Commitments and Reciprocity in Trust Situations

Commitments and Reciprocity in Trust Situations

Experimental Studies on Obligation, Indignation, and Self-Consistency

Commitments en reciprociteit in vertrouwenssituaties.
Experimentele studies naar verplichting, verontwaardiging en consistentie
(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. J.C. Stoof,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op maandag 22 juni 2009
des middags te 12.45 uur

door

Manuela Dorothea Vieth

geboren op 4 maart 1977
te Salzkotten (Duitsland)

Promotor:	Prof. dr. W. Raub
Co-promotoren:	Dr. J. Weesie
	Dr. ir. V. Buskens

This thesis received financial support from the Netherlands Organization for Scientific Research (NWO) for the project "Commitments and Reciprocity" (MaGW Open Competition, project no. 400–05–89).

Manuscript commission: Prof. dr. A. Diekmann
Prof. dr. H. Flap
Prof. dr. P.G.M. van der Heijden
Prof. dr. T. Voss

Typesetting: LaTeX2e
Cover design: Inspired by Jon Clark "James S. Coleman" (1996);
technical realization by Wim Wennekes.
Printing: Ponsen & Looijen b.v., Ede

ISBN 978-90-393-50836

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The "Big" Problem

1.1.1 The Problem of Social Order

Explaining social order is one of the main problems of social theories (Hobbes, 1651/1966; Parsons, 1937). Fundamental questions of how cohesion in a society can be maintained and of how conflicts can be avoided are concerned. Exchange of goods in markets or in a barter economy, specialization and division of labor, education and innovation, stratification and power, prosperity, legal and political systems (including the realization of democracy), production and transportation systems, organizations (economic, governmental, etc.), peace within and between states, culture and arts, and so on—all of these ingredients of human societies require a certain degree of social order.

The problem of social order arises because many social and economic interactions in everyday life between individuals or organizations involve incentives for taking advantage of situations at the costs of others. For instance, sellers can benefit from providing low quality products for high prices. Specialists can take advantage of our lack of knowledge by offering advice that benefits them, but not us. Members of work teams are often tempted to reduce their own effort and to let others do the work. Similarly, social and political movements depend on the mobilization and the engagement of people. Moreover, environmental protection requires that a sufficient number of people reduces their resource consumption, while we all benefit from people who sustain natural resources irrespective of our contribution. In negotiations between labor unions and employer associations, compromises mitigate conflicts, but every party prefers the others to concede. Similar problems occur in negotiations between or within political parties and companies, as well as between friends, couples, and family members. Furthermore, rather than standing up for someone, we are tempted to wait for others to volunteer, if helping the other person requires sacrificing some of our own well-being or resources. Such incentives for "opportunistic behavior" (Williamson, 1985) cause conflicts of interests, which can result in a sub-optimal outcome. This gives rise to cooperation problems and distribution problems (Harsanyi, 1977) that constitute social dilemmas. Cooperation problems are characterized by common interests to improve joint outcomes, while parties in distribution problems have opposed interests concerning the allocation of shares. An interaction situation can involve both cooperation and distribution problems. For instance, if we ask a specialist for advice and pay for the service, it can yield joint benefits (cooperation problem). The specialist decides whether to share the benefits by providing proper

advice (distribution problem). Given the specialist's temptation to take advantage of us, we might expect to be misled and therefore abstain from investing in the service (sub-optimal outcome).

1.1.2 Social Norms and Sanctions

Social order inevitably depends on solving or mitigating cooperation problems and distribution problems (see also Voss, 1982, 1985; Binmore, 1994). This creates a "demand for social norms" (Coleman, 1990: ch. 10). Roughly speaking, social norms are behavioral regularities so that people overcome opportunism in social interactions (supplemented below). In cooperation problems, "conjoint social norms" require *all* parties involved to strive for improving joint outcomes, whereas in distribution problems, "disjoint social norms" require *some* parties to improve the outcomes of others (Coleman, 1990: 247–248). Coleman describes this dimension as a continuum, such that mixtures are also possible. For instance, since the interaction with a specialist involves a cooperation problem and a distribution problem, a social norm would be both conjoint because the specialist and we would benefit from it and disjoint because only the specialist has an incentive to behave opportunistically. The desirable social norm would induce the specialist to refrain from betraying. However, the mere desirability (or societal functionality) of a social norm does not restrict incentives for opportunistic behavior. Rather, the "realization of social norms" (Coleman, 1990: ch. 11) requires support by sanctions (i.e., reward or punishment) that create sufficient incentives to conform to social norms. Thus, more precisely, social norms are behavioral regularities in recurrent interactions in a population of actors who expect that deviant behavior will be punished (Voss, 2001: 108) or that conformity will be rewarded.

One possibility for sanctioning arises in repeated interactions with the same partner, i.e., in sufficiently stable relationships (Taylor, 1987/1976; Axelrod, 1984) or in interactions with sufficient exchange of information among people about past performance (e.g., for reputation in social networks, see Weesie, 1988: ch. 5; Coleman, 1990; Raub and Weesie, 1990; Ellickson, 1991; Buskens, 2002; Buskens and Raub, 2002). The prospect of gains from future cooperative interactions allows for control opportunities by conditional cooperation (e.g., "Trigger" strategies or "Tit for Tat" strategies, Axelrod, 1984; see also "reciprocal altruism", Trivers, 1971). For instance, the specialist might refrain from betrayal because he expects us to repeatedly invest in his service or because he fears a negative reputation that discourages others from a deal. Another possibility for sanctioning is a direct sanctioning option (e.g., Voss,

1998a, 2001; Fehr and Gächter, 2000, 2002). Especially in single encounters, social norms can be supported by possibilities to punish others for their opportunistic behavior. Voss (2001) concludes that the prospect of punishment will induce cooperation if punishment is effective in removing incentives for opportunistic behavior, but that the implicit threat is only credible if punishment requires no sacrifice from the person who performs it. Similar conclusions could be drawn for rewards. As soon as sanctioning becomes costly, solving the problem of opportunistic behavior is shifted to the sanctioning level creating a "second-order dilemma" (Taylor, 1987/1976; Coleman, 1990).

In contrast to these ideas, many examples in everyday life provide evidence that people are willing to incur substantial own sacrifices in order to retaliate against others' wrong-doings or to seek revenge when betrayed. Similarly, people also show their gratitude for received favors, and people behave cooperatively or generously if there is no punishment to fear for opportunism. For instance, people leave tips in restaurants despite visiting only one time, incur sacrifices in order to help others, and tend to keep their promises. Experimental research on social dilemmas shows that a substantial number of people behaves cooperatively and divides resources generously, even if no sanctions are possible or if sanctioning is costly (for reviews see, e.g., Ledyard, 1995; Roth, 1995; Kollock, 1998; Kopelman et al., 2002; Camerer, 2003: ch. 2; Ostrom and Walker, 2003; Shinada and Yamagishi, 2008). The findings also reveal that cooperative behavior is unstable and declines in subsequent interactions if no suitable sanctioning possibilities are available. Opportunities to sanction others' opportunistic behavior, even if costly, can help enforce and maintain social norms. Nevertheless, people engage in cooperative or generous behavior at a relatively high level at any beginning of a series of interactions, even if sanctioning opportunities are absent. This indicates that people also intrinsically follow some notion of a social norm and attempt to realize it.

The real-life observations and experimental findings cast doubt on the assumptions employed for the theoretical analyses of social dilemmas. The underlying actor model is known as "homo economicus" which is based on two core assumptions (see also the discussion by Weesie, 1994a): rationality and selfishness. Rational actors have consistent preferences and beliefs and process all available information that is necessary for optimal decision-making. This assures that decisions are made without systematic errors. Some approaches propose to relax the assumption of full rationality by some concept of "bounded rationality" (Simon, 1957), such as simple heuristics or framing (e.g., Simon, 1957; Tversky and Kahneman, 1981; Lindenberg, 1998, 2001; Gigeren-

zer and Selten, 2001; and for reviews of anomalies, e.g., Thaler, 1992; Camerer, 1995, 2003). It can be argued that suitable heuristics or (normative) frames give rise to cooperative and generous behavior. However, this also requires specifying a mechanism that determines what heuristic or frame is activated in what situations. Moreover, this approach typically implies that people would not actively and consciously make decisions but behave in accordance with the activated script. There is no doubt that various psychological mechanisms help people to simplify decision-making in order to save cognitive resources and to cope with complexities in life, and that people are subject to various cognitive biases. The challenge at this stage of research is to develop parsimonious general models of bounded rationality with better predictive power and without attributing any deviation in people's decision-making to cognitive limitations.

In the context of social dilemmas, evolutionary game-theoretical models have been employed in order to demonstrate the development of social norms through underlying learning processes (e.g., Boyd and Richerson, 1985; Samuelson, 1997; Fudenberg and Levine, 1998; Young, 1998; Binmore, 1994, 1998, 2005). This provides an opportunity to show how social norms can evolve based on the assumption that people follow certain heuristics in the sense of strategies. Binmore and Samuelson (1994) link the evolutionary approach to framing by arguing that people perceive an interaction situation with a certain probability in a way that induces them to behave more generously and to retaliate for other's greediness. From this perspective, cooperation and generosity as well as retaliation and reward would be the result of random (mis)perceptions, rather than the result of consciously motivated decision-making (for this argument and an alternative model, also see, e.g., Vieth, 2003). A more fruitful interpretation of error terms in theoretical models might be that components other than the explicitly modeled utility components motivate people's behavior (for this argument, also see the discussion in Chapter 3 on possible application of random utility models, McFadden, 1973, in combination with quantal response equilibrium analysis, McKelvey and Palfrey, 1998). However, given this interpretation, it would be desirable to investigate alternative motivations in order to study underlying processes and mechanisms rather than lumping everything together as a non-understood error.

For the purpose of explaining norm compliance and, thus, the existence of social norms, it seems therefore fruitful to also study people's motivations to behave cooperatively or generously. This shifts our attention to the selfishness assumption. Selfish actors are exclusively concerned with their own objective outcome. In addition to

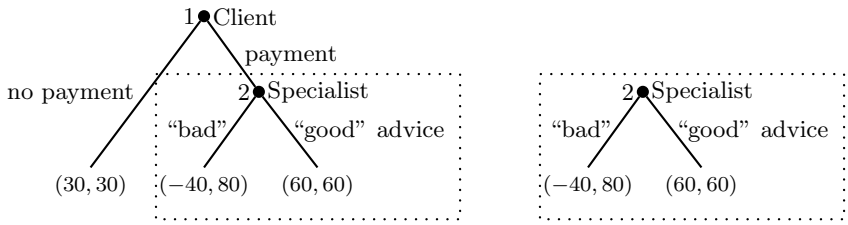
objective outcomes, intrinsic factors also play a role. In practice, assuming selfishness in lab experiments only implies that people should prefer receiving more money to receiving less money for themselves. This leaves some room for other motivations than purely monetary ones, provided an actor's own outcome is not reduced (e.g., costless sanctioning). However, the findings mentioned previously provide ample evidence that not even such a weak selfishness assumption holds empirically. Among the most powerful motivational forces are emotions. We tend to become angry about low offers in negotiations and reject them. We can enjoy supporting someone despite our limited time. We feel guilt if we do not volunteer to help someone or if we betray someone. Similarly, we are inclined to suffer from distress when even considering telling a lie. Emotions create intrinsic incentives that also serve as internal sanctions and thereby support social norms. Internalized social norms (e.g., Parsons, 1937; Coleman, 1990) are particularly powerful because deviant behavior is typically discovered and punished automatically. The same holds for rewards. Intrinsic sanctions do not result in a second-order dilemma that arises in the case of costly extrinsic sanctions (Taylor, 1987/1976; Coleman, 1990), and intrinsic sanctions can even help to render costly sanctions credible (Frank, 1988). Thus, it seems reasonable to first seek to understand what motivations drive people to behave cooperatively or generously before attempting to explain sanctioning behavior. The motivations underlying various behaviors, such as cooperating and sanctioning, might even be the same.

1.2 Behavioral Processes in Social Dilemmas

1.2.1 Reciprocity as Implication of Other-Regarding Motivations and Self-Consistency

Two basic types of motivations can be distinguished: outcome-based motivations and process-based motivations (see also the glossary for definitions of key terms). *Outcome-based motivations* are rooted in social (value) orientations (Messick and McClintock, 1968). The basic idea is that people also take into account others' outcomes. In doing so, people's own well-being is to some extent influenced by others' objective outcomes (social comparison). Preferences concerning the distributions of actors' own and others' objective outcomes transform objective outcomes into subjective utilities. This can induce people to, e.g., behave cooperatively or to punish opportunistic behavior. For instance, consider the example of the specialist and us in the role of the client. The specialist might resist the objective temptation to take advantage of us simply because he feels uncomfortable (e.g., due to feelings of

Figure 1.1: Paid and free advice



a) Paid advice **b)** Free advice

guilt) abaut betraying us and gaining more from the deal than we do. Such fairness concerns, e.g., based on inequality aversion, are one example of social orientations.

However, outcome-based motivations can only explain differences in behavior if the objective outcomes are different. From this perspective, the two stylized decision situations presented in Figure 1.1 would be completely identical concerning the specialist's decision. The first decision situation (Figure 1.1a) describes the example in which the specialist decides whether or not to take advantage of our lack of knowledge by providing "bad" advice. If he does so, we lose 40 Euro while he earns 80 Euro (e.g., because we have been advised to buy equipment that we do not need). If the specialist provides "good" advice, both of us benefit. The joint gain amounts to 60 Euro for each of us. This is more than if we did not invest in the advice yielding only 30 Euro for each of us (suboptimal outcome). The second decision situation (Figure 1.1b) describes the specialist's decision of whether to provide "bad" or "good" advice without our preceding investment. We might consider a situation in which the specialist sees a problem that we did not see.

It seems reasonable to assume that the specialist would be more inclined to provide "good" advice after we have asked and paid for it, given that our decision is favorable to him as he gains while we risk a loss. In this sense, our decision to ask and pay for the specialist's advice involves elements of kindness. The specialist knows, of course, that we also hope to benefit from "good" advice. This leaves some ambiguity concerning the kindness of our investment. However, given that we risk incurring a loss when we invest and receive "bad" advice, while the specialist would gain in any case, we might assume that our investment is appreciated and perceived as friendly. Preceding behavior can activate intention-based motivations. The basic assumption is that people take into account the behavioral process of how

certain outcomes are obtained. Others' preceding decisions are evaluated in terms of kind or unkind behavior. Kind behavior induces a *feeling of obligation* to return the favor, while unkind behavior inflicts a *feeling of indignation* that triggers a thirst for revenge (Gouldner, 1960; Coleman, 1990: ch. 12; Cialdini, 2001: ch. 2). Intention-based motivations are one variant of *process-based motivations*. In addition to such inter-personal process-based motivations, people's own preceding behavior gives rise to intra-personal motivations. People have a basic *desire for self-consistency* (Cialdini, 2001: ch. 3; Kunda, 2002) in order to avoid cognitive dissonance (Festinger, 1957). For instance, due to self-consistency, a promise can intrinsically serve as a commitment in the sense of a voluntary "strategic move" that creates a bond (Schelling, 1960).

Other-regarding outcome-based motivations, intention-based motivations, and the desire for self-consistency can imply reciprocal behavior. *Reciprocity* is a fundamental behavioral pattern of returning favors and retaliating for losses (Gouldner, 1960). The principle of reciprocity can also be observed when people sanction others. This suggests that sanctions are not necessarily only motivated by people's concern about objective outcomes, but that mere acts of preceding behavior can motivate people to sanction others.

1.2.2 Aim and Approach of Four Studies

This book contains four studies that are based on the idea that process-based motivations intrinsically give rise to reciprocal behavior and, therefore, allow for the explanation of both norm conformity and sanctioning. In order to test this assumption empirically, the influence of process-based motivations is studied by comparing people's decision-making in different endogenously generated behavioral contexts while controlling for influences of outcome-based motivations.

How do process-based motivations affect people's behavior in social dilemmas?

The focus of the four studies is on single encounters in two-person trust situations (such as Figure 1.1a) and related sharing situations (such as Figure 1.1b). Single encounters allow influences of non-selfish motivations to be studied without the confounding influences that arise from the prospect of future gains in durable relationships. The *Trust Game* (Figure 1.1a) is the core decision situation that is enriched by adding further options before and after it. Two types of such additional behavioral options are studied: commitments (before the core decision situation) and sanctions (after the core decision situation).

Commitments are voluntary strategic actions with the purpose of "reducing one's freedom of choice" or of changing the outcomes (Schelling, 1960). Commitments involve intrinsic bonds due to the desire for self-consistency and can also be combined with objective incentives. For instance, a commitment can be to some extent objectively binding. This is the case when the commitment is accompanied by something of value to the committed person (binding value) that is lost if the person deviates from the action to which the person is committed. Commitments can also provide an objective compensation to the other person in case the committed person deviates. Moreover, incurring a commitment can be associated with transaction costs. Examples are contracts, warranties, or guaranties, but also mere promises and threats. Promises and threats are announced intentions to perform a certain action that yields a gain to the other person, whereas threats involve the perspective of a loss. Thus, promises are friendly actions, whereas threats are unfriendly. Receiving a promise therefore creates feelings of obligation to return the favor, whereas being subject to a threat triggers feelings of indignation. Thus, promises and threats activate process-based motivations and can influence subsequent behavior. Announcements addressed in the four studies collected in this book are promises of trustworthiness by the trustee and announcements of sanctions by the trustor (i.e., reward promises or punishment threats).

Sanctions are behavioral options by which people express pleasure about others' good conduct or disapproval about others' misbehavior. Such rewarding and punishing often also affects objective outcomes, and thereby creates extrinsic incentives for norm-conform behavior. People then reward others with gratification or inflict a fine as punishment. Sanctioning itself can be costly such that it requires carrying an outlay. Rewarding others is an expression of gratitude that can be motivated by feelings of obligation to return a received favor, whereas indignation feelings motivate punishment as an expression of revengefulness. In the case of announced sanctions, the desire for self-consistency can also motivate people to perform the reward or the threatened punishment.

The four studies reported in this book investigate the influence of process-based motivations (i.e., obligation, indignation, and self-consistency) on trustfulness, trustworthiness, and sanctioning behavior (Studies 1, 3, and 4). Moreover, process-based motivations activated by preceding behavior can also moderate the influence of outcome-based motivations (Study 2). The following list provides an overview of the four studies and the respective research questions.

Study 1: Trust and Promises as Friendly Advances. Experimental Evidence on Reciprocated Kindness
How do making and omitting a promise of trustworthiness influence trustfulness? How do such promise decisions and trustfulness affect trustworthiness?

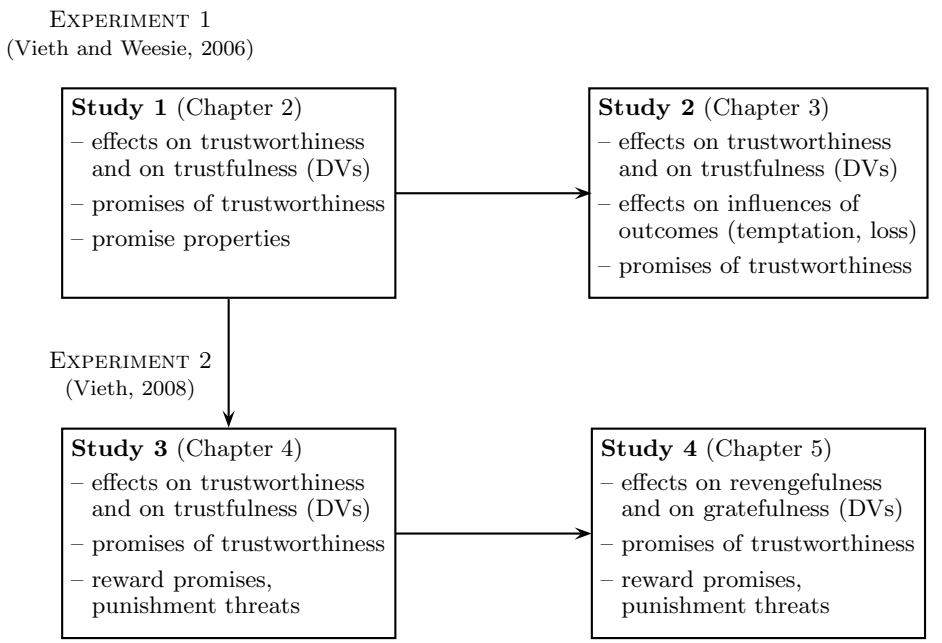
Study 2: Temptation, Loss, and Promises of Trustworthiness. Experimental Evidence on Context-Dependency of Outcome-Based Motivations
How does the behavioral context resulting from kind and unkind behavior moderate the effects of outcome-based motivations on trustfulness and trustworthiness?

Study 3: Influences of Promises and Threats on Trust and Trustworthiness. Experimental Evidence on Reciprocated Behavioral Advances
How do promises and threats shape trustfulness and trustworthiness?

Study 4: Revenge and Gratitude in Trust Situations Involving Promises and Threats. Experimental Evidence on Reciprocity by Intention-Based Sanctioning
How does preceding behavior affect subsequent sanctioning decisions?

The focus of each study and the relations between the four studies are displayed in a simplified manner in Figure 1.2. Studies 1 and 2 involve trust situations (Trust Games, similar to Figure 1.1a), sharing situations (Dictator Games, similar to Figure 1.1b), and trust situations in which the trustee decides whether or not to promise trustworthiness prior to the trustor's choice (Hostage Trust Games). In Study 1 (Chapter 2), the influence of trustfulness on trustworthiness is examined. Moreover, effects of omitted promises and made promises to honor trust on trustworthiness and on trustfulness are studied. Furthermore, it is investigated how these effects are moderated by the properties of the promise, specifically, by the size of the binding value and by the size of transaction costs. Thus, Study 1 focuses on the influence of process-based motivations that are triggered by preceding decisions on trustfulness and trustworthiness. In Study 2 (Chapter 3), the analyses of Study 1 are extended by analyzing how the influence of process-based motivations moderates effects of outcome-based motivations on trustfulness and on trustworthiness. For this purpose, a classical altruism model has been informally applied in order to specify outcome-based motivations. This model allows the influences of the trustee's outcomes (temptation) on trustworthiness and trustfulness to be separated from the influences of the trustor's outcomes (loss). The properties of the promise are not considered for the hypotheses, but are included as control variables in the data analyses.

Figure 1.2: Focus of each study and relations between the four studies



Dependent variables are indicated with "DVs".

In Studies 3 and 4 sanctioning options for trustors and announcements of sanctions by trustors are incorporated. Therefore, Studies 3 and 4 involve the three decision situations of Studies 1 and 2, but supplemented with sanctioning options for trustors after the trustee's decision of whether or not to honor trust. Depending on the trustee's decision, the trustor either decides whether or not to reward the trustee for honored trust or whether or not to punish the trustee for abused trust. In these studies, sanctioning is associated with objective properties: the size of the reward given to the trustee (gratification), the size of punishment inflicted upon the trustee (fine), and the cost of sanctioning incurred by the trustor (outlay). Thereby, sanctioning is always costly and not always effective in removing the trustee's objective incentives to abuse trust. In addition to the three decision situations that are involved in Studies 1 and 2 with sanctioning options for the trustor, two further decision situations are involved. The first additional decision situation is a trust situation in which the trustor decides about combining his decision to place trust with a reward promise or with a sanctioning announcement. In contrast to Studies 1 and 2, promises and

threats are not associated with objective incentives in Studies 3 and 4, but are purely "cheap-talk". This holds for announcements by both trustees (promise of trustworthiness) and trustors (reward promise or punishment threat). Second, Studies 3 and 4 also involve a distribution situation in which an allocator decides whether or not to make an investment in order to either increase or reduce the other's outcome (Allocation Game).

Study 3 (Chapter 4) focuses again on the influence of process-based motivations activated by preceding behavior on trustfulness and trustworthiness. In doing so, two extensions of Study 1 are explored. First, it is investigated whether the findings from Study 1 also hold for the respective decision situations in Study 3, in which the trustor has sanctioning options. Second, the influence of reward promises and punishment threats on trustworthiness is investigated. Study 4 (Chapter 5) builds upon Study 3 in analyzing how the trustor's revengefulness (punishing behavior) and gratefulness (rewarding behavior) are influenced by generously or greedily shared gains, by honored or abused trust, and by the various announcement decisions (promise of trustworthiness, reward promise, punishment threat).

In order to explore the influences of process-based motivations, two lab experiments have been conducted. Studies 1 and 2 are based on the data from the first experiment (conducted in November 2006 at the ELSE lab of the Sociology Department at Utrecht University in The Netherlands). Studies 3 and 4 use the data from the second experiment (conducted in April 2008 at the CeDEx lab of the Nottingham School of Economics at Nottingham University in Great Britain). The experiments are designed as *within-subject sets of structurally identical (sub)games* resulting from kind or unkind actual behavior in single encounters (for details, see Vieth and Weesie, 2006; Vieth, 2008). Structurally identical (sub)games constitute decision situations in which objective outcomes and available options are the same. These structurally identical decision situations only differ with respect to the behavioral context, i.e., with respect to the preceding decisions that have been made. In combination with the within-subject design employed in the two experiments, this approach allows for statistical analyses of decision-making in different behavioral contexts while controlling for individual heterogeneity and for influences of various outcome-based motivations without making assumptions about specific representations of outcome-based motivations. An exception is Study 2, in which influences of specific outcome-based motivations are studied. In order to analyze the "pure" effects of behavioral advances, the decisions of each subject that have been made in structurally identical (sub)games are grouped into *subject-payoff response sets*. Again, an exception is Study 2, for which subject

response sets are constructed (for reasons of restricted sample size). To analyze the data, logistic regression models with *fixed effects for response sets* (Rasch, 1960/1980) are employed in Studies 1 and 2. This allows minimal assumptions to be made about differences between subjects and outcomes. Due to the specific decisions that participants made, logistic regression models with *random effects for response sets* are used in Studies 3 and 4. Note that within-subject designs have advantages and disadvantages. The major disadvantage is that decisions are to some degree influenced by practice effects and carryover effects that are difficult or even impossible to control properly. However, for the type of studies reported in this book, a within-subject design appears to be more suitable than a between-subjects design because it allows influences of motivations to be analyzed at the individual level while controlling for (additive) individual heterogeneity and for influences of objective outcomes without making assumptions about specific outcome-based motivations.

The four studies reported in this book are written as separate articles and can therefore be read independently from one another. This also implies some degree of overlap and of similar text parts which is especially the case for the sections on the experimental design and the statistical model of Studies 1 and 2, using data from Experiment 1, and of Studies 3 and 4, using data from Experiment 2. Moreover, all four studies share the same basic theoretical arguments about influences of process-based motivations (i.e., obligation, indignation, and self-consistency). Therefore, some hypotheses of Study 1 and the respective arguments occur again as (parts of) hypotheses in Study 2 and to some extent in Study 3. In the reports of the studies, these repeated parts are summarized and the reader is referred to Study 1 for details.

Chapter 2

Trust and Promises as Friendly Advances

Experimental Evidence on Reciprocated Kindness

This chapter is a revised version of a working paper co-authored with Jeroen Weesie (Vieth and Weesie, 2007). I am grateful for his support and for the wonderful experience of thinking and working together with him.

We thank Vincent Buskens for comments, assistance during the experiment, and improvements in the Dutch version of the instructions used in our experiment. For assistance during the experiment, we also thank Rense Corten, Dennie van Dolder, and Richard Zijdeman. We acknowledge comments made by Ozan Aksoy, Davide Barrera, Ben Jann, Wojtek Przepiòrka, and Werner Raub, as well as by participants at the LMU seminar in Venice 2006, at the Japanese-German meeting of the DGS section "Modellbildung und Simulation" 2007 in Zurich, at the "Behavioral Studies" colloquium at ETH Zurich in 2007, and at the IIS 2008 World Congress in Budapest.

Abstract

People evaluate others' behavior and reciprocate kind and unkind actions. In doing so, people's decision-making is influenced not only by their own and others' outcomes, but also by mere preceding choice of a behavioral option. This can be due to feelings of obligation to return a favor and the desire for self-consistency. We explore the impact of trustfulness and of promising trustworthiness on subsequent decisions using experimental data from various Trust Games, Hostage Trust Games, and Dictator Games. Our lab experiment is designed as within-subject sets of structurally identical (sub)games resulting from friendly or unfriendly actual behavior in single encounters. This allows us to analyze the "pure" effects of decisions without making specific assumptions about actors' outcome preferences. We find evidence that both friendly and unfriendly advances are reciprocated. Trustors reward trustees' promises and punish omitted promises, controlling for changes in objective outcomes induced by a promise. Trustees tend to reciprocate trustfulness and, more strongly, to keep promises.

2.1 Introduction

Social and economic situations with interdependencies between actors are often characterized by incentives for opportunistic behavior Williamson (1985), i.e., actors are tempted to take advantage at the expense of their partner. An example in everyday life is trust. Trust is involved in situations in which people make a "risky advance" in the sense that they provide others an opportunity for exploitation. For instance, if we lend a book to a person with whom we have little contact, the other person has an incentive to keep the book. Without this risk, we might even have an interest ourselves in lending the book because we might like the other person to make use of some insights gained from the book and to refer to them in their own work. However, being aware of the possibility that we might not receive our book back, we might refuse to lend it. Similarly, if we buy something second-hand, we often cannot be sure about the product quality. Buying something online involves the additional risk that the seller might not deliver. In all of these examples, we would like the other persons to convince us that they can be trusted. For example, we wish sellers to provide safeguards such as guarantees or warranties for products we buy. Similarly, we are also frequently asked to provide a safeguard, e.g., a deposit for using certain facilities or for borrowing something. However, many safeguards do not completely remove the temptation to abuse trust. Some safeguards are of symbolic value and invoke an intrinsic commitment. For instance, informal promises can be cheap-talk, i.e., they neither involve objective costs for the person making the promise nor objective benefits for the person receiving the promise. Nevertheless, people tend to respond in kind to such actions, even when dealing with strangers. Particularly with regard to non-business relationships, it can be perceived as impolite not to accept an imperfect safeguard, especially if presented as a gift. This also holds for unwanted gifts and is exploited as an advertisement strategy (e.g., free sample products, methods of door-to-door salesmen, and Hare Krishna missionaries approaching passers-by with flowers, as described by Cialdini, 2001).

Placing trust and providing a safeguard or simply promising to be trustworthy can be perceived as a "friendly advance" and can create feelings of obligation to reciprocate. Moreover, a desire for self-consistency plays a role, inducing an incentive to keep promises. In turn, an omitted promise can inflict feelings of indignation that drive people to seek revenge. Thus, two powerful social-psychological forces are at work: feelings of obligation or indignation and self-consistency (Gouldner, 1960; Coleman, 1990: ch. 12; Cialdini, 2001: chs. 2–3). Both forces can induce behavioral patterns of reciprocity. In this study, we explore the effects of these forces in trust situations:

How do making and omitting a promise of trustworthiness influence trustfulness? And how do such promise decisions and trustfulness affect trustworthiness? In order to investigate these questions, we conducted a game-theoretical lab experiment, designed as within-subject sets of structurally identical (sub)games. This allows us to control for effects of outcomes and of individual characteristics and thereby gives maximal room for studying reciprocity resulting from motivations that are triggered by preceding behavior rather than induced by changes in objective outcomes.

In doing so, our work adds substantively and methodologically to previous studies. In recent years, research has explored effects of preceding behavior on subsequent decision-making. The experiments presented by Snijders (1996), McCabe, Rigdon, and Smith (2003), Cox (2004), and the analyses by Gautschi (2000) are most closely related to our approach. Originally inspired by Snijders (1996), we contribute to previous research by improving and extending analyses of behavioral advances in trust situations, both theoretically and empirically. For instance, in previous research, experimental conditions were distributed across subjects and promise options were not included (McCabe et al., 2003; Cox, 2004) or participants did not make actual and sequential decisions (Snijders, 1996). In fact, few studies employ a within-subject design for studying individual motivations, and few studies analyze effects of promise making on mitigating opportunistic behavior in social dilemmas, particularly in trust situations. Moreover, we apply insights from sociological and social-psychological research as a theoretical foundation and show how reciprocity results from motivations induced by mere behavioral processes and not by outcome concerns.

2.2 Trustfulness and Promises of Trustworthiness as Friendly Advances

2.2.1 Reciprocity Based on Obligation and Self-Consistency

Numerous studies have shown that people are not motivated only by their own objective outcomes when making decisions (for reviews see, e.g., Pruitt and Kimmel, 1977; Messick and Brewer, 1983; van Lange et al., 1992; Ledyard, 1995; Komorita and Parks, 1996; Kollock, 1998; Kopelman et al., 2002; Camerer, 2003: ch. 2; Ostrom and Walker, 2003). For instance, people reciprocate others' kind and unkind actions, even if they have to incur costs for rewarding or punishing and even if the target is a stranger (Fehr and Gächter, 2000). *Reciprocity is a behavioral pattern of returning favors and retaliating unkind actions.* The principle of reciprocity has been studied in various disciplines and in a wide range of fields (for reviews, see Fehr and Schmidt, 2006; Hann, 2006; Kolm, 2006; Lévy-Garboua et al., 2006). The idea of reciprocity has

roots in Scottish moral philosophy (Hume, 1739/1978; Smith, 1759/1976) and in theories of social exchange in sociology, social-psychology, and anthropology (Malinowski, 1922; Mauss, 1950; Thibaut and Kelley, 1959; Gouldner, 1960; Blau, 1964; Homans, 1974; Emerson, 1976; Coleman, 1990). These classical studies and related ones on social exchange, solidarity, and social capital typically focus on repeated interactions with sufficient exchange of information between partners. In such relationships, reciprocal behavior can be motivated by an actor's interest in his own expected objective outcomes of future interactions. More recent experimental studies specifically focus on reciprocity in single encounters ("one-shot situations") and reveal strong evidence for systematic reciprocal behavior. The results show that actors derive some utility or avoid disutility from rewarding others' friendly behavior and from retaliating against others' unfriendly behavior. Such patterns of reciprocal behavior can be implied by other-regarding motivations. Two kinds of *other-regarding motivations* have been distinguished that give rise to reciprocity: outcome-based and intention-based motivations (for a review, see Fehr and Schmidt, 2006).

