld "...in dit onderzoek verklaren we gedrag van onze respondenten in een serie 'Trust Games' met een ongelijkheidsaversie–rationele-verwachtingenmodel" boven "...in dit onderzoek verklaren we gedrag van onze respondenten in een serie 'Trust Games' met een ongelijkheidsaversiemodel".

Toekomstig onderzoek

Verklaring van variatie in sociale preferenties

Onderzoek naar de verklaring van het bestaan van en de variatie in sociale preferenties is een voor de hand liggende vervolgstap. We behandelen sociale preferenties gedeeltelijk als individuele eigenschappen ('traits'), een aspect van de psychische gesteldheid van een individu, en gedeeltelijk als toestanden ('states') beïnvloed door de context. Er is ook heterogeniteit in sociale preferenties als eigenschappen en toestanden. Hoe kunnen we het bestaan van en de variatie in sociale preferenties als eigenschappen en toestanden verklaren?

Er is een debat gaande over de vraag of prosociale preferenties tegenover vreemdelingen in een evolutionair proces kunnen ontstaan aangezien het zelf zich met prosociale preferenties kosten op de hals haalt ten gunste van genetisch niet-gerelateerde anderen. Diverse evolutionaire modellen laten zien dat zulke preferenties zich inderdaad kunnen ontwikkelen en dat heterogeniteit in preferenties onder bepaalde condities kan voortbestaan. Deze modellen omvatten diverse mechanismen zoals groepsselectie (e.g, Bowles & Gintis,

2004) of het gebruik van waarneembare kenmerken die een betrouwbaar signaal zijn voor sociale preferenties van anderen (e.g., Frank, 1988), genen-cultuur-interactie (e.g., Henrich & Henrich, 2007), sociale preferenties als bijproducten van vaardigheden in sociale contacten (De Waal, 2010), enzovoort. We verwachten dat verder onderzoek naar de evolutionaire oorsprong van sociale preferenties een theoretische basis kan leveren voor de aannames van het spel-theoretisch framework dat in dit proefschrift ontwikkeld is. Naast de evolutionaire oorsprong van sociale preferenties is het belangrijk om verder sociologisch onderzoek te doen naar de bijkomende determinanten van sociale preferenties. In dit proefschrift onderzoeken we al één factor, namelijk iemands cultuur, in hoofdstuk 7, en we laten zien dat sociale preferenties inderdaad variëren tussen sociale groepen. Dit wijst erop dat sociale preferenties vergelijkbaar zijn met geïnternaliseerde normen en beïnvloed worden door macro-sociologische condities. Een grootschalig vervolgonderzoek dat sociale preferenties terugvoert op macro-variabelen, zoals indicatoren van macro-economische condities en sociale instituties (e.g., Delhey & Newton, 2005; Buchan et al., 2009), micro-variabelen, zoals leeftijd, opleiding, religie, sekse (e.g., Gheorghiu et al., 2009; Bekkers & Wiepking, 2011) en socialisatie-variabelen (Gattig, 2002) zou helpen om een sociologische verklaring te ontwikkelen voor heterogene sociale preferenties.

We toonden ook aan dat contextuele factoren sociale preferenties kunnen beïnvloeden (zie β_j in (1)). Bovendien is er interactie tussen contextuele en dispositionele factoren (zie γ_{ij} in (1)). We hebben evenwel niet systematisch geanalyseerd welk kenmerk van de context precies invloed heeft op welk type sociale preferenties en wat de aard is van de interactie tussen context en dispositie. Een systematisch onderzoek naar sociale preferenties in een reeks paradigmatische spellen, waaronder verschillende 2×2 games, 'Trust Games', 'Dictator Games', 'Ultimatum Games', etc. met gebruik van een 'within subject design' (e.g., Guyer & Rapoport, 1972; Blanco et al., 2009, 2011) kan bijdragen aan begrip van de aard van het effect van spelstructuur en relatiegeschiedenis op sociale preferenties als ook van de aard van de interactie tussen contextuele en dispositionele determinanten van sociale preferenties.

Andere spellen

In dit proefschrift analyseerden we alleen niet-ingebedde interacties. Bovendien kregen proefpersonen in onze experimenten geen feedback over de keuzes van hun interactiepartners of van andere actoren.[8] Dus onze proefpersonen hadden niet veel informatie om te leren en hun verwachtingen aan te passen. Er zijn verschillende manieren om ons werk uit te breiden naar meer dynamische situaties.

Één uitbreiding betreft de analyse van herhaalde spellen of sequentiële spellen met verschillende beslispunten. In zulke spellen zouden spelers, in elk geval theoretisch, hun verwachtingen over sociale preferenties van de ander aanpassen op basis van eerdere keuzes van de ander en er ook rekening mee houden dat het eigen gedrag van de actor effect heeft op verwachtingen van de ander. In de analyse van zulke gevallen kunnen sociale preferenties op basis van uitkomsten en sociale preferenties op basis van processen in één model gecombineerd worden. Op dit moment werken we bijvoorbeeld aan de volgende benadering van een model voor reciprociteit. Stel dat actoren effectieve sociale preferenties hebben. Deze effectieve sociale preferenties zijn preferenties die actoren gebruiken wanneer ze een besluit nemen. Effectieve preferenties zijn een functie van iemands 'echte' sociale preferenties en de verwachtingen over sociale preferenties van de ander. Bijvoorbeeld, een 'echt' coöperatieve actor, i.e., een actor die groot gewicht toekent aan de uitkomst van de ander, kan haar effectieve gewicht naar beneden bijstellen als zij verwacht dat de ander zelfzuchtig is, i.e., geen enkel 'echt' gewicht toekent aan de uitkomsten van de ander. Merk op dat dit model ook toegepast kan worden op niet-dynamische settingen, zoals 'Dictator Games' of het 'Prisoner's Dilemma'. Maar in dynamische spellen, mede afhankelijk van geobserveerd gedrag van de ander, veranderen verwachtingen van mensen over sociale preferenties van de ander. Daarom veranderen ook hun 'effectieve' sociale preferenties.

Een andere uitbreiding van ons werk kan gaan over het leren van sociale

[8] Behalve hoofdstuk 4 waar we het sequentiële PD analyseren. In het sequentiële PD neemt de tweede speler een beslissing afhankelijk van de hoe de beslissing van de eerste speler uitvalt. Behalve dat we naar deze conditionele beslissingen vragen, hebben we in hoofdstuk 4 geen feedback gegeven over de feitelijke keuzes van anderen voordat alle proefpersonen hun keuzes hadden voltooid.

preferenties van de ander. Literatuur over het leren van preferenties van de ander is te vinden bij Dawes (1989); Iedema & Poppe (1994b); Engelmann & Strobel (2000); Hyndman et al. (2012) en Danz et al. (2012). In al die studies kregen proefpersonen feedback over keuzes van de ander en konden proefpersonen hun verwachtingen updaten op basis van de feedback. In de meeste van de in dit proefschrift beschreven experimenten krijgen proefpersonen geen feedback over keuzes van de ander. Dus wat we meten in onze experimenten zijn de 'prior'-verwachtingen van de proefpersoon over preferenties van de ander. Hoewel we 'prior'-verwachtingen meten, levert ons onderzoek duidelijke voorspellingen over hoe actoren met verschillende sociale preferenties hun verwachtingen verschillend kunnen updaten. In de hoofdstukken 2 en 3 laten we bijvoorbeeld zien dat 'prior'-verwachtingen van zelfzuchtige actoren informatiever zijn—de variantie in hun verwachtingen over preferenties van de ander is kleiner—dan de 'prior'-verwachtingen van coöperatieve of competitieve actoren. Uit deze bevinding kan men de volgende hypothese afleiden: bij herhaaldelijke feedback zullen zelfzuchtige actoren hun verwachtingen langzamer bijstellen dan coöperatieve of competitieve actoren omdat hun 'prior' informatiever is. Maar als mensen hun verwachtingen op een rationele manier bijstellen, in lijn met de theorie van Bayes, zal het effect van dit verschil in 'prior'-verwachtingen afnemen naarmate de hoeveelheid nieuwe informatie—feedback over keuzes van de ander—toeneemt. Uiteindelijk zullen verschillende typen verwachtingen convergeren tot een gedeelde verwachting, die zal corresponderen met de werkelijke verdeling van preferenties en zal het Bayesiaans-Nash evenwicht prevaleren.[9] Mensen kunnen hun verwachtingen echter ook niet-rationeel bijstellen, e.g., ze kunnen informatie systematisch negeren (Iedema & Poppe, 1994b). In dat geval zullen hun verwachtingen zelf na regelmatige feedback niet convergeren tot de correcte en gedeelde verwachting.

In dit proefschrift laten we zien dat de rationele-verwachtingen-aanname die ten grondslag ligt aan het Bayesiaans-Nash evenwichtsbegrip duidelijk wordt weerlegd. Meer onderzoek is echter nodig om duidelijk te documenteren wat de *gedragsconsequenties* zijn van verkeerde aannames over verwach-

[9] Vooropgesteld dat 'prior'-verwachtingen van actoren aan regulariteitsaannamen voldoen, zoals het toekennen van kans 0 aan bepaalde typen preferenties in de populatie.

tingen van de actor ten aanzien van preferenties van anderen: onder welke condities leiden inaccurate aannames over verwachtingen tot inaccurate voorspellingen van gedrag? In de hoofdstukken 2 en 3 toetsen we de rationele-verwachtingen-aanname direct, zonder de gevolgen van het veronderstellen van rationele verwachtingen voor beslissingen van de actors te analyseren. In hoofdstuk 4 breiden we de analyse uit naar de beslissing van de eerste speler in het sequentiële 'Prisoner's Dilemma', die direct beïnvloed wordt door de verwachtingen van de eerste speler over preferenties van de tweede speler. Hoofdstuk 4 laat zien dat—hoewel rationele verwachtingen weerlegd worden en een model dat ego-centrische 'biases' in verwachtingen aanneemt een betere 'fit' geeft; de aanname van rationele verwachtingen levert geen dramatisch kwaliteitsverlies op van voorspellingen van de beslissing van de eerste speler in het sequentiële PD. Hoofdstuk 5 analyseert een simultaan gespeeld 'Prisoner's Dilemma', en weerlegt opnieuw de rationele-verwachtingen-aanname. Echter, hoofdstuk 5 focust op een speciaal type 'Prisoner's Dilemma', het parallelle 'Prisoner's Dilemma', waarbij de invloed van verwachtingen op gedrag beperkt is. Experimentele analyse van gevallen waarin verwachtingen het gedrag waarschijnlijk sterker beïnvloeden, kunnen bijdragen aan het documenteren van de gevolgen van verkeerde aannames over verwachtingen van de actor. Zulke gevallen zijn de beslissing van de 'proposer' in het 'Ultimatum Game' (Güth et al., 1982), de 'trustor's beslissing in het 'Trust Game' (e.g., Berg et al., 1995; Snijders, 1996), bijdragen in 'Public Good'-spellen met niet-lineaire productiefuncties (Erev & Rapoport, 1990), en niet-parallelle versies van het 'Prisoner's Dilemma' (Zie Aksoy & Weesie, 2013b).