Outcome-based other-regarding motivations are commonly linked to social (value) orientations which are rooted in social comparisons. The basic idea is that actors take into account the objective outcomes of their interaction partners (for reviews, see McClintock and van Avermaet, 1982; Au and Kwong, 2004). Thus, actors' utility is determined by some combination of their own and others' objective outcomes (Messick and McClintock, 1968; McClintock, 1972; Liebrand, 1984; Weesie, 1993, 1994b). Various social values have been distinguished (Knight and Dubro, 1984), e.g., that actors minimize the difference between their own and others' objective outcomes (Kelley and Thibaut, 1978), also known as "equalitarian" orientation (MacCrimmon and Messick, 1976). This idea has been employed in models of inequality aversion (e.g., Weesie, 1994a; Ledyard, 1995; Fehr and Schmidt, 1999; van Lange, 1999; Bolton and Ockenfels, 2000). With some equal outcome as a reference point, deviations from this "fair" outcome are assumed to induce emotional disutility that complements an actor's own objective gains. Such emotional disutility can result from feelings of guilt or envy (Fehr and Schmidt, 1999; for a guilt model, see Snijders, 1996). Inequality aversion can promote reciprocal behavior such that deviations from equal outcomes are retaliated and movements towards outcome equality are rewarded. Note that outcome-based motivations can also affect people's decisions in a non-reciprocal way, especially in situations without a decision for the other person (e.g., in Dictator Games or in Ultimatum Games with an additional passive receiver).

Intention-based motivations are activated by the perceived kindness and unkindness of interaction partners. As experimental studies indicate that information about others' behavioral options serves as an indication of others' kindness (e.g., Snijders, 1996; Gallucci and Perugini, 2000; Gautschi, 2000; Brandts and Solà, 2001; Falk et al., 2003; McCabe et al., 2003; Cox, 2004; Charness and Rabin, 2005). Actors observe others' behavior, evaluate how friendly it is, form expectations about others' kindness, and reciprocate to some extent. In doing so, actors consider not only the final outcome, but also the behavioral process of how a certain outcome is obtained. When someone provides a favor to us, we feel indebted to that person. We are then driven to give something in return in order to remove this "shadow of indebtedness" (Gouldner, 1960: 174), even if it is an unwanted favor (Coleman, 1990: ch. 12; Cialdini, 2001: ch. 2). Omitting or delaying the obligation to return a favor causes intrinsic distress and emotional tension to a person. Similarly, inflicted harm demands retaliation, especially if the harm was avoidable or unjustified. For instance, people become unfriendly toward others who behave impolitely without justified reasons. Thus, the driving forces are *feelings of obligation to return a favor* (Gouldner, 1960; Coleman, 1990: ch. 12; Cialdini, 2001: ch. 2) and *feelings of indignation that induce people to retaliate for inflicted losses* (Gouldner, 1960). Various theoretical models have been developed to account for intention-based motivations (e.g., Rabin, 1993; Levine, 1998; Charness and Rabin, 2002; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). Basically, others' choice of a specific option and information about non-chosen alternatives indicate others' kindness. The evaluation of others' kindness seems to be mainly based on three ingredients: the direction and extent to which *an actor's own outcomes* and *others' outcomes* are shaped by *others' intentional decisions* (see also Falk and Fischbacher, 2006). First, actors benefit from others' friendly actions and suffer from others' unfriendly actions. Thus, an actor perceives others' behavior as more friendly, the more benefits it yields to the actor or the more it helps the actor to avoid losses. In this sense, feelings of obligation to return a favor or feelings of indignation that drive people to retaliate for unkindness can motivate people even regardless of others' outcomes. Second, others' outcomes can also influence intention-based motivations. For instance, others' actions might be perceived as particularly kind if others incur sacrifices in order to behave in a friendly manner. Third, however, received gains and others' sacrifices are only perceived as friendly if the other could have chosen a less friendly alternative. If an actor has no other choice, empirical evidence suggests that only outcome-based motivations are relevant (Falk et al., 2003), e.g., inequality aversion as suggested by Fehr and Schmidt (1999). Falk

and Fischbacher (2006) propose a theoretical model combining such outcome-based motivations and intention-based motivations. In contrast to contemporary theoretical models that account for intention-based motivations, preceding behavior can also influence subsequent decision-making without inducing changes in objective outcomes. In general, perceiving kindness gives rise to positive feelings toward the other person, while unkindness triggers negative feelings. Emotional utility (e.g., happiness, satisfaction, relief) and disutility (e.g., envy, guilt, anger) caused by others' friendly or unfriendly behavior constitute the basis for feelings of obligation and indignation. Based on such emotions, actors are motivated to retaliate if others could have done something friendly, but chose another option. Similarly, actors tend to reward others who have chosen a friendly alternative or omitted an unfriendly option.

Self-consistency is another process-based motivation that plays a role in situations in which an actor makes several related decisions. People seek to behave consistently with their beliefs, attitudes, and previous choices (for reviews see, e.g., Webster, 1975; Cialdini, 2001: ch. 3; Kunda, 2002; Gass and Seiter, 2007: ch. 3). For instance, people tend to behave according to agreements they made, even if they discover hidden costs or feel uneasy when rethinking the agreement (for examples of studies on persuasion and on salesman practices, see Cialdini, 2001; Gass and Seiter, 2007). Similar evidence for self-consistency has been reported concerning public statements or announcements people make, even if people were instructed or forced to talk or write in a way that was against their original attitudes. Moreover, after choosing among several alternatives, people favor the chosen option, despite initial indifference or even opposite preferences (e.g., Brehm, 1956). In all of these examples, people for instance, have been found to change their beliefs and opinions in order to maintain an impression of self-consistency. Inconsistent behavior causes cognitive dissonance, which inflicts internal tension and distress that people seek to avoid (Heider, 1944, 1958; Festinger, 1957; Akerlof and Dickens, 1982; Aronson, 1992). In addition to attitude change, various other methods for reducing cognitive dissonance have been identified, e.g., bolstering (coming up with good reasons supporting a certain decision) or denial (denying or ignoring issues causing inconsistencies) (for an overview see, e.g., Gass and Seiter, 2007: 58). People use these methods, because they benefit in different ways from behaving consistently. First, self-consistency is crucial for keeping up a self-schema, i.e., "an integrated set of memories, beliefs, and generalizations about one's behavior in a given domain" (Kunda, 2002: 452). Such self-knowledge in specific areas is the basis for people's general self-evaluation and self-esteem. Second, self-consistency helps to create an image of self-competence in a complex world. By

behaving consistently, people maintain the impression that they control events in their life. Third, self-consistency reduces decision costs as people do not have to rethink all aspects of the same or of a similar situation (Cialdini, 2001).

The desire for self-consistency is an intra-personal process-based motivation that can result in *reciprocal behavior without other-regarding motivations*. For instance, if others reciprocated our friendly behavior, we might in turn continue to be friendly not because we feel obliged, but simply in order to behave consistently with our previous behavior. However, the desire for self-consistency can also *moderate the feelings of obligation or indignation*. If we induce someone to give a favor to us, we share some responsibility for the other's decision which increases obligation feelings. Similarly, if self-consistent behavior conflicts with induced obligation feelings, the various mechanisms of reducing cognitive dissonance can undermine obligation feelings by legitimizing the decision not to return a favor.

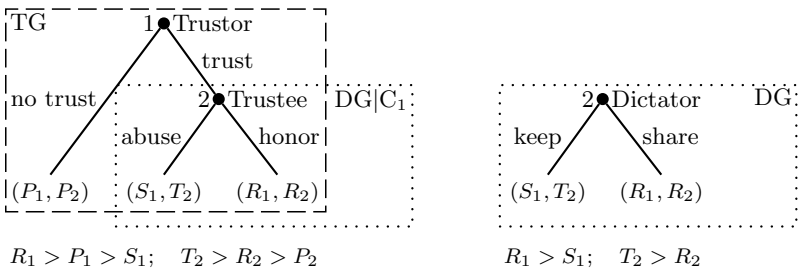
2.2.2 Obligation and Self-Consistency in Trust Situations

The Problem of Trustworthiness for Trustfulness

Feelings of obligation and the desire for self-consistency have implications for trust situations. The standard game-theoretical model describing trust situations is the *Trust Game* (TG) (Dasgupta, 1988; Kreps, 1990) (Figure 2.1a). It highlights the core features of trust situations (see also Coleman, 1990: ch. 5). First, two actors are involved: a trustor and a trustee. Both actors are better off with honored trust than with no trust at all ($R_i > P_i$, with $i = 1, 2$). Second, the trustee makes a decision after the trustor has placed trust. Despite the collective advantage arising from honored trust, the trustee has incentives to abuse trust ($T_2 > R_2$), while the trustor has something to lose if trust is abused ($S_1 < P_1$). The outcomes displayed in the game tree are "objective" outcomes, e.g., in monetary terms. If actors are motivated largely by their own objective outcomes and assume similar motivations on the part of others, trust will not be placed because placed trust would be abused. However, numerous studies have shown that people often do place trust and do honor trust, indicating that other-regarding motivations play a role (Snijders, 1996; and for reviews, see Camerer, 2003: ch. 2; Ostrom and Walker, 2003).

Given outcome-based motivations such as inequality aversion, trustees honor trust or share gains generously if guilt feelings are strong enough (Snijders, 1996; McCabe et al., 2003). Now consider that the trustee's decision of whether or not to honor trust constitutes a distribution decision in the TG because to honor trust means to return some benefit. Separating this subgame yields a dichotomous *Dictator Game* (DG)

Figure 2.1: Trust Game (TG) and dichotomous Dictator Game (DG)



a) Trust Game **b)** Dictator Game

with the trustee in the role of the dictator that represents the trustee's sharing decision without behavioral context (Figure 2.1b). We use the term "behavioral context" in the sense that an actor makes a decision in a subgame as a part of a larger game, i.e., the behavioral context consists of decisions made earlier in that game. For instance, the trustor's decision to place trust is the behavioral context for the trustee's choice of whether or not to honor trust. We can now compare the trustee's decision in the TG with the dictator's decision in the DG. McCabe, Rigdon, and Smith (2003) and Cox (2004) followed a similar reasoning of separating subgames. However, Cox (2004) studies the kindness of risky investments in the Investment Game (Berg et al., 1995), i.e., actors decide what amount to invest and to return. Compared to binary decisions of whether or not to invest or to return a given amount, the interpretation of behavior in terms of kindness is more ambiguous and likely to vary with actors' own preferences and expectations. For instance, investing half of the available endowment can already be perceived as kind by some trustees, while other trustees might perceive this as an unkind indication of distrust or even of a lack of generosity and benevolence (possibly also depending on experiences with other trustors in previous encounters).

If the objective outcomes in the TG and the DG are identical, outcome-based motivations do not predict a difference in behavior between the two situations (see also McCabe et al., 2003). However, the trustee will only be in the favorable position in the TG if the trustor has placed trust. The increase in the trustee's objective outcome due to honored trust ($R_2 > P_2$) can be perceived as a "gift". Gifts are usually seen as kind actions and create a feeling of obligation to return the favor of giving. Of course, in a trust situation both the trustee and the trustor benefit from honored trust ($R_1 > P_1$). However, the trustor still faces the risk of trust being abused, which then inflicts a loss upon the trustor ($S_1 < P_1$). We argued above

that an actor's behavior is particularly kind if that actor incurs actual or potential sacrifices while providing benefits to others. Based on these arguments and given the alternative decision to withhold trust, trustors behave in a friendly manner by placing trust. Considering intention-based motivations, trustees feel an obligation to return the favor of placed trust. Thus, the motivation to share gains is stronger for trustees in the TG than for dictators in the DG. Note that the effect of the perceived kindness of placed trust on trustworthiness can be outweighed by the effect of the trustee's temptation ($T_2 - R_2$). However, by comparing decisions in behavioral contexts with identical objective outcomes, objective outcomes only become relevant as moderating effects, but are ignored for the purpose of the study presented here.

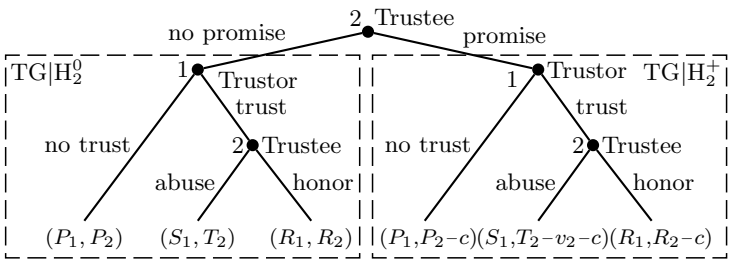
Hypothesis 2.1: Kindness of placed trust
Compared to honoring trust in the TG, gains are *less* likely to be shared in the DG.

Promises of Trustworthiness

Trustfulness depends on the possibility of trustworthiness. Thus, trustees might seize opportunities that decrease the trustor's concern about abused trust. One way for trustees to assure their trustworthiness to the trustor is to make a promise. *Promises are expressed intentions to perform a certain action that yields a gain to the other person.* Promises intrinsically demand fulfillment. As we will argue, this is based on the desire for self-consistency and on feelings of obligation to return the favor received because of the promise. In order to increase credibility, objective incentives can be attached to a promise. For instance, a trustee's promise of trustworthiness can be combined with a guarantee (cf. sellers providing warranties for technical products). In the case of abused trust, a guarantee provides compensation to the trustor or inflicts costs upon the trustee that reduce the trustee's temptation to renege on his promise. Moreover, making a promise can be associated with costs on the part of the trustee. For instance, sending a message, making a phone call, visiting the other person, or designing a contract are activities involving (transaction) costs. In this sense, a promise serves as a *commitment*, i.e., a *"voluntary strategic action", costly or not, with the purpose of "reducing one's freedom of choice" or changing the outcomes* (Schelling, 1960).

Raub (1992) proposes the *Hostage Trust Game* (HTG) in which the trustee has a commitment option prior to the TG (see also Weesie and Raub, 1996). Posting a commitment is a "strategic move" involving a "hostage" in the sense of a bond (Schelling, 1960). In our context, a commitment represents the trustee's choice of whether or

Figure 2.2: Hostage Trust Game (HTG)



$$R_1 > P_1 > S_1; \quad T_2 > R_2 > P_2; \quad v_2 \geq 0; \quad c \geq 0$$

not to promise trustworthiness (Figure 2.2). As argued above, making a promise can be associated with transaction costs (c) that the trustee loses irrespective of subsequent choices, i.e., even if trust is subsequently withheld. Moreover, the promise can be combined with an objective bond, i.e., the promise can have a value (v_2) for the trustee that is more or less binding. The trustee loses the value of the bond if he abuses trust after making the promise. Note that we do not consider promises that provide an objective compensation to the trustor, such as warranties. In case the trustee promises his trustworthiness, the initial outcomes are modified by the properties of the promise, i.e., trustees choose between playing the initial TG (TG|H$_2^0$) and a TG with modified outcomes (TG|H$_2^+$). The trustor receives information prior to his decision about the properties of the available promise and whether or not the trustee made the promise.

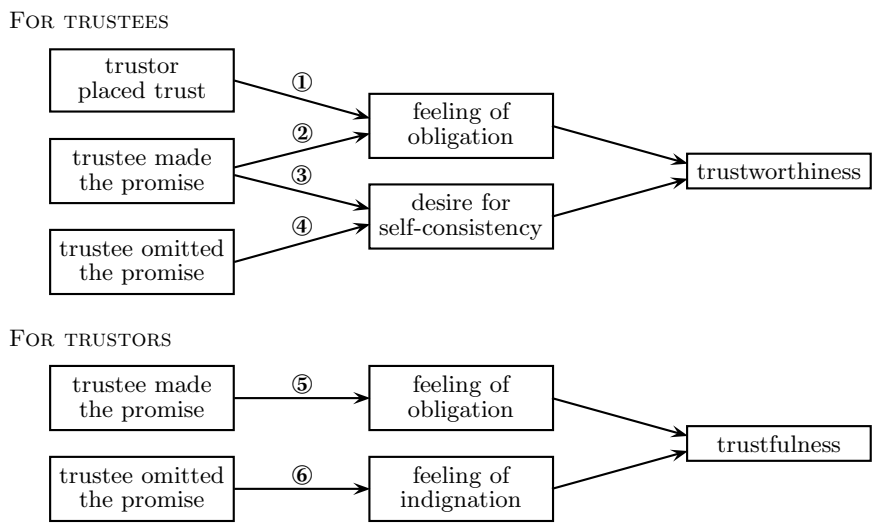
First, consider actors who are largely motivated by their own objective outcomes. In this case, trustees promise to behave in a trustworthy manner, trustors place trust and trustees honor trust if the value of the bond completely removes the trustee's temptation to abuse trust (perfectly binding: $v_2 > T_2 - R_2$) and if the transaction costs are low enough (affordable: $c < R_2 - P_2$). (For formal game-theoretical analyses of commitments in the TG and, closely related, in the Prisoner's Dilemma based on standard selfishness assumptions, see Raub and Keren, 1993; Weesie and Raub, 1996; Voss, 1998b; Raub and Weesie, 2000; Raub, 2004; and accounting for other-regarding motivations, Snijders, 1996.) Experiments using the HTG or the Prisoner's Dilemma with commitment option show that imperfectly binding or imperfectly compensating commitments also promote placing and honoring trust, while minimal transaction costs already hamper commitment posting (Yamagishi, 1986; Raub and Keren, 1993;

Mlicki, 1996; Snijders, 1996; for negotiation problems, also see Prosch, 2006). Even free communication that does not influence objective outcomes ("cheap-talk") promotes trustfulness and trustworthiness (for reviews see, e.g., Sally, 1995; Crawford, 1998; Bicchieri, 2002; Kopelman et al., 2002; Shankar and Pavitt, 2002; Ostrom and Walker, 2003; Brosig, 2006). One major insight is that the content of communication is relevant, specifically that people explicitly make commitments (Dawes et al., 1977). In contrast to largely uncontrolled discussions (face-to-face or by written messages), evidence for the promoting effects of cheap-talk promises is provided by experiments using promise options with predefined content while assuring anonymity (Brandts and Charness, 2003; Bochet and Putterman, 2007).

Outcome-based motivations (e.g., inequality aversion) induce trustees with sufficiently strong guilt feelings to honor trust in the TG (Snijders, 1996; McCabe et al., 2003). In the HTG, the binding value v_2 and the transaction costs c reduce the trustee's outcome after abused trust. Thus, in addition to the outcome-based guilt feelings, the reduction of the trustee's temptation also promotes trustworthiness. However, in a symmetric HTG (where $R_1 = R_2$ and $P_1 = P_2$), promising trustworthiness combined with binding properties and transaction costs usually reduces the outcome inequality after abused trust. Since advantageous inequality is the basis for guilt feelings, a more binding promise hampers the promoting effect of guilt feelings on trustworthiness. Thus, accounting only for outcome-based motivations, trustworthiness is increased in the HTG after the promise has been made only because the binding value of the promise reduces the trustee's temptation. Guilt-based inequality aversion of trustees then has an even smaller impact in the HTG than in the TG (in symmetric decision situations). Moreover, trustees' outcome-based motivations do not predict a difference in trustworthiness between the TG and the HTG after the trustee has not made the promise ($\mathrm{TG}|\mathrm{H}_2^0$). The same holds in general for comparing structurally identical decision situations that only differ with respect to the behavioral context, e.g., comparing the behavior in the decision situation that arises after the trustee has promised trustworthiness ($\mathrm{TG}|\mathrm{H}_2^+$) with the behavior in a separated TG with the same modified payoffs.

Considering intention-based motivations, the motivational influences are more complex (Figure 2.3). For the trustee, both inter-personal and intra-personal process-based motivations become relevant. We argued already that trustees feel an obligation to return the favor of placed trust and, thus, to behave in a trustworthy manner (arrow 1 and Hypothesis 2.1). Promising trustworthiness before the trustor's decision to place trust likewise influences the feeling of obligation (arrow 2). The in-

Figure 2.3: Basic assumptions about motivational influences of preceding decisions



Interaction effects are omitted in this figure, but are addressed in the accompanying text (i.e., moderating influences of promise properties and influences of self-consistency on obligation feelings). For trustees, feelings of obligation after promising trustworthiness are not due to the promise itself, but due to the combination of having made the promise and having subsequently received trust.

fluence of the felt obligation is then also moderated by the properties of the promise (arrow omitted). Moreover, consider that the trustee decides twice: whether or not to make the promise and, if trust has been placed, whether or not to honor trust. Thus, independent of feelings of obligation, the desire for self-consistency becomes relevant (arrows 3 and 4). The properties of the promise shape the desire for self-consistency. The desire for self-consistency can also promote the influence of obligation feelings after trustworthiness has been promised and undermine the influence of obligation feelings after the promise has not been made (arrows omitted). Concerning the trustor, a promise generally involves something friendly as it provides positive perspectives and serves as an indication of the trustee's kindness. Thus, received promises create a feeling of obligation, inducing the trustor to place trust in return (arrow 5). In contrast, an omitted promise causes feelings of indignation (arrow 6). The properties of the promise are assumed to influence feelings of obligation and feelings of indignation. In the following, we explain in more detail our predictions for effects of behavioral advances and for moderating effects of promise properties on trustworthiness and on trustfulness.

Effects of Obligation and Self-Consistency on Trustworthiness

We start by analyzing trustworthiness in the decision situation that arises after trustee has made the promise to honor trust (TG|H_2^+). First, similar to comparing trustworthiness in the TG with generosity in the DG (Hypothesis 2.1), a feeling of obligation to return a favor becomes relevant. As argued above, placed trust can be perceived as a friendly advance because the trustor provides gains to the trustee and risks losses himself. Thus, trustfulness motivates the trustee to respond in kind by honoring trust. Second, the desire for self-consistency induces the trustee to keep his promise. Note that lying in order to exploit others' trustfulness also causes distress. Thus, some trustees who would abuse trust in the TG tend to honor trust just because they promised trustworthiness. Considering the two arguments, trustees should be more likely to honor trust after they have promised to behave in a trustworthy manner (TG|H_2^+) compared to the decision situation in which no promise is possible (TG). This implies that trustworthiness is also more likely after the promise has been made than sharing gains in the DG (Hypothesis 2.1).

The feeling of obligation to return a favor requires a more detailed analysis because it is modified by the properties of the promise. First, recall that a promise can serve as an intrinsic commitment complementing the trustee's temptation ($T_2 - R_2$) and thereby reducing the risk of abused trust. Thus, placing trust could be perceived as less kind after a promise has been made than in the TG. The same reasoning applies to promises with a high binding value that reduces the trustee's temptation to abuse trust. One could argue that the feeling of obligation therefore is weakened. This might also hold because the trustee's favorable position is not solely due to the trustor's kindness, but also due to the trustee's own initiative of promising trustworthiness. However, the trustee has made the promise in order to induce the trustor to place trust. Therefore, the trustee shares some responsibility for the trustor's trustfulness. The desire for self-consistency then fosters the influence of the felt obligation to return the favor of placed trust, given that the trustor still risks a loss. Moreover, placed trust after the trustee has made the promise indicates that the trustor believes that the trustee will indeed keep the promise. This likewise boosts feelings of obligation the less binding the promise is in objective terms. Thus, given suitably low binding values v_2, the trustee should be less likely to honor trust the more binding the promise that has been made. Second, making a promise can be associated with transaction costs c. If trust would not be placed after making the promise the incurred transaction costs were wasted. Therefore, trustees might perceive placed trust as a reward for

the sacrificed transaction costs. This increases the trustee's feelings of obligation to return the favor. Thus, trustworthiness should increase with transaction costs.

> **Hypothesis 2.2: Kindness of placed trust after the promise has been made**
> Compared to the TG (i.e., without promise opportunity), trust is *more* likely to be honored after trustworthiness has been promised ($\mathrm{TG}|\mathrm{H}_2^+$). Moreover, the effect of placed trust on trustworthiness becomes *less promoting* with increasing binding value v_2, but *more promoting* with increasing transaction costs c.

We now turn to the decision situation that arises after the trustee has not made the promise to behave in a trustworthy manner ($\mathrm{TG}|\mathrm{H}_2^0$). It is fruitful to distinguish implications of not making the promise for the desire for self-consistency from implications for actually deciding whether to honor placed trust. In general, trustees who abuse trust after refusing to promise trustworthiness avoid internal distress that would otherwise be due to the desire for self-consistency. This argument is based on the following reasoning about why a trustee might not have made the promise. First, a trustee might perceive the transaction costs as overly high and anticipate that trust would be placed despite omitting the promise. In this case, some trustees would honor trust and some trustees would abuse trust. Note that trustees who tend to behave in an untrustworthy manner in a certain decision situation might be more concerned with their own objective outcome and, thus, about saving transaction costs, than more trustworthy trustees are. This would result in a selection effect of untrustworthy trustees. Second, a trustee might assume that trust will not be placed anyway, e.g., because of a high temptation ($T_2 - R_2$) or a large loss ($P_1 - S_1$). If placed trust is rewarding for trustees after a promise has been made (Hypothesis 2.2), withheld trust after a promise has been made is disappointing and causes emotional disutility. Therefore, trustees can avoid such internal distress if they manage to convince themselves that they would abuse trust anyway given the specific temptation and the specific loss for the trustor. Note that trustees, who omit making the promise because of transaction costs and expect trustfulness only after the promise has been made, likewise have an incentive to convince themselves to abuse trust in order to reduce emotional disutility (e.g., disutility due to regret).

Now consider that trustees decide whether to honor trust after they have received trust despite having omitted the promise. In general, the trustor's trustfulness invokes a feeling of obligation to return the favor of placed trust. However, we argued that

trustees not promising trustworthiness reduce intrinsic distress if they are convinced they would abuse trust. Thus, after a withheld promise, feelings of obligation compete with implications of the desire for self-consistency: self-consistency induces the trustee to abuse trust, while obligation feelings induce the trustee to honor trust. Therefore, one would expect less trustworthiness after the promise has been omitted than in the TG. Moreover, trustees, especially those with a strong desire for self-consistency, might even perceive placed trust negatively after the promise has been omitted. First, placed trust might confuse trustees who did not expect trustfulness. Given that trustees explicitly refused to promise trustworthiness, they might abuse trust because they feel puzzled or even irritated by the caused intrinsic conflict between obligation and self-consistency. Second, trustees might even perceive placed trust as unintelligent behavior not worthy of reward in the given decision situation. Trustees can also become irritated if the explicitly omitted promise indicates that they did not want the trustor to place trust. Unwanted gifts trigger punishment rather than reward if the gifts are perceived as manipulation attempts (Cialdini, 2001). Third, trustees might abuse trust because it seems more legitimate after they have omitted the promise of trustworthiness. Based on these arguments, the desire for self-consistency undermines feelings of obligation to return the favor and thereby prevents trustees who abuse trust from feeling the unease of cognitive dissonance. Thus, trustworthiness should generally be lower after the promise has not been made ($\mathrm{TG}|\mathrm{H}_2^0$) than if no promise is possible (TG).

Again, the properties of the promise deserve additional attention. First, recall that the binding value v_2 becomes relevant for trustees who abuse trust after they have made the promise. Thus, trustees who do not promise trustworthiness because of a high binding value show their intention to abuse trust. The higher the binding value of an omitted promise, the more likely trustees are to perceive trustfulness as unintelligent behavior. Therefore, the feeling of obligation is undermined more strongly the higher the binding value, and trustworthiness decreases. Second, we argued already that trustees might refrain from making the promise because of high transaction costs c. Trustees cannot easily neglect this fact. The higher the transaction costs, the more difficult it becomes for trustees, who would honor trust after they have promised trustworthiness, to reduce cognitive dissonance by becoming convinced that they would abuse trust anyway. In fact, a trustee might hope that trustors believed and accepted that the trustee refrained from making the promise just because of high transaction costs. Thus, the higher the transaction costs, the stronger the feelings of obligation to behave in a trustworthy manner in return and the less likely

it becomes that these feelings will be outweighed by the desire for self-consistency. High binding values of a withheld promise can also foster a selection of trustees who do not experience strong feelings of obligation. In contrast, high transaction costs also induce trustees with strong feelings of obligation to withdraw from making the promise.

Summarizing the arguments, the hampering influence of self-consistency might be stronger than the promoting influence of obligation feelings after the promise has been omitted. Moreover, the binding value promotes self-consistency, while transaction costs support obligation feelings. The promoting effect of transaction costs on the influence of obligation feelings can outweigh the hampering impact of self-consistency.

> **Hypothesis 2.3: Unkindness of placed trust after the promise has been omitted**
> Compared to the TG (i.e., without promise opportunity), trust is *less* likely to be honored after a possible promise of trustworthiness has not been made ($\mathrm{TG}|\mathrm{H}_2^0$). Moreover, the effect of placed trust on trustworthiness becomes *more hampering* with increasing binding value v_2, but *less hampering* with increasing transaction costs c.

Evidently, the effects of the kindness of placed trust on trustworthiness in the HTG can hardly be disentangled from selection effects due to making the promise and from the desire for self-consistency inducing trustees to keep their promise. This difficulty occurs in the case in which a possible promise has been omitted ($\mathrm{TG}|\mathrm{H}_2^0$) and also in the case in which the promise has been made ($\mathrm{TG}|\mathrm{H}_2^+$). We address this issue in the discussion.

Effects of Obligation and Anticipated Self-Consistency on Trustfulness

We now analyze the trustor's decision of whether or not to place trust. Promising trustworthiness ($\mathrm{TG}|\mathrm{H}_2^+$) involves a prospect for the trustor to receive increased outcomes ($R_1 > P_1$) from honored trust. In this sense, a voluntary promise of trustworthiness is a friendly advance that invokes feelings of obligation to return the favor by placing trust. Moreover, trustors might anticipate the general desire for self-consistency that induces trustees to keep their promise (Hypothesis 2.2). Thus, trustfulness should in general be increased after a promise has been made.

This promoting effect of received promises varies with the properties of the promise. First, by making a promise with a high binding value v_2, a trustee reduces his temptation ($T_2 - R_2$) to abuse trust. This indicates the trustee's willingness to

bind himself and thereby helps the trustor to place trust. Second, a trustee sacrifices transaction costs c that are an irreversible investment. The higher the transaction costs, the more a trustee indicates his interest in honored trust. As previously addressed, the trustor might anticipate that a selection of more honest trustees would more likely sacrifice high transaction costs. Thus, after receiving a promise of trustworthiness, trustors might believe that the risk of abused trust is reduced. In addition, trustors might feel obliged to show some kindness in return and reward the trustee for the sacrificed transaction costs. Even trustors who are reluctant to place trust, because they perceive the risk of abused trust as overly high, can be induced to place trust, because they feel obliged to return the favor of receiving the promise despite high transaction costs.

> **Hypothesis 2.4: Kindness of promising trustworthiness**
> Compared to the TG (i.e., without promise opportunity), trust is *more* likely to be placed after trustworthiness has been promised ($\mathrm{TG}|\mathrm{H}_2^+$). Moreover, the effect of receiving the promise of trustworthiness on placing trust becomes *more promoting* with increasing binding value v_2 and *more promoting* with increasing transaction costs c.

Next, recall the arguments that the trustees' desire for self-consistency in general reduces trustworthiness if the promise has not been made ($\mathrm{TG}|\mathrm{H}_2^0$) (Hypothesis 2.3). Trustors might anticipate the hampering effect of an omitted promise and become more reluctant to place trust. Moreover, consider that the trustee explicitly chose not to promise his trustworthiness and thereby explicitly chose not to provide the trustor with the prospect of a gain. Thus, trustors might perceive it as unfriendly that a possible promise has not been made. The omitted promise then gives rise to feelings of indignation that drive the trustor to retaliate by withholding trust.

The felt indignation might increase with the binding value v_2 of the omitted promise because having omitted such a promise indicates that the trustee has considered to abuse trust. Trustfulness then decreases with higher binding values of the omitted promise. Concerning the transaction costs c, recall that trustees might increasingly refrain from making the promise the higher the transaction costs (Hypothesis 2.3). As previously argued, a trustee's feeling of obligation to return the favor of placed trust is less likely to be undermined by his desire for self-consistency if transaction costs are high. Thus, trustors might indeed accept that the promise has not been made and might be encouraged to place trust. Moreover, we have argued that not choosing a friendly option can be perceived as unfriendly. However, omitting a friendly choice becomes justified and less unfriendly the greater the sacrifices

would otherwise be. Since transaction costs inflict sacrifices upon the trustee, higher transaction costs should mitigate the hampering effect of not making a promise on trustfulness.

> **Hypothesis 2.5: Kindness of promising trustworthiness**
> Compared to the TG (i.e., without promise opportunity), trust is *less* likely to be placed after a possible promise of trustworthiness has not been made ($\mathrm{TG}|\mathrm{H}_2^0$). The effect of an omitted promise on placing trust becomes *more hampering* with increasing binding value v_2, but *less hampering* with increasing transaction costs c.