N-personen-spellen

In dit proefschrift onderzoeken we alleen 2-persoonsinteracties. Zoals we bespreken in de hoofdstukken 2 en 3, zijn in het geval van twee personen veerschillende manieren om op uitkomsten gebaseerde sociale preferenties te modelleren mathematisch bijna equivalent, vooropgesteld dat het aantal onbekende parameters hetzelfde is. Bijvoorbeeld, het ongelijksheid-aversie-model Fehr & Schmidt (1999) met gewichten voor, vanuit ego-perspectief, voordelige en onvoordelige ongelijkheid, is in wiskunding opzicht bijna equivalent met een

sociaal-orientatie model met gewichten voor de uitkomst van de ander en voor het absolute verschil tussen ego and alter.[10] Dit heeft een keerzijde: de keuze van een specifieke nutsfunctie beïnvloedt de fit niet substantieel. Wanneer men echter verschillende soorten sociale preferenties tegelijk wil onderzoeken moet men spellen overwegen met meer dan twee spelers. Daarnaast zullen gevallen met N personen ons helpen begrijpen hoeveel mensen geven om de verdeling van uitkomsten tussen anderen, niet alleen in verhouding tot henzelf, maar in verhouding tot elkaar. 'Dictator Games' voor N personen zijn een goed startpunt om sociale preferenties te bestuderen in spellen met meer dan twee spelers. Engelmann & Strobel (2007) bieden een overzicht van experimenten met 'Dictator Games' voor N personen en concluderen dat er niet genoeg dictatorexperimenten zijn met meer dan twee spelers (Engelmann & Strobel, 2007, pp. 290). Studies die 'Dictator Games' voor N personen experimenteel analyseren zijn Stahl & Haruvy (2006); Engelmann & Strobel (2004); Charness & Rabin (2002) en er is beslist ruimte voor meer onderzoek naar dit onderwerp.

[10] De modellen zijn niet geheel equivalent door de stochastische term in de nutsvergelijking.

References

Aksoy, O., & Weesie, J. (2009). Inequality and procedural justice in social dilemmas. *Journal of Mathematical Sociology*, *33*(4), 303–322.

Aksoy, O., & Weesie, J. (2012a). Altruisme en ongelijkheidsaversie in intra- en intergroupinteracties [Altruism and inequality aversion in intra- and intergroup interactions]. In V. Buskens, & I. Maas (Eds.) *Samenwerking in sociale dilemma's: voorbeelden van Nederlands onderzoek. Boekaflevering Mens en Maatschappij 2012*. Amsterdam: Amsterdam University Press.

Aksoy, O., & Weesie, J. (2012b). Beliefs about the social orientations of others: A parametric test of the triangle, false consensus, and cone hypotheses. *Journal of Experimental Social Psychology*, *48*(1), 45–54.

Aksoy, O., & Weesie, J. (2012c). Simultaneous analysis of ones own social orientation and ones beliefs about the social orientations of others using Mplus: a supplementary reseach note to Aksoy & Weesie (2012b).

Aksoy, O., & Weesie, J. (2013a). Hierarchical Bayesian analysis of biased beliefs and distributional social preferences. *Games*, *4*(1), 66–88.

Aksoy, O., & Weesie, J. (2013b). Social motives and expectations in one-shot asymmetric Prisoner's Dilemmas. *Journal of Mathematical Sociology*, *37*(1), 24–58.

Al-Ubaydli, O., Lee, M. S., Gneezy, U., & List, J. A. (2010). Towards an understanding of the relative strengths of positive and negative reciprocity. *Judgment and Decision Making*, *5*(7), 524–539.

Allport, G. W. (1954). *The nature of prejudice*. Cambridge, MA: Perseus Books.

Andersen, E. B. (1980). *Discrete statistical models with social science applications*. Amsterdam: North Holland.

Anderson, L. R., DiTraglia, F. J., & Gerlach, J. R. (2011). Measuring altruism in a public goods experiment: a comparison of U.S. and Czech subjects. *Experimental Economics*, *14*(3), 426–437.

Anderson, S. P., Goeree, J. K., & Holt, C. A. (1998). A theoretical analysis of altruism and decision error in public goods games. *Journal of Public Economics*, *70*, 297–323.

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

Asch, S. (1951). Effects of group pressure on the modification and distortion of judgements. In H. Guetzkow (Ed.) *Groups, leadership, and men*, (pp. 177–190). Pittsburgh, PA: Carnegie Press.

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.

Bacharach, M., & Gambetta, D. (2001). Trust in signs. In K. Cook (Ed.) *Trust in society*, vol. 2, (pp. 148–184). New York: Russell Sage.

Bahry, D., Kosolapov, M., Kozyreva, P., & Wilson, R. K. (2005). Ethnicity and trust: evidence from Russia. *American Political Science Review*, *99*(4), 521–532.

Balliet, D. (2010). Communication and cooperation in social dilemmas: A meta-analytic review. *Journal of Conflict Resolution*, *54*(1), 39–57.

Banfield, E. C. (1958). *The moral basis of a backward society*. New York: Free Press.

Becker, G. S. (1993). *A treatise on the family*. Cambridge, MA: Harvard University Press.

Bekkers, R., & Wiepking, P. (2011). Who gives? A literature review of predictors of charitable giving part one: Religion, education, age and socialisation. *Voluntary Sector Review*, *2*(3), 337–365.

Bellemare, C., Kröger, S., & Van Soest, A. (2008). Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. *Econometrica*, *76*(4), 815–839.

Benz, M., & Meier, S. (2008). Do people behave in experiments as in the field? evidence from donations. *Experimental Economics*, *11*(3), 268–281.

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

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

Blanco, M., Engelmann, D., Koch, A. K., & Normann, H.-T. (2008). Belief elicitation in experiments: Is there a hedging problem? IZA Discussion Papers 3517, Institute for the Study of Labor (IZA).
URL `http://ideas.repec.org/p/iza/izadps/dp3517.html`

Blanco, M., Engelmann, D., Koch, A. K., & Normann, H.-T. (2009). Preferences and beliefs in a sequential social dilemma: a within-subject analysis. IZA Discussion Papers 4624, Institute for the Study of Labor (IZA).
URL `http://ideas.repec.org/p/iza/izadps/dp4624.html`

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

Blau, P. M. (1964). *Exchange and power in social life*. New York: Wiley.

Blau, P. M., & Schwartz, J. E. (1984). *Crosscutting social circles. Testing a macrosutructural theory of intergroup relations*. New York: Academic Press.

Bogaert, S., Boone, C., & Declerck, C. (2008). Social value orientation and cooperation in social dilemmas: a review and conceptual model. *British Journal of Social Psychology*, *47*(3), 453–480.

Bolton, G. E., & Ockenfels, A. (2000). ERC: a theory of equity, reciprocity, and competition. *American Economic Review*, *100*, 166–193.

Bornhorst, F., Inchino, A., Kirchkamp, O., Schlag, K. H., & Winter, E. (2010). Similarities and differences when building trust: the role of cultures. *Experimental Economics*, *13*, 260–283.

Bouckaert, J., & Dhaene, G. (2004). Inter-ethnic trust and reciprocity: results of an experiment with small businessmen. *European Journal of Political Economy*, *20*, 869–886.

Bowles, S., & Gintis, H. (2004). The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology*, *65*(1), 17–28.

Buchan, N. R., Grimalda, G., Wilson, R., Brewer, M., Fatas, E., & Foddy, M. (2009). Globalization and human cooperation. *Proceedings of the National Academy of Sciences*, *106*(11), 4138–4142.

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

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

Burton-Chellew, M., & West, S. (2013). Prosocial preferences do not explain human cooperation in public-goods games. *Proceedings of the National Academy of Sciences*, *110*(1), 216–221.

Buskens, V., & Raub, W. (2002). Embedded trust: control and learning. *Advances in Group Processes*, *19*, 167–2002.

Buskens, V., & Raub, W. (2013). Rational choice research on social dilemmas. In R. Wittek, T. Snijders, & V. Nee (Eds.) *Handbook of rational choice social research*, (pp. 113–150). New York: Russell Sage.

Camerer, C. (2003). *Behavioral game theory: experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

Caplan, B. (2003). Stigler–Becker versus Myers–Briggs: why preference-based explanations are scientifically meaningful and empirically important. *Journal of Economic Behavior and Organization*, *50*(4), 391–405.

Carvalho, J.-P. (2007). An interview with Thomas Schelling. *Oxonomics*, *2*(2), 1–8.

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

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics*, *14*, 47–83.

Chuah, S.-H., Hoffmann, R., Jones, M., & Williams, G. (2007). Do cultures clash? Evidence from cross-national ultimatum game experiments. *Journal of Economic Behavior and Organization*, *64*, 35–48.

Coleman, J. (1990). *Foundations of social theory*. Cambridge, MA: Belknap Press.

Coleman, J. S. (1981). *Longitudinal data analysis*. New York: Basic Books.

Costa-Gomes, M. A., & Weizsäcker, G. (2008). Stated beliefs and play in normal-form games. *Review of Economic Studies*, *75*(3), 729–762.

Crawford, S. E. S., & Ostrom, E. (1995). A grammar of institutions. *Political Science Review*, *89*(3), 582–600.

Croson, R. T. A. (2000). Thinking like a game theorist: factors affecting the frequency of equilibrium play. *Journal of Economic Behavior and Organization*, *41*, 299–314.

Croson, R. T. A. (2007). Theories of commitment, altruism and reciprocity: evidence from linear public good games. *Economic Inquiry*, *45*(2), 199–216.

Danz, D. N., Fehr, D., & Kuebler, D. (2012). Information and beliefs in a repeated normal-form game. *Experimental Economics*, *15*(4), 622–640.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology*, *31*, 169–193.

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology*, *25*, 1–17.

De Cremer, D., & Van Vugt, M. (1999). Social identification effects in social dilemmas: a transformation of motives. *European Journal of Social Psychology*, *29*, 871–893.

De Jasay, A., Güth, W., Kliemt, H., & Ockenfels, A. (2004). Take or leave? Distribution in asymmetric one-off conflict. *Kyklos*, *57*(2), 217–235.

De Waal, F. (2010). *The age of empathy: nature's lessons for a kinder society*. New York: Three Rivers Press.

Delhey, J., & Newton, K. (2005). Predicting cross-national levels of social trust: global pattern or Nordic exceptionalism. *European Sociological Review*, *12*(4), 311–327.

Dufwenberg, M., & Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior*, *30*, 163–182.

Eagly, A. H., & Crowley, M. (1986). Gender and helping behavior: a meta-analytic review of the social psychological literature. *Psychological Bulletin*, *100*(3), 283–308.

Eek, D., Biel, A., & Gärling, T. (1998). The effect of distributive justice on willingness to pay for municipality child care: An extension of the GEF hypothesis. *Social Justice Research*, *11*(2), 121–142.

Elberfeld, W. (1997). Incentive monotonicity and equilibrium selection in 2-by-2 games. *Journal of Economics*, *3*, 279–290.

Ellingsen, T., Johannesson, M., Tjotta, S., & Torsvik, G. (2010). Testing guilt aversion. *Games and Economic Behavior*, *68*, 95–107.

Elster, J. (1989). *The cement of society: a survey of social order*. Cambridge: Cambridge University Press.

Engelmann, D., & Strobel, M. (2000). The false consensus effect disappears if representative information and monetary incentives are given. *Experimental Economics*, *3*, 241–269.

Engelmann, D., & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review*, *94*(4), 857–869.

Engelmann, D., & Strobel, M. (2007). Preferences over income distributions: experimental evidence. *Public Finance Review*, *35*(2), 285–310.