Summarizing the hypotheses highlights the pattern of reciprocal behavior that result from intention-based motivations and from self-consistency (Table 2.1). Kind advances are replied in kind, and unkind behavior triggers unkind responses. Following this principle, receiving a promise of trustworthiness ($\mathrm{TG}|\mathrm{H}_2^+$) promotes trustfulness due to obligation feelings (Hypothesis 2.4), while an omitted promise ($\mathrm{TG}|\mathrm{H}_2^0$) causes indignation feelings, resulting in withheld trust (Hypothesis 2.5). Similarly, in the DG, the dictator is less generous than the trustee in the TG because the kind decision by the trustor to place trust is not preceding the sharing decision in the DG such that obligation feelings are not activated (Hypothesis 2.1). Next, by promising to honor trust ($\mathrm{TG}|\mathrm{H}_2^+$), the desire for self-consistency drives the trustee to keep his promise (Hypothesis 2.2). Moreover, self-consistency fosters the impact of obligation feelings after the promise has been made because the trustee shares some responsibility for the trustor's subsequent decision to place trust. In contrast, after the promise has been omitted in the $\mathrm{TG}|\mathrm{H}_2^0$, the impact of self-consistency results in reduced trustworthiness (Hypothesis 2.3). Self-consistency supports feelings of obligation after the promise has been made, but undermines obligation feelings after the promise has been explicitly omitted.

The properties of the promise moderate the influence of these processes. Transaction costs promote trustfulness and trustworthiness by increasing the positive influences of made promises (Hypotheses 2.2 and 2.4 for the $\mathrm{TG}|\mathrm{H}_2^+$) and by mitigating the negative influences of omitted promises (Hypotheses 2.3 and 2.5 for the $\mathrm{TG}|\mathrm{H}_2^0$). However, as objective bonds of made promises increase ($\mathrm{TG}|\mathrm{H}_2^+$), the influence of process-based motivations on behaving in a trustworthy manner is reduced (Hypothesis 2.2). After the promise has been omitted ($\mathrm{TG}|\mathrm{H}_2^0$), increasing objective bonds aggravate the hampering influence of self-consistency on trustworthiness and facilitate that obligation feelings are undermined (Hypothesis 2.3). Note again that the binding

Table 2.1: Overview of hypotheses and notation

	Placing Trust	Honoring Trust	
Behavioral contexts			
DG		–	Dictator Game (no placed trust)
TG	(ref.)	(ref.)	Trust Game (no promise option)
$TG\|H_2^+$	+	+	TG after a made promise to honor trust
$TG\|H_2^0$	–	–	TG after an omitted promise to honor trust
Binding value			Change of the effects of made and
in $TG\|H_2^+$	+	–	omitted promises of trustworthiness
in $TG\|H_2^0$	–	–	with increasing binding value v_2
Transaction costs			Change of the effects of made and
in $TG\|H_2^+$	+	+	omitted promises of trustworthiness
in $TG\|H_2^0$	+	+	with increasing transaction costs c

The hypotheses for effects of behavioral contexts are formulated in terms of differences toward the TG.

value promotes self-consistency after the trustee has omitted the promise ($TG|H_2^0$), which reduces trustworthiness, whereas transaction costs strengthen obligation feelings, which increase trustworthiness. Concerning trustfulness, higher binding values likewise aggravate the negative influence of an omitted promise (Hypothesis 2.5), but promote the positive impact of a received promise (Hypothesis 2.4).

2.3 Design of the Experiment, Data, and Statistical Method

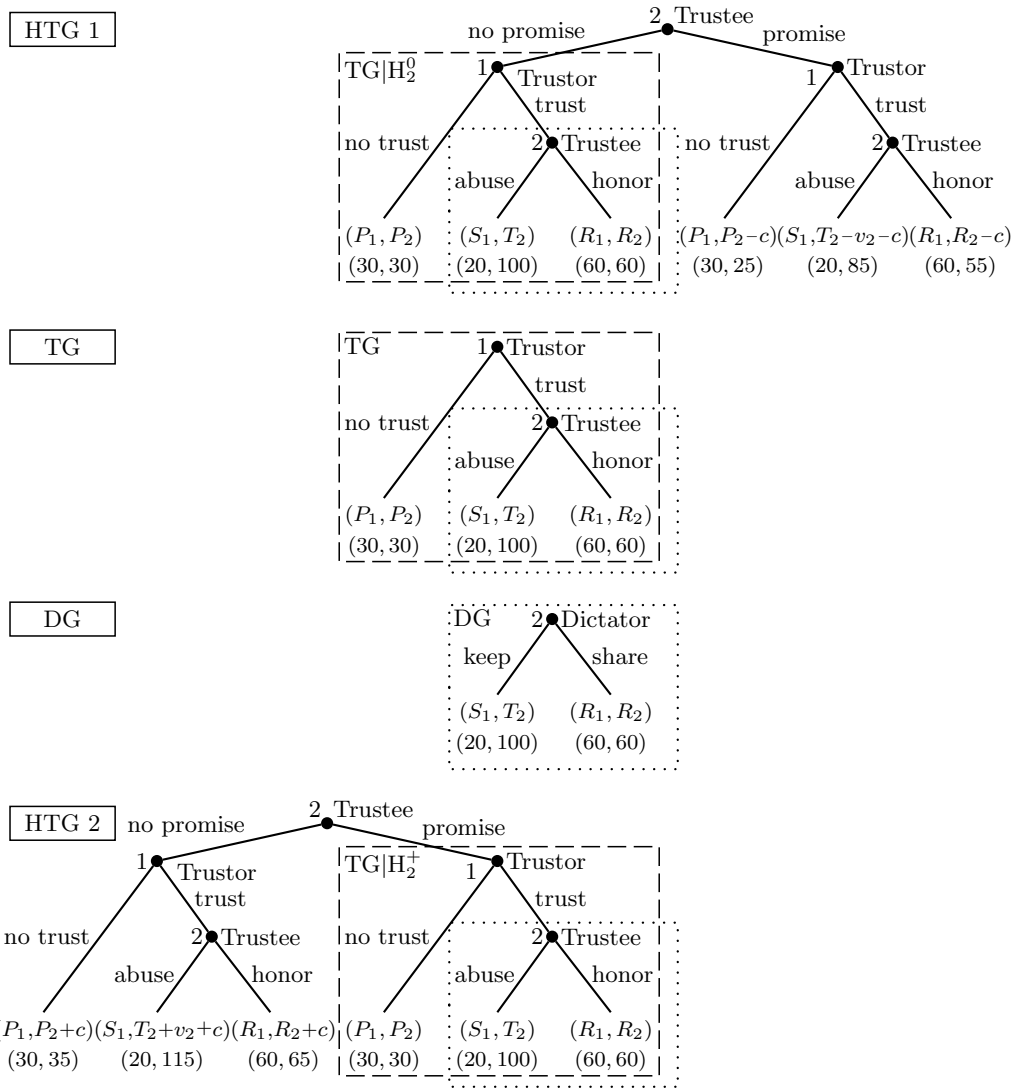
2.3.1 Experimental Design: Sets of (Sub)Games

Based on the theoretical reasoning, the aim of our experiment is to analyze effects of preceding decisions on subsequent behavior in trust situations while controlling for diverse outcome-based motivations and for general personal characteristics of participants. For this purpose, we designed our lab experiment as sets of games (TGs, HTGs, and DGs), in which (sub)games have identical extensive forms (i.e., identical choice structure and payoff structure). Each game constituted a single encounter (one-shot game). This experimental design allowed for within-subject comparisons of trustfulness and trustworthiness in (sub)games of identical extensive form, but with different behavioral contexts created endogenously by preceding kind or unkind actual decisions.

We improve in two main respects on previous studies that constructed (sub)games with identical payoffs (Snijders, 1996; McCabe et al., 2003; Cox, 2004). First, in most game-theoretical experiments different decision situations are distributed across subjects (for exceptions, e.g., Snijders, 1996; Charness and Rabin, 2002, 2005; Blanco et al., 2006; Brosig et al., 2007; Sandbu, 2007). This has the advantage that decisions in one experimental condition are not affected by experience with another experimental condition. However, any comparison of behavior is only possible on an aggregate level, while effects can even be the opposite on the individual level (see an experiment by Blanco et al., 2006; and on the "ecological fallacy", see Robinson, 1950). Moreover, controlling for diverse outcome-based motivations and individual heterogeneity is hardly possible. Second, in many studies employing a within-subject design, participants are asked to indicate their choices for all possible states of the decision situation. Using this "strategy method" (Selten, 1967), decisions remain hypothetical (even if a randomly chosen one is paid), which undermines influences of emotions and creates artificial consistency (for critical remarks, see also Roth, 1995: 322-323; McCabe et al., 2003). Therefore, we employ the "actual response method" for eliciting participants' actual decisions. Empirical evidence on differences in behavior due to the two elicitation methods is mixed (e.g., for found differences, see Brosig et al., 2003; Casari and Cason, 2009; for no support for differences, see Brandts and Charness, 2000; Oxoby and McLeish, 2004). Differences seem to depend on the type of decision situation involved. For instance, the strategy method implies simultaneous decision-making transforming sequential decision situations into strategic form (for this argument, also see McCabe et al., 2003; on differences McKelvey and Palfrey, 1998; McCabe et al., 2000).

For our experiment, we constructed sets of (sub)games such that the payoffs in games without a behavioral context were exactly the same as in the corresponding subgame of the TG or the HTG. The HTG contains two TGs as subgames resulting from the trustee's decision of whether or not to make the promise of trustworthiness. Each TG contains a DG as a subgame for the trustee's decision to return some benefit. We constructed different HTGs such that making the promise in one HTG resulted in a subgame with identical payoffs as in the subgame of another HTG after the promise was not made (Figure 2.4). This was reached by subtracting and adding the absolute values of promise properties at the beginning of some HTGs. We then completed our design with separate TGs and DGs for different payoff combinations (for details, see Vieth and Weesie, 2006).

Figure 2.4: Sets of games with identical subgames



The design allows for the comparison of the trustor's behavior in (sub)games indicated by *dashed boxes* and of the trustee's behavior in (sub)games indicated by *dotted boxes*. These sets of (sub)games constitute "subject-payoff response sets" used in the statistical analyses. Numerical example: $S_1^{\text{high}} = 20$, $T_2^{\text{high}} = 100$, $R_1 = R_2 = 60$, $P_1 = P_2 = 30$, $v_2^{\text{low}} = 10$, $c^{\text{low}} = 5$.

Figure 2.5: Outcome parameters of the experimental design

DESIGN PARAMETERS:
$S_1(2) \times T_2(2) \times v_2(3) \times c(3)$

Payoff parameters: $S_1(2) \times T_2(2)$

$S_1^{\text{low}} = 0 \qquad T_2^{\text{low}} = 80$
$S_1^{\text{high}} = 20 \qquad T_2^{\text{high}} = 100$
$R_1 = R_2 = 60$
$P_1 = P_2 = 30$

Promise properties: $v_2(3) \times c(3)$

$v_2^{\text{no}} = 0$
$v_2^{\text{low}} = \frac{1}{4}(T_2 - R_2) = \{5, 10\}$
$v_2^{\text{high}} = \frac{3}{4}(T_2 - R_2) = \{15, 30\}$
$c^{\text{no}} = 0$
$c^{\text{low}} = \frac{1}{6}(R_2 - P_2) = 5$
$c^{\text{high}} = \frac{4}{6}(R_2 - P_2) = 20$

For instance, consider the following baseline payoffs used in a set of (sub)games: $R_1 = R_2 = 60$, $P_1 = P_2 = 30$, $S_1 = 20$, and $T_2 = 100$ (see the numerical example in Figure 2.4). These payoffs constituted the outcomes of the subgame after the promise was not made (TG|H_2^0 of HTG1). The same payoffs were used in a separate TG and in a separate DG. In a HTG, making the promise changes the trustee's payoffs because the promise properties are subtracted. In order to get a subgame after the promise was made (TG|H_2^+) with payoffs identical to the decision situation after the promise was not made (TG|H_2^0), we constructed a second HTG by adding the transaction costs ($c = 5$) and the binding value ($v_2 = 10$) at the beginning to the respective trustee's payoffs. This yielded $R_2 = 60 + 5$, $P_2 = 30 + 5$ and $T_2 = 100 + 5 + 10$ after the promise was not made (TG|H_2^0 of HTG2). In the subgame after the promise was made (TG|H_2^+ of HTG2), the initially added promise properties were subtracted again. Thus, making the promise in HTGs with promise properties added in the beginning (HTG2) results in a subgame with payoffs identical to payoffs in the HTG after the promise is not made starting with the baseline payoffs (HTG1). Similarly, if the promise properties were subtracted at the beginning (HTG3, not included in Figure 2.4), the subgame after the promise was not made has exactly the same payoffs as the subgame after the promise was made in the HTG starting with the baseline payoffs (HTG1). These implicit shifts of payoffs in HTGs, TGs, and DGs on the scale of the promise properties were not explicit to participants and were hidden by variations of outcome parameters and by mixing sets of (sub)games (as described below).

In order to achieve different sets of sub(games) with identical payoffs, we varied some outcome parameters (Figure 2.5). These variations were included in the design

for methodological reasons mentioned above (for details, see Vieth and Weesie, 2006) and for further analyses (e.g., Chapter 3). Four baseline payoff combinations were distinguished by varying the payoffs resulting from abused trust (S_1 and T_2) at two levels each (low, high). As baseline payoffs, we chose 0 or 20 for S_1 and 80 or 100 for T_2. The baseline payoffs after no trust (P_i) and after honored trust (R_i) were fixed at $P_1 = P_2 = 30$ and $R_1 = R_2 = 60$. The two promise properties were varied at three levels each (no, low, high). "No" indicates $v_2 = 0$ or $c = 0$. Low binding values (v_2) were defined as $\frac{1}{4}(T_2 - R_2)$, and high binding values as $\frac{3}{4}(T_2 - R_2)$. For instance, consider $T_2 = 100$ and $R_2 = 60$. In this case, the possible binding values in our design are $v_2^{\text{no}} = 0$, $v_2^{\text{low}} = \frac{1}{4}(T_2 - R_2) = 10$, and $v_2^{\text{high}} = \frac{3}{4}(T_2 - R_2) = 30$. Levels for transaction costs (c) are defined as $\frac{1}{6}$ (for "low") and $\frac{4}{6}$ (for "high") on the scale $R_2 - P_2$, i.e., the "gain of cooperation". In all baseline payoff combinations the upper limit for transaction costs was $R_2 - P_2 = 30$ such that $c^{\text{low}} = \frac{1}{6}(R_2 - P_2) = 5$ and $c^{\text{high}} = \frac{4}{6}(R_2 - P_2) = 20$. Binding values v_2 and transaction costs c resulted in nine combinations of promise properties, yielding 36 combinations of baseline payoffs and promise properties. We explained above that the promise properties were added to or subtracted from the baseline payoffs to design sets of (sub)games. For initially added promise properties, we selected only combinations in which both promise properties were positive ($c > 0$ and $v_2 > 0$). Since the cheap-talk case ($c = 0$ and $v_2 = 0$) does not change the payoffs, this design yields 80 different combinations of total payoffs that can occur in sets of (sub)games. Note that in some HTGs with promise properties initially subtracted, the promise is perfectly binding ($v_2 > T_2 - R_2$).

Each participant played two sets of (sub)games in the role of player 1 (trustor, receiver) and two in the role of player 2 (trustee, dictator). For each encounter, participants were randomly and anonymously matched with another participant (stranger matching whereby the probability of re-matching was minimized within each type of game, see Vieth and Weesie, 2006). The sets of (sub)games were mixed by clustering the types of games. First, 12 TGs were played, then 14 HTGs, and thereafter 10 DGs. In two of the TGs and in two of the HTGs, trustees had no objective incentive to abuse trust ($T_2 < R_2$). These games are not involved in the reported analyses, but were included in the design in order to check participants' attention. In these decision situations, we observed 82.1% trustfulness and 95.3% trustworthiness, which indicates that participants paid sufficient attention to the objective outcomes. These percentages are significantly higher ($p < 0.0001$) than the highest average levels (in the TG|H_2^+) in the analyses (Table 2.4). Note that in the decision situations in which $T_2 < R_2$, we did not expect full trustfulness or full trustworthiness because of

possible influences of other-regarding outcome-based motivations (e.g., aggressive or competitive tendencies). A brief questionnaire about participants' socio-demographic characteristics (e.g., gender, age, education) separated the TGs from the HTGs. Other questions about personal attitudes and opinions followed the DGs. Analyses of questionnaire items are not reported here. In each game cluster, player roles were changed after half of the periods. In addition to randomly changing interaction partners, payoffs and promise properties (in HTGs) also changed from one period to the next. The combinations and sequences of payoffs and promise properties were varied across experimental sessions employing a factorial design.

The experiment was computer-assisted, employing the software package "z-Tree" (Fischbacher, 2007) (for an example of the decision screens, see Appendix A.1). In addition to general information on paper, participants received on-screen instructions and a tutorial before each game cluster. Outcomes were displayed as points in tables representing monetary gains (one Euro cent for each point). Participants were paid anonymously and immediately after the experiment. On average, participants earned 16 EUR. The experiment was conducted in November 2006 at the ELSE lab at Utrecht University. Using "ORSEE" (Greiner, 2004), 156 persons were recruited from the ELSE participant pool and took part in nine groups of 16 to 20 participants. Nearly all of the participants were students enrolled in various fields at Utrecht University.

2.3.2 Data and Statistical Method

The 156 subjects made 1716 "placing trust" decisions in the role of the trustor and 1389 "honoring trust" decisions as a trustee or dictator (Table 2.2). Of the 80 possible different payoff combinations that could occur in the (sub)games, 76 were realized for "placing trust" decisions of trustors. Despite withheld trust in some combinations, trustees decided in 71 (sub)games with different payoffs. For our analyses, we construct *"subject-payoff response sets"*, i.e., we group the *decisions of each subject that were made in (sub)games of identical extensive form* (Figure 2.4). Note again that the combinations of total payoffs are counted, i.e., transaction costs and the binding value are subtracted after the promise has been made, which thus separates the two subgames of HTGs into different groups. This yields 929 subject-payoff response sets of TGs or subgames of HTGs for "placing trust" decisions. For "honoring trust" decisions, we have 877 subject-payoff response sets of DGs or subgames of TGs or HTGs. Each subject-payoff response set involved 1–5 decisions. The reason for this variation is mainly that we elicited decisions in actually realized subgames. For instance, a trustee cannot decide whether to honor trust if the trustor did not place

Table 2.2: Number of cases and units of analyses

	Placing trust		Honoring trust	
Number of ...	all data	analyses	all data	analyses
subjects	156	118	156	70
total payoffs	76	48	71	35
subject-payoff response sets	929	212	877	101
decisions in total	1716	560	1389	248

Total payoffs are combinations of payoffs and promise properties.

trust. Whether a subject makes a decision in a certain subgame thus depends on previous decisions made in that game.

This grouping in subject-payoff response sets is reflected by a "fixed effects" statistical model in which we make minimal assumptions about differences between subjects and outcomes in order to analyze the "pure" effects of behavioral advances. In such a fixed effects approach, only subject-payoff response sets in which decisions vary carry statistical information. For instance, a trustor always withholding trust regardless of whether trustworthiness has been promised ($\mathrm{TG}|\mathrm{H}_2^+$) or a promise is not possible (TG) can neither support nor reject our hypothesis that promises increase trustfulness. The reason for the trustor's decisions can be anything but a reaction to behavioral advances. Therefore, a number of subject-payoff response sets are excluded in our fixed effects approach. We provide a more detailed overview in Table 2.3 in order to explain the composition of response sets involved in our analyses. First, consider rows 1–4. Subject-payoff combinations involving only *one single decision* cannot contain any variation in decisions across (sub)games. This is the case for 397 "placing trust" decisions ($248 + 149$) and for 467 "honoring trust" decisions ($97 + 370$). Note that each TG includes the DG for trustees' decisions and response sets without TG but with DG are impossible for trustors. Therefore, some (sub)game combinations can only occur in response sets for trustors, others only for trustees. Second (rows 5–15), subject-payoff response sets that consist of more than one decision, but *always the same decision*, are likewise excluded. In 320 subject-payoff response sets ($469 + 248 - 397$), 759 "placing trust" decisions are either always cooperative (trust placed, all x) or always non-cooperative (trust withheld, all $\bar{\mathrm{x}}$). The same holds for 674 "honoring trust" decisions in 309 subject-payoff response sets ($629 + 147 - 467$). In our analyses, we thereore have 212 subject-payoff response sets with 560 "placing trust" decisions and 101 subject-payoff response sets with 248 "honoring trust"

Table 2.3: Number of subject-payoff response sets per combination of (sub)games

	(Sub)game combinations				Placing trust (x)				Honoring trust (z)			
No.	per response set				all $\bar{x}$	all x	mix	Σ	all $\bar{z}$	all z	mix	Σ
1	DG								26	71		97
2	TG				164	84		248	$\sim$	$\sim$		$\sim$
3	H_2^+				52	97	$\sim$	149	327	43	$\sim$	370
4	H_2^0				$\sim$	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$
5	DG	TG							152	11	34	197
6	DG	H_2^+							57	16	27	100
7	DG	H_2^0							23	3	5	31
8	TG	H_2^+			64	35	73	172	$\sim$	$\sim$	$\sim$	$\sim$
9	TG	H_2^0			135	19	57	211	$\sim$	$\sim$	$\sim$	$\sim$
10	DG	TG	H_2^+						30	3	22	55
11	DG	TG	H_2^0						9	0	4	13
12	DG	H_2^+	H_2^0						1	0	6	7
13	TG	H_2^+	H_2^0		54	13	82	149	$\sim$	$\sim$	$\sim$	$\sim$
14	DG	TG	H_2^+	H_2^0					4	0	3	7
15	H_2^+	H_2^0			$\sim$	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$
				Σ	469	248	**212**	929	629	147	**101**	877

Blank cells indicate decision situations that are logically impossible, and "$\sim$" indicates (sub)game combinations that did not occur either by design or because of the decisions made by subjects. "Placing trust" decisions are denoted by "$\bar{x}$" for withheld trust and by "x" for placed trust. Similarly, "honoring trust" decisions are denoted by "$\bar{z}$" for abused trust and by "z" for honored trust.

decisions. Note that the selection of informative cases is a strength of the powerful design we employed in order to explore the effect of behavioral advances. We have sufficient decisions and response sets for our statistical analyses.

The decisions and response sets in the data involved in the analyses are summarized per subgame in Table 2.4. Usually, each response set consists of one decision per (sub)game that is involved in the response set. Since in some HTGs the made or omitted promise was cheap-talk ($c = 0$ and $v_2 = 0$), the total payoffs in the resulting subgame were the same as the total payoffs in the same subgame of another HTG in which the promise was not cheap-talk. Thus, some response sets involve two decisions for the same subgame. For instance, this is the case for 49 "placing trust" decisions ($204 - 155$) after the trustee has promised trustworthiness ($TG|H_2^+$). On average

Table 2.4: Summary of data in the analyses per (sub)game

	Placing trust (x)			Honoring trust (z)		
	N(dec)	N(sets)	%x	N(dec)	N(sets)	%z
DG				101	101	28.7
TG	212	212	57.1	63	63	52.4
TG|H_2^+	204	155	59.8	66	63	84.8
TG|H_2^0	144	139	18.1	18	18	50.0
Σ	560		48.0	248		51.2

Only mixed response sets are in the analyses. The percentages of placed trust (%x) and honored trust (%z) are calculated for the respective number of decisions.

across all (sub)games, subjects decided in 269 of the 560 cases (48.0%) to place trust and in 127 of the 248 cases (51.2%) to honor trust. Note that the extent of placing trust and of honoring trust is somewhat lower in the complete data because the excluded non-mixed response sets consisting of always withheld trust or always abused trust occur more frequently than those with always placed trust or always honored trust. The frequency of placed trust and of honored trust seems to differ considerably between the behavioral contexts. However, consider that behavioral contexts are created endogenously by the specific decisions made. For instance, trustees might have had the chance to honor trust and then might have also done so more often in some behavioral contexts than in others just because the outcomes were perceived as favorable. Testing our hypotheses about effects of behavioral advances on subsequent decisions therefore requires controlling for influences of outcome-based motivations and for individual heterogeneity.

Each subject-payoff response set is constituted by the decisions of one subject in (sub)games of identical extensive form, but in different behavioral contexts (including the "empty context"). Thus, decisions constitute the observations nested in a certain subject-payoff response set. We use logistic regression models with fixed effects to analyze the decisions in the subject-payoff response sets. Models are fitted by conditional maximum likelihood. Concerning this approach, see the Rasch program (Fischer and Molenaar, 1995; Rasch, 1960/1980), which is known as the fixed effects estimator for binary panel data in econometrics (Chamberlain, 1980). The baseline models can be described as follows:

$$\text{Prob}(y_{ijk}|\sigma_{ij}) = \sigma_{ij} + \eta'_{ijk}\beta$$

The model specifies the probability of trustfulness or trustworthiness of a subject i in the behavioral context of a (sub)game k that has a total payoff combination j, where σ_{ij} represents the fixed effects for subject-payoff combinations, η_{ijk} are attributes of the behavioral contexts k (and of controls) that vary within subject-payoff combinations, and β are parameters. Our analysis assumes neither that all subjects have the same responses for all payoff combinations nor that the difference in the probability of trustfulness or trustworthiness between total payoff combinations is the same for all subjects. In fact, subjects and payoffs may fully interact. We do however assume that the effect of behavioral context on behavior is the same for all subject-payoff combinations (see the discussion for further remarks).

2.4 Results

2.4.1 Analyses for Trustworthiness

Trustees decide whether or not to honor trust after trust has been placed. Now recall that this decision can take place in the TG or after the trustee's decision of whether or not to promise trustworthiness in the HTG (TG$|\mathrm{H}_2^+$ and TG$|\mathrm{H}_2^0$). Furthermore, the trustee decides whether or not to share gains without behavioral context in the DG. Thus, four differently embedded DGs can be distinguished. Since our hypotheses are formulated as comparisons of behavioral contexts towards the TG, we use the TG as the reference category in our analyses (Table 2.5). The *first model for trustworthiness* (model TW1) contains dummy variables for the behavioral contexts (Panel A). We present coefficients rather than marginal effects or unit effects because response probabilities can only be estimated at the cost of making specific assumptions about the distribution of fixed effects (Hoijtink and Boomsma, 1995). Wald tests for the differences between the coefficients of (sub)game dummies are reported at the bottom of the table (Panel C). We control for the period in which a decision is made (i.e., the number of past periods per game) because subject-payoff response sets are composed of decisions made in different periods. The period is counted for each type of game (i.e., 1–12 for the TG, 1–14 for the HTG, and 1–10 for the DG). In the *second model for trustworthiness* (model TW2), we include the properties of the promise, i.e., transaction costs c and the binding value v_2. We distinguish these effects for the two HTG subgames resulting from the trustee's decision about making the promise. Note that changes in objective payoffs due to promise properties are captured by subject-payoff response sets. Therefore, coefficients for behavioral contexts and for promise properties represent effects that are not based on objective outcomes. More-

Table 2.5: Logistic regression of trustworthiness with fixed effects for subject-payoff response sets

(A) Regression coefficients

		TW1		TW2	
	Hyp.	b	se	b	se
Behavioral contexts					
DG	H_1: –	-0.54°	0.29	-0.55°	0.29
TG		(ref.)		(ref.)	
TG|H_2^+	H_2: +	1.92^{***}	0.53	1.74^{***}	0.60
TG|H_2^0	H_3: –	0.18	0.57	1.01	1.34
Binding value					
in TG|H_2^+	H_2: –			−0.03	0.04
in TG|H_2^0	H_3: –			-0.24°	0.13
Transaction costs					
in TG|H_2^+	H_2: +			0.12°	0.07
in TG|H_2^0	H_3: +			0.18°	0.10
Past periods per game		−0.15	0.11	−0.16	0.12

(B) Likelihood-ratio tests

	χ^2	df	χ^2	df
LR test (control)	36.20^{***}	3	47.18^{***}	7
LR test (TW1)			10.98^{*}	4

(C) Pairwise comparisons (Wald tests)

	Δb	se	Δb	se
TG|H_2^0 – TG|H_2^+	-1.75^{**}	0.64	−0.73	1.35
DG – TG|H_2^+	-2.47^{***}	0.55	-2.29^{**}	0.61
DG – TG|H_2^0	−0.72	0.55	−1.57	1.35

N(response sets) = 101, N(decisions) = 248, N(subjects) = 70;
two-sided p-values: *** p = 0.001, ** p = 0.01, * p = 0.05, ° p = 0.1; (sub)games (0/1), past periods per game (1 . . . 10/12/14), binding value $v_2 = (0, 5, 10, 15, 30)$, transaction costs $c = (0, 5, 20)$. Likelihood-ratio tests are reported against the null model with period control and against model TW1.

over, the coefficients for the two HTG subgames represent the effects of making and omitting the promise in cases in which the promise is cheap-talk ($v_2 = 0$ and $c = 0$).

The likelihood-ratio test (Panel B) for model TW1 against the null model with period control shows that trustworthiness in general differs significantly between behavioral contexts (LR $\chi^2_{3\,\mathrm{df}} = 36.20$ with $p < 0.0001$). Moreover, properties of the promise in general significantly moderate the influences that the trustee's promise decision exerts on trustworthiness. This is indicated by the likelihood-ratio test for model TW2 against model TW1 (LR $\chi^2_{4\,\mathrm{df}} = 10.98$ with $p = 0.0268$). Separate likelihood-ratio tests likewise show that the binding value (LR $\chi^2_{2\,\mathrm{df}} = 5.84$ $p = 0.0538$) and the transactions costs (LR $\chi^2_{2\,\mathrm{df}} = 8.39$ with $p = 0.0150$) significantly moderate the influences of making and omitting the promise in the model TW2. Separate likelihood-ratio tests for the two HTG subgames in the model TW2 show that only the influence of omitted promises is significantly influenced by the properties of the promise (LR $\chi^2_{2\,\mathrm{df}} = 4.50$ with $p = 0.1057$). No support for such moderating influences can be found for made promises (LR $\chi^2_{2\,\mathrm{df}} = 6.77$ with $p = 0.0339$). In the following, we describe and discuss the results for specific behavioral contexts.

We argued that trustees perceive placed trust as a friendly advance that invokes feelings of obligation that increase trustworthiness (Hypothesis 2.1). Although the coefficient is only marginally significant, we find a tendency that trustees are indeed less likely to share gains in the DG than after trust has been placed in the TG (Table 2.5) (see also Gautschi, 2000; McCabe et al., 2003; Cox, 2004).

Next, we hypothesized that promising to honor trust (TG$|\mathrm{H}_2^+$) promotes trustworthiness by increasing feelings of obligation and activating influences of self-consistency (Hypothesis 2.2). This reasoning is likewise supported (Table 2.5). The results show a strongly positive and highly significant coefficient of having made the promise. This holds in general, including influences of promise properties (model TW1), and in cases in which the promise is cheap-talk (model TW2). In fact, the positive effect of making the promise is also significantly stronger than the negative effect of no placed trust in the DG (test of differences between absolute coefficients: Wald $\chi^2_{1\,\mathrm{df}} = 4.50$ with $p = 0.0338$ in model TW1 and Wald $\chi^2_{1\,\mathrm{df}} = 2.84$ with $p = 0.0918$ for cheap-talk promises in model TW2). This might suggest that placed trust after the promise has been made motivates trustworthiness through both obligation feelings and self-consistency. As previously argued, self-consistency might also boost obligation feelings because the trustee shares some responsibility for the trustor's decision to place trust (Hypothesis 2.2). We cannot ascertain whether the influence of self-consistency complements the impact of obligation feelings after the promise has

been made or whether self-consistency even fosters the feeling of obligation. Moreover, recall that we also mentioned that making the promise might reduce the kindness of placed trust as perceived by the trustee and, thereby, the trustee's obligation feelings. It would then be a strong impact of promising trustworthiness that motivates the trustee due to the desire for self-consistency rather than the trustor's decision to place trust motivating the trustee due to obligation feelings. Thus, it is possible that we have found the strong increase in trustworthiness reported here due to a very strong impact of self-consistency despite possibly reduced feelings of obligation (see the discussion for further remarks).

Now consider that the properties of the promise can moderate the positive influence that making the promise has on trustworthiness. The binding value of the promise increases the trustor's belief that trust might be honored. Thus, the binding value should hamper the promoting influence of having promised to honor trust because placed trust becomes a smaller favor as the binding value increases (Hypothesis 2.2). In our analyses, the coefficient of the binding value after the promise has been made is not significant, though indeed negative (model TW2 in Table 2.5). Thus, we do not find support for the idea that the promoting influence of promising to honor trust would depend on the binding value of the promise. However, the results show that the influence of making the promise tends to be more promoting with increasing transaction costs. This suggests that trustees might perceive placed trust as more kind and as a reward for having sacrificed high transaction costs (Hypothesis 2.2).

If the trustee explicitly omits the promise of trustworthiness ($\mathrm{TG}|\mathrm{H}_2^0$), but nevertheless gets the chance to decide about honoring trust, the influence of self-consistency competes with obligation feelings (Hypothesis 2.3). We discussed above that the mechanisms of cognitive dissonance reduction might undermine the feeling of obligation induced by placed trust and thereby decrease trustworthiness. In contrast to our Hypothesis 2.3, the results show that the coefficient for having omitted the promise is positive, though not significant (Table 2.5). This holds both for included influences of promise properties (model TW1; and see Snijders, 1996) and for the cases in which the omitted promise is cheap-talk (model TW2). The positive sign of the coefficient indicates the strength of the obligation feelings that are undermined by the desire for self-consistency. That the coefficient is not significant suggests that the two motivations indeed compete with one another. Nevertheless, due to the lack of significance, we cannot reject the part of our Hypothesis 2.3, which states that self-consistency would generally undermine the promoting influence of obligation feelings on trustworthiness after the promise has been omitted. However, our analysis reveals that

this lack of support only holds for cheap-talk promises and for the analysis in which the opposing influences of promise properties are not controlled. Considering the properties of the omitted promise, our results provide support for both the hampering influence of self-consistency and the promoting influence of obligation feelings.

We argued above that the binding value of the omitted promise promotes self-consistency, while transaction costs that would have been associated with making the promise strengthen obligation feelings (Hypothesis 2.3). If trust has been placed after a promise associated with high transaction costs has not been made, trustees might assume that trustors show their understanding. This would increase the trustee's feeling of obligation to return the favor of placed trust. We indeed find a tendency that the effect of placed trust after an omitted promise promotes trustworthiness as transaction costs increase that the trustee would have sacrificed. The influence of the omitted promise on trustworthiness is significantly positive for transaction costs $c \geq 7$ (b $= 2.29$ ($= 1.74 + 7 \cdot 0.18$), se $= 1.37$, Wald $\chi^2_{1\,\mathrm{df}} = 2.77$ with $p = 0.0961$). Moreover, we also find a tendency that the omitted promise actually significantly hampers trustworthiness for binding values $v_2 \geq 18$ (b $= -3.23$ ($= 1.74 - 18 \cdot 0.24$), se $= 1.95$, Wald $\chi^2_{1\,\mathrm{df}} = 2.74$ with $p = 0.0976$). This indicates that mechanisms of cognitive dissonance reduction might have been successful. As argued above, trustees might have convinced themselves that they would abuse trust and might even perceive placed trust negatively after having omitted the promise. The desire for self-consistency then undermines the unwanted feeling of obligation induced by placed trust. Thus, in line with Hypothesis 2.3, the findings indicate that the binding value hampers trustworthiness due to self-consistency, while transaction costs promote obligation feeling indicating that the trustee is indeed delighted about the trustors understanding. However, recall our discussion that trustees who actually intended to abuse trust might be more likely to refrain from promising trustworthiness the higher the binding value. Therefore, the hampering tendency of the binding value after the promise has not been made could also reflect a selection effect. Similarly, the positive moderating effects of transaction costs can likewise indicate a selection of more trustworthy trustees.

2.4.2 Analyses for Trustfulness

For the trustor's decision of whether or not to place trust, we distinguish three differently embedded TGs: the TG itself without behavioral context and the two subgames in the HTG after the trustee has decided whether or not to promise trustworthiness (TG$|\mathrm{H}_2^+$ and TG$|\mathrm{H}_2^0$). The results for trustfulness (Table 2.6) are presented in a way

similar to the results for trustworthiness (as described for Table 2.5). The period in which the trustor decides whether to place trust has a strong and highly significant negative effect on trustfulness (Table 2.6). Recall that no effect of the decision period on trustworthiness has been found (Table 2.5). The reason for this difference might be that trustors have experienced abused trust in previous encounters and therefore become more reluctant to place trust, whereas trustees only need to react to behavioral advances. In contrast to the analyses for trustworthiness, we also do not find support for the idea that properties of the promise would moderate effects of behavioral context on trustfulness (LR $\chi^2_{4\,\mathrm{df}} = 2.61$ with $p = 0.6257$; also see the results of model TF2 in Table 2.6). Separate likelihood-ratio tests of joint significance as reported for the analyses for trustworthiness also do not provide support for such moderating influences of promise properties (analyses not reported). Concerning the influence of the trustee's promise decision, we find that trustfulness in general differs significantly between the behavioral contexts (LR $\chi^2_{2\,\mathrm{df}} = 54.99$ with $p < 0.0001$).

We argued above that the trustee's promise to behave in a trustworthy manner is a friendly gesture, because it provides the trustor with the perspective of a gain (Hypothesis 2.4). Thus, trustors should feel an obligation to reward trustees by placing trust if trustworthiness has been promised ($\mathrm{TG}|\mathrm{H}_2^+$). Moreover, trustors might anticipate the increase in trustworthiness after the promise has been made (Table 2.5 and Hypothesis 2.2). Our results indeed show that trustors are significantly more likely to place trust after the promise has been made (Table 2.6). Note that the effect is not very strong. In a sense, this contrasts with the previous finding that trustees seek to keep their promises, irrespective of objective bonds (Table 2.5). As previously mentioned, we do not find support for the idea that the impact of receiving the promise on trustfulness would become more promoting with increasing binding value or with increasing transaction costs (Table 2.6).

Whereas receiving a promise of trustworthiness promotes trustfulness, the trustor should be more reluctant to place trust if a possible promise has not been made ($\mathrm{TG}|\mathrm{H}_2^0$) (Hypothesis 2.5). Our reasoning has been that trustors might perceive an omitted promise as unfriendly and retaliate by withholding trust. Moreover, trustors might anticipate reduced trustworthiness (given our reasoning for Hypothesis 2.3). We find indeed that trustfulness is strongly reduced after the promise has not been made (Table 2.6) (see also Snijders, 1996; Gautschi, 2000). In fact, the hampering impact of the omitted promise on trustfulness is in general significantly larger than the promoting influence of having received the promise (test of differences between absolute coefficients: Wald $\chi^2_{1\,\mathrm{df}} = 6.16$ with $p = 0.0131$ in model TF1 and Wald $\chi^2_{1\,\mathrm{df}} = 2.40$

Table 2.6: Logistic regression of trusfulness with fixed effects for subject-payoff response sets

(A) Regression coefficients

		TF1		TF2	
	Hyp.	b	se	b	se
Behavioral contexts					
TG		(ref.)		(ref.)	
TG|H_2^+	H_4: +	0.46^{*}	0.19	0.49^{*}	0.23
TG|H_2^0	H_5: –	-1.29^{***}	0.24	1.31^{**}	0.44
Binding value					
in TG|H_2^+	H_4: +			0.01	0.02
in TG|H_2^0	H_5: –			0.03	0.02
Transaction costs					
in TG|H_2^+	H_4: +			−0.02	0.02
in TG|H_2^0	H_5: +			−0.02	0.03
Past periods per game		-0.16^{***}	0.04	-0.16^{***}	0.05

(B) Likelihood-ratio tests

	χ^2	df	χ^2	df
LR test (control)	54.99^{***}	2	57.60^{***}	6
LR test (TF1)			2.61^{*}	4

(C) Pairwise comparisons (Wald tests)

	Δb	se	Δb	se
TG|H_2^0 – TG|H_2^+	-1.75^{**}	0.26	−1.80	0.47

N(response sets) = 560, N(decisions) = 212, N(subjects) = 118;
two-sided p-values: *** p = 0.001, ** p = 0.01, * p = 0.05, ° p = 0.1; (sub)games (0/1), past periods per game (1 . . . 12/14), binding value $v_2 = (0, 5, 10, 15, 30)$, transaction costs $c = (0, 5, 20)$. Likelihood-ratio tests are reported against the null model with period control and against model TF1.

with $p = 0.1214$ for cheap-talk promises in model TF2). The difference between the coefficients is not significant in the case of cheap-talk promises. This might be due to the small number of cases in which trustees have omitted a cheap-talk promise (9 of 79 cases (11.4%) in our analyses for trustfulness). The strong withdrawal of trustfulness suggests that trustors punish trustees for omitted promises.

An alternative reasoning mentioned above is that trustors believe that trustees become more reluctant to honor trust after the promise has been omitted. However, we do not believe this. Recall that participants in the experiments made decisions in some sets of (sub)games in the role of the trustor and in others with different payoffs in the role of the trustee. Thus, the analyses involve decisions of the same subject in both roles. Therefore, it is rather unlikely that a false consensus effect would be responsible for the generally and strongly hampering effect of omitted promises on trustfulness. In fact, one would expect that trustors then also anticipate the promoting impact of made promises on trustworthiness and the moderating influences of the properties of the omitted promise. However, the findings in our analyses suggest that anticipated influences hardly play a role. First, the negative coefficient of omitted promises is about three times larger than the positive coefficient of received promises (model TF1 in Table 2.6, test reported above), whereas it is the coefficient of given promises that is significantly more positive than the coefficient of omitted promises in the analyses for trustworthiness (model TW1 in Table 2.5, test reported in the text). Second, in contrast to the findings for trustworthiness (Table 2.5), we do not find support for moderating influences of the promise properties on trustfulness (Table 2.6). Given our findings, we conclude that it might be a strong feeling of indignation that induces trustors to seek revenge for omitted promises by withholding trust rather than the influence of an anticipated abuse of trust.

2.5 Summary and Perspectives

2.5.1 Summary of Basic Ideas, Approach, and Contributions

In this paper, we analyzed trustfulness and trustworthiness in different behavioral contexts created by preceding friendly and unfriendly decisions. We argued that two powerful social-psychological forces become relevant and give rise to process-based motivations. First, a feeling of obligation to return a favor is invoked by perceived kindness of others' decisions or, in the case of perceived unkindness of other's decisions, a feeling of indignation (Gouldner, 1960; Coleman, 1990: ch. 12; Cialdini, 2001: ch. 2). Second, the desire for self-consistency helps reduce cognitive dissonance (Festinger, 1957; Heider, 1958; Cialdini, 2001: ch. 3). Obligation and indignation constitute in-

tention-based motivations, while self-consistency is an intra-personal process-based motivation. Both forces can result in behavioral patterns of reciprocity. Placing trust can be perceived as a friendly advance that creates an obligation to be friendly in return. Concerning decision situations in which the trustee can promise trustworthiness, we expected that trustees would seek to behave consistently. This can strengthen the effect of felt obligations after the promise has been made or undermine obligation feelings. Concerning trustfulness, received promises induce obligation feelings, while explicitly omitted promises inflict indignation feelings. Finally, we expected that binding properties of promises and transaction costs associated with making a promise moderate the perceived kindness of behavioral advances.

In order to avoid specific assumptions about actors' outcome-based motivations (e.g., fairness or equity considerations), we designed our experiment as sets of one-shot (sub)games of identical extensive form. Decisions about placing trust were analyzed in three differently embedded Trust Games (TGs): the TG without behavioral context and twice as a subgame of Hostage Trust Games (HTGs), i.e., a TG after the trustee promised trustworthiness and a TG after the trustee omitted the promise. These TGs contain the decision to honor trust which constitutes a dichotomous Dictator Game (DG). Thus, we distinguished four differently embedded DGs: the DG without behavioral context and a DG as a subgame in each of the three TGs. Employing a within-subject design, each subject made decisions in such sets of (sub)games. In order to analyze our data, we used logistic regression with fixed effects for subject-payoff response sets (i.e., each set consists of decisions that one subject made in (sub)games with identical extensive form in different behavioral contexts). In doing so, we controlled for influences of various outcome-based motivations and for (additive) individual heterogeneity. Thus, differences in trustfulness and in trustworthiness between the behavioral contexts indicate the "pure" influence of preceding behavior on subsequent decision-making.

By combining several ingredients, we improve upon and extend previous research on behavioral advances (e.g., Snijders, 1996; McCabe et al., 2003; Cox, 2004). First, we suggested two kinds of process-based motivations as a theoretical basis that give rise to behavioral patterns of reciprocity: feelings of obligation or indignation and self-consistency. Second, focusing on trust situations, we incorporated an explicit promise option for trustees (more or less binding and costly) rather than less controlled discussions or exchanges of written messages. In order to reduce the ambiguity of decisions, we employed single encounters of binary-choice trust situations. This also conveniently reduces the number of (sub)games that represent the different

behavioral contexts. Third, we constructed sets of structurally identical (sub)games in order to study the "pure" effects of behavioral advances by controlling for various outcome-based motivations. We employed a within-subject design, which allows conclusions to be drawn on the individual level while controlling for unobserved (additive) individual heterogeneity. By eliciting actual and sequential decisions rather than hypothetical strategies we provide a more direct test and account for methodological insights gained from research on decision-making (e.g., concerning the activation of emotions and biases induced by probability perceptions). Fourth, we grouped our data into subject-payoff response sets and used logistic regression models with fixed effects for these subject-payoff response sets in order to analyze our data. In doing so, we take advantage of the power and specific features of our experimental design.

Our study provides evidence for reciprocity that is due to influences of preceding behavior on subsequent decision-making irrespective of changes in objective outcomes. The results support the theoretical reasoning that people are motivated by feelings of obligation or indignation and by the desire for self-consistency. Making or omitting a cheap-talk promise of trustworthiness even affects trustworthiness and trustfulness (except for influences of omitted promises on trustworthiness summarized below). Note that this finding contrasts with theoretical models in which perceived kindness is based on changes of objective outcomes induced by preceding behavior (for a discussion of a theoretical extension, see Chapter 4). The properties of the promise tend to moderate effects of behavioral contexts on trustworthiness, while we do not find support for such moderating influences concerning trustfulness.

Specifically, we found that trustfulness tends to promote trustworthiness. Similarly, making a promise to honor trust increases both trustfulness and, particularly strongly, trustworthiness. These findings suggest that trustors and trustees indeed feel obliged to return others' favors. For trustees, the desire for self-consistency also induces them to keep their promise. Moreover, the strong impact of having promised to honor trust on actually behaving in a trustworthy manner tends to become even more positive with increasing transaction costs associated with making the promise. This indicates that placed trust is perceived as a reward for the sacrificed transaction costs, and thereby increases the feeling of obligation. If a possible promise has not been made, trustfulness is strongly decreased, suggesting that trustors punish trustees for their unkindness. In the few cases in which trust nevertheless has been placed despite an omitted promise, we have not found support for the idea that having omitted the promise would generally hamper trustworthiness. However, the results show that trustees are motivated by both obligation and self-consistency. The influences of these

two motivations seem to cancel each other out in the cheap-talk case and if the opposing influences of promise properties are not controlled. Supporting the idea of such opposing effects, we found that promise properties tend to moderate the influence of omitted promises on trustworthiness. The binding value of the omitted promise promotes self-consistency such that the impact of the omitted promise on trustworthiness becomes hampering. Transaction costs that would have been associated with making the promise increase feelings of obligation to return the trustor's favor of placed trust despite the having omitted the promise.

2.5.2 Further Discussion and Perspectives

Some aspects of our study require further remarks concerning (1) statistical issues, (2) experimental design, (3) social-psychological assumptions, and (4) moderating effects of outcomes. First, in some behavioral contexts, only a small number of decisions is available in our data. For instance, few trustees who omitted the promise have the chance to decide whether to honor trust. This is due to the strongly negative effect that an omitted promise exerts on trustfulness. Changing the objective outcomes (e.g., reducing the trustor's possible losses and the trustee's temptation) typically yields a higher level of placed trust, which might give more room for finding effects of omitted promises. However, if trust is placed anyway, promising trustworthiness loses importance. Of course, using the fixed effects approach requires excluding response sets without variation which reduces the number of observations involved in the analyses. However, recall that the excluded observations provide no statistical within-subject information for testing our hypotheses about effects of behavioral advances anyway. Using other statistical models requires more assumptions, e.g., specifying outcome-based motivations (Chapter 3) or statistical assumptions about the distribution of unobserved effects.

Furthermore, we treat subject-payoff response sets as independent observations, which are, strictly speaking, nested in subjects and nested in experimental groups (sessions). Thus, a multilevel model with four levels would be desirable, combining fixed effects for subject-payoff response sets and random effects for subjects and for sessions. To our knowledge, efficient estimation procedures combining fixed and random effects do not yet exist. Moreover, additional assumptions about the distribution of random effects would be required, which we avoid in our fixed effects approach. However, it has been mentioned that the effect of behavioral context on behavior is assumed to be the same for all subject-payoff combinations in the statistical approach taken here. This is a strong homogeneity assumption and could be relaxed by allow-

ing the (sub)game coefficients to vary randomly with subject-payoff response sets. However, such analyses would require more observations.

Second, the design of our experiment involves a relatively large number of variations, including asymmetric payoff structures. To some extent, the results might be seen as more general because the results are based on decision situations with various parameters. However, the reported effects might differ between payoff combinations (see also the discussion point 4). We ignored this issue for reasons of restricted sample size and controlled for additive effects of objective outcomes. To some extent, fewer variations seem advisable for further experiments. However, fixing for instance the payoff parameters would have induced participants to focus exclusively on behavioral advances and choice options. This might reduce decision noise, but at the cost of a higher risk of response biases (e.g., participant awareness biases). In this respect, an advantage of our experiment is that participants also paid attention to payoff variations (reported in the section on the experimental design). In this sense, the relatively large number of variations in our experiment might strengthen the reported results.

Next, consider that the sets of (sub)games were based on outcome changes induced by properties of the promise. Thus, the only variation in trustors' payoffs was a high or low loss (based on S_1). This might have induced trustors to focus more on the trustees' payoffs than trustees might have been induced to take into account the trustors' outcomes. For further experiments, it seems advisable to vary the trustor's payoffs in a similar way. Such variation can be achieved by incorporating compensating values attached to a promise of trustworthiness. In addition to the variation argument, the moderating effect of compensation for trust and trustworthiness allows for further tests. For instance, trustworthiness might actually be reduced after a promise with a high compensating value has been made because placed trust might be perceived as a smaller favor given that the trustor risks a smaller loss.

Furthermore, we clustered the types of games in our experiment with fixed ordering, i.e., first TGs, then HTGs, and finally DGs. A short questionnaire assured a break between the TGs and the HTGs, and the relationship between HTGs and DGs is less obvious. Although decisions from different periods and games are compared within response sets, it is possible that the specific ordering also had an effect, e.g., if participants made increasingly selfish decisions in the course of the experiment. However, we chose this fixed ordering for of two reasons. First, we feared that decisions in DGs could most strongly affect subsequent play because outcome-based preferences are revealed. Therefore, we scheduled the DGs at the end of the experiment. Second, decision situations should not become simpler as it would have been the case with

an ordering like HTG—TG—DG. Among other methodological reasons, this would have revealed the nesting of (sub)games.

Some of the discussed issues concerning the experimental design arise specifically because we employed a within-subject design, but they would not be relevant in a between-subjects design (see also Keren, 1993; Putt, 2005). For the purpose of our study, a within-subject design appears to be more suitable because it allows us to analyze influences of motivations on the individual level and to control for (additive) individual heterogeneity and for influences of objective outcomes without making assumptions about specific outcome-based motivations. However, between-subjects designs have the advantage that practice effects and carryover effects can be ruled out. In fact, experiences in preceding decision situations might subsequently influence people's behavior toward other persons, even in a series of single encounters (e.g., due to indirect reciprocity, changes of mood and of beliefs induced by positive and negative experiences, or influences of group dynamics). For the study presented here, it would be of particular interest to control for positive and negative experiences of trustors and trustees in previous encounters. However, assessing perceived kindness of preceding encounters requires separate analyses (e.g., depending on outcomes and promise properties). Therefore, we decided to leave this issue for a future study and to control for the number of past decisions made without distinguishing the content.

Third, considering social-psychological research, we have simplified our arguments about the influence of feelings of obligation or indignation and about the influence of the desire for self-consistency on people's decision-making. People can also feel an obligation to return a favor because of normative expectations rather than perceived kindness. For instance, we mentioned that receiving unwanted gifts do not necessarily trigger positive feelings toward the giver (Cialdini, 2001). This might hold as well for requested or even enforced promises, which is an issue for further research. Moreover, it is not obvious what exactly "self-consistent behavior" is. For instance, we argued that lying generally causes some internal distress, which might be less problematic for trustees who perceive taking advantage of the trustor as legitimate in certain decision situations. Furthermore, recall that the promise properties have been helpful in disentangling the opposing effects of obligation and self-consistency on trustworthiness, whereas we cannot separate these influences in the decision situation that arises after the promise has been made. Thus, we do not yet know whether it is self-consistency, obligation, or a combination of the two that gives rise to the strongly promoting impact of making the promise on actually behaving in a trustworthy manner.

Investigating underlying motivations and psychological processes requires measuring whether an action is indeed perceived as kind or unkind, what emotions are triggered, what beliefs actors have, and how these beliefs are updated. In our experiment, we consciously decided to omit such questions because asking participants to consider the other person can influence decision-making. Some experiments report a bias towards other-regarding behavior (Gächter and Renner, 2006; Hoffman et al., 2008), whereas in other experiments indications that measuring beliefs promotes selfish behavior have been found (Croson, 2000). Thus, measuring beliefs influences people's decision-making, but the direction, the extent, and the conditions for such biases are still open questions that require further research. Note that separating the motives underlying self-consistent behavior is difficult in trust situations, but easier in other games (e.g., a sequential Prisoner's Dilemma). Furthermore, negotiation problems allow for the distinction between omitting a promise to behave in a friendly manner, which we argued to be unfriendly, and actually promising to behave in an unfriendly manner (for the Chicken Game, see Prosch, 2006). This distinction can also be made concerning announcements of sanctions, i.e., reward promises (friendly) and punishment threats (unfriendly) (Chapters 4 and 5).

Fourth, we focused on effects of the behavioral contexts and left out interactions with objective outcomes. As mentioned above (discussion point 2), extending our analyses by incorporating context-outcome interactions would require a larger number of observations for the fixed effects approach we employed. However, our theoretical reasoning suggests that the extent of perceived kindness depends on the size of outcome changes that are due to the specific choice of a behavioral option. Future research could shed more light on this. Further experiments could be also conducted in order to increase the sample size. This would allow us to keep the fixed effects approach for analyzing how the influence of the behavioral context depends on objective outcomes. Alternatively, other statistical approaches could be employed at the cost of making more assumptions (see discussion point 1). Moreover, other statistical models do not allow moderating effects of outcomes to be distinguished from effects of outcome-based motivations moderated by behavioral advances. Obviously, it is also worth studying how the influence of outcome-based motivations depends on the specific behavioral context. Support for such context-dependency contrasts with theoretical models which typically treat outcome-based motivations as individual constants (for this argument and empirical support, see Chapter 3). In analyses with subject-subgame response sets, context-payoff interactions represent how the influence of outcome-based motivations is moderated by the behavioral context.

Chapter 3

Temptation, Loss, and Promises of Trustworthiness

Experimental Evidence on Context-Dependency of Outcome-Based Motivations

This paper is based on previous work with Jeroen Weesie and benefited from his feedback and from discussions with him. I also thank Vincent Buskens for discussions and comments, for improvements in the Dutch version of the instructions used in the experiment and, together with Rense Corten, Dennie van Dolder, and Richard Zijdeman, for helping during the experiment. I am grateful for discussions with Michał Bojanowski and for a small statistical tool from Ben Jann and acknowledge comments made by Werner Raub and by participants at the "Behavioral Studies" colloquium at ETH Zurich in 2007.

Abstract

Other-regarding motivations influence people's behavior and form the basis for reciprocity. People respond to others' kind and unkind behavior due to feelings of obligation to return favors and of indignation about losses. People also derive emotional utility from others' outcomes, e.g., benevolence and spite are the basis for cooperative and competitive social orientations. Contrary to the idea of stable individual traits employed in formal models, the current study investigates interactions of outcome-based motivations with behavioral contexts. Data from a game-theoretical lab experiment are used that is designed as within-subject sets of single encounters in structurally identical Trust Games, Hostage Trust Games, and Dictator Games. This design allows for the comparison of influences of objective outcomes on trustfulness and on trustworthiness in different behavioral contexts while controlling for individual heterogeneity. The results provide evidence that outcome-based motivations differ between player roles and behavioral contexts.

3.1 Introduction

Trust is a basic ingredient in everyday life. Long-term relationships, such as friendships, family relations, or alliances between firms, are typically based on trust. Moreover, trust also plays an important role in single encounters with strangers and in situations in which sufficient and reliable exchange of information is lacking. For example, when traveling by train, we might ask another passenger to keep an eye on our luggage while we leave our seat. Later, we are expected to do the same in return when the other passenger leaves for some time. We might also wish to place our suitcase on the luggage rack above our heads in order to have more space to ourselves or to free the neighboring seat for another passenger. If the suitcase is too heavy for us alone, the other passenger might help to heave the suitcase onto the luggage rack. We then trust that he will also help us get our suitcase down again. Similarly, we have to trust specialists, such as doctors, mechanics, or lawyers, not to take advantage of our lack of knowledge or abilities. When making a purchase, we are required to trust the seller that the product quality is as advertised. If the product must be delivered to us, we also need to trust that we will receive it within a reasonable time. In these and many other social and economic interactions, people have to trust others in order to achieve some benefit. Trust can improve the situation for both parties involved. Still, trustfulness provides those who are trusted with an opportunity to take advantage of the situation, which inflicts harm on those who have trusted. However, people's decision-making is motivated not only by their interest in their own outcomes but also by feelings of happiness or spitefulness concerning others' outcomes. This can limit incentives for "opportunistic behavior" (Williamson, 1985).

For instance, consider the "train example" again. Assume that the other passenger has to change trains before we leave the train. The other passenger is then tempted to escape the trouble of helping to get our heavy suitcase down again. This holds especially, if there is only little time for him to catch his next train. In addition, he might be concerned about us ending up with the suitcase still on the luggage rack. His concern about us conflicts with his self-interest. Now imagine that he promised to help us. He might then feel bound to help us, both because he prefers to behave consistently with his promise and because he feels obliged toward us. His promise can therefore reduce the impact of his self-interest and increase his concern about the consequences for us if he reneged on his promise. This is especially true if he thinks that he shares some responsibility, given that we agreed to put our suitcase up only because he promised to help us get it down again. In turn, imagine that the other passenger explicitly told us that he does not promise to help us with the suitcase

again. We might think that he will help if he has time but in any case, we are then less inclined to agree to place our suitcase on the luggage rack, unless our altruistic tendencies drive us to do him a favor at our costs. Typically, the omitted promise involves unkindness and would aggravate our concern about the situation in which we have to find someone else who has the time to help us. Moreover, we might become displeased about the other's advantage, given that he seems to care little about the consequences for us.

The example shows that not only outcomes, but also preceding kind and unkind decisions can influence trustfulness and trustworthiness. As in the train example, people can indicate their trustworthiness in order to increase the chance of being trusted. Safeguards (e.g., warranties for products or fines for delivery delays), certain signals that require some effort (e.g., certificates), and even promises without an objective bond can be perceived as indications for a person's integrity. In Chapter 2, sociological and social-psychological insights about feelings of obligation to return others' favors or feelings of indignation about others' unkindness, and about the desire for self-consistency have been applied to trust situations (Gouldner, 1960; Coleman, 1990: ch. 12; Cialdini, 2001: chs. 2–3). These process-based motivations induce people to behave consistently and to reciprocate kind and unkind behavior. Moreover, the behavioral context that is created by preceding decisions can moderate people's concern about their own outcomes and about others' outcomes. *How does the behavioral context resulting from kind and unkind behavior moderate the effects of outcome-based motivations on trustfulness and trustworthiness?*

Data from the game-theoretical lab experiment conducted by (Vieth and Weesie, 2006; and see Chapter 2) are used in order to explore this question. The experiment is designed as within-subject sets of structurally identical (sub)games. This allows for the analysis of effects of outcomes on actors' decision-making in (sub)games that only differ with respect to the behavioral context created by kind and unkind preceding behavior. The present study contributes to previous research in several respects. First, most theoretical and empirical game-theoretical research on other-regarding motivations has focused on people's concerns with their own and others' outcomes, e.g., models with social orientations such as altruism or inequality aversion (e.g., Brew, 1973; Weesie, 1994a,b; Snijders, 1996; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). In theoretical models, such outcome-based motivations have primarily been treated as individual constants, stable over time and across decision contexts. In contrast to this assumption, social-psychological research has revealed that the influence of people's traits differs especially between decision situations and that people's

behavior is hardly correlated between different decision situations (for a review, see Kunda, 2002). The study presented here investigates the extent to which influences of outcome-based motivations differ between decision situations.

Second, an experiment testing inequality aversion as modeled by Fehr and Schmidt (1999) in structurally different decision situations reveals that the theoretical model lacks explanatory power at the individual level (Blanco et al., 2006). While Blanco, Engelmann, and Normann (2006) attribute these discrepancies to individual heterogeneity of (outcome-based) motivations, their results can be interpreted as indications for context-dependency of outcome-based motivations. Moreover, the authors only show that discrepancies between theoretical predictions and actual behavior exist, without investigating possible reasons. The present study focuses on structurally identical decision situations that are generated by kind and unkind preceding behavior. This focus aids in explaining differences in the influence of outcome-based motivations between behavioral contexts as implications of process-based motivations (i.e., obligation, indignation, and self-consistency). Moreover, a classical altruism model is informally applied in order to specify assumptions about outcome-based motivations, which allows influences of people's own outcomes to be separated from influences and of others' outcomes.

Third, theoretical models have been proposed in order to account for intention-based motivations that are triggered by evaluations of other's kindness (e.g., Rabin, 1993; Levine, 1998; Charness and Rabin, 2002; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). Nearly all of these theoretical models exclusively focus on intention-based motivations. The model proposed by Falk and Fischbacher (2006) incorporates outcome-based motivations, but these motivations are assumed to influence people's behavior only in decision situations in which intention-based motivations are not activated. In contrast, the present study investigates whether both outcome-based and intention-based motivations (or more generally, process-based motivations) simultaneously influence people's decision-making in a given decision situation.

Fourth, in the theoretical models on intention-based motivations, others' kindness is assessed by changes in objective outcomes. However, numerous experiments provide evidence that even cheap-talk promises can promote cooperative behavior (see also Chapter 2). The present study includes such cheap-talk promises and investigates how process-based motivations activated by making or omitting promises interact with motivations that are based on objective outcomes.

3.2 Reciprocity, Trust, and Promises of Trustworthiness

3.2.1 Reciprocal Behavior as an Implication of Other-Regarding Motivations and Self-Consistency

One fundamental principle in human interaction is reciprocity. *Reciprocity is a behavioral pattern of returning favors and retaliating for harm* (for reviews, see Fehr and Schmidt, 2006; Hann, 2006; Kolm, 2006; Lévy-Garboua et al., 2006). People help others who have helped them, and people become unfriendly toward others who have misbehaved toward them. The principle of reciprocity has been stressed in Scottish moral philosophy (Hume, 1739/1978; Smith, 1759/1976), cultural-anthropology (Malinowski, 1922; Mauss, 1950), social-psychology (Thibaut and Kelley, 1959), and in sociological theories of social exchange (Gouldner, 1960; Blau, 1964; Homans, 1974; Emerson, 1976; Coleman, 1990). Numerous studies have shown that people cooperate, reward others for cooperation and generosity, and punish others for non-cooperation and greediness, even if the other person is a stranger and even at people's own expense (for reviews, see Camerer, 2003: ch. 2; Ostrom and Walker, 2003; Kopelman et al., 2002; Kollock, 1998; Komorita and Parks, 1996; Ledyard, 1995; van Lange et al., 1992; Messick and Brewer, 1983; Pruitt and Kimmel, 1977). Reciprocal behavior can result from two types of *other-regarding motivations* that influence people's decision-making: outcome-based motivations and intention-based motivations (Fehr and Schmidt, 2006).

Outcome-based motivations shape the influence of objective outcomes on actors' decision-making. This idea is known as social (value) orientations, which are rooted in social comparisons, i.e., distributive preferences that transform objective outcomes into subjective utilities (e.g., Messick and McClintock, 1968; McClintock, 1972; Liebrand, 1984; also see Iedema, 1993). Various social orientations have been suggested and empirically identified (for reviews, see Au and Kwong, 2004; McClintock and van Avermaet, 1982). The basic assumption is that actors are not only interested in their own objective outcomes but also derive emotional utility from others' objective outcomes. In numerous studies, an actor's utility is modeled as the sum of the actor's own objective outcome and the other's outcome, which is individually weighted by an altruism parameter (Brew, 1973; Taylor, 1987/1976; Weesie, 1993, 1994a,b; Snijders, 1996). The altruism parameter is positive for reflecting pro-social orientations and negative for anti-social orientations. A positive altruism parameter indicates that people, e.g., enjoy others' well-being and suffer guilt from inflicting harm on others. A negative altruism parameter reflects feelings such as spite and

envy and that people, e.g., find joy in others' misfortunes. Selfishness or individualistic orientations arise as a special case from the absence of concerns with others' outcomes, which implies an exclusive focus on one's own outcomes.

The altruism parameter is often constrained in a way that actors are assumed to be at most equally interested in others' outcomes, which restricts self-destructive behavior. In this case, the altruism model allows for the distinction between cooperative, selfish, and competitive social orientations. Cooperators not only aim to maximize their own outcomes, but to some extent they also prefer to maximize others' outcomes. Fully cooperative actors equally value their own and others' outcomes and therefore seek to maximize the joint outcome. Actors who are competitive derive disutility from others' outcomes, such that they to some extent prefer minimizing others' outcomes while maximizing their own outcomes. Fully competitive actors dislike others' outcomes as much as they enjoy their own outcomes and seek to maximize the advantageous difference between their own outcomes and others' outcomes.