Erev, I., & Rapoport, A. (1990). Provision of step-level public goods the sequential contribution mechanism. *Journal of Conflict Resolution*, *34*(3), 401–425.

Erlei, M. (2008). Heterogeneous social preferences. *Journal of Economic Behavior and Organization*, *65*(3), 436–457.

Ermisch, J., & Gambetta, D. (2010). Do strong family ties inhibit trust? *Journal of Economic Behavior and Organization*, *75*(3), 365–376.

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

Falk, A., & Zhender, C. (2007). Discrimination and in-group favoritism in a citywide trust experiment.
URL `http://ideas.repec.org/p/zur/iewwpx/318.html`

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, *13*(1), 1–25.

Fehr, E., & Gintis, H. (2007). Human motivation and social cooperation: experimental and analytical foundations. *Annual Review of Sociology*, *33*, 43–64.

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

Fehr, E., & Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism—experimental evidence and new theories. In S. C. Kolm, & J. M. Ythier (Eds.) *Handbook of the economics of giving, altruism and reciprocity*, vol. 1. Amsterdam: North Holland.

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

Fischbacher, U., & Gaechter, S. (2010). Social preferences, beliefs and the dynamics of free riding in public goods experiments. *American Economic Review*, *100*(1), 541–556.

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: foundations, recent developments, and applications*. New York: Springer.

Fox, J.-P. (2010). *Bayesian item response modeling: theory and applications*. New York: Springer.

Frank, R. (1988). *Passions within reason: the strategic role of the emotions*. New York: W. W. Norton.

Friedman, D., & Cassar, A. (2004). *Economics lab*. London: Routledge.

Friedman, D., & Massaro, D. W. (1998). Understanding variability in binary and continuous choice. *Psychonomic Bulletin and Review*, *5*, 370–389.

Friedman, M. (1953). The methodology of positive economics. In M. Friedman (Ed.) *Essays in positive economics*, (pp. 3–43). Chicago: University of Chicago Press.

Fudenberg, D., & Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, *54*(3), 533–554.

Gaechter, S., Hermann, B., & Thoeni, C. (2010). Culture and cooperation. *Philosophical Transactions of the Royal Society B*, *365*(1553), 2651–2661.

Gaechter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B*, *364*, 791–806.

Gaechter, S., & Renner, E. (2010). The effects of (incentivized) belief elicitation in public good experiments. *Experimental Economics*, *13*(3), 364–377.

Gambetta, D. (2009). *Codes of the underworld: how criminals communicate*. Princeton, NJ: Princeton University Press.

Ganzeboom, H. B., Treiman, D. J., & Ultee, W. C. (1991). Comparative intergenerational stratification research: three generations and beyond. *Annual Review of Sociology*, *17*, 277–302.

Gattig, A. L. W. (2002). *Intertemporal decision making: studies on the working of myopia*. Groningen: ICS Doctoral Dissertation.

Gautschi, T. (2000). History effects in social dilemma situations. *Rationality and Society*, *12*, 131–162.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis, second edition*. Boca Raton, FL: Chapman and Hall, 2nd ed.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel and hierarchical models*. New York: Cambridge University Press.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–807.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457–472.

Gheorghiu, M. A., Vignoles, V. L., & Smith, P. B. (2009). Beyond the United States and Japan: testing Yamagishi's emancipation theory of trust across 31 nations. *Social Psychology Quarterly*, *72*(4), 365–383.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Gil-White, F. J. (2003). Ultimatum game with an ethnicity manipulation. Results from Khovdiin Bulgan Sum, Mongolia. In J. Henrich, R. Boyd, S. Bowles, & H. Gintis (Eds.) *Foundations of human sociality: ethnography and experiments in 15 small-scale societies*. Oxford University Press.

Gintis, H., Henrich, J., Bowles, S., Boyd, R., & Fehr, E. (2008). Strong reciprocity and the roots of human morality. *Social Justice Research*, *21*, 241–253.

Goethals, G. R., Messick, D. M., & Allison, S. T. (1991). The uniqueness bias: studies of constructive social comparison. In J. Suls, & T. A. Wills (Eds.) *Social comparison: contemporary theory and research*, (pp. 149–176). Hillsdale, NJ: Lawrence Erlbaum.

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

Goette, L., Huffman, D., & Meier, S. (2012). The impact of social ties on group interactions: Evidence from minimal groups and randomly assigned real groups. *American Economic Journal: Microeconomics*, *4*(1), 101–115.

Greenberg, J. (1987). A taxonomy of organizational justice theories. *Academy of Management Review*, *12*(1), 9–22.

Greiner, B. (2004). The online recruitment system ORSEE 2.0 a guide for the organization of experiments in economics. *Working Paper Series in Economics*, *10*, 1–15. Mimeo, University of Cologne.

Griesinger, D. W., & Livingston, J. J. (1977). Toward a model of interpersonal orientation in experimental games. *Behavioral Science*, *18*, 173–188.

Grzelak, J. L., Iwinski, T. B., & Radzicki, J. J. (1977). Motivational components of utility. In H. Jungermann, & G. de Zeeuw (Eds.) *Decision making and change in human affairs*. Dordrecht: Reidel.

Güth, W., Levati, M. V., & Ploner, M. (2008). Social identity and trust: an experimental investigation. *Journal of Socio-Economics*, *34*, 1293–1308.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, *3*(4), 367–388.

Guyer, M. J., & Rapoport, A. (1972). 2×2 games played once. *Journal of Conflict Resolution*, *16*(3), 409–431.

Habyarimana, J., Humphreys, M., Posner, D., & Weinstein, J. M. (2007). Why does ethnic diversity undermine public goods provision? *American Political Science Review*, *101*(4), 709–725.

Hagendoorn, L., Drogendijk, R., Tumanov, S., & Hraba, J. (1998). Inter-ethnic preferences and ethnic hierarchies in the former Soviet Union. *International Journal of Intercultural Relations*, *11*(3), 372–492.

Hamamura, T. (2012). Social class predicts generalized trust but only in wealthy societies. *Journal of Cross-Cultural Psychology*, *43*(3), 498–509.

Harsanyi, J. (1980). Rule utilitarianism, rights, obligations and the theory of rational behavior. *Theory and Decision*, *12*, 115–133.

Harsanyi, J. C. (1968). Games with incomplete information played by Bayesian players. *Management Science*, *14*, 468–502.

Harsanyi, J. C. (1977). *Rational behavior and bargaining equilibrium in games and social situations*. Cambridge: Cambridge University Press.

Harsanyi, J. C., & Selten, R. (1988). *A general theory of equilibrium selection*. Cambridge, MA: MIT Press.

Heap, S. P. H., & Patrick, S. (2010). Out-group favoritism. In D. Fetchenhauer, J. Pradel, & E. Hoelzl (Eds.) *A boat trip through economic change. Proceedings of the IAREP/SABE/ICABEEP 2010 Conference Cologne*, (pp. 40–41). Lengerich: PABST Science Publishers.

Heckathorn, D. D. (1990). Collective sanctions and compliance norms: a formal theory of group-mediated social control. *American Sociological Review*, *55*(3), 366–384.

Heckathorn, D. D. (1996). The dynamics and dilemmas of collective action. *American Sociological Review*, *61*(2), 250–277.

Hedström, P. (2005). *Dissecting the social: on the principles of analytical sociology*. Cambridge: Cambridge University Press.

Hegselmann, R. (1989). Wozu könnte Moral gut sein? oder: Kant, das Gefangenendilemma und die Klugheit contexts. *Grazer Philosophische Studien*, *31*, 1–28.

Henrich, J., & Henrich, N. (2007). *Why humans cooperate: a cultural and evolutionary explanation*. New York: Oxford University Press.

Hermann, B., Thoeni, C., & Gaechter, S. (2008). Antisocial punishment across societies. *Science*, *319*, 1362–1367.

Hertel, G., & Fiedler, K. (1994). Affective and cognitive influences in social dilemma game. *European Journal of Social Psychology*, *24*(1), 131–145.

Hertel, G., & Fiedler, K. (1998). Fair and dependent versus egoistic and free: effects of semantic and evaluative priming on the ring measure of social values. *European Journal of Social Psychology*, *28*(1), 49–70.

Hoffman, E., McCabe, K., Shachat, K., & Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, *7*(3), 346–380.

Hoijtink, H., Klugkist, I., & Boelen, P. A. (2008). *Bayesian evaluation of informative hypotheses*. New York: Springer.

Hong, K., & Bohnet, I. (2007). Status and distrust: the relevance of inequality and betrayal aversion. *Journal of Economic Psychology*, *28*, 197–213.

Howard, N. (1966). The theory of meta-games. *General Systems*, *11*(Part V), 167–86.

Hyndman, K., Oezbay, E. Y., Schotter, A., & Ehrblatt, W. (2012). Belief formation: an experiment with outside observers. *Experimental Economics*, *15*, 176–203.

Iedema, J. (1993). *The perceived consensus of one's social value orientation*. Tilburg University: Doctoral Dissertation.

Iedema, J., & Poppe, M. (1994a). Causal attribution and self-justification as explanations for the consensus expectations of one's social value orientation. *European Journal of Personality*, *8*, 395–408.

Iedema, J., & Poppe, M. (1994b). Effects of social value orientation on expecting and learning others' orientations. *European Journal of Social Psychology*, *24*(5), 565–579.

Iedema, J., & Poppe, M. (1999). Expectations of others' social value orientations in specific and general populations. *Personality and Social Psychology Bulletin*, *25*, 1443–1450.

Kass, R. E., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kelley, H. H., & Stahelski, A. J. (1970). Social interaction basis of cooperators' and competitors' beliefs about others. *Journal of Personality and Social Psychology*, *16*(1), 66–91.

Kelley, H. H., & Thibaut, J. W. (1978). *Interpersonal relations: a theory of interdependence*. New York: Wiley.

Klugkist, I., Laudy, O., & Hoijtink, H. (2010). Bayesian evaluation of inequality and equality constrained hypotheses for contingency tables. *Psychological Methods*, *15*(3), 281–299.

Kohler, S. (2012). Incomplete information about social preferences explains equal division and delay in bargaining. *Games*, *3*(3), 119–137.

Kollock, P. (1998). Social dilemmas: the anatomy of cooperation. *Annual Review of Sociology*, *24*, 183–214.

Kuhlman, D. M., Brown, C., & Teta, P. (1992). Judgements of cooperation and defection in social dilemmas: the moderating role of judge's social orientation. In W. Liebrand, , D. M. Messick, & H. A. M. Wilke (Eds.) *Social dilemmas, theoretical issues, and research findings*. Oxford: Pergamon.

Kuhlman, D. M., Camac, C. R., & Cunha, D. A. (1986). Individual differences in social orientation. In H. A. M. Wilke, & D. M. Messick (Eds.) *Experimental social dilemmas*. New York: Lang.

Kuhlman, D. M., & Wimberley, D. L. (1976). Expectations of choice behavior held by cooperators, competitors, and individualists across four classes of experimental game. *Journal of Personality and Social Psychology*, *34*(1), 69–81.

Kunreuther, H., Silvasi, G., Bradlow, E. T., & Small, D. (2009). Bayesian analysis of deterministic and stochastic prisoner's dilemma games. *Statistical Science*, *4*(5), 363–384.

Kuwabara, K. (2005). Nothing to fear but fear itself: Fear of fear, fear of greed and gender effects in two-person asymmetric social dilemmas. *Social Forces*, *84*(2), 1257–1272.

Kuwabara, K., Willer, R., Macy, M. W., Mashima, R., Terai, S., & Yamagishi, T. (2007). Culture identity, and structure in social exchange: a web based trust experiment in the United States and Japan. *Social Psychology Quarterly*, *70*(4), 461–479.

Laudy, O., & Hoijtink, H. (2007). Bayesian methods for the analysis of inequality constrained contingency tables. *Statistical Methods in Medical Research*, *16*, 123–138.

Laury, S. (2005). Pay one or pay all: Random selection of one choice for payment. *Andrew Young School of Policy Studies Research Paper Series*, (06-13).

Leventhal, G. S. (1970). Industrialization and social stratification. In D. J. Treiman (Ed.) *Social stratification: research and theory for the 1970s*, (pp. 207–234). Indianapolis: Bobbs-Merrill.

Leventhal, G. S. (1976). Fairness in social relationship. In J. W. Thibaut, J. T. Spence, & R. C. Carson (Eds.) *Contemporary topics in social psychology*, (pp. 211–239). Morristown, NJ: General Learning Press.

Leventhal, G. S. (1980). What should be done with equity theory? In K. J. Gergen, M. S. Greenberg, & R. W. Willis (Eds.) *Social exchange: advances in theory and research*, (pp. 27–55). New York: Plenum.

Lewin, K. (1936). *Principles of topological psychology*. New York: McGraw-Hill.

Liberman, V., Samuels, S. M., & Ross, L. (2004). The name of the game: predictive power of reputations versus situational labels in determining prisoners dilemma game moves. *Personality and Social Pychology Bulletin*, *30*(9), 1175–1185.

Liebe, U., & Tutic, A. (2010). Status groups and altruistic behavior in dictator games. *Rationality and Society*, *22*(3), 353–380.

Liebrand, W., & McClintock, C. (1988). The ring measure of social values: a computerized procedure for assessing individual differences in information processing and social value orientation. *Europoean Journal of Personality*, *2*, 217–230.

Lindenberg, S. (2001). Intrinsic motivation in a new light. *Kyklos*, *54*(2-3), 317–342.

Lindenberg, S. (2008). Social rationality, semi-modularity and goal-framing: What is it all about? *Analyse & Kritik*, *30*(2), 669–687.

Lindenberg, S., & Steg, L. (2007). Normative, gain and hedonic goal frames guiding environmental behavior. *Journal of Social Issues*, *63*(1), 117–137.

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, *92*(368), 805–824.

Lunn, D., Jackson, C., Best, N., Spiegelhalter, D. J., & Thomas, A. (2013). *The BUGS Book: a Practical Introduction to Bayesian Analysis*. New York: CRC.

Lunn, D., Spiegelhalter, D. J., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions (with discussion). *Statistics in Medicine*, *28*, 3049–3082.

Macy, M. M., & Skvoretz, J. (1998). The evolution of trust and cooperation between strangers: A computational model. *American Sociological Review*, *63*(5), 638–660.

Mandel, R. (1990). Shifting centres and emergent identities: Turkey and Germany in the lives of Turkish Gastarbeiter. In D. F. Eickelman, & J. Piscatori (Eds.) *Muslim Travellers. Pilgrimage, migration, and the religious imagination*, (pp. 153–171). London: Routlege.

Marks, G., & Miller, N. (1987). Ten years of research on the false-consensus effect: an empirical and theoretical review. *Psychological Bulletin*, *102*(1), 72–90.

McCabe, K., Rigdon, M., & Smith, V. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization*, *52*(2), 267–275.

McClintock, C. G. (1972). Social motivation—a set of propositions. *Behavioral Science*, *31*, 1–28.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.) *Frontiers in econometrics*, (pp. 105–142). New York: Academic Press.

McKelvey, R., & Palfrey, T. (1992). An experimental study of the centipede game. *Econometrica*, *60*(4), 803–836.

McKelvey, R. D., & Palfrey, T. R. (1995). Quantal Response Equilibria for normal form games. *Games and Economic Behavior*, *10*, 6–38.

Messe, L. A., & Sivacek, J. M. (1979). Predictions of others' responses in a mixed-motive game: self-justification of false consensus? *Journal of Personality and Social Psychology*, *37*(4), 602–607.

Miller, I., & Miller, M. (2004). *John E. Freund's mathematical statistics with applications*. Upper Saddle River, NJ: Pearson Prentice Hall.

Morishima, Y., Schunk, D., Bruhin, A., Ruff, C. C., & Fehr, E. (2012). Linking brain structure and activation in the temporoparietal junction to explain the neurobiology of human altruism. *Neuron*, *75*(1), 73–79.

Morris, S. (1995). The common prior assumption in economic theory. *Economics and Philosophy*, *11*, 227–227.

Mulder, J. (2010). *Bayesian model selection for constrained multivariate normal linear models*. Utrecht: Doctoral Dissertation.

Muthén, L., & Muthén, B. (1998–2010). *Mplus user's guide*. Los Angeles, CA: Muthén and Muthén., 6th ed.

Neugebauer, T., Perote, J., Schmidt, U., & Loos, M. (2009). Selfish-biased conditional cooperation: on the decline of contributions in repeated public goods experiments. *Journal of Economic Psychology*, *30*, 52–60.

Nikiforakis, N., & Normann, H.-T. (2008). A comparative statics analysis of punishment in public-good experiments. *Experimental Economics*, *11*(4), 358–369.

Nilsson, H., Rieskamp, J., & Wagenmakers, E.-J. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*, *55*, 84–93.

Oechssler, J. (Forthcoming). Finitely repeated games with social preferences. *Experimental Economics*, (pp. 1–10).

Offerman, T., Sonnemans, J., & Schram, A. (1996). Value orientations, expectations and voluntary contributions in public goods. *Economic Journal*, *106*, 817–845.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., & Rakow, T. (2006). *Uncertain judgements: eliciting experts' probabilities*. New York: Wiley.

Olson, M. (1965). *The logic of collective action: public goods and the theory of groups.*. Cambridge: Harvard University Press.

Orbell, J. M., & Dawes, R. M. (1993). Social welfare, cooperators' advantage and the option of not playing the game. *American Sociological Review*, *58*(6), 787–800.

Osterberk, H., Sloof, R., & van de Kuilen, G. (2004). Cultural differences in ultimatum game experiments: evidence from a meta analysis. *Experimental Economics*, *7*(2), 171–188.

Palfrey, T. R., & Prisbrey, J. E. (1997). Anomalous behavior in public good experiments: how much and why? *American Economic Review*, *87*(5), 829–846.

Palfrey, T. R., & Wang, S. W. (2009). On eliciting beliefs in strategic games. *Journal of Economic Behavior and Organization*, *71*(2), 98–109.

Putnam, R. (2001). *Bowling alone: the collapse and revival of American community*. New York: Simon and Schuster.

Putnam, R. D. (1993). *Making democracy work: civic traditions in modern Italy*. Princeton, NJ: Princeton University Press.

Quattrone, G. A., & Tversky, A. (1986). Self-deception and the voter's illusion. In J. Elster (Ed.) *The multiple self*, (pp. 35–58). Cambridge: Cambridge University Press.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata Journal*, *2*(1), 1–21.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, *83*(5), 1281–1301.

Raiffa, H. (1970). *Decision analysis: introductory lectures on choices under uncertainty: (2. Printing)*. Oxford: Addison-Wesley.

Rapoport, A. (1988). Provision of step-level public goods: effects of inequality in resources. *Journal of Personality and Social Psychology*, *54*(3), 432–440.

Rapoport, A., Bornstein, G., & Erev, I. (1989). Intergroup competition for public goods: effects of unequal resources and relative group size. *Journal of Personality and Social Psychology*, *56*(5), 748–756.

Raub, W., Buskens, V., & Corten, R. (2014). Social dilemmas and cooperation. In N. Braun, & N. J. Saam (Eds.) *Handbuch Modellbildung und Simulation in den Sozialwissenschaften*. Wiesbaden: Springer.

Raub, W., & Snijders, C. (1997). Gains, losses, and cooperation in social dilemmas and collective action: The effects of risk preferences. *Journal of Mathematical Sociology*, *22*(3), 263–302.

Raub, W., & Weesie, J. (1990). Reputation and efficiency in social interactions: An example of network effects. *American Journal of Sociology*, *96*(3), 626–654.

Rey-Biel, P. (2009). Equilibrium play and best response to (stated) beliefs in normal form games. *Games and Economic Behavior*, *65*(2), 572–585.

Rodriguez-Lara, I., & Moreno-Garrido, L. (2012). Modeling inequity aversion in a Dictator Game with production. *Games*, *3*(4), 138–149.

Ross, L., Greene, D., & House, P. (1977). The false consensus effect: an egocentric bias in social perception and attribution processes. *Journal of Personality and Social Psychology*, *13*, 279–301.

Sally, D. (1995). Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, *7*(1), 58–92.

Savage, L. J. (1951). The theory of statistical decision. *Journal of the American Statistical Association*, *46*(253), 55–67.

Schenkler, B. R., & Goldman, H. J. (1978). Cooperators and competitors in conflict: a test of the triangle model. *Journal of Conflict Resolution*, *22*(3), 393–410.

Schuessler, R. (1989). Exit threats and cooperation under anonymity. *Journal of Conflict Resolution*, *33*(4), 728–749.

Schulz, U., & May, T. (1989). The recording of social orientations with ranking and pair comparison procedures. *European Journal of Social Psychology*, *19*, 41–59.

Self, S. G., & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*, 605–610.

Selten, R. (1967). Die Strategiemethode zur Erforschung des Eingeschränkt Rationalen Verhaltens im Rahmen eines Oligopolexperiments. In H. Sauerman (Ed.) *Beiträge zur Experimentellen Wirtschaftsforschung*, (pp. 136–168). Tübingen: JCB Mohr.

Sherif, M., Harvey, O. J., White, B. J., Hood, W. R., & Sherif, C. W. (1961). *Intergroup conflict and cooperation: the Robber's Cave experiment*. Norman: University of Oklahoma Book Exchange.

Sidanius, J., & Pratto, F. (1999). *Social dominance*. Cambridge: Cambridge University Press.

Simpson, B. (2003). Sex, fear, and greed: a social dilemma analysis of gender and cooperation. *Social Forces*, *82*(1), 35–52.

Simpson, B. (2004). Social values, subjective transformations, and cooperation in social dilemmas. *Social Psychology Quartely*, *67*(4), 385–395.

Simpson, B. (2006). Social identity and cooperation in social dilemmas. *Rationality and Society*, *18*(4), 443–470.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, *439*(26), 466–469.

Smith, E. R., Jackson, J. W., & Sparks, C. W. (2003). Effects of inequality and reasons for inequality on group identification and cooperation in social dilemmas. *Group Processes and Intergroup Relations*, *6*(2), 201–220.

Sniderman, P. M., & Hagendoorn, L. (2007). *When ways of life collide*. Princeton, NJ: Princeton University Press.

Snijders, C. (1996). *Trust and commitments*. Amsterdam: Thela Thesis.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. London: Sage, 2nd ed.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, *64*(4), 583–639.

Stahl, D. O., & Haruvy, E. (2006). Other-regarding preferences: egalitarian warm glow, empathy, and group size. *Journal of Economic Behavior and Organization*, *61*(1), 20–41.

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*, *13*(5), 389–408.

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

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American*, *223*, 96–102.