Now consider that sanctioning behavior can likewise have "self-destructive" elements if it inflicts costs upon the person who rewards or punishes others. Rewarding others can require altruistic inclinations beyond cooperation. Altruistic actors value others' outcomes more than their own outcomes. In contrast, punishing others can be based on aggressive tendencies beyond competition. Aggressive actors aim to minimize others' outcomes even if doing so does not increase their own outcome. The restrictive assumption about the range of the altruism parameter excludes that such forms of sanctioning can result from outcome-based motivations. Therefore, it seems more fruitful to allow for an unconstrained altruism parameter.

In formal theoretical models, social orientations are typically incorporated as individual constants that represent personal traits that are individually stable across time and contexts and thus define different types of actors (e.g., cooperators, individualists, and competitors). However, social-psychological research indicates some degree of temporal stability of personal traits, but it reveals that cross-situational consistency is minimal (for a review, see Ross and Nisbett, 1991: ch. 4). Measures of social traits have been found to be predictive on an aggregate level, i.e., a person's behavior is measured in a variety of situations in order to predict that person's average behavior across situations. This aggregation approach is useful for inferring behavioral trends, but it does a poor job in predicting people's behavior in a specific situation. Moreover, the aggregation approach "completely overlooked the fact that different situations could draw out different behaviors from different people, and made it impossible to assess each individual's unique pattern of behavior as it varied from

one situation to another" (Kunda, 2002: 422). This sheds light on two puzzles that are related to game-theoretical research. First, people's answers to survey questions about attitudes and opinions typically do not explain the behavior elicited in specific experimental decision situations (e.g., Burks et al., 2003) or indicate an influence of people's general world view (e.g., a positive relationship between self-reported general trustfulness in a questionnaire and actual trustworthy behavior in an experiment has been found, see Glaeser et al., 2000).

Second, game-theoretical models involving outcome-based motivations seem to predict behavior quite well on an aggregate level, but they fail on the individual level. This has been revealed in an experiment by Blanco, Engelmann, and Normann (2006) that tested the predictive power of inequality aversion as modeled by Fehr and Schmidt (1999). The results indicate that a person behaves differently in different games (Ultimatum Game, Dictator Game, Sequential Prisoner's Dilemma, and Public Goods Game) and that for the most part a person's behavior in one game cannot be predicted by the person's behavior in another game. Blanco, Engelmann, and Normann (2006: 37) suggest "that the success of the inequality aversion model at the aggregate level could be based on an ability to qualitatively capture different important motives in different games but that the low predictive power of the model at an individual level is driven by the low correlation of these motives within subjects". It seems reasonable to assume that the influence of outcome-based motivations (e.g., inequality aversion parameters in their experiment or the altruism parameter described above) not only differs individually but also depends on the specific context in which a decision is made. Obviously, structurally different decision situations constitute different decision contexts (also see an experiment by McClintock and Liebrand, 1988). However, previous behavior within a decision situation likewise changes the context in which a certain decision is made. For instance, the first-mover's choices in the Ultimatum Game and in the Sequential Prisoner's Dilemma generate a behavioral context for the second mover's decision. Depending on such behavioral contexts, further motivations can be activated (e.g., intention-based motivations) that can alter the influence of an actor's outcome preferences.

Intention-based motivations are directly rooted in the principle of reciprocity. The basic idea is that intention-based motivations are process-based motivations that are invoked by others' behavior. Actors consider information about others' behavioral options, i.e., about behavioral processes of how certain outcomes are obtained (for experimental studies see, e.g., Snijders, 1996; Gallucci and Perugini, 2000; Gautschi, 2000; Brandts and Solà, 2001; Falk et al., 2003; McCabe et al., 2003; Cox, 2004;

Charness and Rabin, 2005; and see Chapter 2). Received favors create a "shadow of indebtedness" until the favor is repaid (Gouldner, 1960: 174; also see Coleman, 1990: ch. 12). Outstanding obligations intrinsically demand repayment by causing emotional tension in a person omitting or delaying to fulfill these obligations. People experience a feeling of obligation to return a favor, even if the received favor is unwanted (Cialdini, 2001: ch. 2; Coleman, 1990: ch. 12). Similarly, inflicted harm causes feelings of indignation that induce people to retaliate if the other could have avoided the harmful action. For instance, people become unfriendly toward others who behave impolitely, especially if others' impoliteness is perceived as unjustified. People evaluate others' behavior in terms of kindness, which triggers positive and negative emotions toward others. Perceived kindness and unkindness depend on intentionality and on the size of outcome changes caused by others' behavior (e.g., Rabin, 1993; Levine, 1998; Charness and Rabin, 2002; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). For instance, the larger the benefit that others provide, the more their behavior is perceived as kind. Perceived kindness increases as others incur more sacrifices in order to provide the benefit. Next, others' behavior can only be kind or unkind if alternatives resulting in different outcomes were available. Falk and Fischbacher (2006) combined outcome-based and intention-based motivations in their model, proposing that outcome-based motivations shape people's behavior in cases in which the other person has no alternative option that would allow for intentionally kind or unkind decision-making. By now, models that account for intention-based motivations are based on influences of outcomes that shape the perception of others' kindness. However, even a promise without changing objective outcomes can be perceived as a kind advance (Cialdini, 2001) that is reciprocated (see Chapter 2) and influences people's motivations. The felt obligation to reciprocate others' kind behavior and the felt indignation about others' unkind behavior can directly influence people's decisions and can also moderate the effect of motivations based on objective outcomes. For instance, people might feel guilt about taking advantage of another person and might feel even more guilt if they previously received a favor from that person (Cialdini, 2001; Gass and Seiter, 2007).

In situations that involve multiple sequential decisions by an actor, the *desire for self-consistency* plays a special role. In contrast to intention-based motivations (e.g., obligation or indignation), self-consistency is an intra-personal process-based motivation. People suffer from cognitive dissonance (Heider, 1944, 1958; Festinger, 1957; Aronson, 1992; Akerlof and Dickens, 1982) if their behavior is inconsistent with their beliefs, attitudes, or previous decisions (for reviews see, e.g., Webster, 1975; Cialdini,

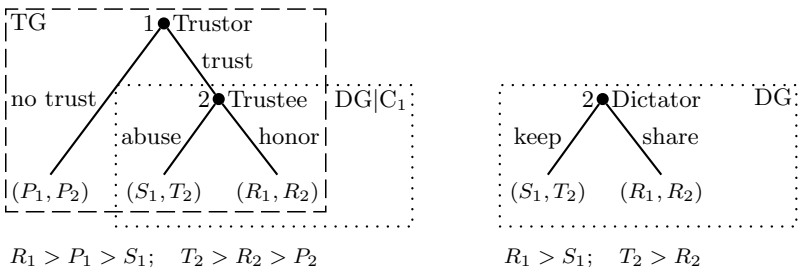
2001: ch. 3; Kunda, 2002; Gass and Seiter, 2007). For instance, salesmen increase the sales rate of energy-saving electrical equipment by first asking potential customers whether they are concerned about environmental protection. People then feel bound to prove that they do indeed care about environmental issues by buying the new product. When people have agreed to do something, they tend to later behave in accordance with their agreement, even if they discover hidden costs (for examples, see Cialdini, 2001; Gass and Seiter, 2007). Results of experiments on post-decisional attitude change indicate that people adjust their beliefs and opinions in a way that they favor the chosen alternative (e.g., Brehm, 1956). People use different methods in order to reduce cognitive dissonance and to maintain the impression of self-consistency (for an overview see, e.g., Gass and Seiter, 2007: 58). Among the identified methods are, e.g., attitude change, bolstering (coming up with good reasons supporting a certain decision), and denial (denying or ignoring issues causing inconsistencies). An instructive example is the tale of the fox that persuaded himself that the grapes that he could not reach looked sour, rather than to continue longing for them. If the self-persuasion is successful, the fox would ignore grapes falling down in front of his nose. Some foxes would even angrily crush the counter-evidence that threatens their peace of mind. If attitudes change, the fox will begin to dislike grapes altogether. When dealing with cognitive dissonance, people to some extent refuse to accept evidence that their belief about something is wrong. When people have negative feelings or prejudices toward another person, they tend to interpret even kind advances of the other in a negative way. In turn, this also holds for positive feelings: a beloved person can do no wrong. Note that the evidence of how people strive to maintain an impression of self-consistency does not conflict with the findings that people are quite inconsistent as far as general traits are concerned. Rather, by using various rationalization and justification methods in order to avoid or reduce cognitive dissonance, the desire for self-consistency alters influences of outcome-based motivations and intention-based motivations.

3.2.2 Effects of Outcomes and Behavioral Context in Trust Situations

Trustworthiness and Trustfulness in the Trust Game

The basic structure of trust situations can be described by the *Trust Game* (TG) (Dasgupta, 1988; Kreps, 1990) (Figure 3.1a). Two actors are involved: the trustor (player 1) and the trustee (player 2). Both actors would receive a greater benefit from honored trust than from no trust placed ($R_i > P_i$, with $i = 1, 2$). However, the trustee has an incentive to abuse trust ($T_2 > R_2$), while the trustor incurs a loss in

Figure 3.1: Trust Game (TG) and dichotomous Dictator Game (DG)



a) Trust Game **b)** Dictator Game

Figure repeated from Chapter 2.

that case ($P_1 > S_1$). Therefore, trustors would withhold trust because trustees would abuse placed trust. This is the result of the classical analyses assuming that actors' utility coincides with the outcomes. If outcomes are defined in objective terms (e.g., certain amounts of money), such an analysis implies the assumption that actors are selfish in the sense that they only care about their own objective outcome (i.e., the altruism parameter equals zero). Experimental studies report systematical evidence against the selfishness assumption: trustors often do place trust, and trustees often do honor trust (e.g., Snijders, 1996; Camerer, 2003: ch. 2; Ostrom and Walker, 2003).

Non-selfish trustees are not only concerned with their own outcomes, but also with the trustors' outcomes. Trustees receive the full benefit of T_2 from abusing trust or the shared benefit R_2 from honoring trust. Thus, the trustee's own relative gain ($T_2 - R_2$) constitutes the trustees' temptation to abuse trust, which hampers trustworthiness. The trustee's other-regarding utility component is the weighted objective outcome of the trustor. Trustees with a cooperative motivation (positive altruism parameter) derive some additional utility from S_1 after abused trust and some more from R1 after honored trust. If the weighted relative loss of the trustor ($R_1 - S_1$) that would be inflicted by abused trust outweighs the trustee's temptation, the trustee honors trust. Thus, the trustor's loss should have a positive influence on trustworthiness if trustees are sufficiently cooperative. Considering that trustees' other-regarding motivations vary individually, the hampering impact of the trustee's temptation might be stronger than the promoting impact of the trustor's loss. Note that Snijders (1996) uses the label "temptation" for a guilt index (i.e., the trustee's gain from abusing trust relative to the advantageous outcome inequality caused by abusing trust). The altruism model

underlying the approach taken here allows influences of people's own outcomes to be separated from influences of others' outcomes.

> **Hypothesis 3.1: Honoring trust and outcome-based motivations**
> Trust is *less* likely to be honored, the higher the trustee's temptation ($T_2 - R_2$), and trust is *more* likely to be honored, the larger the trustor's loss ($R_1 - S_1$).

The trustee's decision of whether or not to honor trust in the TG means sharing some benefit. Separating this part of the TG yields a *dichotomous Dictator Game* (DG), with the trustee in the role of the dictator and the trustor in the role of the receiver (Figure 3.1b). Trustees with purely outcome-based motivations do not take into account that the trustor placed trust in the TG. For these trustees, it makes no difference whether they honor trust in the TG or share benefits in the DG (see Chapter 2; McCabe et al., 2003; Cox, 2004). However, as previously argued, intention-based motivations arise from the specific behavioral context of a decision situation and can give rise to differences between a trustee's choice and a dictator's choice. Trustees are in the advantageous position solely due to the trustor's decision to place trust. In a sense, the trustor gave the power to control the outcomes to the trustee and made himself dependent on the trustee (Coleman, 1990: ch. 5). Whereas the trustor can lose from placing trust, the trustee only gains from it ($P_2 < R_2 < T_2$). Thus, placed trust can be perceived as a friendly advance and feelings of obligation to return the favor can induce trustees to respond kindly. Experimental results indeed show that the mere act of placing trust tends to promote trustworthiness (see Chapter 2; McCabe et al., 2003; Cox, 2004).

The feeling of obligation can also indirectly influence the trustee's decision by moderating effects of outcome-based motivations. Perceived favors cause positive feelings toward the other person (i.e., "warm-glow") that increase the concern about the other's well-being. Trustees might therefore become more motivated to avoid inflicting a loss upon the trustor (i.e., the altruism parameter becomes more positive for cooperative trustees and less negative for competitive trustees). Moreover, warm-glow shifts the focus away from one's own outcomes and towards others' outcomes. People feel less tempted to strive for their own objective gain when there is an outstanding obligation to fulfill. Therefore, the temptation to abuse trust might become less important for trustees. Note that these emotional processes also reduce cognitive dissonance, which would arise from regret about the foregone gain from abused trust.

Hypothesis 3.2: Honoring trust and kindness of placed trust
Compared to honoring trust in the TG, gains are *less* likely to be shared in the DG. Moreover, the effect of the dictator's temptation $(T_2 - R_2)$ on generous sharing in the DG is more hampering than the effect of the trustee's temptation on honoring trust in the TG. Similarly, the effect of the receiver's loss $(R_1 - S_1)$ on generous sharing in the DG is less promoting than the effect of the trustor's loss on honoring trust in the TG.

Now consider the decision of trustors. Typically, people are assumed to become more reluctant to trust, the more they can lose or the more the other can gain from abusing trust. The first component is the trustor's loss $(R_1 - S_1)$ that would be inflicted by abused trust, and the second component is the trustee's temptation $(T_2 - R_2)$ to abuse trust. These two components are the basis for the trustor's selfish and other-regarding outcome-based motivations. The trustor's interest in his own outcome motivates him to avoid losses. This decreases trustfulness as the trustor's possible loss increases. The trustor's social component of outcome-based motivation is based on the trustee's temptation. Since the trustor incurs a loss if trust is abused, he might feel spiteful and angry about the trustee's gain (competitive motivation with negative altruism parameter). Trustfulness then decreases with the trustee's temptation. Trustors can also have a cooperative motivation (positive altruism parameter) and enjoy the trustee's gain despite the trustor's own loss. For instance, when the trustee could gain considerably if the trustor sacrificed a small loss, a cooperative motivation seems quite likely. However, given a substantial loss for trustors and moderate gains for trustees in the case of abused trust, it seems reasonable to assume that the majority of trustors feel spite rather than benevolence. The positive impact of some cooperatively motivated trustors is unlikely to outweigh the hampering effect of many competitively motivated trustors. Therefore, it seems reasonable to assume that the influence of the trustee's temptation in the trustor's other-regarding motivation might hamper trustfulness.

In addition to the direct influences, loss and temptation also shape the trustor's belief about trustworthiness, i.e., about the trustee's motivations. First, consider the trustor's loss. As previously argued, the trustor's loss constitutes the basis for the trustee's other-regarding component of his outcome-based motivations. The influence of the trustor's loss on trustworthiness is promoting only if trustees are sufficiently cooperative (Hypothesis 3.1). Since the trustee's motivations can be heterogeneous, the trustors' beliefs about the influence of their loss are presumably likewise diverse.

Therefore, it seems unlikely that a trustor's belief about a promoting effect of the trustor's loss on trustworthiness outweighs the hampering impact that arises from the trustor's interest in his own outcome. Now consider the trustee's temptation. The trustee's temptation hampers trustworthiness motivated by the trustee's interest in his own outcomes (Hypothesis 3.1). This influences the trustor's belief such that trustfulness likewise decreases with an increase in the trustee's temptation. Thus far, the impact of the temptation on trustor's beliefs is much less ambiguous than the impact of the loss. However, consider that trustors' beliefs are not only restricted to outcome effects but are also shaped by expected intention-based motivations. Trustors also know that placing trust can invoke feelings of obligation in trustees to return the favor by honoring trust. As argued above, this can reduce the hampering impact of the temptation on trustworthiness and motivate the trustee to honor trust (Hypothesis 3.2). Therefore, the trustor's worries about the hampering effect of the trustee's temptation are likewise mitigated if they believe in the power of reciprocity.

The direct influence of the trustor's outcome-based motivations cannot be disentangled from the indirect influence resulting from beliefs about the trustee's outcome-based motivations. However, direct influences are typically stronger than those that are indirect. The reason is that beliefs about the trustee's motivations require more cognitive steps of reasoning and that trustors are more certain about the evaluation of their own motivations. This especially holds for other-regarding motivations, because they can take various forms. Moreover, the trustor's belief about the impact of the trustee's temptation on trustworthiness is ambiguous due to induced feelings of obligation. Thus, it seems reasonable to assume that the trustor's loss $(R_1 - S_1)$ hampers trustfulness motivated by the trustor's interest in his own outcome, and that the trustee's temptation $(T_2 - R_2)$ hampers trustfulness because the trustor is spiteful about the trustee's gain in the case of abused trust rather than motivated by beliefs about reduced trustworthiness.

Hypothesis 3.3: Placing trust and outcome-based motivations
Trust is *less* likely to be placed, the larger the trustor's loss $(R_1 - S_1)$, and *less* likely to be placed, the higher the trustee's temptation $(T_2 - R_2)$.

Note that assuming spitefulness on the part of the trustor rather than benevolence, which is assumed for the trustee, implies that outcome-based other-regarding motivations also depend on the player role. The position in which an actor makes a decision constitutes a certain decision context that shapes people's motivations and normative expectations (for a review of sociological role theories, see Biddle, 1986).

Promises of Trustworthiness

The trustor's loss and the trustee's temptation are assumed to hamper trustors to place trust, whereas the trustor's loss can positively influence the trustee's decision to honor trust based on guilt feelings. Accounting for other-regarding motivations, the impact of trustors' beliefs and of how beliefs are influenced by outcomes is somewhat unclear (see the argumentation for Hypothesis 3.3). After all, trustfulness depends on the possibility of trustworthiness. Since trustees gain from placed trust, they have an incentive to behave in a way that promotes trustfulness by influencing trustors' beliefs and motivations. One possibility is to promise trustworthiness. *Promises are expressed intentions to perform a certain action that yields a gain to the other person.* In social interactions (as in the train example outlined in the introduction), people typically make promises without objective bonds and others rely on such promises. For instance, if the other passenger in the train promises to later also help us get our suitcase down, we tend to be even happy to lift our heavy suitcase onto the luggage rack so that it is out of the way. However imperfect objectively, a promise is an indication of trustworthiness and powerful in promoting trustfulness and trustworthiness (see Chapter 2; Cialdini, 2001; Snijders, 1996). Given intrinsic bonds that arise from the desire for self-consistency, a promise serves as a *commitment*, i.e., as a *"voluntary strategic action", costly or not, with the purpose of "reducing one's freedom of choice" or of changing the outcomes* (Schelling, 1960). Objective incentives can enhance the credibility of a promise. For instance, sellers who provide guarantees for products are less tempted to sell bad quality products and increase the buyers' willingness to buy an expensive product. Moreover, sellers bound to pay a fine for delivery delays have an increased incentive to deliver within the stipulated time. In turn, buyers are more willing to engage in expensive transactions, even if the fine would not be paid to them, but to an external agency. Furthermore, the mere action of making a promise or of making objectively binding or compensating agreements (e.g., formal contracts) can be associated with (transaction) costs on the part of the trustee.

The *Hostage Trust Game* (HTG) proposed by Raub (1992) describes trust situations in which the trustee chooses whether or not to post a commitment prior to the TG (Figure 3.2) (see also Weesie and Raub, 1996). A commitment is a "strategic move" by which an actor voluntarily offers a "hostage" in the sense of a bond (Schelling, 1960). Promises of trustworthiness are commitments posted by a trustee in order to promote trustfulness. Making the promise can be associated with transaction costs (c) that the trustee loses even if the promise would not induce the trustor to place trust. The promise can also be combined with something valuable to the

Figure 3.2: Hostage Trust Game (HTG)

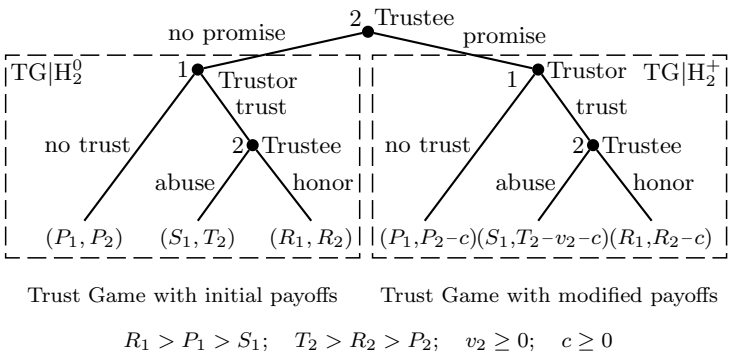


Figure repeated from Chapter 2.

trustee that will be lost if the trustee abuses trust. Such a value (v_2) of the promise to some extent binds the trustee to honor trust. The transaction costs and the binding value are properties of the promise and modify the subsequent outcomes. By deciding whether or not to make a promise, the trustee chooses the subsequent context of the interaction. If the promise is associated with objective properties (e.g., with a binding value or with transaction costs), the trustee chooses between the initial TG (TG|H$_2^0$) and a TG with modified outcomes (TG|H$_2^+$). In the HTG, the trustor is informed of the properties of the promise and of the trustee's decision about making the promise.

Formal game-theoretical analyses of commitments in trust situations are primarily based on the classical selfishness assumption (for TG and Prisoner's Dilemma, see Weesie and Raub, 1996; Voss, 1998b; Raub and Weesie, 2000; Raub, 2004; and including other-regarding motivations, Snijders, 1996). Assuming that actors are largely motivated by their own objective outcomes, trustees promise trustworthiness, trustors subsequently place trust, and trustees honor trust if the commitment is perfectly binding ($v_2 > T_2 - R_2$) and commitment posting is affordable ($c < R_2 - P_2$). Experimental studies indicate that other motivations than selfishness also play a role. First, imperfectly binding commitments promote placing and honoring trust and even small transaction costs hamper commitment posting (Yamagishi, 1986; Raub and Keren, 1993; Mlicki, 1996; Snijders, 1996). Second, even free communication without an impact on objective outcomes ("cheap-talk") promotes trustworthiness and trustfulness (Sally, 1995; Crawford, 1998; Kopelman et al., 2002; Bicchieri, 2002; Shankar and Pavitt, 2002; Ostrom and Walker, 2003; Brosig, 2006). Most experiments employed open face-to-face discussions. One main finding from these studies is that communication

promotes cooperation if the decision situation is discussed (rather than socializing by talking about an unrelated topic) and if people explicitly promise to perform a certain behavior (Dawes et al., 1977). In addition to uncontrolled exchanges (in face-to-face discussions or by written messages), pre-defined promise options have been studied in order to isolate effects of cheap-talk promises from other influences transmitted in face-to-face discussion. The findings provide evidence for strongly promoting influences of mere cheap-talk promises on cooperative behavior (Brandts and Charness, 2003; and on repeated interactions Bochet and Putterman, 2007).

Recall that outcome-based motivations are based on the trustee's temptation to abuse trust and on the trustor's loss caused by abused trust. As argued above, the trustee's temptation might hamper trustworthiness and trustfulness, whereas the trustor's loss might promote trustworthiness and hamper trustfulness (Hypotheses 3.1 and 3.3). Making a promise associated with some objective binding value (v_2) reduces the trustee's temptation to abuse trust ($T_2 - v_2 - R_2$). Thus, even imperfectly binding commitment properties should promote trustworthiness and trustfulness. Transaction costs (c) are subtracted from the trustee's outcomes irrespective of subsequent decisions and therefore only hamper promise-making as far as outcome-based motivations are concerned. These arguments are based on changes of objective outcomes that are induced by the properties of the promise. In the approach taken here, the trust situation after the promise has been made ($\mathrm{TG}|\mathrm{H}_2^+$) is compared to the trust situation in which no promise opportunity is available (TG), but objective outcomes are identical ($T_2^{'} - R_2 = T_2 - v_2 - R_2$). Thus, outcome-based arguments can explain neither a difference in behavior nor a difference in effects of outcomes on behavior. However, promises can affect trustfulness and trustworthiness by activating intention-based motivations and the desire for self-consistency (see Chapter 2). As previously argued, these motivations can also interact with outcome-based motivations.

Moderating Effects of Promises and Trustfulness on Trustworthiness

In Chapter 2, evidence has been provided that trustworthiness is increased after a promise has been made ($\mathrm{TG}|\mathrm{H}_2^+$) compared to the decision situation in which no promise option is available (TG). The arguments are based on the two social-psychological processes of self-consistency and obligation to return a favor. First, after the trustee has promised trustworthiness, the desire for self-consistency drives the trustee towards keeping his promise because lying inflicts intrinsic distress. Second, placed trust can be perceived as a friendly advance because the trustee always benefits from it, whereas the trustor risks a loss. This creates a feeling of obligation to

return the favor, which induces the trustee to honor trust. Third, by promising trustworthiness, the trustee shares some responsibility for the trustor's decision to place trust. Therefore, the trustee's desire for self-consistency strengthens the promoting influence of obligation feelings.

The positive influence of promising trustworthiness and subsequently being trusted on trustworthiness can also interact with the effects of outcome-based motivations. The desire for self-consistency requires the trustee to live up to his promise and to forgo the gain from abusing trust. Therefore, self-consistency should undermine the extent to which the trustee is concerned with his own outcome after he has promised to honor trust. This mitigates the hampering impact of the trustee's temptation on trustworthiness. Similarly, feelings of obligation also undermine the hampering impact of the trustee's temptation. Moreover, recall that feelings of obligation to return the favor of placed trust can be associated with warm-glow feelings that likewise mitigate the hampering influence of the trustee's temptation and foster the promoting impact of the trustor's loss (see the arguments for Hypothesis 3.2). The same argument applies to the decision situation considered here in which the trustor decides after the trustee has promised to honor trust. Moreover, due to warm-glow feelings, the desire for self-consistency might enhance the positive influence of obligation feelings. Note that making a promise might also give rise to spitefulness, especially if the trustee incurred high transaction costs. However, it seems unlikely that trustees become spiteful because this would conflict with the desire for self-consistency. Spite would increase cognitive dissonance rather than facilitate behaving consistently with the promise made. Alternatively, it is also possible that the promise of trustworthiness shifts the trustee's focus on behaving consistently and thereby undermines outcome-based motivations in general. Thus, self-consistency could also reduce the concern about the trustor's outcomes and, thereby, any impact of the trustor's loss. However, given that warm-glow facilitates promise-keeping, it seems reasonable to assume that the impact of the trustor's loss on trustworthiness is more promoting after the promise has been made.

Hypothesis 3.4: Honoring trust after the promise has been made
Compared to the TG (i.e., without promise opportunity), trust is *more* likely to be honored after trustworthiness has been promised ($\mathrm{TG}|\mathrm{H}_2^+$). Moreover, the effect of the trustee's temptation ($T_2' - R_2$) on honoring trust is *less hampering*, and the effect of the trustor's loss ($R_1 - S_1$) on honoring trust is *more promoting*, after trustworthiness has been promised ($\mathrm{TG}|\mathrm{H}_2^+$) than in the TG.

Now consider the decision situation in which the trustor decides whether to place trust after the trustee has omitted to promise trustworthiness ($TG|H_2^0$). In Chapter 2, it has been argued that self-consistency should hamper trustworthiness, whereas obligation feelings induced by the trustor's decision to nevertheless place trust promote trustworthiness. Thereby, self-consistency undermines feelings of obligation because mechanisms of cognitive dissonance reduction might induce the trustee to perceive placed trust in a negative way after the promise has been explicitly omitted (e.g., as an unintelligent mistake rather than as a kind advance, or even as an attempt to induce an unwanted feeling of obligation). Due to the competing influences of the two motivations, trustworthiness should be decreased after the promise has not been made compared to the decision situation in which the trustee does not have an opportunity to make a promise. Empirically, no support has been found that omitting the promise would generally hamper trustworthiness (see Chapter 2). It has been argued in Chapter 2 that this indicates the power of induced obligation such that the opposing influences of self-consistency and obligation feelings cancel each other out. In fact, the properties of the promise revealed the opposing influences of both self-consistency and obligation.

Self-consistency and obligation feelings can also moderate the trustee's outcome-based motivations. Self-consistency allows for increased selfishness, which induces the trustee to focus on his own outcomes. This would aggravate the hampering effect of the trustee's temptation on trustworthiness. As previously indicated, trustees might have "bolstered" their decision to omit the promise with good reasons. This convinced them to abuse trust or to believe that trust would not be placed (see Chapter 2). Moreover, trustees might perceive abused trust as legitimate after they have omitted the promise. Trustees might also feel confused or even irritated that the trustor nevertheless placed trust (see Chapter 2). This can induce trustees to abuse trust because they are annoyed by the cognitive dissonance caused by placed trust despite the promise has been explicitly omitted. The same holds for trustees who have been convinced that trust would be withheld. Denying the counter-evidence is one method of reducing cognitive dissonance (see the section on self-consistency in the general theory part). Negative emotions also undermine a positive concern about the trustor's outcome, which weakens a promoting impact of the trustor's loss ($R_1 - S_1$). However, the negative influence of self-consistency competes with a positive influence that arises from the obligation feelings induced by placed trust.

Given the finding in Chapter 2 that the opposing influence of both motivations simultaneously affects trustworthiness, self-consistency and obligation feelings can also

moderate the trustee's outcome-based motivations in opposite ways. Self-consistency mainly influences the trustee's selfish outcome-based motivation and thereby could also reduce the positive impact of the other-regarding outcome-based motivation. In contrast, obligation feelings mainly influence the trustee's other-regarding outcome-based motivation. As a side-effect, the influence of the trustee's selfish motivation could become less hampering. Thus, concerning the moderating effect of the two process-based motivations on the impact of the trustee's temptation, a strong influence of self-consistency competes with a weaker influence of obligation feelings. Concerning the trustor's loss, a strong influence of obligation feelings competes with a weaker influence of self-consistency. Thus, the impact of the trustee's temptation on trustworthiness might be aggravated due to the desire for self-consistency, whereas the impact of the trustor's loss might be more promoting due to feelings of obligation.

Hypothesis 3.5: Honoring trust after the promise has been omitted

Compared to the TG (i.e., without promise opportunity), trust is *less* likely to be honored after a possible promise of trustworthiness has not been made ($\mathrm{TG}|\mathrm{H}_2^0$). Moreover, the effect of the trustee's temptation ($T_2 - R_2$) on honoring trust is *more hampering*, but the effect of the trustor's loss ($R_1 - S_1$) on honoring trust is *more promoting*, after trustworthiness has not been promised ($\mathrm{TG}|\mathrm{H}_2^0$) than in the TG.

Moderating Effects of Promises on Trustfulness

The trustee's decision of whether or not to promise trustworthiness also generates a behavioral context for the trustor's subsequent decision whether to place or to withhold trust. First, consider again the decision situation that arises after the trustee has made the promise to honor trust ($\mathrm{TG}|\mathrm{H}_2^+$). As argued in Chapter 2, making a promise involves indications of kindness because of positive prospects. Trustworthiness provides the trustor with gains ($R_1 > P_1$). Therefore, promising trustworthiness invokes feelings of obligation to place trust. Moreover, promises of trustworthiness imply a concession by trustees to forego some larger gains that would inflict harm upon trustors. In Chapter 2, empirical evidence has been provided that trustfulness indeed increases due to the trustee's promise of trustworthiness. Of course, trustors might not be entirely convinced by promises that lack objective credibility. However, even doubts might be compensated by obligation feelings. Moreover, trustors might anticipate the promoting effect of promising trustworthiness on honoring trust (see the arguments for Hypothesis 3.4).

Concerning the trustor's outcome-based motivations, recall that the trustor's loss ($R_1 - S_1$) might hamper trustfulness in the TG because of the trustor's self-interest, and that the trustee's temptation ($T_2' - R_2$) might likewise be hampering because of spite on the part of the trustor (Hypothesis 3.3). Compared to the TG, the trustee promised to honor trust, which is a kind advance that inflicts obligation feelings upon trustors and promotes trustworthiness (see Chapter 2). As previously argued, positive feelings toward the other person ("warm-glow") facilitate returning a favor (Hypothesis 3.2). Such warm-glow feelings would also reduce the trustor's spite over the trustee's gain and the trustor's concern with his own loss. Warm-glow feelings would increase with the trustee's sacrifices, i.e., the binding value reducing the trustee's temptation and the transaction costs the trustee incurred to make the promise. However, whereas warm-glow feelings in the TG enhance the trustee's positive emotions, arising from the trustee's concern for the trustor's well-being, warm-glow feelings would have to be much stronger in order to neutralize the influence of negative emotions arising from the trustor's spite over the trustee's temptation.

Next, consider the trustor's own motivations. The promise also shapes the trustor's beliefs about the trustee's motivations (Hypotheses 3.2, 3.3, and 3.4). The promise is meant to facilitate trustfulness, and the trustee shares some responsibility for placed trust (see the argument for Hypothesis 3.4). Therefore, trustors might assume that the trustee has some interest in the trustor's well-being and is less concerned with his own outcomes. This also enhances the promoting influence of the trustor's warm-glow feelings. However, in contrast to the kind advance of placed trust after which the trustee controls the outcomes, the trustor has to take into account that trust might be abused. Therefore, promises that are not perfectly binding also induce trustors to consider that the trustee might exploit placed trust. Trustors might become more suspicious as the trustee's temptation increases. This undermines the perceived kindness of the promise and, thus, the feeling of obligation. Moreover, if the trustee abuses trust, he reneged on his promise and misled the trustor. It seems reasonable to assume that the trustor might be more disappointed or even angry about a trustee who has reneged on his promise than about a trustee who has abused trust in a decision situation in which no promise has been possible. This increases the trustor's spitefulness concerning the trustee's gain from abused trust and thus increases the hampering impact of the trustee's temptation. The trustor's loss then likewise becomes more hampering because warm-glow feelings are undermined. Alternatively, the trustor might take into account the power of the desire for self-consistency. As previously argued, the trustee might become solely concerned with behaving consistently

by keeping his promise, which reduces effects of outcome-based motivations (see the argument for Hypothesis 3.4). This reduces the trustor's belief about a hampering impact of the trustee's temptation on trustworthiness. However, a similar reduction holds for the trustor's belief about promoting influences of his loss on trustworthiness.