Tajfel, H., & Billig, M. (1974). Familiarity and categorization in intergroup behavior. *Journal of Experimental Social Psychology*, *10*, 159–170.

Taylor, M. (1987). *The possibility of cooperation*. Cambridge: Cambridge University Press.

Thibaut, J. W., & Walker, L. (1975). *Procedural justice: a psychological analysis*. Hillsdale, NJ: Erlbaum.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286.

Tutic, A., & Liebe, U. (2009). A theory of status-mediated inequity aversion. *Journal of Mathematical Sociology*, *33*(3), 157–195.

Van der Heijden, E. C. M., Nelissen, J. H. M., Potters, J. J. M., & Verbon, H. A. A. (2001). Social dilemmas. *Economic Inquiry*, *39*(2), 280–297.

Van Dijk, E., & Wilke, H. (1995). Coordination rules in asymmetric social dilemmas: a comparison between public good dilemmas and resource dilemmas. *Journal of Experimental Social Psychology*, *31*(1), 1–27.

Van Lange, P. A. M. (1991). *The rationality and morality of cooperation*. Groningen: Doctoral Dissertation.

Van Lange, P. A. M. (1992). Confidence in expectations: a test of the triangle hypothesis. *European Journal of Personality*, *6*, 371–379.

Van Lange, P. A. M. (1999). The pursuit of joint outcomes and equaility of outcomes: an integrative model of social value orientation. *Journal of Personality and Social Psychology*, *73*, 337–349.

Van Lange, P. A. M. (2000). Beyond self-interest: a set of propositions relevant to interpersonal orientations. *European Review of Social Psychology*, *11*(1), 297–331.

Van Lange, P. A. M., Agnew, C. R., Harinck, F., & Steemers, G. E. (1997). From game theory to real life: how social value orientation affects willingness to sacrifice in ongoing close relationships. *Journal of Personality and Social Psychology*, *73*(6), 1330–1344.

Van Lange, P. A. M., Bekkers, R., Schuyt, T. N. M., & Vugt, M. V. (2007). From games to giving: social value orientation predicts donations to noble causes. *Basic and Applied Social Psychology*, *29*(4), 375–384.

Van Lange, P. A. M., & Liebrand, W. B. G. (1991). Social value orientation and intelligence: A test of the goal prescribes rationality principle. *European Journal of Social Psychology*, *21*, 273–292.

Veit, S., & Koopmans, R. (2010). Cooperation in the shadow of ethnic diversity and otherness. In D. Fetchenhauer, J. Pradel, & E. Hoelzl (Eds.) *A boat trip through economic change. Proceedings of the IAREP/SABE/ICABEEP 2010 conference Cologne*, (pp. 44–45). Lengerich: PABST Science Publishers.

Verkuyten, M., & Yildiz, A. A. (2006). The endorsement of minority rights: the role of group position, national context, and ideological beliefs. *Political Psychology*, *27*(4), 527–548.

Verkuyten, M., & Zaremba, K. (2005). Inter-ethnic relations in a changing political context. *Social Psychology Quarterly*, *68*, 375–386.

Vieth, M. (2009). *Commitments and reciprocity*. Utrecht: ICS Doctoral Dissertation.

Vogt, S. (2008). *Heterogeneity in social dilemmas: the case of social support*. Utrecht: ICS Doctoral Dissertation.

Weesie, J. (1994). Social orientations in symmetric 2×2 games. *ISCORE Discussion Paper Series*, *17*, 1–29.

Weesie, J., & Raub, W. (1996). Private ordering: a comparative institutional analysis of hostage games. *Journal of Mathematical Sociology*, *21*(3), 201–240.

White, M. D. (2004). Can homo economicus follow Kant's categorical imperative? *Journal of Socio-Economics*, *33*, 89–106.

Whitt, S., & Wilson, R. K. (2007). The Dictator Game, fairness and ethnicity in postwar Bosnia. *American Journal of Political Science*, *51*(3), 655–668.

Wilke, H. A. M. (1991). Greed, efficiency and fairness in resource management situations. *European Review of Social Psychology*, *2*(1), 165–187.

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Motivation and Emotion*, *18*, 129–166.

Acknowledgments

Many people have contributed to this thesis, directly or indirectly. The lion's share of my gratitude, however, has to go to my daily supervisor Jeroen Weesie. I have known Jeroen for eight years—since I started the Sociology and Social Research (SaSR) master's programme in Utrecht. I remember clearly the day we chatted for the first time to decide which statistics track I should follow. I also remember clearly that I, quite deservedly, received my lowest grade during my entire master's and PhD studies from Jeroen's first statistics assignment. Still, his presence was one of the biggest reasons for me to stay in Utrecht after SaSR. Over the last eight years, Jeroen and I designed projects and experiments, wrote proposals and papers, analyzed data, taught courses, and always had fun doing these things together. Obviously, I learned a lot from him. But no matter how hard I tried, my knowledge in statistics (and in quite some other topics) always remained a subset of Jeroen's. A quote attributed to Albert Einstein reads: "As our circle of knowledge expands, so does the circumference of darkness surrounding it". Yet, as my circle of knowledge expanded during the last eight years, so did the circumference of Jeroen's knowledge surrounding it. He will likely remain as the single person who has had the biggest influence in my academic intellect.

I also owe much to Werner Raub. First as the coordinator of SaSR, then as my promotor, Werner had a significant impact on my metamorphosis from a fresh business graduate to a sociologist. I wouldn't have come to the Netherlands in the first place, had I not seen Werner's e-mail advertising the SaSR programme in 2005. His extensive feedback on several chapters of the manuscript improved and helped us publish them. His input in those publica-

tions could easily make him a co-author, a role he declined when we explicitly offered. His feedback on the three grant proposals that we prepared during my master's and PhD studies not only helped me learn the art of writing a proposal but also helped us secure two of these three applications. He guided me to reach places that I could have never reached otherwise. His schedule was quite busy but he always appeared in crucial moments as in the quote "a wizard is never late, nor is he early, he arrives precisely when he means to."

Each empirical chapter includes a separate acknowledgments note listing the people who contributed to that chapter. I would like to recite those people once again and thank Vincent Buskens, Rense Corten, Manuela Vieth, Rene Torenvlied, Michael Maes, Rens van de Schoot, I. Ercan Alp and the Psychology Department of Bogazici University, Tobias Stark, Andreas Flache, Edwin Poppe, Ali Aslan Yildiz, Axel Ockenfels, Siegfried Berninghaus, and several anonymous referees from *Journal of Mathematical Sociology*, *Journal of Experimental Social Psychology*, and *Games*. I especially appreciate Vincent Buskens' help in programming in z-tree and conducting the experiments. Additionally, those experiments would be very difficult to carry out without the facilities of the ELSE lab which was established by Vincent and Stephanie Rosenkranz. I would like to thank the Netherlands Organization for Scientific Research (NWO) for financial support, which did not cease even after completing my PhD. I owe many thanks to Helma van der Mijl for her editorial help. Diego Gambetta was a great host and mentor during my visit at Nuffield College, University of Oxford. His scientific approach provided new insights and his guidance opened up new doors. I also thank him for his willingness to read the manuscript. I am grateful to Herbert Hoijtink, Vincent Buskens, and Stephanie Rosenkranz for their time to read the manuscript and for their comments.

I should also thank the administrative team at the Department of Sociology, especially Tineke Edink, Miranda Jansen, and Ellen Janssen who very efficiently helped me solve all of my odd problems, including finding the right person to consult for during my *several* visa applications.

I would like to thank my colleagues and friends who made life abroad easier, especially Dominik (who particularly "endured" our pool sessions),

Michal, Vincenz (my paranymph), Wiebke, Martijn, Esen, Baris (my other paranymph), Inan, and Ali Aslan. I am also grateful to my two brothers, parents, and my close friends in Turkey who stood by me during my prolonged stay abroad.

Finally, I thank my wife Gulcin for her love and unconditional support. She witnessed all of my ups and downs during the years I have been working on this thesis. She accepted my decision to study and work abroad, *twice*, each-time moving with me, risking (but hopefully not ruining) her bright career. I cannot give enough credit to Gulcin's contribution with any line of acknowledgement here.

Curriculum Vitae

Ozan Aksoy was born on October 4, 1980 in Kayseri, Turkey. He obtained a bachelor's degree in Business Administration from Bogazici University in Istanbul in 2004. From 2004 to 2005, he worked as a specialist assistant at the Master of Business Administration (MBA) programme at Istanbul Bilgi University and completed the remedial class of the Social Psychology master's programme at Bogazici University. In 2005 and 2006, he received the Utrecht Excellence Scholarship of Utrecht University and the HSP Huygens scholarship to study at the 'Sociology and Social Research' master's programme. After graduating *cum laude* from this master's programme in 2007, he became a PhD candidate at the Interuniversity Center for Social Science Theory and Methodology (ICS) in Utrecht. His PhD project was financed by the Netherlands Organization for Scientific Research (NWO) and completed in 2013. In 2010 and 2012, he was a visiting scholar for two and three months respectively at the Department of Economics, University of Zurich in Switzerland and at Nuffield College, University of Oxford in England. As of Fall 2013, he is employed as a postdoctoral researcher at the Department of Sociology and Nuffield College, University of Oxford through an NWO grant.

ICS Dissertation Series

The ICS-series presents dissertations of the Interuniversity Center for Social Science Theory and Methodology. Each of these studies aims at integrating explicit theory formation with state-of-the-art empirical research or at the development of advanced methods for empirical research. The ICS was founded in 1986 as a cooperative effort of the universities of Groningen and Utrecht. Since 1992, the ICS expanded to the University of Nijmegen. Most of the projects are financed by the participating universities or by the Netherlands Organization for Scientific Research (NWO). The international composition of the ICS graduate students is mirrored in the increasing international orientation of the projects and thus of the ICS-series itself.

1. C. van Liere, (1990), *Lastige Leerlingen. Een empirisch onderzoek naar sociale oorzaken van probleemgedrag op basisscholen*, Amsterdam: Thesis Publishers.
2. Marco H.D. van Leeuwen, (1990), *Bijstand in Amsterdam, ca. 1800–1850. Armenzorg als beheersings- en overlevingsstrategie*, ICS-dissertation, Utrecht.
3. I. Maas, (1990), *Deelname aan podiumkunsten via de podia, de media en actieve beoefening. Substitutie of leereffecten?*, Amsterdam: Thesis Publishers.
4. M.I. Broese van Groenou, (1991), *Gescheiden Netwerken. De relaties met vrienden en verwanten na echtscheiding*, Amsterdam: Thesis Publishers.
5. Jan M.M. van den Bos, (1991), *Dutch EC Policy Making. A Model-Guided Approach to Coordination and Negotiation*, Amsterdam: Thesis Publishers.
6. Karin Sanders, (1991), *Vrouwelijke Pioniers. Vrouwen en mannen met een 'mannelijke' hogere beroepsopleiding aan het begin van hun loopbaan*, Amsterdam: Thesis Publishers.
7. Sjerp de Vries, (1991), *Egoism, Altruism, and Social Justice. Theory and Experiments on Cooperation in Social Dilemmas*, Amsterdam: Thesis Publishers.
8. Ronald S. Batenburg, (1991), *Automatisering in bedrijf*, Amsterdam: Thesis Publishers.
9. Rudi Wielers, (1991), *Selectie en allocatie op de arbeidsmarkt. Een uitwerking voor de informele en geïnstitutionaliseerde kinderopvang*, Amsterdam: Thesis Publishers.
10. Gert P. Westert, (1991), *Verschillen in ziekenhuisgebruik*, ICS-dissertation, Groningen.
11. Hanneke Hermsen, (1992), *Votes and Policy Preferences. Equilibria in Party Systems*, Amsterdam: Thesis Publishers.
12. Cora J.M. Maas, (1992), *Probleemleerlingen in het basisonderwijs*, Amsterdam: Thesis Publishers.
13. Ed A.W. Boxman, (1992), *Contacten en carrière. Een empirisch-theoretisch onderzoek naar de relatie tussen sociale netwerken en arbeidsmarktposities*, Amsterdam: Thesis Publishers.
14. Conny G.J. Taes, (1992), *Kijken naar banen. Een onderzoek naar de inschatting van arbeidsmarktkansen bij schoolverlaters uit het middelbaar beroepsonderwijs*, Amsterdam: Thesis Publishers.

15. Peter van Roozendaal, (1992), *Cabinets in Multi-Party Democracies. The Effect of Dominant and Central Parties on Cabinet Composition and Durability*, Amsterdam: Thesis Publishers.
16. Marcel van Dam, (1992), *Regio zonder regie. Verschillen in en effectiviteit van gemeentelijk arbeidsmarktbeleid*, Amsterdam: Thesis Publishers.
17. Tanja van der Lippe, (1993), *Arbeidsverdeling tussen mannen en vrouwen*, Amsterdam: Thesis Publishers.
18. Marc A. Jacobs, (1993), *Software: Kopen of Kopiëren? Een sociaal-wetenschappelijk onderzoek onder PC-gebruikers*, Amsterdam: Thesis Publishers.
19. Peter van der Meer, (1993), *Verdringing op de Nederlandse arbeidsmarkt. Sector- en sekseverschillen*, Amsterdam: Thesis Publishers.
20. Gerbert Kraaykamp, (1993), *Over lezen gesproken. Een studie naar sociale differentiatie in leesgedrag*, Amsterdam: Thesis Publishers.
21. Evelien Zeggelink, (1993), *Strangers into Friends. The Evolution of Friendship Networks Using an Individual Oriented Modeling Approach*, Amsterdam: Thesis Publishers.
22. Jaco Berveling, (1994), *Het stempel op de besluitvorming. Macht, invloed en besluitvorming op twee Amsterdamse beleidsterreinen*, Amsterdam: Thesis Publishers.
23. Wim Bernasco, (1994), *Coupled Careers. The Effects of Spouse's Resources on Success at Work*, Amsterdam: Thesis Publishers.
24. Liset van Dijk, (1994), *Choices in Child Care. The Distribution of Child Care Among Mothers, Fathers and Non-Parental Care Providers*, Amsterdam: Thesis Publishers.
25. Jos de Haan, (1994), *Research Groups in Dutch Sociology*, Amsterdam: Thesis Publishers.
26. K. Boahene, (1995), *Innovation Adoption as a Socio-Economic Process. The Case of the Ghanaian Cocoa Industry*, Amsterdam: Thesis Publishers.
27. Paul E.M. Ligthart, (1995), *Solidarity in Economic Transactions. An Experimental Study of Framing Effects in Bargaining and Contracting*, Amsterdam: Thesis Publishers.
28. Roger Th. A.J. Leenders, (1995), *Structure and Influence. Statistical Models for the Dynamics of Actor Attributes, Network Structure, and their Interdependence*, Amsterdam: Thesis Publishers.
29. Beate Völker, (1995), *Should Auld Acquaintance Be Forgot...? Institutions of Communism, the Transition to Capitalism and Personal Networks: the Case of East Germany*, Amsterdam: Thesis Publishers.
30. A. Cancrinus-Matthijsse, (1995), *Tussen hulpverlening en ondernemerschap. Beroepsuitoefening en taakopvattingen van openbare apothekers in een aantal West-Europese landen*, Amsterdam: Thesis Publishers.
31. Nardi Steverink, (1996), *Zo lang mogelijk zelfstandig. Naar een verklaring van verschillen in oriëntatie ten aanzien van opname in een verzorgingstehuis onder fysiek kwetsbare ouderen*, Amsterdam: Thesis Publishers.
32. Ellen Lindeman, (1996), *Participatie in vrijwilligerswerk*, Amsterdam: Thesis Publishers.

33. Chris Snijders, (1996), *Trust and Commitments*, Amsterdam: Thesis Publishers.
34. Koos Postma, (1996), *Changing Prejudice in Hungary. A Study on the Collapse of State Socialism and Its Impact on Prejudice Against Gypsies and Jews*, Amsterdam: Thesis Publishers.
35. Jooske T. van Busschbach, (1996), *Uit het oog, uit het hart? Stabiliteit en verandering in persoonlijke relaties*, Amsterdam: Thesis Publishers.
36. René Torenvlied, (1996), *Besluiten in uitvoering. Theorieën over beleidsuitvoering modelmatig getoetst op sociale vernieuwing in drie gemeenten*, Amsterdam: Thesis Publishers.
37. Andreas Flache, (1996), *The Double Edge of Networks. An Analysis of the Effect of Informal Networks on Cooperation in Social Dilemmas*, Amsterdam: Thesis Publishers.
38. Kees van Veen, (1997), *Inside an Internal Labor Market: Formal Rules, Flexibility and Career Lines in a Dutch Manufacturing Company*, Amsterdam: Thesis Publishers.
39. Lucienne van Eijk, (1997), *Activity and Well-being in the Elderly*, Amsterdam: Thesis Publishers.
40. Róbert Gál, (1997), *Unreliability. Contract Discipline and Contract Governance under Economic Transition*, Amsterdam: Thesis Publishers.
41. Anne-Geerte van de Goor, (1997), *Effects of Regulation on Disability Duration*, ICS-dissertation, Utrecht.
42. Boris Blumberg, (1997), *Das Management von Technologiekooperationen. Partnersuche und Verhandlungen mit dem Partner aus Empirisch-Theoretischer Perspektive*, ICS-dissertation, Utrecht.
43. Marijke von Bergh, (1997), *Loopbanen van oudere werknemers*, Amsterdam: Thesis Publishers.
44. Anna Petra Nieboer, (1997), *Life-Events and Well-Being: A Prospective Study on Changes in Well-Being of Elderly People Due to a Serious Illness Event or Death of the Spouse*, Amsterdam: Thesis Publishers.
45. Jacques Niehof, (1997), *Resources and Social Reproduction: The Effects of Cultural and Material Resources on Educational and Occupational Careers in Industrial Nations at the End of the Twentieth Century*, ICS-dissertation, Nijmegen.
46. Ariana Need, (1997), *The Kindred Vote. Individual and Family Effects of Social Class and Religion on Electoral Change in the Netherlands, 1956–1994*, ICS-dissertation, Nijmegen.
47. Jim Allen, (1997), *Sector Composition and the Effect of Education on Wages: an International Comparison*, Amsterdam: Thesis Publishers.
48. Jack B.F. Hutten, (1998), *Workload and Provision of Care in General Practice. An Empirical Study of the Relation Between Workload of Dutch General Practitioners and the Content and Quality of their Care*, ICS-dissertation, Utrecht.
49. Per B. Kropp, (1998), *Berufserfolg im Transformationsprozeß, Eine theoretisch-empirische Studie über die Gewinner und Verlierer der Wende in Ostdeutschland*, ICS-dissertation, Utrecht.

50. Maarten H.J. Wolbers, (1998), *Diploma-inflatie en verdringing op de arbeidsmarkt. Een studie naar ontwikkelingen in de opbrengsten van diploma's in Nederland*, ICS-dissertation, Nijmegen.
51. Wilma Smeenk, (1998), *Opportunity and Marriage. The Impact of Individual Resources and Marriage Market Structure on First Marriage Timing and Partner Choice in the Netherlands*, ICS-dissertation, Nijmegen.
52. Marinus Spreen, (1999), *Sampling Personal Network Structures: Statistical Inference in Ego-Graphs*, ICS-dissertation, Groningen.
53. Vincent Buskens, (1999), *Social Networks and Trust*, ICS-dissertation, Utrecht.
54. Susanne Rijken, (1999), *Educational Expansion and Status Attainment. A Cross-National and Over-Time Comparison*, ICS-dissertation, Utrecht.
55. Mérove Gijsberts, (1999), *The Legitimation of Inequality in State-Socialist and Market Societies, 1987–1996*, ICS-dissertation, Utrecht.
56. Gerhard G. Van de Bunt, (1999), *Friends by Choice. An Actor-Oriented Statistical Network Model for Friendship Networks Through Time*, ICS-dissertation, Groningen.
57. Robert Thomson, (1999), *The Party Mandate: Election Pledges and Government Actions in the Netherlands, 1986–1998*, Amsterdam: Thela Thesis.
58. Corine Baarda, (1999), *Politieke besluiten en boeren beslissingen. Het draagvlak van het mestbeleid tot 2000*, ICS-dissertation, Groningen.
59. Rafael Wittek, (1999), *Interdependence and Informal Control in Organizations*, ICS-dissertation, Groningen.
60. Diane Payne, (1999), *Policy Making in the European Union: an Analysis of the Impact of the Reform of the Structural Funds in Ireland*, ICS-dissertation, Groningen.
61. René Veenstra, (1999), *Leerlingen – Klassen – Scholen. Prestaties en vorderingen van leerlingen in het voortgezet onderwijs*, Amsterdam: Thela Thesis.
62. Marjolein Achterkamp, (1999), *Influence Strategies in Collective Decision Making. A Comparison of Two Models*, ICS-dissertation, Groningen.
63. Peter Mühlau, (2000), *The Governance of the Employment Relation. A Relational Signaling Perspective*, ICS-dissertation, Groningen.
64. Agnes Akkerman, (2000), *Verdeelde vakbeweging en stakingen. Concurrentie om leden*, ICS-dissertation, Groningen.
65. Sandra van Thiel, (2000), *Quangocratization: Trends, Causes and Consequences*, ICS-dissertation, Utrecht.
66. Rudi Turksema, (2000), *Supply of Day Care*, ICS-dissertation, Utrecht.
67. Sylvia E. Korupp (2000), *Mothers and the Process of Social Stratification*, ICS-dissertation, Utrecht.
68. Bernard A. Nijstad (2000), *How the Group Affects the Mind: Effects of Communication in Idea Generating Groups*, ICS-dissertation, Utrecht.
69. Inge F. de Wolf (2000), *Opleidingsspecialisatie en arbeidsmarktsucces van sociale wetenschappers*, ICS-dissertation, Utrecht.
70. Jan Kratzer (2001), *Communication and Performance: An Empirical Study in Innovation Teams*, ICS-dissertation, Groningen.

71. Madelon Kroneman (2001), *Healthcare Systems and Hospital Bed Use*, ICS/NIVEL-dissertation, Utrecht.

72. Herman van de Werfhorst (2001), *Field of Study and Social Inequality. Four Types of Educational Resources in the Process of Stratification in the Netherlands*, ICS-dissertation, Nijmegen.

73. Tamás Bartus (2001), *Social Capital and Earnings Inequalities. The Role of Informal Job Search in Hungary*, ICS-dissertation, Groningen.

74. Hester Moerbeek (2001), *Friends and Foes in the Occupational Career. The Influence of Sweet and Sour Social Capital on the Labour Market*, ICS-dissertation, Nijmegen.