In order to draw a conclusion, recall the argument that trustors' own motivations tend to have a stronger impact than their beliefs about possible motivations of the trustee (see the argument for Hypothesis 3.3). First, consider the impact of the trustor's own loss. The perceived increase in trustworthiness reduces the trustor's concern with his loss. Therefore, it might be reasonable to assume that the hampering impact of the trustor's loss is reduced after the promise has been made. This can be enhanced by the trustor's feelings of obligation. Thus, the impact of the trustor's loss should be less hampering due to the received promise. Second, consider the impact of the trustee's temptation on trustfulness. Recall that the trustor is unsure about the outcomes because he is dependent on the trustee's decision, and that the trustor is unsure about the trustee's motivations. Therefore, it seems unlikely that the promise would give rise to warm-glow feelings that would be strong enough to outweigh the trustor's worry over the trustee's temptation. As argued above, the trustee's temptation might increase the trustor's suspicion and, thereby, also increase the trustor's spitefulness and anger over the gain that the trustee would receive by reneging on the promise. Therefore, the negative influence of the trustee's temptation is assumed to be more hampering after the promise has been made.

> **Hypothesis 3.6: Placing trust after the promise has been made**
> Compared to the TG (i.e., without promise opportunity), trust is *more* likely to be placed after trustworthiness has been promised ($\mathrm{TG}|\mathrm{H}_2^+$). Moreover, the effect of the trustor's loss ($R_1 - S_1$) on placing trust is *less hampering*, but the effect of the trustee's temptation ($T_2' - R_2$) on placing trust is *more hampering*, after trustworthiness has been promised ($\mathrm{TG}|\mathrm{H}_2^+$) than in the TG.

Finally, consider the trustor's decision of whether or not to place trust after the trustee has omitted the promise of trustworthiness ($\mathrm{TG}|\mathrm{H}_2^0$). Omitting a kind action involves unkindness and triggers punishment. The trustor might perceive the omitted promise as unfriendly because the trustee could have supported the trustor to place trust, but explicitly refused to do so (see Chapter 2). This gives rise to feelings of indignation that motivate the trustor to withhold trust. Experimental evidence indeed reveals that omitted promises hamper trustfulness (see Chapter 2; Snijders,

1996). Trustors might expect a detrimental effect of omitted promises on the trustee's trustworthiness due to the trustee's desire for self-consistency (see the arguments for Hypothesis 3.4). As a result of an expected decrease in trustworthiness, trustfulness would also be reduced.

Indignation feelings and an unpleasant anticipation of reduced trustworthiness due to beliefs about the trustee's desire for self-consistency after the promise has been omitted also affect the influence of the trustor's outcome-based motivations. Due to unpleasant anticipations, the trustor's loss $(R_1 - S_1)$ becomes more salient and its hampering impact on trustfulness increases. Note that indignation feelings also induce the trustor to care less about the possibility of receiving a gain due to eventually honored trust. The perceived unkindness of a refused promise also boosts the trustor's spite over the trustee's gains from abused trust. Therefore, the trustee's temptation should become more hampering. It is possible that the trustor understands that the trustee omitted a promise because of high transaction costs (see Chapter 2). Then, the trustor might place trust because he believes in the power of obligation feelings. This would mitigate the hampering impact of the trustor's loss on trustfulness. However, the trustor might also anticipate that the trustee strives to behave consistently (Hypothesis 3.5), which aggravates the hampering impact of the trustor's loss on trustfulness. Moreover, it seems unlikely that positive beliefs about the trustee's motivations would outweigh the hampering impact of the trustor's own motivations. Therefore, it seems more reasonable to assume that omitted promises aggravate the hampering impact that both the trustor's loss and the trustee's gain from abusing trust exert on trustfulness.

Hypothesis 3.7: Kindness of promising trustworthiness
Compared to the TG (i.e., without promise opportunity), trust is *less* likely to be placed after a possible promise of trustworthiness has not been made $(\mathrm{TG}|\mathrm{H}_2^0)$. Moreover, the effect of the trustor's loss $(R_1 - S_1)$ on placing trust is *more hampering*, and the effect of the trustee's temptation $(T_2 - R_2)$ on placing trust is also *more hampering*, after trustworthiness has not been promised $(\mathrm{TG}|\mathrm{H}_2^0)$ than in the TG.

An overview of the hypotheses is provided in Table 3.1. The upper part of Table 3.1 summarizes the hypotheses of behavioral contexts as studied in Chapter 2. In the DG, the trustee's generosity is lower than in the TG because feelings of obligation to return the favor of placed trust are not activated (Hypothesis 3.2). If the trustee promised to honor trust $(\mathrm{TG}|\mathrm{H}_2^+)$, both trustfulness and trustworthiness are increased

Table 3.1: Overview of hypotheses and notation

	Placing Trust	Honoring Trust	
Behavioral contexts			
DG		–	Dictator Game (no placed trust)
TG	(ref.)	(ref.)	Trust Game (no promise option)
$TG\|H_2^+$	+	+	TG after a made promise to honor trust
$TG\|H_2^0$	–	–	TG after an omitted promise to honor trust
Temptation	–	–	Trustee's Temptation ($T_2' - R_2$)
DG		–	Differences between the effect of the trustee's temptation in the TG and the effect of the trustee's temptation in each of the other behavioral contexts
TG	(ref.)	(ref.)	
$TG\|H_2^+$	–	+	
$TG\|H_2^0$	–	–	
Loss	–	+	Trustor's Loss ($R_1 - S_1$)
DG		–	Differences between the effect of the trustor's loss in the TG and the effect of the trustor's loss in each of the other behavioral contexts
TG	(ref.)	(ref.)	
$TG\|H_2^+$	+	+	
$TG\|H_2^0$	–	+	

The hypotheses are formulated in terms of differences toward the TG.

compared to the TG without promise option (Hypotheses 3.4 and 3.6). This is due to the promoting influence of obligation feelings and, on the part of the trustee, also due to the desire for self-consistency. In contrast, an omitted promise ($TG|H_2^0$) creates feelings of indignation that result in reduced trustfulness (Hypothesis 3.7). Trustworthiness might likewise decrease because the desire for self-consistency induced by explicitly omitting the promise undermines the positive influence of obligation feelings that result from trust that has been placed regardless of the omitted promise (Hypothesis 3.5).

The middle and lower parts of Table 3.1 display the hypotheses for outcome-based motivations. People's own outcomes represent the selfish utility component, whereas outcome-based other-regarding motivations are derived from the other's outcomes. The selfish motivation generally hampers trustfulness and trustworthiness (Hypotheses 3.1 and 3.3). Concerning the other-regarding outcome-based motivation, trustfulness decreases with the trustee's temptation because the trustor is spiteful due to the trustee's gain from abused trust (Hypothesis 3.3). In contrast, trustworthiness increases with the trustor's loss if the trustee tends to be cooperative (Hypothesis 3.1).

Next, the influence of outcome-based motivations is assumed to differ between behavioral contexts, i.e., intention-based motivations and self-consistency moderate the effect of outcome-based motivations on trustfulness and trustworthiness. Compared to the TG, the impact of selfish outcome-based motivations is less hampering after trustworthiness has been promised ($TG|H_2^+$), but more hampering after the promise has been omitted ($TG|H_2^0$) (see Hypotheses 3.4 to 3.7). Concerning trustworthiness, obligation feelings and self-consistency mitigate the hampering impact of the trustee's temptation after the promise has been made (Hypothesis 3.4). In contrast, if the promise has been omitted self-consistency aggravates the impact impact of the trustee's temptation but thereby competes with some feeling of obligation (Hypothesis 3.5). Concerning trustfulness, obligation feelings and pleasant anticipations mitigate the hampering impact of the trustor's loss after the trustor has received the promise (Hypothesis 3.6), whereas unpleasant anticipations increase the trustor's concern over his possible loss (Hypothesis 3.6).

The influence of other-regarding outcome-based motivations appears more favorable for trustworthiness (Hypotheses 3.4 and 3.5), but less favorable for trustfulness (Hypotheses 3.6 and 3.7). This holds for the decision situations after the trustee has made the promise ($TG|H_2^+$) and after the trustee has omitted the promise ($TG|H_2^0$). Concerning trustworthiness, obligation feelings boost the concern for the trustor's after both made and omitted promises ($TG|H_2^+$; see Hypotheses 3.4 and 3.5). In the case of omitted promises, the obligation feeling competes with some influence of self-consistency. Note again that self-consistency can undermine the trustee's concern for outcome-based motivations in general, i.e., also the positive impact of the trustor's loss after the promise has been made. Concerning trustfulness, the trustor is assumed to become more suspicious and more spiteful over the trustee's gain that would result from reneging on the promise (Hypothesis 3.6), while indignation feelings increase the trustor's spite after the promise has been omitted (Hypothesis 3.7).

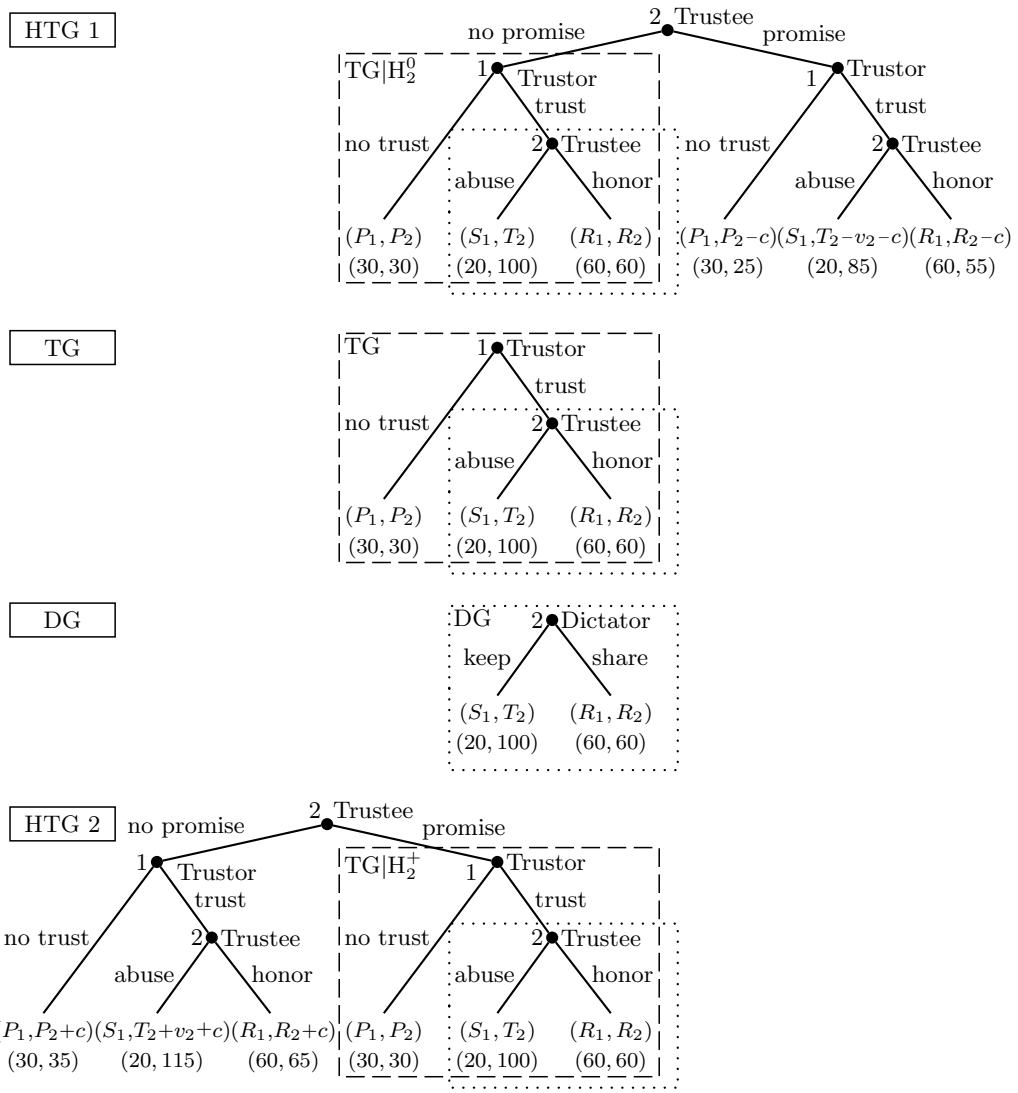
The influence of the temptation on generosity is more hampering in the DG than in the TG, and the impact of the other's loss is less promoting (Hypothesis 3.2). The reason is that the lack of the other's preceding kindness rules out the promoting impact that the obligation feelings have in the TG.

3.3 Design of the Experiment, Data, and Statistical Method

3.3.1 Experimental Design: Sets of (Sub)Games

In order to investigate the effects of outcomes in various behavioral contexts, data from the experiment conducted by Vieth and Weesie (2006) are used (for a description, also

Figure 3.3: Sets of games with identical subgames



The design allows for the comparison of the trustor's behavior in (sub)games indicated by *dashed boxes* and of the trustee's behavior in (sub)games indicated by *dotted boxes*. These sets of (sub)games constitute "subject-payoff response sets" used in the statistical analyses. Numerical example: $S_1^{\text{high}} = 20$, $T_2^{\text{high}} = 100$, $R_1 = R_2 = 60$, $P_1 = P_2 = 30$, $v_2^{\text{low}} = 10$, $c^{\text{low}} = 5$. Figure repeated from Chapter 2.

Figure 3.4: Outcome parameters of the experimental design

DESIGN PARAMETERS:
$S_1(2) \times T_2(2) \times v_2(3) \times c(3)$

Payoff parameters: $S_1(2) \times T_2(2)$		*Promise properties:* $v_2(3) \times c(3)$
$S_1^{\text{low}} = 0$	$T_2^{\text{low}} = 80$	$v_2^{\text{no}} = 0$
$S_1^{\text{high}} = 20$	$T_2^{\text{high}} = 100$	$v_2^{\text{low}} = \frac{1}{4}(T_2 - R_2) = \{5, 10\}$
$R_1 = R_2 = 60$		$v_2^{\text{high}} = \frac{3}{4}(T_2 - R_2) = \{15, 30\}$
$P_1 = P_2 = 30$		$c^{\text{no}} = 0$
		$c^{\text{low}} = \frac{1}{6}(R_2 - P_2) = 5$
		$c^{\text{high}} = \frac{4}{6}(R_2 - P_2) = 20$

Figure repeated from Chapter 2.

see Chapter 2). The description provided here briefly summarizes the main features. The experiment was designed as within-subject sets of single encounters in different games (TGs, HTGs, and DGs). Thereby, sets of (sub)games have identical extensive form, i.e., identical choice structure and payoff structure (for similar designs of sets of (sub)games, see Snijders, 1996; McCabe et al., 2003; Cox, 2004).

An HTG contains two TGs as subgames that result from the trustee's decision of whether or not to make the promise. These TGs contain two DGs for the trustee's decision to return some benefit. In order to construct sets of identical (sub)games that only differ with respect to the behavioral context, the HTGs were taken as the starting point (Figure 3.3). Omitting the promise in one HTG resulted in a subgame ($\text{TG}|\text{H}_2^0$ of HTG1) in which payoffs were identical to payoffs in the subgame of another HTG after the promise was made ($\text{TG}|\text{H}_2^+$ of HTG2). This was reached by implicitly subtracting or adding the absolute values of promise properties at the beginning of some HTGs, i.e. the payoffs of different HTGs were shifted on the scale of promise properties (see the numerical example for the HTG2 in Figure 3.3 for the case of initially added promise properties). For each payoff combination in subgames of HTGs, separate TGs (each containing a DG) and separate DGs were included in the experiment. The implicit shifts of payoffs in HTGs, TGs, and DGs on the scale of the promise properties were not explicit to participants and were hidden by variations of outcome parameters and by mixing sets of (sub)games (as described below).

By varying some outcome parameters, Vieth and Weesie (2006; and see Chapter 2) created various sets of (sub)games with identical payoffs (Figure 3.4). Specifically, the payoffs resulting from abused trust (S_1 and T_2) varied at two levels each (low, high),

while the payoffs after no trust (P_i) and after honored trust (R_i) were fixed. This resulted in four baseline payoff combinations that were modified by nine combinations of promise properties. Binding values (v_2) and transaction costs (c) were each varied on three levels (no, low, and high). The binding values were defined as a share of the trustee's temptation ($T_2 - R_2$) and the transaction costs as a share of the trustee's "gain of cooperation" ($R_2 - P_2$) (for details, see Figure 3.4). As explained above, the 36 combinations of baseline payoffs and promise properties were then modified by implicitly adding or subtracting promise properties. This resulted in 80 different combinations of total payoffs that can occur in the (sub)games. In some HTGs with promise properties initially subtracted, the promise is perfectly binding ($v_2 > T_2 - R_2$).

As described by Vieth and Weesie (2006; and see Chapter 2), each participant played two sets of games in the role of player 1 (trustor, receiver) and two sets of game in the role of player 2 (trustee, dictator) For each encounter, participants were randomly and anonymously matched with another participant (stranger matching whereby the probability of re-matching was minimized within each type of game, see Vieth and Weesie, 2006). The sets of games were mixed by clustering the types of games. First, 12 TGs were played, then 14 HTGs, and thereafter 10 DGs. In order to check for participants attention, in two of the TGs and in two of the HTGs the objective incentive for trustees to abuse trust was removed ($T_2 < R_2$). These games are not involved in the reported analyses. As reported in Chapter 2, 82.1% trustfulness and 95.3% trustworthiness have been observed in these decision situations. This allows for the conclusion that participants paid sufficient attention to the objective outcomes. As mentioned in Chapter 2, in the decision situations in which $T_2 < R_2$, neither full trustfulness nor full trustworthiness is expected because other-regarding outcome-based motivations (e.g., aggressive or competitive tendencies) might have an influence. A brief questionnaire about participants' socio-demographic characteristics (e.g., gender, age, education) separated TGs and HTGs. Other questions about personal attitudes and opinions followed the DGs. Analyses of questionnaire items are not reported here. In each game cluster, player roles were changed after half of the periods. In addition to randomly changing interaction partners, payoffs and promise properties (in HTGs) also changed from one period to the next. Employing a factorial design, the combinations and sequences of payoffs and promise properties were varied across experimental sessions.

The experiment was computer-assisted, employing the software package "z-Tree" (Fischbacher, 2007) (for an example of the decision screens, see Appendix A.1). In addition to general information on paper, participants received on-screen instructions

Table 3.2: Number of subjects and decisions

	Placing trust		Honoring trust	
Number of ...	all data	analyses	all data	analyses
subjects (response sets)	156	138	156	112
total payoffs	76	76	71	71
decisions in total	1716	1518	1389	1022

Total payoffs are combinations of payoffs and promise properties.

and a tutorial before each game cluster. Outcomes were displayed as points in tables and represented monetary gains (one Euro cent for each point). Participants were paid anonymously and immediately after the experiment. On average, participants earned 16 EUR. The experiment was conducted in November 2006 at the ELSE lab at Utrecht University. Using "ORSEE" (Greiner, 2004), 156 persons were recruited from the ELSE participant pool and took part in nine groups of 16 to 20 participants. Nearly all of the participants were students enrolled in various fields at Utrecht University.

3.3.2 Data and Statistical Method

The 156 subjects made 1716 "placing trust" decisions in the role of the trustor and 1389 "honoring trust" decisions as a trustee or dictator (Table 3.2). Of the 80 possible different payoff combinations that could occur in the (sub)games, 76 were realized for "placing trust" decisions of trustors. Recall that trustees could only decide whether to honor trust if trust was actually placed. Despite withheld trust in some combinations, 71 (sub)games remained with different payoffs in which trustees decided whether to honor trust. Note again that the combinations of total payoffs are counted, i.e., transaction costs and the binding value are subtracted from the trustee's outcomes after the promise has been made (Figure 3.3). Trustors always made 11 "placing trust" decisions (5 in TGs and 6 in HTGs), and trustees made 5–13 "honoring trust" decisions (5 in DGs and 0–5 in TGs and HTGs each). The number of decisions made by trustees depended on the trustors' decision to place trust because only decisions in realized subgames were elicited.

In the data analyses, decisions are grouped per subject. Individual heterogeneity, i.e., differences between subjects, is controlled by fixed effects statistical models (more precisely, additive heterogeneity). In fixed effects models, subjects who always made the same decision in a given player role (e.g., trustors who always placed trust) provide no statistical information. These decisions are excluded from the analyses (Table 3.3).

Table 3.3: Number of decisions within subject response sets per (sub)game

	Placing trust (x)					Honoring trust (z)				
	all $\bar{x}$	all x	mix	Σ	$\%x_{mix}$	all $\bar{z}$	all z	mix	Σ	$\%z_{mix}$
DG						215	5	560	780	17.9
TG	85	5	690	780	38.7	74	1	197	272	23.4
TG|H_2^+	51	5	510	566	53.7	51	2	226	279	64.6
TG|H_2^0	51	1	318	370	17.9	19	0	39	58	30.8
Σ	187	11	**1518**	1716	39.3	359	8	**1022**	1389	29.7

Only mixed response sets are in the analyses. The percentages of placed trust ($\%x_{mix}$) and honored trust ($\%z_{mix}$) are calculated within mixed response sets (data in the analyses). "Placing trust" decisions are denoted by "$\bar{x}$" for withheld trust and by "x" for placed trust. Similarly, "honoring trust" decisions are denoted by "$\bar{z}$" for abused trust and by "z" for honored trust.

For the trustor role, this concerns the 11 "placing trust" decisions of 1 person who always placed trust (all x) and 187 "placing trust" decisions of 17 persons who always withheld trust (all $\bar{x}$). Concerning the trustee role or dictator role, 8 "honoring trust" decisions of 1 person who always shared benefits (all z) and 359 "honoring trust" decisions of 43 persons who always kept the benefits for themselves (all $\bar{z}$) are excluded. Nevertheless, all realized total payoffs are involved in the analyses. In mixed response sets, trustors decided to place trust in 597 of the 1518 cases (39.3%), while trustees decided to honor trust in 304 of the 1022 cases (29.7%). These two percentages are averages across (sub)games. The levels of trustfulness and trustworthiness differ between behavioral contexts, ranging from less than once per five cases to more than half of the cases with placed trust and approximately two-thirds of the cases with honored trust. Note that the extent of trustfulness and of trustworthiness is somewhat lower in the complete data because for given total payoffs, more participants decided to withhold or to abuse trust irrespective of the behavioral context than to always place or honor trust.

For testing the hypotheses, logistic regression models with fixed effects for subjects are used, fitted by conditional maximum likelihood. For the analyses in Chapter 2, the same statistical method has been used, but the data have been grouped in subject-payoff response sets in order to control for additive influences of outcome-based motivations without specifying such motivations. For the present study, the basic model can be described as follows:

$$\text{Prob}(y_{ijk}|\sigma_i) = \sigma_i + \eta'_{ijk}\beta$$

The model specifies the probability of trustfulness or trustworthiness of a subject i in the behavioral context of a (sub)game k that has a total payoff combination j. The fixed part of the model, represented by the vector σ of subject-specific intercepts, allows additive individual heterogeneity to be controlled. The term η'_{ijk} represents attributes of factors that vary within subjects and that are weighted by the parameters β. Specifically, the analyses include attributes of behavioral contexts k, of total payoffs j, and of controls. This model makes strong homogeneity assumptions: The effects of behavioral contexts and of outcomes in the behavioral contexts are the same for all subjects (see the discussion for further remarks).

3.4 Results

3.4.1 Analyses for Trustworthiness

Recall that trustees' outcome-based motivations are assumed to consist of the trustee's temptation $(T_2 - R_2)$ as the selfish component and of the trustor's loss $(R_1 - S_1)$ as the other-regarding component. The influence of intention-based motivations is reflected by the behavioral context. The trustee decides whether or not to honor trust after the trustor has placed trust in the TG or in the HTG. In the HTG, the trustee's decision of whether or not to promise trustworthiness constitutes two different contexts (TG$|\mathrm{H}_2^0$ and TG$|\mathrm{H}_2^+$). Moreover, in the DG the trustee decides whether or not to share gains without behavioral context. Thus, four behavioral contexts can be distinguished for the trustee's decision. Since the hypotheses are formulated as comparisons of behavioral contexts with the TG, the TG is chosen as the reference category for the (sub)game dummies in the statistical models (Table 3.4).

The *first model for trustworthiness* (model TW1) shows effects of behavioral contexts and overall effects of trustee's temptation and trustor's loss on trustworthiness across behavioral contexts (Panel A of Table 3.4). Pairwise comparisons of trustworthiness in the behavioral contexts are reported at the bottom of Table 3.4 (Panel C). The differences between one's own and others' total payoffs are divided by 10 and centered at the mean, i.e., $(x_{ij} - \bar{x})/10$, where x_{ij} represents the trustee's temptation $(T'_2 - R_2)$ or the trustor's loss $(R_1 - S_1)$ of a subject i for a total payoff combination j from which the respective overall mean $\bar{x}$ is subtracted. Note again that total payoffs include changes in objective outcomes induced by the promise properties and that the experimental design involves a TG, a DG, and a TG$|\mathrm{H}_2^0$ with identical total payoffs for each subject. Thus, temptation and loss represent the influences of actual objective outcomes on decision-making, and these influences can be distinguished for the different behavioral contexts while objective outcomes remain the same (model TW2).

Table 3.4: Logistic regression of trustworthiness with fixed effects for subjects

(A) Regression coefficients

	Hyp.	TW1 b	TW1 se	TW2 b	TW2 se
Behavioral contexts					
DG	H_2: −	-0.62^{*}	0.25	-0.84^{**}	0.25
TG		(ref.)		(ref.)	
TG|H_2^+	H_4: +	1.28^{***}	0.35	1.19^{***}	0.35
TG|H_2^0	H_5: −	−0.09	0.91	−0.52	1.01
Outcome components					
Temptation	H_1: −	-0.67^{***}	0.09	-0.49^{***}	0.15
Interactions:					
DG	H_2: −			-0.39^{*}	0.19
TG				(ref.)	
TG|H_2^+	H_4: +			0.27	0.26
TG|H_2^0	H_5: −			-1.51°	0.87
Loss	H_1: +	0.28^{*}	0.14	0.50^{*}	0.23
Interactions:					
DG	H_2: −			−0.17	0.25
TG				(ref.)	
TG|H_2^+	H_4: +			-0.50°	0.29
TG|H_2^0	H_5: +			0.10	0.54
Binding value					
in TG|H_2^+		0.36	0.27	0.97^{*}	0.41
in TG|H_2^0		−0.01	0.33	−0.93	0.69
Transaction costs					
in TG|H_2^+		0.82	0.80	0.85	0.81
in TG|H_2^0		0.39	1.52	1.54	1.77
Past periods per game		−0.02	0.05	−0.02	0.05

(Continued on next page. See next page also for notes.)

(Table 3.4 continued from previous page.)

(B) Likelihood-ratio tests

	TW1		TW2	
	χ^2	df	χ^2	df
LR test (control)	102.90^{***}	5	117.88^{***}	11
LR test (v_2, c)	3.67	4	10.68^{*}	4
LR test (TW1)			14.97^{*}	6

(C) Pairwise comparisons (Wald tests)

	Δb	se	Δb	se
$\text{TG}\vert\text{H}_2^0$ – $\text{TG}\vert\text{H}_2^+$	−1.37	0.93	-1.71°	1.02
DG – $\text{TG}\vert\text{H}_2^+$	-1.90^{***}	0.33	-2.03^{***}	0.34
DG – $\text{TG}\vert\text{H}_2^0$	−0.53	0.92	−0.32	1.01

N(decisions) = 1022, N(subjects) = 112;
two-sided p-values: *** p = 0.001, ** p = 0.01, * p = 0.05, ° p = 0.1; (sub)games (0/1), past periods per game (1 . . . 10/12/14), temptation $(T_2-R_2)/10$ in $[-4.6; 4.4]$, loss $(R_1 - S_1)/10$ in $[-1; 1]$, binding value $v_2/(T_2 - R_2)$ in $([0; 0.75], 3)$, transaction costs $c/(R_2 - P_2)$ in $[0; 0.67]$; likelihood-ratio tests against null model with all controls, against model without promise properties (v_2, c), and against model TW1.

The number of the past periods per game and the promise properties are included as control variables. The period in which a decision is has been made is counted per type of game (1–12 for the TG, 1–14 for the HTG, and 1–10 for the DG). Promise properties are interacted with the two (sub)games of HTG resulting from the trustee's decision of whether or not to make a promise. Thus, the main effects of HTG subgames represent the effect of making or omitting the promise for the cheap-talk case (i.e., binding value $v_2 = 0$ and transaction costs $c = 0$). The properties of the promise are represented as "extent of bonding" $(v_2/(T_2 - R_2))$ and as "extent of costliness" $(c/(R_2 - P_2))$, i.e., the share of the benefit from honored trust that is invested in making the promise. The extent of the transaction costs ranges from 0 to 67 percent, and the "extent of bonding" ranges from 0 to 75 percent. Moreover, in decision situations in which the promise is perfectly binding $(v_2 > T_2-R_2)$, the binding value of the promise is three times larger than the trustee's temptation. Since temptation and loss are defined in terms of total payoffs (i.e., promise properties are incorporated after the promise has been made), coefficients of promise properties can be interpreted as representing effects that do not result from outcome-based motivations. Furthermore,

by incorporating the properties of the promise as control variables, their influences are removed from the estimates of temptation and loss. Thus, influences of the trustee's temptation and of the trustor's loss are only outcome-based and not mixed with influences of other motivations induced by the promise properties.

In the *second model for trustworthiness* (model TW2), hypotheses on moderating effects of behavioral advances on outcome effects are tested. For this purpose, the effects of temptation and loss are interacted with (sub)games. Since the trustee's temptation and the trustor's loss are centered at the mean, the main effects of (sub)games represent the effects of the behavioral contexts for average temptation and for average loss. Differences between the coefficients of temptation and of loss in the (sub)games compared to the TG are reported in Table 3.4. Overall effects of temptation and of loss per (sub)game are presented in Table 3.6. Note that hypotheses are formulated in terms of differences in outcome effects between behavioral contexts (presented in Table 3.4). For increased effects of outcomes (i.e., more negative or more positive), the direction of the overall outcome effect remains the same as in the TG (Hypothesis 3.1). However, if an outcome effect is reduced (i.e., less negative or less positive) the direction of the overall effect depends on the relative magnitude of the reduction, which is not involved in the hypotheses. Therefore, predictions are not indicated in Table 3.6. In the following, the description of results presented in Table 3.4 also includes references to the overall effects reported in Table 3.6.

The main idea underlying this study is that different behavioral contexts influence subsequent decision-making due to process-based motivations, i.e., due to feelings of obligation or indignation and due to the desire for self-consistency. Assuming that placed trust is perceived as a kind advance, sharing gains in the TG (i.e., honoring trust) should be more likely than sharing gains in the DG (Hypothesis 3.2). In the HTG, the desire for self-consistency also becomes relevant, such that trustworthiness should be more likely after the promise has been made (Hypothesis 3.4) and, in general, less likely after an omitted promise (Hypothesis 3.5). In line with the findings reported in Chapter 2, the analyses presented here provide evidence for decreased generosity in the DG compared to the TG (Table 3.4) (see also Gautschi, 2000; McCabe et al., 2003; Cox, 2004) and for a strong increase in trustworthiness after the promise has been made ($\mathrm{TG}|\mathrm{H}_2^+$). The positive effect of making the promise seems to be stronger than the negative effect of no placed trust in the DG. However, this impression cannot be supported (test of differences between absolute coefficients: Wald $\chi^2_{1\,\mathrm{df}} = 1.67$ with $p = 0.1969$ in model TW1 and Wald $\chi^2_{1\,\mathrm{df}} = 0.46$ with $p = 0.4973$ in model TW2). Note that the difference is significant in the analyses

reported in Chapter 2, though only at a borderline level if the promise is cheap-talk. The influence of having made the promise on trustworthiness is more promoting with increasing binding value (in TG|H_2^+ in model TW2). Note again that the coefficients of promise properties represent effects that are not based on objective outcomes. One would assume that the sizable increase in the positive effect that making the promise exerts on trustworthiness would be due to the cases in which the promise is perfectly binding. However, this intuition could not be supported in further analyses (only jointly significant, analyses not reported). The idea that omitted promises (TG|H_2^0) hamper trustworthiness cannot be supported (see also Snijders, 1996; whereas in Chapter 2, the influence is found to depend on the properties of the omitted promise). Overall, the results are largely consistent with the findings reported in Chapter 2. Differences in the results are found in the moderating effects of promise properties. However, in the analyses presented here, promise properties are only incorporated as controls in order to remove the influences from the effects of the HTG subgames and from the effects of the outcome (see the discussion for further remarks).