75. Marcel van Assen (2001), *Essays on Actor Perspectives in Exchange Networks and Social Dilemmas*, ICS-dissertation, Groningen.

76. Inge Sieben (2001), *Sibling Similarities and Social Stratification. The Impact of Family Background across Countries and Cohorts*, ICS-dissertation, Nijmegen.

77. Alinda van Bruggen (2001), *Individual Production of Social Well-Being. An Exploratory Study*, ICS-dissertation, Groningen.

78. Marcel Coenders (2001), *Nationalistic Attitudes and Ethnic Exclusionism in a Comparative Perspective: An Empirical Study of Attitudes Toward the Country and Ethnic Immigrants in 22 Countries*, ICS-dissertation, Nijmegen.

79. Marcel Lubbers (2001), *Exclusionistic Electorates. Extreme Right-Wing Voting in Western Europe*, ICS-dissertation, Nijmegen.

80. Uwe Matzat (2001), *Social Networks and Cooperation in Electronic Communities. A theoretical-empirical Analysis of Academic Communication and Internet Discussion Groups*, ICS-dissertation, Groningen.

81. Jacques P.G. Janssen (2002), *Do Opposites Attract Divorce? Dimensions of Mixed Marriage and the Risk of Divorce in the Netherlands*, ICS-dissertation, Nijmegen.

82. Miranda Jansen (2002), *Waardenoriëntaties en partnerrelaties. Een panelstudie naar wederzijdse invloeden*, ICS-dissertation, Utrecht.

83. Anne Rigt Poortman (2002), *Socioeconomic Causes and Consequences of Divorce*, ICS-dissertation, Utrecht.

84. Alexander Gattig (2002), *Intertemporal Decision Making*, ICS-dissertation, Groningen.

85. Gerrit Rooks (2002), *Contract en Conflict: Strategisch Management van Inkooptransacties*, ICS-dissertation, Utrecht.

86. Károly Takács (2002), *Social Networks and Intergroup Conflict*, ICS-dissertation, Groningen.

87. Thomas Gautschi (2002), *Trust and Exchange, Effects of Temporal Embeddedness and Network Embeddedness on Providing and Dividing a Surplus*, ICS-dissertation, Utrecht.

88. Hilde Bras (2002), *Zeeuwse meiden. Dienen in de levensloop van vrouwen, ca. 1850–1950*, Amsterdam: Aksant Academic Publishers.

89. Merijn Rengers (2002), *Economic Lives of Artists. Studies into Careers and the Labour Market in the Cultural Sector*, ICS-dissertation, Utrecht.

90. Annelies Kassenberg (2002), *Wat scholieren bindt. Sociale gemeenschap in scholen*, ICS-dissertation, Groningen

91. Marc Verboord (2003), *Moet de meester dalen of de leerling klimmen? De invloed van literatuuronderwijs en ouders op het lezen van boeken tussen 1975 en 2000*, ICS-dissertation, Utrecht.

92. Marcel van Egmond (2003), *Rain Falls on All of Us (but Some Manage to Get More Wet than Others): Political Context and Electoral Participation*, ICS-dissertation, Nijmegen.

93. Justine Horgan (2003), *High Performance Human Resource Management in Ireland and the Netherlands: Adoption and Effectiveness*, ICS-dissertation, Groningen.

94. Corine Hoeben (2003), *LETS' Be a Community. Community in Local Exchange Trading Systems*, ICS-dissertation, Groningen.

95. Christian Steglich (2003), *The Framing of Decision Situations. Automatic Goal Selection and Rational Goal Pursuit*, ICS-dissertation, Groningen.

96. Johan van Wilsem (2003), *Crime and Context. The Impact of Individual, Neighborhood, City and Country Characteristics on Victimization*, ICS-dissertation, Nijmegen.

97. Christiaan Monden (2003), *Education, Inequality and Health. The Impact of Partners and Life Course*, ICS-dissertation, Nijmegen.

98. Evelyn Hello (2003), *Educational Attainment and Ethnic Attitudes. How to Explain their Relationship*, ICS-dissertation, Nijmegen.

99. Marnix Croes en Peter Tammes (2004), *Gif laten wij niet voortbestaan. Een onderzoek naar de overlevingskansen van joden in de Nederlandse gemeenten, 1940–1945*, Amsterdam: Aksant Academic Publishers.

100. Ineke Nagel (2004), *Cultuurdeelname in de levensloop*, ICS-dissertation, Utrecht.

101. Marieke van der Wal (2004), *Competencies to Participate in Life. Measurement and the Impact of School*, ICS-dissertation, Groningen.

102. Vivian Meertens (2004), *Depressive Symptoms in the General Population: a Multifactorial Social Approach*, ICS-dissertation, Nijmegen.

103. Hanneke Schuurmans (2004), *Promoting Well-Being in Frail Elderly People. Theory and Intervention*, ICS-dissertation, Groningen.

104. Javier Arregui (2004), *Negotiation in Legislative Decision-Making in the European Union*, ICS-dissertation, Groningen.

105. Tamar Fischer (2004), *Parental Divorce, Conflict and Resources. The Effects on Children's Behaviour Problems, Socioeconomic Attainment, and Transitions in the Demographic Career*, ICS-dissertation, Nijmegen.

106. René Bekkers (2004), *Giving and Volunteering in the Netherlands: Sociological and Psychological Perspectives*, ICS-dissertation, Utrecht.

107. Renée van der Hulst (2004), *Gender Differences in Workplace Authority: An Empirical Study on Social Networks*, ICS-dissertation, Groningen.

108. Rita Smaniotto (2004), *'You Scratch My Back and I Scratch Yours' Versus 'Love Thy Neighbour'. Two Proximate Mechanisms of Reciprocal Altruism*, ICS-dissertation, Groningen.

109. Maurice Gesthuizen (2004), *The Life-Course of the Low-Educated in the Netherlands: Social and Economic Risks*, ICS-dissertation, Nijmegen.

110. Carlijne Philips (2005), *Vakantiegemeenschappen. Kwalitatief en Kwantitatief Onderzoek naar Gelegenheid- en Refreshergemeenschap tijdens de Vakantie*, ICS-dissertation, Groningen.

111. Esther de Ruijter (2005), *Household Outsourcing*, ICS-dissertation, Utrecht.

112. Frank van Tubergen (2005), *The Integration of Immigrants in Cross-National Perspective: Origin, Destination, and Community Effects*, ICS-dissertation, Utrecht.

113. Ferry Koster (2005), *For the Time Being. Accounting for Inconclusive Findings Concerning the Effects of Temporary Employment Relationships on Solidary Behavior of Employees*, ICS-dissertation, Groningen.

114. Carolien Klein Haarhuis (2005), *Promoting Anti-Corruption Reforms. Evaluating the Implementation of a World Bank Anti-Corruption Program in Seven African Countries (1999–2001)*, ICS-dissertation, Utrecht.

115. Martin van der Gaag (2005), *Measurement of Individual Social Capital*, ICS-dissertation, Groningen.

116. Johan Hansen (2005), *Shaping Careers of Men and Women in Organizational Contexts*, ICS-dissertation, Utrecht.

117. Davide Barrera (2005), *Trust in Embedded Settings*, ICS-dissertation, Utrecht.

118. Mattijs Lambooij (2005), *Promoting Cooperation. Studies into the Effects of Long-Term and Short-Term Rewards on Cooperation of Employees*, ICS-dissertation, Utrecht.

119. Lotte Vermeij (2006), *What's Cooking? Cultural Boundaries among Dutch Teenagers of Different Ethnic Origins in the Context of School*, ICS-dissertation, Utrecht.

120. Mathilde Strating (2006), *Facing the Challenge of Rheumatoid Arthritis. A 13-year Prospective Study among Patients and Cross-Sectional Study among Their Partners*, ICS-dissertation, Groningen.

121. Jannes de Vries (2006), *Measurement Error in Family Background Variables: The Bias in the Intergenerational Transmission of Status, Cultural Consumption, Party Preference, and Religiosity*, ICS-dissertation, Nijmegen.

122. Stefan Thau (2006), *Workplace Deviance: Four Studies on Employee Motives and Self-Regulation*, ICS-dissertation, Groningen.

123. Mirjam Plantinga (2006), *Employee Motivation and Employee Performance in Child Care. The effects of the Introduction of Market Forces on Employees in the Dutch Child-Care Sector*, ICS-dissertation, Groningen.

124. Helga de Valk (2006), *Pathways into Adulthood. A Comparative Study on Family Life Transitions among Migrant and Dutch Youth*, ICS-dissertation, Utrecht.

125. Henrike Elzen (2006), *Self-Management for Chronically Ill Older People*, ICS-dissertation, Groningen.

126. Ayşe Güveli (2007), *New Social Classes within the Service Class in the Netherlands and Britain. Adjusting the EGP Class Schema for the Technocrats and the Social and Cultural Specialists*, ICS-dissertation, Nijmegen.

127. Willem-Jan Verhoeven (2007), *Income Attainment in Post-Communist Societies*, ICS-dissertation, Utrecht.

128. Marieke Voorpostel (2007), *Sibling support: The Exchange of Help among Brothers and Sisters in the Netherlands*, ICS-dissertation, Utrecht.

129. Jacob Dijkstra (2007), *The Effects of Externalities on Partner Choice and Payoffs in Exchange Networks*, ICS-dissertation, Groningen.

130. Patricia van Echtelt (2007), *Time-Greedy Employment Relationships: Four Studies on the Time Claims of Post-Fordist Work*, ICS-dissertation, Groningen.

131. Sonja Vogt (2007), *Heterogeneity in Social Dilemmas: The Case of Social Support*, ICS-dissertation, Utrecht.

132. Michael Schweinberger (2007), *Statistical Methods for Studying the Evolution of Networks and Behavior*, ICS-dissertation, Groningen.

133. István Back (2007), *Commitment and Evolution: Connecting Emotion and Reason in Long-term Relationships*, ICS-dissertation, Groningen.

134. Ruben van Gaalen (2007), *Solidarity and Ambivalence in Parent-Child Relationships*, ICS-dissertation, Utrecht.

135. Jan Reitsma (2007), *Religiosity and Solidarity - Dimensions and Relationships Disentangled and Tested*, ICS-dissertation, Nijmegen.

136. Jan Kornelis Dijkstra (2007) *Status and Affection among (Pre)Adolescents and Their Relation with Antisocial and Prosocial Behavior*, ICS-dissertation, Groingen.

137. Wouter van Gils (2007), *Full-time Working Couples in the Netherlands. Causes and Consequences*, ICS-dissertation, Nijmegen.

138. Djamila Schans (2007), *Ethnic Diversity in Intergenerational Solidarity*, ICS-dissertation, Utrecht.

139. Ruud van der Meulen (2007), *Brug over Woelig Water: Lidmaatschap van Sportverenigingen, Vriendschappen, Kennissenkringen en Veralgemeend Vertrouwen*, ICS-dissertation, Nijmegen.

140. Andrea Knecht (2008), *Friendship Selection and Friends' Influence. Dynamics of Networks and Actor Attributes in Early Adolescence*, ICS-dissertation, Utrecht.

141. Ingrid Doorten (2008), *The Division of Unpaid Work in the Household: A Stubborn Pattern?*, ICS-dissertation, Utrecht.

142. Stijn Ruiter (2008), *Association in Context and Association as Context: Causes and Consequences of Voluntary Association Involvement*, ICS-dissertation, Nijmegen.

143. Janneke Joly (2008), *People on Our Minds: When Humanized Contexts Activate Social Norms*, ICS-dissertation, Groningen.