Next, considering the outcome-based motivations, it has been argued that trustees are less likely to honor trust as their temptation increases ($T_2' - R_2$), whereas the trustor's loss ($R_1 - S_1$) promotes trustworthiness if trustees are sufficiently cooperative (Hypothesis 3.1). The predictions about the general impact of both the trustee's temptation and the trustor's loss are supported (Table 3.4). The trustee's temptation has a highly significant hampering impact, whereas the trustor's loss promotes trustworthiness. The positive coefficient of the trustor's loss on trustworthiness is significantly smaller than the negative coefficient of the trustee's temptation (Wald $\chi^2_{1\,\mathrm{df}} = 5.16$ with $p = 0.0231$ for the difference between absolute coefficients in model TW1). Thus, across behavioral contexts, trustees are generally more concerned with their own outcomes than about the trustor's outcomes (see the discussion for further remarks). This supports the assumption that would be implied by a theoretical model with a restricted altruism parameter as far as trustworthiness is concerned.

In the second model for trustworthiness (model TW2), temptation and loss are interacted with the behavioral contexts. As argued above, the feeling of obligation to return the favor of placed trust might reduce the hampering impact of the trustee's temptation on trustworthiness and increase the promoting influence of the trustor's loss (Hypothesis 3.2). The analyses show that the temptation is indeed more hampering in the DG than in the TG (Table 3.4). In fact, the temptation in the DG is almost twice as hampering as the temptation in the TG (Table 3.6). The trustor's loss is promoting in the TG and also tends to be supportive in the DG, although the

coefficient is reduced by approximately one-third in the DG (Table 3.6). However, the difference between the coefficients for the trustor's loss in the DG compared to the TG is not significant (Table 3.4).

In the HTG, promising to honor trust (TG|H_2^+) is hypothesized to mitigate the hampering effect of the trustee's temptation on trustworthiness and to increase the promoting effect of the trustor's loss (Hypothesis 3.4). Compared to the TG, the temptation indeed seems to be somewhat less hampering in the TG|H_2^+, which would be due to obligation feelings or self-consistency, but the difference is not significant (Table 3.4). Nevertheless, the negative impact of temptation after the promise has been made is less than half of the impact in the TG and no longer significant (Table 3.6). Concerning the trustor's loss, the positive influence on trustworthiness tends to be decreased rather than increased after the promise has been made (Table 3.4). Thus, the prediction for the trustor's loss in Hypothesis 3.4 is rejected. In fact, whereas the trustor's loss promotes trustworthiness in the TG, no support for an influence of the trustor's loss has been found after the promise has been made (Table 3.6). As a result, neither the trustee's temptation nor the trustor's loss significantly influences trustworthiness after the promise has been made. This indicates that making the promise influences trustworthiness directly rather than by modifying the influences of outcomes. Trustees seem to be mainly concerned with behaving consistently by keeping their promise and thereby neglect objective outcomes (see the arguments for Hypothesis 3.4).

After the promise has been omitted (TG|H_2^0), the trustee's temptation should have a more hampering impact on trustworthiness than in the TG because of the desire for self-consistency, which conflicts with obligation feelings that might strengthen the positive effect of the trustor's loss (Hypothesis 3.5). In the analyses, the temptation is indeed considerably more hampering after the promise has been omitted (Table 3.4). This effect is only marginally significant, possibly, because only 35 trustees in mixed response sets could decide whether to honor trust after they omitted the promise (Table 3.3). Therefore, the test of the effects of outcome variations in the TG|H_2^0 has little statistical power. Nevertheless, the negative coefficient of the temptation is approximately four times larger after the promise has not been made (Table 3.6). In contrast, the effect of the trustor's loss does not differ significantly between the TG|H_2^0 and the TG (Table 3.4). Thus, no support can be found for the reasoning that the influence of the trustor's loss on trustworthiness might be more promoting after trust has been placed despite the omitted promise (Hypothesis 3.5). Note that the coefficient has a positive sign, which suggests some influences of obligation feelings

that compete with influences of self-consistency. However, the slightly increased total effect of the trustor's loss after the promise has been omitted is likewise not significant (Table 3.6). As mentioned above, statistical power is low in the TG|H_2^0 due to the small number of cases. Moreover, it is possible that more selfish trustees omitted the promise. Such a selection effect might also be responsible for the lack of support for changes in the loss effect after the promise has been omitted.

In general, the moderating effects of preceding behavior on the effects of outcomes are sizable, although mostly only marginally significant or non-significant (Table 3.4). The likelihood-ratio test comparing model TW1 and model TW2 shows that the interactions of outcomes and (sub)games are jointly significant (LR $\chi^2_{6\,\mathrm{df}} = 14.97$ with $p = 0.0205$). This indicates that influences of outcome-based motivations on trustworthiness in general differ significantly between behavioral contexts. Testing the joint influence of temptation and of loss separately shows that only the influence of trustee's temptation on trustworthiness is found to differ significantly across contexts (LR $\chi^2_{3\,\mathrm{df}} = 11.52$ with $p = 0.0092$), while the analysis do not provide support for differences in the effect of the trustor's loss (LR $\chi^2_{3\,\mathrm{df}} = 3.59$ with $p = 0.3093$). At first sight, one might conclude that the influence of the trustee's selfish utility component is found to be context-dependent, but that the evidence is not sufficient to conclude context-dependency for the trustee's other-regarding utility component. This might be due to influences of beliefs or the specific representation of outcome-based motivations in the underlying altruism model (see the discussion for further remarks). However, another reason might be that certain behavioral contexts influence the trustee's concern about his own outcome (DG and TG|H_2^0), while other behavioral contexts affect the trustee's concern about the trustor's outcome (TG|H_2^+). In addition, even within a behavioral context, the moderating impact of self-consistency can affect outcome effects differently than the moderating influence of obligation feelings, and can also affect the influence of temptation differently from the influence of loss (Hypotheses 3.4 and 3.5, and Table 3.4). In this sense, the tests of the joint influence of different behavioral contexts on the effects of outcomes only indicate that the influence of the trustee's temptation on trustworthiness differs more strongly between the behavioral contexts considered in this study than the influence of the trustor's loss on trustworthiness.

The coefficients of promise properties indicate how the influences of behavioral advances are moderated in the HTG subgames. In Chapter 2, evidence has been provided for such moderating impacts by using statistical models in which various representations of outcome-based motivations have been controlled in addition to

controlling for additive individual heterogeneity. In the analyses presented here, only the binding value of a made promise is found to promote trustworthiness (Table 3.4). Note again that the sizable influence is not based on outcome changes. The coefficient is only significant when allowing for context-dependency of outcome effects. This might suggest that the specific representation of outcome-based motivation employed here does not fully capture all relevant heterogeneity, i.e., that outcome-based motivations are better controlled in the study reported in Chapter 2 (see the discussion for further remarks). Likelihood-ratio tests (Panel B of Table 3.4) also show that controlling for promise properties significantly improves the model that includes the context interactions (model TW2), but the joint influence of promise properties is not significant in model TW1.

3.4.2 Analyses for Trustfulness

Since trustors make no decisions in the DG, the DG is not included in the analyses on trustfulness. Trustors decide whether or not to place trust in the TG or in one of the two subgames of the HTG after the trustee has decided either to promise trustworthiness ($TG|H_2^+$) or to omit the promise ($TG|H_2^+$). The TG serves again as the reference category. The results of the statistical analyses on trustfulness (Table 3.5) are presented in the same way as described for the analyses on trustworthiness (see above for the description of variables and the setup of Table 3.4 and Table 3.6).

Similar to trustworthiness, the findings for trustfulness concerning the effects of motivations that arise from behavioral contexts are likewise in line with previous findings (see Chapter 2). Since promising trustworthiness can be perceived as friendly behavior ($TG|H_2^+$), receiving a promise should promote trustfulness (Hypothesis 3.6). In contrast, by omitting the promise ($TG|H_2^0$), trustees refuse to support the trustor to place trust. Not making the promise might therefore be perceived as unfriendly and should hamper trustfulness (Hypothesis 3.7). In line with previous studies (see Chapter 2; Snijders, 1996), both parts of Hypotheses 3.6 and 3.7 are supported in the analyses (Table 5). Trustfulness increases in the $TG|H_2^+$ due to the received promise. This indicates that trustors feel obliged to return the favor and might anticipate increased trustworthiness. Moreover, trustfulness is reduced in the $TG|H_2^0$ because the trustee chose to omit the promise (see also Gautschi, 2000). As discussed in Chapter 2, the strong hampering impact of omitted promises suggests that feelings of indignation drive trustors to take revenge rather than that trustors would only withhold trust because trustors might be convinced of reduced trustworthiness.

Concerning the impact of outcome-based motivations on trustfulness, various influences of trustors' own motivations and of their possible beliefs about the trustee's motivations have been discussed. It has been argued that the trustor's loss and the trustee's temptation hamper trustfulness because of the trustor's interest in avoiding losses and because the trustor is spiteful concerning the trustee's possible gain from abused trust (Hypothesis 3.3). The analyses indeed show that trustfulness significantly decreases with the trustee's temptation and with the trustor's loss (Table 3.5). The coefficient of the trustee's temptation seems to be somewhat smaller than the coefficient of the trustor's loss. However, this difference is not significant (Wald $\chi^2_{1\,\mathrm{df}} = 1.23$ with $p = 0.2666$). Recall that the analyses on trustworthiness provide support for a smaller general influence of the other-regarding outcome component compared to the influence of the selfish component. That no support for such a difference can be found in the analyses on trustfulness suggests that the influence of the other-regarding component for trustors is relatively strong. It is unlikely that the strongly negative temptation effect would only be due to the trustor's belief about the hampering impact of the temptation on trustworthiness because this would also increase the trustor's concern about his loss. Therefore, the strong effect of the trustees' temptation indicates that trustors might indeed be spiteful concerning the gain that the trustee would receive from abusing trust. This supports the intuition that the influence of actors' motivations is role-dependent and, thus, not individually constant across decision contexts. That each participant made decisions in both player roles in the experiment might strengthen the evidence, although further analyses are required for statistical tests (see the discussion for further remarks).

The second model for trustfulness (model TF2) includes interactions of outcomes and behavioral context. It has been hypothesized that a promise of trustworthiness ($\mathrm{TG}|\mathrm{H}_2^+$) should also reduce the hampering impact of the trustor's loss because the trustor might anticipate an increase in trustworthiness and might feel obliged to return the favor of the trustee's promise (Hypothesis 3.6). Concerning the temptation effect, it has been argued that trustors become more suspicious due to the trustee's temptation and more spiteful about the gain the trustee would receive from reneging on his promise. Therefore, the influence of the trustee's temptation on trustworthiness is assumed to be more hampering after the promise has been made than in the TG (Hypothesis 3.6). The analysis provides neither support for the reasoning about the difference in the influence of the temptation that would be due to the received promise nor support for the reasoning about the difference in the influence of the loss (Table 3.5). First, although the influence of the trustee's temptation after

Table 3.5: Logistic regression of trustfulness with fixed effects for subjects

(A) Regression coefficients

	Hyp.	TF1 b	TF1 se	TF2 b	TF2 se
Behavioral contexts					
TG		(ref.)		(ref.)	
TG|H_2^+	H_6: +	1.28***	0.35	1.19***	0.35
TG|H_2^0	H_7: –	−0.09	0.91	−0.52	1.01
Outcome components					
Temptation	H_3: –	−0.22***	0.05	−0.17**	0.06
Interactions:					
TG				(ref.)	
TG|H_2^+	H_6: –			−0.13	0.12
TG|H_2^0	H_7: –			−0.18	0.16
Loss	H_3: –	−0.34***	0.10	0.28*	0.11
Interactions:					
TG				(ref.)	
TG|H_2^+	H_6: +			−0.00	0.14
TG|H_2^0	H_7: –			−0.43*	0.20
Binding value					
in TG|H_2^+		0.79***	0.17	0.66**	0.22
in TG|H_2^0		0.22	0.17	0.08	0.21
Transaction costs					
in TG|H_2^+		−0.47	0.47	−0.41	0.78
in TG|H_2^0		−0.83	0.60	−0.91	0.63
Past periods per game		−0.15***	0.03	−0.15***	0.03

(Continued on next page. See also next page for notes.)

(Table 3.5 continued from previous page.)

(B) Likelihood-ratio tests

	TF1		TF2	
	χ^2	df	χ^2	df
LR test (control)	52.83^{***}	4	60.18^{***}	8
LR test (v_2, c)	29.32^{***}	4	11.65^{*}	4
LR test (TF1)			7.35	4

(C) Pairwise comparisons (Wald tests)

	Δb	se	Δb	se
$\text{TG}\vert\text{H}_2^0$ – $\text{TG}\vert\text{H}_2^+$	-1.39^{***}	0.34	-0.34^{***}	0.36

N(decisions) = 1518, N(subjects) = 138;
two-sided p-values: *** p = 0.001, ** p = 0.01, * p = 0.05, ° p = 0.1; (sub)games (0/1), past periods per game $(1 \dots 12/14)$, temptation $(T_2 - R_2)/10$ in $[-4.6; 4.4]$, loss $(R_1 - S_1)/10$ in $[-1; 1]$, binding value $v_2/(T_2 - R_2)$ in $([0; 0.75], 3)$, transaction costs $c/(R_2 - P_2)$ in $[0; 0.67]$; likelihood-ratio tests against null model with all controls, against model without promise properties (v_2, c), and against model TF1.

the promise has been made ($\text{TG}|\text{H}_2^+$) is indeed more negative than in the TG, the difference is not significant. This is surprising given that the effect of the trustee's temptation is nearly twice as negative after the promise has been made than in the TG (Table 3.6). Further analyses showed that the negative impact of the temptation becomes significantly more hampering if promise properties are not controlled. However, the temptation effect then also includes the influences of promise properties that are not outcome-based. Second, the difference in the influence of the trustor's loss between the $\text{TG}|\text{H}_2^+$ and the TG is basically zero (Table 3.5). The lack of support for differences in the effect of temptation and of loss that would be caused by the received promise suggests that outcome-based motivations might influence trustfulness largely independent from obligation feelings induced by the received promise.

When a possible promise has been omitted ($\text{TG}|\text{H}_2^0$), the effects of temptation and of loss on trustfulness should be more negative than in the TG (Hypothesis 3.7). The reason is that the trustor might be more concerned with his loss due to unpleasant anticipations, and that the perceived unkindness increases the trustor's spitefulness over the trustee's gain from abusing trust. Although the analysis shows that the impact of the trustee's temptation is indeed somewhat more hampering in the $\text{TG}|\text{H}_2^0$, the coefficient is not significant (Table 3.5). However, the omitted promise considerably

aggravates the hampering influence of the trustor's loss on trustfulness (Table 3.5). The coefficient is approximately two and a half times more negative after the promise of trustworthiness has been explicitly omitted (Table 3.6). This indicates that trustors become more selfish and more focused on their loss after the promise has been omitted. The lack of support for increased spitefulness over the trustee's possible gain might favor this interpretation.

Except for the strong and positive coefficient of the binding value of a promise, no other evidence for effects of promise properties on trustfulness is found (Table 3.5). As mentioned for the analyses of trustworthiness, it is possible that this is due to the specific representation of outcome-based motivations. However, in previous analyses using a more powerful approach of controlling for outcomes-based motivations, support has not even been found for a positive impact of the binding value on trustfulness (see Chapter 2). Next, the coefficient for the period in which the trustor has decided whether to place trust is negative and highly significant (see also Chapter 2). This indicates that trustors might have experienced abused trust in previous encounters and therefore become more reluctant to place trust. However, note again that the coefficient for the decision period is not significant in the analyses for trustworthiness (Table 3.4).

In contrast to the analyses for trustworthiness, less support for context-dependency of outcome-based motivations is found for trustfulness (Table 3.5). Interactions of outcomes with behavioral contexts are also not jointly significant (LR $\chi^2_{4\,\mathrm{df}} = 7.35$ with $p = 0.1187$). Testing the differences in the effects of loss and of temptation separately reveals again that the effect of the trustor's loss on trustfulness differs significantly across contexts (LR $\chi^2_{2\,\mathrm{df}} = 4.98$ with $p = 0.0827$), but no support can be found for differences in the effect of the trustee's temptation (LR $\chi^2_{2\,\mathrm{df}} = 2.00$ with $p = 0.3685$). Similar to trustees, this might also suggest for trustors that the influence of the selfish component is context-dependent, whereas no support for context-dependency of the other-regarding component of the trustor's outcome-based motivation can be found. As discussed for trustworthiness, alternative representations of other-regarding outcome-based motivations might differ between behavioral contexts (see the discussion for further remarks). Moreover, consider that the trustor's outcome from placing trust involves an uncertain element because it depends on the trustee's decision. Therefore, it is possible that the trustee's decision of whether or not to promise his trustworthiness is not sufficient to induce changes in the trustor's other-regarding outcome-based motivations, and that the activated intention-based motivations affect trustfulness directly rather than moderate influences of outcome-based motivations.

Table 3.6: Effects of the trustee's temptation and the trustor's loss per (sub)game

	Trustworthiness		Trustfulness	
Effects ...	$b_o + b_o b_s$	se	$b_o + b_o b_s$	se
in the DG				
Temptation	-0.88^{***}	0.13		
Loss	0.33°	0.17		
in the TG				
Temptation	-0.49^{***}	0.15	-0.17^{**}	0.06
Loss	0.50^{*}	0.23	-0.28^{*}	0.11
in the TG$\vert H_2^+$				
Temptation	-0.22	0.22	-0.30^{**}	0.12
Loss	-0.01	0.23	-0.28^{*}	0.13
in the TG$\vert H_2^0$				
Temptation	-2.00^{*}	0.86	-0.35^{*}	0.16
Loss	0.60	0.52	-0.71^{***}	0.20

The table shows the Wald tests for the sum of coefficients in model TW2 for trustworthiness and in model TF2 for trustfulness (two-sided p-values: *** p = 0.001, ** p = 0.01, * p = 0.05, ° p = 0.1). The outcome coefficient is denoted by b_o (temptation $[-4.6; 4.4]$, loss $[-1; 1]$), and b_s represents the (sub)game coefficient (DG, TG, TG$\vert H_2^+$, TG$\vert H_2^0$).

3.4.3 Comparison of Results for Trustworthiness and Trustfulness

Comparing the effects on trustfulness with the effects on trustworthiness seems to reveal some discrepancies (Table 3.6). Typically, one would expect more similarities due to the influence of beliefs. For instance, the trustor's loss is found to promote trustworthiness in nearly all behavioral contexts (except for the TG$\vert H_2^+$), but hampers trustfulness. This finding does not necessarily suggest that trustors would not anticipate the positive impact of their loss on trustworthiness. One could argue that the hampering influence of the loss on trustfulness would otherwise be even stronger (disregarding unobserved heterogeneity and differences in decision noise). However, the negative influence of the trustee's temptation on trustfulness is not found to be significantly stronger than the negative influence of the trustor's loss (analyses not reported), despite the strongly negative temptation effect on trustworthiness.

At first sight, trustors also seem to hardly anticipate the moderating effects that preceding behavior exerts on the effects of outcomes on trustworthiness. In particular, the influence of the trustee's temptation on trustworthiness is found to differ between behavioral contexts (Table 3.4), whereas no support for such differences could be found for the temptation effect on trustfulness (Table 3.5). Similarly, consider the influence of the trustor's loss in the decision situation after the trustee has made the promise ($TG|H_2^+$). The positive influence of the trustor's loss on trustworthiness is significantly reduced after the promise has been made compared to the TG (Table 3.4). Due to this reduction, no support can be found for a promoting effect of the trustor's loss on trustworthiness after the promise has been made (Table 3.6). In contrast, the effect of the trustor's loss on trustfulness basically does not change after the promise has been made compared to the TG (Table 3.5). Next, consider the decision situation after an omitted promise ($TG|H_2^0$). The trustee's temptation strongly hampers trustworthiness (significantly more than in the TG, see Table 3.4), but the temptation effect on trustfulness is not found to be significantly more negative compared to the TG (Table 3.5).

However, the results show that the effect of the trustor's loss on trustfulness is more negative after the promise has been omitted ($TG|H_2^0$) than in the TG (Table 3.5). This might indicate that the negative temptation effect on trustworthiness is anticipated and affects the trustor's selfish utility component. Trustors then become more reluctant to place trust because they fear losses, and not because they would become more spiteful over the trustee's possible gain from abusing trust. A similar reasoning might hold for the finding that the trustor's loss hampers trustfulness, although it promotes trustworthiness. Thus, these findings indicate that the hampering effect of the trustee's selfish utility component is mirrored on the trustor's selfish utility component, i.e., not mediated by the same outcomes (see the discussion for further remarks). Note that no such "mirroring" is visible in the decision situation after the promise has been made ($TG|H_2^+$).

3.5 Summary and Perspectives

3.5.1 Summary of Basic Ideas, Approach, and Contributions

Different motivations drive people to choose a certain action. One of these motivations is selfishness, i.e., people's focus on their own objective outcome in a decision situation. In addition to selfishness, other-regarding motivations rooted in emotions also play a role. In single encounters with strangers, behavioral patterns of reciprocity are typically implications of other-regarding motivations. First, people are not only

selfish, but they also take into account the objective outcomes of their interaction partners. Such social orientations are outcome-based motivations. While various representations of an actor's outcome concern are possible, a simple model incorporating an altruism parameter as a weight for others' objective outcomes has been a model that has attracted much attention over the past decades. The altruism parameter can account for positive impacts of benevolence (joy from others' well-being) and negative impacts of spite (dislike of others' gains). Second, people also take into account behavioral processes of how outcomes are obtained. Kind and unkind behavior of interaction partners can activate process-based motivations. Fundamental social-psychological processes that arise from feeling an obligation to return favors and from feeling indignation about unkindness give rise to intention-based motivations. In addition to intention-based motivations, people's own behavioral advances trigger a desire for self-consistency, which constitutes an intra-personal motivation. Such process-based motivations arising from specific behavioral contexts can directly influence people's decision-making and can also interact with outcome-based motivations. In contrast to the assumption that is commonly applied in formal models of other-regarding motivations, social-psychological research provides evidence that people's motivational parameters are not individually stable across various decision contexts (for reviews, see Ross and Nisbett, 1991: ch. 4; Kunda, 2002).

In the study presented here, these ideas have been applied to trust situations with and without the opportunity for the trustee to promise trustworthiness. Based on the altruism model, outcome-based utility components have been defined as the trustee's temptation to abuse trust and as the trustor's loss from abused trust. The temptation constitutes the trustee's selfish utility component, and the trustor's loss constitutes the trustee's other-regarding outcome-based component, and vice versa for trustors. Thus, the altruism model allows influences of people's own outcomes to be separated from influences of others' outcomes. In order to test effects of behavioral advances, data from a lab experiment conducted by Vieth and Weesie (2006; and see Chapter 2) were used. The experiment is designed as within-subject sets of structurally identical (sub)games that result from friendly or unfriendly actual behavior in single encounters. The trustor's decision of whether or not to place trust was analyzed in three differently embedded Trust Games (TGs): a TG without context and two TGs resulting as subgames in the Hostage Trust Game (HTG) after the trustee decided whether or not to promise trustworthiness. Each TG contains the trustee's decision of whether or not to honor trust. The trustee's decision of whether or not to honor trust constitutes a dichotomous Dictator Game (DG). Thus, to analyze effects on

trustworthiness, four differently embedded DGs were distinguished: the DG without context and a DG as a subgame in each of the three TGs. Employing a within-subject design, each participant in the experiment made decisions in a mix of such sets of identical (sub)games. Outcomes and properties of the promise were varied across sets of (sub)games. This design allows additive individual heterogeneity to be controlled.

The results provide evidence that outcome-based motivations are role-dependent and differ between behavioral contexts generated by preceding kind and unkind decisions. As shown in Chapter 2, people's kind and unkind behavior strongly affects subsequent decisions. Placed trust and making a promise increase trustworthiness. Trustfulness is promoted if the trustee has promised trustworthiness, and trustfulness is hampered, if a promise has been explicitly omitted. No support can be found for a detrimental influence of omitting a promise on trustworthiness. In addition to influences of preceding behavior, outcomes are also found to affect trustfulness and trustworthiness. In general, across behavioral contexts, trustworthiness is hampered by the temptation and promoted by the trustor's loss, whereas both temptation and loss hamper trustfulness. This indicates that people in the role of the trustee tend to be positively motivated by the trustor's loss, while people in the role of the trustor are spiteful about the trustee's gain from abused trust. Thus, depending on the position in a decision situation, altruistic tendencies are turned into aggressive tendencies.

Moreover, evidence has been found that the influences of temptation and loss on trustworthiness depend on the behavioral context. Specifically, the temptation effect is more negative in the DG than in the TG. This indicates the lack of the promoting influence of obligation feelings induced by placed trust in the TG. After the promise has been made ($\mathrm{TG}|\mathrm{H}_2^+$), the influence of temptation is not significantly mitigated, but the promoting impact of the trustor's loss is nearly removed. This suggests that self-consistency rather than obligation feelings induce the trustee to keep his promise. Similarly, omitting a promise ($\mathrm{TG}|\mathrm{H}_2^0$) is found to considerably aggravate the temptation effect, which likewise indicates the influence of self-consistency legitimating to abuse trust. No support is found for the idea that the impact of the trustor's loss on trustworthiness would become more promoting due to obligation feelings induced by the trustor's decision to place trust despite the omitted promise. Concerning trustfulness, less support has been found for changed influences of outcomes in the different behavioral contexts. No significant influence of received promises on effects of temptation and loss is found after the promise has been made ($\mathrm{TG}|\mathrm{H}_2^+$). However, the analyses do reveal that an omitted promise ($\mathrm{TG}|\mathrm{H}_2^0$) considerably aggravates the hampering influence of the trustor's loss, while no support is found that the effect of

temptation on trustfulness would become more negative. Thus, it cannot be assessed whether this finding indicates that trustors become more selfish rather than spiteful or that the omitted promise increases feelings of indignation about the prospective loss. In general, the results suggest that the hampering impacts of the trustee's selfish motivation (temptation) might affect the trustor's selfish motivations (loss) rather than the trustor's other-regarding motivations ("mirroring effect"). Thus, the influences are not mediated by the same outcome components. These indications of "mirroring" do not appear to hold for the decision situation after the promise has been made.

3.5.2 Further Discussion and Perspectives

Some further remarks concern the following aspects: (1) representation of social orientations, (2) eliciting beliefs and emotions, (3) deducing hypotheses, and (4) refined statistical analyses. First, it has been mentioned that the altruism parameter is typically constrained such that actors are assumed to be at most equally interested in others' outcomes. Since this constraint would rule out that most sanctioning behavior can also be motivated by objective outcomes, this assumption was not employed in this study. The reason for this constraint might be that types of actors with a certain social orientation have traditionally been assessed by eliciting distribution preferences in non-strategic dictator-like decision situations (decomposed games) or in strategic but simultaneous decision situations. Considering that the study presented here provides evidence for context-dependency of social orientations, it might be fruitful to also study how process-based motivations moderate the influence of social orientations on sanctioning behavior. It is possible that such moderating influences are stronger because the emotional basis motivating sanctions might be stronger. For instance, the trustor's spite over the trustee's gain might receive a boost after trust has been abused and the trustor is deciding whether to punish the trustee. Alternatively, strong feelings of obligation or indignation might undermine the influence of outcome-based motivations. Moreover, the influence of outcome-based motivations on sanctioning behavior might not only be role-dependent but also change between "decision points", depending on preceding decisions. For instance, the trustor's spite over the trustee's possible gain from abused trust might turn into benevolence when honored trust is to be rewarded. Similarly, the trustee's benevolence toward the trustor's outcomes might turn into aggressiveness when the trustee is explicitly threatened with punishment for abused trust.

In addition to the altruism model employed in this study, numerous other representations of social orientations are possible. For instance, research on social orientations

has also identified that people minimize the difference between their own and others' objective outcomes (Kelley and Thibaut, 1978; Knight and Dubro, 1984). This idea of "equalitarian orientations" (MacCrimmon and Messick, 1976) has also been proposed in more recent theoretical models that capture fairness in the sense of inequality aversion (Weesie, 1994a; Snijders, 1996; Ledyard, 1995; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; and for reviews see, e.g., Camerer, 2003: ch. 2; Fehr and Schmidt, 2006). Payoff equality constitutes the reference point that is considered as fair. Deviation of an actor's own and others' objective outcomes from the reference point inflicts emotional disutility, e.g., envy form disadvantageous inequality and guilt from advantageous inequality with individual differences in the extent of disutility. Modeling inequality aversion in this way differs mainly in two important respects from the simple altruism model employed here. First, some context-dependency in the sense of advantageous and disadvantageous positions in terms of objective outcomes is already taken into account in inequality aversion models. Second, not only the impact of the other-regarding component is shaped by emotions, but envy and guilt also change the influence of an actor's own outcomes, i.e., an actor's own outcomes are likewise weighted by his fairness concern. Envy increases the impact of an actor's interest in his own outcome, while guilt hampers it. This can foster or undermine outcome effects in certain behavioral contexts. However, even if a significant share of people is concerned with equality, the basic message of the present study still applies: social orientations do not seem to be individually constant but tend to be context-dependent. Testing context-dependency of different representations of social orientations would also provide more insights into what types of social orientations become salient in what sets of contexts. For instance, people seem to care about the community in some situations, but they behave competitively in other sets of contexts. Lab experiments have revealed that this difference in behavior can be triggered by simple "label framing" of an interaction situation (Rege and Telle, 2004). This suggests that normative priming might create focal points and influences people's motivations.

Second, it has been argued that emotions and beliefs about others' motivations influence people's decision-making. For instance, recall that the trustor's loss and the trustee's temptation influence the trustor's own motivations and the trustor's beliefs about the trustee's motivations. Since trustees control the outcomes with their decision of whether or not to honor trust, beliefs about the trustor's motivations only play a role in evaluating the kindness of the trustor's trustfulness. In contrast, trustors also benefit in objective terms if they manage to form appropriate beliefs about the

trustee's motivations and, thus, about his trustworthiness. Moreover, differences in outcome effects between behavioral contexts can be due to changes of an actor's own outcome-based motivations or result from changes of an actor's beliefs about others' motivations. For instance, trustors might become more concerned with their loss after the promise has been omitted due to indignation feelings or because they expect reduced trustworthiness. In a similar way, beliefs can also influence the effect of other-regarding outcome-based motivations. Concerning perceived kindness, the trustee's beliefs about the trustor's beliefs also become relevant. Such confounding might not be problematic, considering that social orientations might be seen as an expression of people's beliefs (see the review by Cook and Cooper, 2003: 219). Empirical findings that people scoring high on survey measures for trustfulness are more trustworthy in decision situations (e.g., Burks et al., 2003; Glaeser et al., 2000) might likewise support the idea of social orientations as an expression of beliefs. Thus, the differences in outcome effects between behavioral contexts might be caused by changes of an actor's own outcome-based motivations induced by changes of his beliefs. However, it might also be that some hypotheses about moderated influences of outcome-based motivations received no support in the data analyses because of various opposing effects. Therefore, disentangling influences of motivations from influences of beliefs would provide helpful insights.

In order to separate influences of beliefs and of motivations, beliefs have to be elicited. Moreover, rather than assuming that certain actions are basically perceived as friendly and other actions as unfriendly, the extent of perceived kindness of behavior in a certain context should be explicitly measured. Likewise, measuring people's emotions would help in the understanding of changes in people's motivations and decisions in various behavioral contexts. Vieth and Weesie (2006; and see Chapter 2) consciously avoided including respective measures in their experiment because commonly used measures change people's decision-making. Previous studies indicate biases towards pro-social behavior (Gächter and Renner, 2006; Hoffman et al., 2008) or towards selfishness (Croson, 2000). Understanding and accounting for such induced biases requires more research on how prompting social aspects affects people's decision-making in a certain interaction situation. Considering the context-dependency of outcome influences on behavior discussed in the study presented here, it is doubtful that questions about beliefs and emotions would only induce a higher level of pro-social behavior without also changing the effects of people's motivations.

Third, the same outcome components are involved in an actor's own motivations and in his beliefs about others' motivations. The same holds for interactions of out-

come components and different motivations (e.g., self-consistency and feelings of obligation or indignation). The discussions of possible influences of the various motivations and of beliefs show that the effects are often opposing and can cancel each other out. Moreover, hypotheses could be derived more clearly for ranges of individual parameters rather than an intuitive aggregation at the population level. For instance, it has been mentioned that the impact of the trustor's outcomes on trustworthiness is not necessarily positive for all trustees. Such differences might then also give rise to differences in how the behavioral context moderates the influence of the trustor's loss. Thus, formalization could help derive hypotheses, e.g., by employing a random utility approach (McFadden, 1973) and calculating quantal response equilibriums (McKelvey and Palfrey, 1998). This would allow for the derivation of comparative statics on how parameter variations influence the probability of a cooperative decision. Selected representations of other-regarding motivations can be explicitly modeled that would otherwise be captured by the random error term. For this purpose, a series of computer simulations could be set up in order to derive results for various values for individual utility weights that shape the influence of social orientations, intention-based motivation, and self-consistency. One difficulty is that more realistic assumptions would be desirable in order to derive empirically reasonable hypotheses, such as incomplete information about others' motivations and beliefs. This would require specifying distributions of beliefs about the distribution of others' individual parameters. The step from a theoretical model and derived hypotheses to a statistical model is straightforward in such an approach because the assumption made about the distribution of the random error term (typically a logistic or normal distribution) provides a direct link.