144. Margreet Frieling (2008), *'Joint production' als motor voor actief burgerschap in de buurt*, ICS-dissertion, Groningen.

145. Ellen Verbakel (2008), *The Partner as Resource or Restriction? Labour Market Careers of Husbands and Wives and the Consequences for Inequality Between Couples*, ICS-dissertation, Nijmegen.

146. Gijs van Houten (2008), *Beleidsuitvoering in gelaagde stelsels. De doorwerking van aanbevelingen van de Stichting van de Arbeid in het CAO-overleg*, ICS-dissertation, Utrecht.

147. Eva Jaspers (2008), *Intolerance over Time. Macro and Micro Level Questions on Attitudes Towards Euthanasia, Homosexuality and Ethnic Minorities*, ICS-dissertation, Nijmegen.

148. Gijs Weijters (2008), *Youth delinquency in Dutch cities and schools: A multilevel approach*, ICS-dissertation, Nijmegen.

149. Jessica Pass (2009), *The Self in Social Rejection*, ICS-dissertation, Groningen.

150. Gerald Mollenhorst (2009), *Networks in Contexts. How Meeting Opportunities Affect Personal Relationships*, ICS-dissertation, Utrecht.

151. Tom van der Meer (2009), *States of freely associating citizens: comparative studies into the impact of state institutions on social, civic and political participation*, ICS-dissertation, Nijmegen.

152. Manuela Vieth (2009), *Commitments and Reciprocity in Trust Situations. Experimental Studies on Obligation, Indignation, and Self-Consistency*, ICS-dissertation, Utrecht.

153. Rense Corten (2009). *Co-evolution of Social Networks and Behavior in Social Dilemmas: Theoretical and Empirical Perspectives.* ICS-dissertation, Utrecht.

154. Arieke J. Rijken (2009). *Happy Families, High Fertility? Childbearing Choices in the Context of Family and Partner Relationships.* ICS-dissertation, Utrecht.

155. Jochem Tolsma (2009). *Ethnic Hostility among Ethnic Majority and Minority Groups in the Netherlands. An Investigation into the Impact of Social Mobility Experiences, the Local Living Environment and Educational Attainment on Ethnic Hostility.* ICS-dissertation, Nijmegen.

156. Freek Bucx (2009). *Linked Lives: Young Adults' Life Course and Relations With Parents.* ICS-dissertation, Utrecht.

157. Philip Wotschack (2009). *Household Governance and Time Allocation. Four studies on the combination of work and care.* ICS-dissertation, Groningen.

158. Nienke Moor (2009). *Explaining Worldwide Religious Diversity. The Relationship between Subsistence Technologies and Ideas about the Unknown in Pre-industrial and (Post-)industrial Societies.* ICS-dissertation, Nijmegen.

159. Lieke ten Brummelhuis (2009). *Family Matters at Work. Depleting and Enriching Effects of Employees Family lives on Work Outcomes.* ICS-dissertation, Utrecht.

160. Renske Keizer (2010). *Remaining Childless. Causes and Consequences from a Life Course Perspective.* ICS-dissertation, Utrecht.

161. Miranda Sentse (2010). *Bridging Contexts: The interplay between Family, Child, and Peers in Explaining Problem Behavior in Early Adolescence.* ICS-dissertation, Groningen.

162. Nicole Tieben (2010). *Transitions, Tracks and Transformations. Social Inequality in Transitions into, through and out of Secondary Education in the Netherlands for Cohorts Born Between 1914 and 1985.* ICS-dissertation, Nijmegen.

163. Birgit Pauksztat (2010). *Speaking up in Organizations: Four Studies on Employee Voice.* ICS-dissertation, Groningen.

164. Richard Zijdeman (2010). *Status Attainment in the Netherlands, 1811-1941. Spatial and Temporal Variation Before and During Industrialization.* ICS-dissertation, Utrecht.

165. Rianne Kloosterman (2010). *Social Background and Children's Educational Careers. The Primary and Secondary Effects of Social Background over Transitions and over Time in the Netherlands.* ICS-dissertation, Nijmegen.

166. Olav Aarts (2010). *Religious Diversity and Religious Involvement. A Study of Religious Markets in Western Societies at the End of the Twentieth Century.* ICS-dissertation, Nijmegen.

167. Stephanie Wiesmann (2010). *24/7 Negotiation in Couples Transition to Parenthood.* ICS-dissertation, Utrecht.

168. Borja Martinovic (2010). *Interethnic Contacts: A Dynamic Analysis of Interaction Between Immigrants and Natives in Western Countries.* ICS-dissertation, Utrecht.

169. Anne Roeters (2010). *Family Life Under Pressure? Parents' Paid Work and the Quantity and Quality of Parent-Child and Family Time.* ICS-dissertation, Utrecht.

170. Jelle Sijtsema (2010). *Adolescent Aggressive Behavior: Status and Stimulation Goals in Relation to the Peer Context.* ICS-dissertation, Groningen.

171. Kees Keizer (2010). *The Spreading of Disorder.* ICS-dissertation, Groningen.

172. Michael Ms (2010). *The Diversity Puzzle. Explaining Clustering and Polarization of Opinions.* ICS-dissertation, Groningen.

173. Marie-Louise Damen (2010). *Cultuurdeelname en CKV. Studies naar effecten van kunsteducatie op de cultuurdeelname van leerlingen tijdens en na het voortgezet onderwijs.* ICS-dissertation, Utrecht.

174. Marieke van de Rakt (2011). *Two generations of Crime: The Intergenerational Transmission of Convictions over the Life Course.* ICS-dissertation, Nijmegen.

175. Willem Huijnk (2011). *Family Life and Ethnic Attitudes. The Role of the Family for Attitudes Towards Intermarriage and Acculturation Among Minority and Majority Groups.* ICS-dissertation, Utrecht.

176. Tim Huijts (2011). *Social Ties and Health in Europe. Individual Associations, Cross-National Variations, and Contextual Explanations.* ICS-dissertation, Nijmegen.

177. Wouter Steenbeek (2011). *Social and Physical Disorder. How Community, Business Presence and Entrepreneurs Influence Disorder in Dutch Neighborhoods.* ICS-dissertation, Utrecht.

178. Miranda Vervoort (2011). *Living Together Apart? Ethnic Concentration in the Neighborhood and Ethnic Minorities Social Contacts and Language Practices.* ICS-dissertation, Utrecht.

179. Agnieszka Kanas (2011). *The Economic Performance of Immigrants. The Role of Human and Social Capital.* ICS-dissertation, Utrecht.

180. Lea Ellwardt (2011). *Gossip in Organizations. A Social Network Study.* ICS-dissertation, Groningen.

181. Annemarije Oosterwaal (2011). *The Gap between Decision and Implementation. Decision making, Delegation and Compliance in Governmental and Organizational Settings.* ICS-dissertation, Utrecht.

182. Natascha Notten (2011). *Parents and the Media. Causes and Consequences of Parental Media Socialization.* ICS-dissertation, Nijmegen.

183. Tobias Stark (2011). *Integration in Schools. A Process Perspective on Students Interethnic Attitudes and Interpersonal Relationships.* ICS-dissertation, Groningen.

184. Giedo Jansen (2011). *Social Cleavages and Political Choices. Large-scale Comparisons of Social Class, Religion and Voting Behavior in Western Democracies.* ICS-dissertation, Nijmegen.

185. Ruud van der Horst (2011). *Network Effects on Treatment Results in a Closed Forensic Psychiatric Setting.* ICS-dissertation, Groningen.

186. Mark Levels (2011). *Abortion Laws in European Countries between 1960 and 2010. Legislative Developments and Their Consequences for Women's Reproductive Decision-making.* ICS-dissertation, Nijmegen.

187. Marieke van Londen (2012). *Exclusion of ethnic minorities in the Netherlands. The effects of individual and situational characteristics on opposition to ethnic policy and ethnically mixed neighbourhoods.* ICS-dissertation, Nijmegen.

188. Sigrid M. Mohnen (2012). *Neighborhood context and health: How neighborhood social capital affects individual health.* ICS-dissertation, Utrecht.

189. Asya Zhelyazkova (2012). *Compliance under Controversy: Analysis of the Transposition of European Directives and their Provisions.* ICS-dissertation, Utrecht.

190. Valeska Korff (2012). *Between Cause and Control: Management in a Humanitarian Organization.* ICS-dissertation, Groningen.

191. Maike Gieling (2012). *Dealing with Diversity: adolescents' support for civil liberties and immigrant rights.* ICS-dissertation, Utrecht.

192. Katya Ivanova (2012). *From Parents to Partners: The Impact of Family on Romantic Relationships in Adolescence and Emerging Adulthood.* ICS-dissertation, Groningen.

193. Jelmer Schalk (2012). *The Performance of Public Corporate Actors: Essays on Effects of Institutional and Network Embeddedness in Supranational, National, and Local Collaborative Contexts.* ICS-dissertation, Utrecht.

194. Alona Labun (2012). *Social Networks and Informal Power in Organizations.* ICS-dissertation, Groningen.

195. Michał Bojanowski (2012). *Essays on Social Network Formation in Heterogeneous Populations: Models, Methods, and Empirical Analyses.* ICS-dissertation, Utrecht.

196. Anca Minescu (2012). *Relative Group Position and Intergroup Attitudes in Russia.* ICS-dissertation, Utrecht.

197. Marieke van Schellen (2012). *Marriage and crime over the life course. The criminal careers of convicts and their spouses.* ICS-dissertation, Utrecht.

198. Mieke Maliepaard (2012). *Religious Trends and Social Integration: Muslim Minorities in the Netherlands.* ICS-dissertation, Utrecht.

199. Fransje Smits (2012). *Turks and Moroccans in the Low Countries around the year 2000: determinants of religiosity, trend in religiosity and determinants of the trend.* ICS-dissertation, Nijmegen.

200. Roderick Sluiter (2012). *The Diffusion of Morality Policies among Western European Countries between 1960 and 2010. A Comparison of Temporal and Spatial Diffusion Patterns of Six Morality and Eleven Non-morality Policies.* ICS-dissertation, Nijmegen.

201. Nicoletta Balbo (2012). *Family, Friends and Fertility.* ICS-dissertation, Groningen.

202. Anke Munniksma (2013). *Crossing ethnic boundaries: Parental resistance to and consequences of adolescents' cross-ethnic peer relations.* ICS-dissertation, Groningen.

203. Anja Abendroth (2013). *Working Women in Europe. How the Country, Workplace, and Family Context Matter.* ICS-dissertation, Utrecht.

204. Katia Begall (2013). *Occupational Hazard? The Relationship between Working Conditions and Fertility.* ICS-dissertation, Groningen.

205. Hidde Bekhuis (2013). *The Popularity of Domestic Cultural Products: Cross-national Differences and the Relation to Globalization.* ICS-dissertation, Utrecht.

206. Lieselotte Blommaert (2013). *Are Joris and Renske more employable than Rashid and Samira? A study on the prevalence and sources of ethnic discrimination in recruitment in the Netherlands using experimental and survey data.* ICS-dissertation, Utrecht.

207. Wiebke Schulz (2013). *Careers of Men and Women in the 19th and 20th Centuries.* ICS-dissertation, Utrecht.

208. Ozan Aksoy (2013). *Essays on Social Preferences and Beliefs in Non-embedded Social Dilemmas.* ICS-dissertation, Utrecht.