Fourth, it has been mentioned above that opposing effects can result not only from people's own motivations and beliefs but also from individually varying motivations (see also the previous discussion points). For instance, it is possible that not all trustees are positively concerned with the trustors gain in the TG, and for some trustees, self-consistency might not remove this positive concern after the promise has been made. Such individual heterogeneity affects analyses for trustors as well as for trustees with respect to both the direct influences of people's motivations and the way in which these influences are moderated by the behavioral context that activates further motivations. The statistical models employed in this study are based on the assumption that the effects of outcomes and behavioral contexts are the same for all subjects. Given the theoretical arguments about individually varying influences, statistical models would be preferred in which individual variations of coefficients for outcomes and for behavioral contexts are estimated. The results of multilevel models

with random coefficients would also provide estimates of the proportion of subjects for whom the effect is positive or negative, i.e., when altruism turns into aggression. This would allow for the analysis of how these proportions vary in different experimental conditions. Moreover, to analyze personal characteristics and to separate influences of beliefs from influences of motivations, multilevel structural equation models might be fruitful. Such models can include measurement components in order to estimate influences of personal characteristics. For instance, the stronger people's pro-social and moral orientation, the more consistent their behavior might be over time and across various decision situations (e.g., Smeesters et al., 2002).

Chapter 4

Influences of Promises and Threats on Trust and Trustworthiness

Experimental Evidence on Reciprocated Behavioral Advances

This study follows an approach developed together with Jeroen Weesie. The experiment was prepared and conducted while visiting Simon Gächter at Nottingham University. I am grateful for all the support both of them provided. I thank members of CeDEx at the Nottingham School of Economics for support and comments, in particular, Michail Drouvelis, Maria Montero, and Ping Zhang for help during pre-tests, Ruslan Kabalin for technical support, and Jo Morgan for checking language of instruction texts. Comments are acknowledged which have been made by members of the CREED/CeDEx/UEA meeting 2008 in Amsterdam.

Abstract

Promises involve the prospect of a favor that creates an obligation for repayment, whereas threats cause indignation and might trigger revenge. A game-theoretical lab experiment has been conducted in order to study the effects of behavioral advances in trust situations with sanctioning options by trustors and with announcement options for sanctions by trustors or for trustworthiness by trustees. Announcements are cheap-talk. Sanctions are costly and not always effective in removing objective incentives to abuse trust. The experiment is designed as within-subject sets of structurally identical (sub)games resulting from kind and unkind actual behavior in single encounters. This allows effects of objective outcomes and of individual heterogeneity to be controlled. The results show that promises strongly promote trustfulness and trustworthiness, despite the fact that promises are made most of the time. Due to the frequent decisions of making promises, no support could be found for a detrimental impact of punishment threats on trustworthiness.

4.1 Introduction

Trust is an important ingredient in social interactions as it enables improvements not reachable for one person alone (Coleman, 1990). However, placing trust is a "risky advance" in the sense that it provides others with an opportunity to take advantage of the situation. Such opportunistic behavior (Williamson, 1985) inflicts harm on those who trusted. Buying a used car is a classical example of a trust problem (Akerlof, 1970; Dasgupta, 1988; Buskens and Weesie, 2000). There are good cars and bad cars ("lemons"). Buyers typically cannot tell them apart after a short test drive and have to trust that the seller does not hide important information, e.g., about hidden damages. Buyers and sellers benefit if the buyer accepts a deal and receives a good car compared to no transaction. However, the seller is tempted to increase his profit by selling "lemons" at the price of a good car. While a car dealer's temptation is limited by his interest in ongoing business and good reputation, the trust problem is even more accentuated when sellers are private persons, e.g., incidentally offering goods and services by small advertisements in newspapers or on special internet platforms. Such incentive problems requiring trust are involved in many economic transactions and other social situations. Various mechanisms help mitigate these incentive problems, among them is communication (Coleman, 1990). For instance, people promise to deliver good quality and to deliver in time, to return a borrowed book, to share work, and to return gains from others' investments. Moreover, people also promise to reward good conduct or threaten to punish misbehavior. Typically, the purpose of promises and threats is to create an incentive for others to behave in a desired way. However, the mere act of making a promise or a threat can motivate or discourage certain behavior, even without changing objective incentives. Moreover, whereas promises are kind, omitted promises and especially threats involve unkindness and can trigger adverse effects. *How do promises and threats shape trustfulness and trustworthiness?*

Sociological and social-psychological research suggests that feelings of obligation to return favors and the desire for self-consistency can induce motivations to reciprocate (Cialdini, 2001: chs. 2–3; on obligation, see also Gouldner, 1960; Coleman, 1990: ch. 12). Similarly, people also feel indignation about perceived unkindness inducing a thirst for revenge (Gouldner, 1960). In Chapter 2, these arguments have been applied to trust situations. Experimental evidence shows that the favor of trustfulness is returned by trustworthiness (see also Gautschi, 2000; McCabe et al., 2003; Cox, 2004), that promises of trustworthiness promote trustfulness and induce people to behave consistently by keeping their promise, and that explicitly omitted promises are punished by withheld trust (see also Snijders, 1996; Gautschi, 2000). The present

study investigates whether and to what extent these findings concerning influences of trustfulness and of making or omitting promises of trustworthiness also apply to trust situations in which the trustor can sanction the trustee. Related experimental studies that include cooperation promises (Brandts and Charness, 2003; and in repeated interactions Bochet et al., 2006; Bochet and Putterman, 2007) provide interesting insights, but give rise to methodological concerns. For instance, a within-subject design and control for outcome-based motivations is desirable for addressing such questions (for a discussion, see Chapter 2).

Next, the present study also explores influences of punishment threats and of reward promises on trustworthiness. Concerning punishment threats, some studies stress detrimental influences (Fehr and Rockenbach, 2003; Fehr and List, 2004; Houser et al., 2008), whereas other studies find no support (Fehr and Schmidt, 2007) or a promoting impact (Voss and Vieth, 2006; Bochet and Putterman, 2007). Promises of reward have been found to increase trustworthiness (Fehr and Schmidt, 2004; Fehr et al., 2007). However, previous experiments involve confounding factors (e.g., they do not allow for punishment without a preceding explicit threat) and, again, the controlling for outcome-based motivations and individual heterogeneity should be improved. Next to such methodological issues, the study presented here focuses on process-based motivations resulting from obligation, indignation, and self-consistency. For these purposes, an experiment has been conducted following the approach employed by Vieth and Weesie (2006; and see Chapter 2). The experiment is designed as within-subject sets of structurally identical (sub)games resulting from kind and unkind actual behavior in single encounters. Announcements are "cheap-talk" (i.e., costless and non-binding) without reply option, and sanctions are costly and not always effective in objective terms. The cheap-talk character of announcements allows the question to be studied whether perceptions of kindness and unkindness depend on forgone outcomes of non-chosen options (see also Chapters 2 and 3). This is assumed in contemporary theoretical models that account for intention-based motivations, whereas previous studies reveal that mere communication can promote cooperative behavior.

4.2 Reciprocity, Announced Intentions, and Trust

4.2.1 Reciprocal Behavior as an Implication of Other-Regarding Motivations

Experimental research on social dilemmas provides ample evidence that people return favors and retaliate for others' unkind actions, even if it is against their objective self-interest (for reviews see, e.g., Camerer, 2003: ch. 2; Ostrom and Walker, 2003;

Kopelman et al., 2002; Kollock, 1998; Komorita and Parks, 1996; Ledyard, 1995; van Lange et al., 1992; Messick and Brewer, 1983; Pruitt and Kimmel, 1977). This fundamental behavioral pattern is known as reciprocity (Fehr and Schmidt, 2006; Hann, 2006; Kolm, 2006; Lévy-Garboua et al., 2006). Reciprocity can arise from other-regarding motivations that are rooted in emotions and complement utility that an actor derives from his own objective outcomes. People feel an *obligation* to return a favor, even if the received favor is unwanted (Gouldner, 1960; Coleman, 1990: ch. 12; Cialdini, 2001: ch. 2). Omitting or delaying to fulfill this obligation inflicts intrinsic distress and emotional tension ("shadow of indebtedness", Gouldner, 1960: 174). Similarly, "sentiments of retaliation" (Gouldner, 1960: 172), e.g., anger or irritation, due to experienced harm induces a thirst for revenge, especially if the harm could have been avoided. Such feelings of *indignation* can also be invoked by threats, especially if perceived as unfair. Even omitting something kind without inflicting or threatening objective harm can demand retaliation (e.g., for omitting a promise, see Chapter 2).

Other people's kind and unkind behavior activates *intention-based motivations*, i.e., people take into account the process of how certain outcomes are obtained (for experimental studies, see Snijders, 1996; Gallucci and Perugini, 2000; Gautschi, 2000; Brandts and Solà, 2001; Falk et al., 2003; McCabe et al., 2003; Cox, 2004; Charness and Rabin, 2005; also see Chapters 2 and 3). Falk and Fischbacher (2006) propose a theoretical model in which perceived kindness depends on intentionality and on the size of outcome changes caused by others' behavior (for other models accounting for intentions, see Rabin, 1993; Levine, 1998; Charness and Rabin, 2002; Dufwenberg and Kirchsteiger, 2004). This perspective of modeling intention-based motivations is based on, and improves upon, research devoted to outcome-based motivations. *Outcome-based* motivations are implications of social (value) orientations rooted in social comparisons (Messick and McClintock, 1968; McClintock, 1972; Liebrand, 1984), i.e., preferences concerning the distribution of actors' own and others' outcomes (for reviews, see Au and Kwong, 2004; McClintock and van Avermaet, 1982). Two prominent examples are models of altruism based on benevolence and spite (e.g., Brew, 1973; Taylor, 1987/1976; Weesie, 1993, 1994b; Snijders, 1996) and models of fairness in terms of inequality aversion induced by guilt and envy (e.g., Kelley and Thibaut, 1978; MacCrimmon and Messick, 1976; Weesie, 1994a; Ledyard, 1995; van Lange, 1999; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000).

In addition to other-regarding motivations, people are also driven by a *desire for self-consistency*. For instance, social-psychological studies report that people are found to act according to prior agreements even when discovering costs, to adopt

opinions they announced publicly even when forced, and to start favoring chosen alternatives despite prior indifferent or even opposite preferences (for reviews see, e.g., Webster, 1975; Cialdini, 2001: ch. 3; Kunda, 2002; Gass and Seiter, 2007). By behaving consistently with their own preceding decisions that resulted in a specific decision situation, people avoid cognitive dissonance (Heider, 1944, 1958; Festinger, 1957; Aronson, 1992; Akerlof and Dickens, 1982). As argued in Chapter 2, self-consistency can result in reciprocal behavior even without an interest in reciprocating the other person's decisions, but can also increase or undermine obligation feelings. The increase is based on shared responsibility for others' subsequent decisions. Reasons for the undermining influence are provided by social-psychological research on mechanisms that reduce cognitive dissonance. This research shows that people convince themselves that they made appropriate decisions and even deny objective counter-evidence that threatens their peace of mind (Cialdini, 2001; Gass and Seiter, 2007). In situations in which rewarding others' kindness conflicts with the desire for self-consistency, people tend to develop excuses and justifications for omitting the reward and they might even frame others' kindness as unfriendly.

4.2.2 Kindness of Promises and Unkindness of Threats

Previous studies reveal a strongly positive influence of communication on cooperative behavior (for reviews see, e.g., Sally, 1995; Crawford, 1998; Kopelman et al., 2002; Bicchieri, 2002; Shankar and Pavitt, 2002; Ostrom and Walker, 2003; Brosig, 2006). Most studies focus on face-to-face communication in small groups. The promoting impact has been found to arise if the decision situation is discussed (rather than socializing by talking about an unrelated topic) and if people make explicit promises to perform a certain behavior (Dawes et al., 1977). However, face-to-face communication undermines anonymity between interaction partners, such that participants' concerns about their reputation after the experiment has ended are likely to influence their decision-making during the experiment. Therefore, various other means of communication have been investigated. For instance, exchanging messages in written form while maintaining anonymity has been found to be similarly influential (e.g., Brosig et al., 2003; Bochet et al., 2006). In contrast to exchanges of messages with largely uncontrolled content, other experiments used pre-defined message options (Brandts and Charness, 2003; Bochet and Putterman, 2007). This helps isolate promoting effects of cheap-talk promises from influences of other contents that are communicated in self-composed messages and difficult to control.

Promises are expressed intentions to perform a certain action that yields a gain to the other person. Due to the prospect of gains, promises are kind advances and demand a favor in return (Cialdini, 2001; also see Chapter 2). Moreover, given intrinsic bonds that arise from the desire for self-consistency, a promise serves as a commitment in the sense of a "voluntary strategic action", costly or not, with the purpose of "reducing one's freedom of choice" or of changing the outcomes of choices (Schelling, 1960). In terms of changing outcomes, a commitment is a "strategic move" by which an actor voluntarily offers a "hostage" in the sense of a bond (Schelling, 1960). Such commitments associated with objective incentives (binding values, compensating values, transaction costs) have been studied in trust situations and cooperation problems theoretically (Weesie and Raub, 1996; Voss, 1998b; Raub and Weesie, 2000; Raub, 2004; and including other-regarding motivations Snijders, 1996) as well as experimentally (Raub and Keren, 1993; Snijders, 1996; Mlicki, 1996 also see Chapter 2; and for negotiation problems, also see Prosch, 2006). Empirical results show that imperfectly binding commitments also promote cooperative behavior and that even small transaction costs hamper commitment posting.

In addition to promising cooperation, another mechanism that can mitigate incentive problems is an opportunity for sanctions (for reviews, see Roth, 1995; Camerer, 2003: ch. 2; Shinada and Yamagishi, 2008). For instance, Fehr and Gächter (2000, 2002) show that an opportunity for informal peer punishment is used and increases contribution to public goods among strangers in single encounters, even if punishment is costly for the punisher (see also Gächter et al., 2008, on long-run efficiency of punishment ; and for findings on costly reward, see Sefton et al., 2007). Other studies reveal limitations and reveal detrimental effects of punishment (e.g., Fehr and Rockenbach, 2003; Nikiforakis and Normann, 2008; Voss and Vieth, 2006; Carpenter, 2007; Egas and Riedl, 2008; Herrmann et al., 2008) and of reward (Gürerk et al., 2004). Nevertheless, previous studies on repeated interactions report a particularly promoting influence on cooperation of a combination of both communication and sanctioning opportunities, because this combination allows for the punishment of lies (Ostrom et al., 1992; Bochet et al., 2006; Bochet and Putterman, 2007).

The studies on communication and on commitments primarily focus on promises of cooperation. In the presence of sanctions, reward for cooperation can also be promised. In some previous experiments, reward promises have been addressed in terms of incentive schemes such as bonus contracts (e.g., Fehr and Schmidt, 2004, 2007; Fehr et al., 2007) or third party enforced side payments (Andreoni and Varian, 1999). The results of these experiments indicate that reward promises promote

cooperative behavior, although promises are almost always made and much less frequently kept. In addition to promising a reward, punishment can also be threatened in advance. Threats are expressed intentions to perform a certain action that inflicts a loss upon the other person. Since the desire for self-consistency induces people to perform the threatened action, a threat without objective grounds can also serve as a commitment (in the sense of Schelling, 1960). However, while promises induce obligation feelings and motivate friendly responses, threats invoke feelings of indignation that trigger retaliation and reactance (i.e., doing the opposite of what is demanded by others, see Brehm, 1966).

Some experimental studies suggest support for detrimental influences of punishment threats on cooperative behavior (e.g., measured as back-transfer in an Investment Game by Fehr and Rockenbach, 2003; Fehr and List, 2004; Houser et al., 2008). Other studies do not find support for such influences (Fehr and Schmidt, 2007) or find a promoting impact (Yamagishi, 1986; Voss and Vieth, 2006; Bochet and Putterman, 2007). Fehr and Rockenbach (2003) argue that threats are punished that are associated with unfair claims. However, their results actually show that omitting such threats is rewarded by cooperative behavior, while threatening punishment does not significantly reduce cooperative behavior. Note the difference: The underlying motivation is thus not indignation, but a positive feeling of obligation or gratitude. Moreover, in the experimental design of these studies, the influence of a threat is confounded with the availability of punishment (third party enforced, i.e., participants could not choose to carry out punishment). As previously mentioned, punishment can also undermine cooperation. Thus, a possible positive impact of an explicit punishment threat can be hidden by a stronger hampering influence induced by the mere possibility of punishment. The "detrimental effect" emphasized in these studies might be due to the combination of both threat and punishment. Supporting this intuition, Voss and Vieth (2006) find evidence that making threats promotes cooperation, but that the mere availability of both threats and punishment hampers cooperation. Further confounding factors in previous experiments include, e.g., that the effectiveness of punishment depends on the extent of cooperative behavior because the extent of possible punishment has been fixed while participants have chosen a level of cooperation. Some studies involve both reward promises and punishment threats. As mentioned above, these studies show that the majority of participants chooses to promise rewards rather than threaten punishment (Fehr and Schmidt, 2007; Fehr et al., 2007). However, this finding is ambiguous as choosing to

threaten punishment required an investment, whereas promising a reward was not associated with transaction costs.

In addition to the specific issues discussed above, nearly all of the previous experiments mentioned above do not employ a within-subject design for investigating influences of individual motivations and do not sufficiently control for influences of outcome-based motivations (for a more detailed discussion, see Chapter 2). The experiment presented here has been designed to study the "pure" influence of threats and promises on subsequent decisions. The focus is on trust situations and on employing the social-psychological insights introduced above on obligation, indignation, and self-consistency.

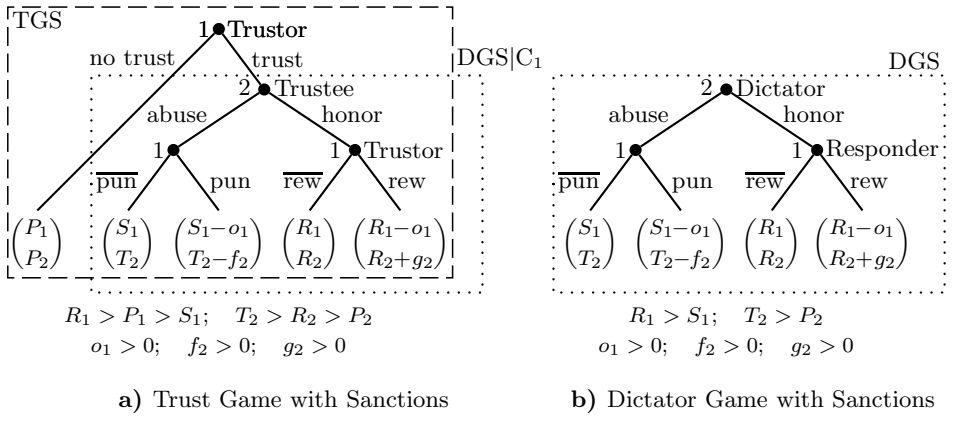
4.2.3 Promises and Threats in Trust Situations with Sanctions

Informal Sanctions in Trust Situations

In trust situations, a trustor chooses to place or to withhold trust and a trustee decides whether to honor or to abuse placed trust. The core features of such interaction situations are described by the Trust Game (Dasgupta, 1988; Kreps, 1990). For the purpose of this study, the Trust Game is supplemented with sanctioning options such that the trustor can reward the trustee for honoring trust or punish him for abused trust (TGS, Figure 4.1a). Outcomes are represented in objective terms, e.g., money. While both actors benefit from honored trust ($R_i > P_i$, with $i = 1, 2$), the trustee is tempted to abuse trust ($T_2 > R_2$) which inflicts a loss upon the trustor ($S_1 < P_1$). The trustee's temptation is reduced by the fine (f_2) and by the gratification (g_2) while the trustor incurs the outlay (o_1) for sanctioning. Punishment and reward are effective in objective terms if the trustee's temptation to abuse trust is removed ($f_2 + g_2 > T_2 - R_2$). However, sanctions are only credible if they do not involve objective costs for the trustor ($o_1 \leq 0$). Given that sanctions are costly ($o_1 > 0$) and that actors largely care about their own objective outcome, trustors neither reward nor punish the trustee and therefore withhold trust because trust would be abused.

Fairness orientations can create an incentive for trustors to punish abused trust, while altruistic tendencies or cooperative preferences can motivate trustors to reward honored trust. Similarly, the trustee might suffer from guilt concerning the loss inflicted upon the trustor. Trustees with strong guilt feelings honor trust even without expecting punishment or reward. For trustees with weaker guilt feelings, the possibility of sanctions promotes honoring trust. However, if people's decision-making was only driven by outcome-based motivations, it would not make a difference whether or not the gains the trustee can share resulted from the trustor's trustfulness. Removing

Figure 4.1: Trust Game with Sanctions (TGS) and dichotomous Dictator Game with Sanctions (DGS)



Decision labels are abbreviated with "pun" for punishment and "rew" for reward; bar for "no".

the trustor's decision of whether or not to place trust from the TGS yields a dichotomous Dictator Game with Sanctions (DGS, Figure 4.1b). The trustee is then in the position of a dictator, but restricted by an active responder (formerly, the trustor) who can reward if the dictator shared the gains and punish if the dictator kept the gains. The subgame of the TGS starting with the trustee's decision and the DGS consist of identical choice options and identical outcomes (i.e., structurally identical subgames). Only the behavioral context differs, i.e., the preceding choice options and actual decisions made by which the alternative decision parts and outcomes are excluded. In contrast to outcome-based motivations, intention-based motivations induce actors to respond to previous decisions and, thus, to discriminate between the decision situations.

In Chapter 2 it has been argued that trustfulness is a kind advance that creates an obligation for trustees to return the favor. The experiment presented in Chapter 2 is concerned with trust situations without sanctions. Is has been found that people indeed tend to be more generous as a trustee than as a dictator (see also McCabe et al., 2003; Cox, 2004). In the TGS, the trustee likewise is in the favorable position solely because the trustor placed trust. Moreover, in placing trust, the trustor risks a loss that can be inflicted by the trustee's decision to abuse trust. Furthermore, trustfulness improves the outcomes of the trustee, who decides whether to share the benefits. Since the trustor's decision of whether or not to place trust is absent in the

DGS, the trustee has no obligation to return a favor, but only focuses on outcomes and possible sanctions.

> **Hypothesis 4.1: Kindness of placed trust**
> Compared to honoring trust in the TGS, gains are *less* likely to be shared in the DGS.

Promises of Trustworthiness in Trust Situations with Sanctions

Communication possibilities enable trustees to promise trustworthiness before the trustor decides whether to place trust in the TGS (H_2TGS). This creates two behavioral contexts for the TGS: one after the trustee has promised trustworthiness (TGS$|H_2^+$) and one after the trustee has omitted the promise (TGS$|H_2^0$). Decisions in each of the two behaviorally embedded TGS can be compared to decisions made in the TGS without promise option. Sanctioning options also allow lies to be punished and kept promises to be rewarded. The prospect of punishment or reward can provide additional incentives that increase the promoting impact of promises, especially if promises are cheap-talk. However, such extrinsic incentives can also undermine the impact of communication such that the mere word of a stranger loses its impact. Moreover, recall that some studies report limitations of the cooperation enhancing impact of sanctions. If potential sanctions hamper cooperative behavior, the positive influence of promises on cooperative behavior can likewise be lost. Nevertheless, obligation feelings, indignation feelings, and the desire for self-consistency have been found to provide a powerful motivational basis (see Chapter 2; Gouldner, 1960; Coleman, 1990: ch. 12; Cialdini, 2001: chs. 2–3). In line with this argument, previous studies, though on repeated interactions, found that the combination of both sanctions and communication promotes cooperative behavior particularly strongly (Ostrom et al., 1992; Bochet et al., 2006; Bochet and Putterman, 2007).

First, consider again the trustee's decision of whether or not to honor trust. For trust situations without sanctions, it has been argued in Chapter 2 that making the promise promotes honored trust, and experimental support has been provided for this influence. In the presence of sanctioning options, promising trustworthiness (TGS$|H_2^+$) should likewise increase the feeling of obligation to return the favor of trustfulness because the trustee shares some responsibility for the trustor's decision to place trust. Moreover, by honoring trust after having promised to do so, the trustee behaves consistently and avoids intrinsic distress caused by lying or reneging on a promise.

Hypothesis 4.2: Kindness of placed trust after trustworthiness has been promised
Compared to the TGS (i.e., without promise opportunity), trust is *more* likely to be honored after trustworthiness has been promised ($\text{TGS}|\text{H}_2^+$).

In contrast, omitting the promise might hamper trustworthiness ($\text{TGS}|\text{H}_2^0$). In Chapter 2, reasons have been provided (summarized below) and empirical support has been found that the influence of omitted promises on trustworthiness depends on the promise properties. No support has been found for a hampering impact of omitting a promise in the case of cheap-talk promises (see also Snijders, 1996; Gautschi, 2000). The arguments that have been provided in Chapter 2 can also be applied to trust situations involving sanctioning options. Although trustfulness is a friendly advance, it has to be considered that the trustee explicitly omitted to promise trustworthiness. The desire for self-consistency might therefore undermine the feeling of obligation to return the favor of placed trust. Based on social-psychological research on methods that reduce cognitive dissonance (for reviews, see Webster, 1975; Gass and Seiter, 2007), trustfulness can even be perceived as unkind and abusing trust as more legitimate after the promise has been omitted.

Hypothesis 4.3: Unkindness of placed trust after trustworthiness has not been promised
Compared to the TGS (i.e., without promise opportunity), trust is *less* likely to be honored after a possible promise of trustworthiness has not been made ($\text{TGS}|\text{H}_2^0$).

The trustee's decision of whether or not to promise trustworthiness also constitutes a behavioral context for the trustor's choice between placing and withholding trust. At first, imagine the decision situation after the trustee has made the promise ($\text{TGS}|\text{H}_2^+$). It has been argued in Chapter 2 that making the promise promotes trustfulness, and empirical support has been found for this argument. Following their arguments, promises are kind advances that create an obligation to return the favor. The trustor is therefore inclined to place trust in order to fulfill the outstanding obligation. Moreover, the trustor might anticipate the increased chance of trustworthiness (Hypothesis 4.2). Furthermore, trustworthiness can be seen as a reward for trustfulness because the trustor benefits from sharing gains (Hypothesis 4.1). Therefore, making a promise to honor trust can be interpreted as promising a reward and can thus promote trustfulness.

Hypothesis 4.4: Kindness of promising trustworthiness
Compared to the TGS (i.e., without promise opportunity), trust is *more* likely to be placed after trustworthiness has been promised ($TGS|H_2^+$).

If the trustee explicitly omitted to promise his trustworthiness ($TGS|H_2^0$), he omitted a friendly option. Previous experimental evidence suggests that trustors are more reluctant to place trust after the promise has been omitted than in the decision situation in which no promise is possible (see Chapter 2; Gautschi, 2000; Snijders, 1996). One reason is that the trustor might anticipate that trustees are less trustworthy (Hypothesis 4.2). Another reason is based on the assumption that omitting a kind action can be interpreted as behaving in an unkind manner. Thus, even trustors who expect a sufficiently high chance that trust would be honored might retaliate by withholding trust for the trustee's unkindness of an explicitly omitted promise (see Chapter 2).

Hypothesis 4.5: Unkindness of not promising trustworthiness
Compared to the TGS (i.e., without promise opportunity), trust is *less* likely to be placed after a possible promise of trustworthiness has not been made ($TGS|H_2^0$).

Reward Promises and Punishment Threats in Trust Situations with Sanctions

Next, the trustor can seize communication possibilities for announcing sanctions (H_1TGS). In order to separate the influences of announcing sanctions from promising trustworthiness, it is assumed that the trustee has no option to promise trustworthiness. Moreover, consider that there will be no further decision stage if the trustor decides to withhold trust. Thus, announcing sanctions is only meaningful in combination with placing trust. Therefore, the trustor chooses one of four options: withholding trust, placing trust while promising reward ($TGS|H_1^+$), placing trust while threatening punishment ($TGS|H_1^-$), or placing trust without announcing sanctions ($TGS|H_1^0$). Thereafter, the trustee decides whether or not to honor trust, followed by the trustor's decision of whether or not to incur the costs for punishing or rewarding the trustee. The combination of the decision to place trust and the decision about announcing sanctions creates three behavioral contexts for the subsequent decisions. The trustee's decision in each of these three behavioral contexts can be compared to the trustee's decision in the TGS without announcement option.

It has been argued that the trustee feels an obligation to return the favor of placed trust (Hypothesis 4.1). In addition, recall that the trustor has to sacrifice an outlay to reward the trustee and that the trustee's outcome is increased by the gratification. Thus, a reward promise ($\text{TGS}|\text{H}_1^+$) is a friendly advance (see also Hypothesis 4.4). This increases the trustee's feeling of obligation because the trustee received two favors: the favor of placed trust and the favor of a reward promise. Moreover, the trustor's reward promise indicates that the trustor thinks positively about the trustee and focuses on the situation of honored trust rather than fearing abused trust. Such kind indications increase the trustee's feelings of obligation to honor trust. The trustee could also expect that the trustor might be more inclined to incur the outlay for rewarding honored trust once the trustor promised to do so (Chapter 5). Therefore, the trustor's reward promise is expected to promote trustworthiness.

> **Hypothesis 4.6: Kindness of placed trust with a reward promise**
> Compared to the TGS (i.e., without announcement opportunity), trust is *more* likely to be honored after a reward for trustworthiness has been promised ($\text{TGS}|\text{H}_1^+$).

Whereas a promise is friendly, a threat involves unkindness ($\text{TGS}|\text{H}_1^-$). As previously argued, punishing inflicts avoidable harm upon the other person. The trustor's willingness to incur costs in order to reduce the trustee's outcome even involves an aggressive component. Moreover, the threat of punishment shifts the focus to the situation of abused trust and thereby expresses the trustor's distrust (Fehr and Falk, 2002). This undermines the trustee's feeling of obligation to return the favor of placed trust. The trustee's indignation about the negative expectation, which the trustor revealed by the threat, might even motivate the trustee to retaliate by abusing trust (for a similar argument, see Fehr and Rockenbach, 2003). Of course, the trustee might expect an increase in the probability of being punished by the trustor (Chapter 5). However, strong indignation feelings can outweigh the expected loss that is caused by the fine. Note that placed trust itself can even be perceived as unkind if it is combined with a punishment threat. The reason is that the trustor attempts to exert power over the trustee and to reduce the trustee's freedom of choice. Actions taken to dominate another person are typically perceived as unkind. Social-psychological research has shown that a perceived involuntary reduction of personal freedom causes reactance (Brehm, 1966), i.e., the trustee is inclined to protest by doing the opposite of what the trustor demands from him.

Hypothesis 4.7: Unkindness of placed trust with a punishment threat
Compared to the TGS (i.e., without announcement opportunity), trust is *less* likely to be honored after a punishment for abused trust has been threatened ($\text{TGS}|\text{H}_1^-$).

If the trustor placed trust but neither promised reward nor threatened sanctions ($\text{TGS}|\text{H}_1^0$), he omitted a friendly option and an unfriendly option. At first sight, not announcing sanctions might therefore be perceived as neither kind nor unkind. However, if trustors primarily promise reward, the trustee comes to expect a reward promise and will be disappointed not to receive it. The omitted announcement then becomes unfriendly and triggers revenge. Similarly, if the trustee expected to be threatened with punishment, the omitted announcement will be perceived as a favor that demands to be rewarded. Thus, depending on the trustee's beliefs, an omitted sanctioning announcement should promote or hamper trustworthiness.

Hypothesis 4.8: Belief-dependent kindness of placed trust with an omitted sanctioning announcement
Compared to the TGS (i.e., without announcement opportunity), trust is *less* likely to be honored after neither reward has been promised nor punishment has been threatened ($\text{TGS}|\text{H}_1^0$) the more the trustee expects a reward promise, and trust is *more* likely to be honored in this decision situation, the more the trustee expects a punishment threat.

Summarizing the hypotheses highlights the underlying principle of reciprocity (Table 4.1). Kind advances (as in the $\text{TGS}|\text{H}_2^+$ and in the $\text{TGS}|\text{H}_1^+$) should trigger friendly responses based on feelings of obligation to return the favor (Hypotheses 4.4 and 4.6). This should also become visible compared to decision situations in which an option for a kind advance is lacking (as in the DGS; see Hypothesis 4.1). In contrast, unkind advances (as in the $\text{TGS}|\text{H}_1^-$), including omitted kindness (as in the $\text{TGS}|\text{H}_2^0$), are assumed to demand retaliation (Hypotheses 4.5 and 4.7). Moreover, the desire for self-consistency can increase obligation feelings due to shared responsibility (as in the $\text{TGS}|\text{H}_2^+$), but can also undermine perceived kindness (as in the $\text{TGS}|\text{H}_2^0$; see Hypotheses 4.2 and 4.3).

Table 4.1: Overview of hypotheses and notation

	Placing Trust	Honoring Trust	
DGS		–	Dictator Game with Sanctions (no placed trust)
TGS	(ref.)	(ref.)	Trust Game with Sanctions
$TGS\|H_2^+$	+	+	TGS after a made promise to honor trust
$TGS\|H_2^0$	–	–	TGS after an omitted promise to honor trust
$TGS\|H_1^+$		+	TGS after placed trust with a reward promise
$TGS\|H_1^-$		–	TGS after placed trust with a punishment threat
$TGS\|H_1^0$		+/–	TGS after placed trust with an omitted announcement of sanctions

The hypotheses are formulated in terms of differences toward the TGS (i.e., the behavioral context without anouncement options).

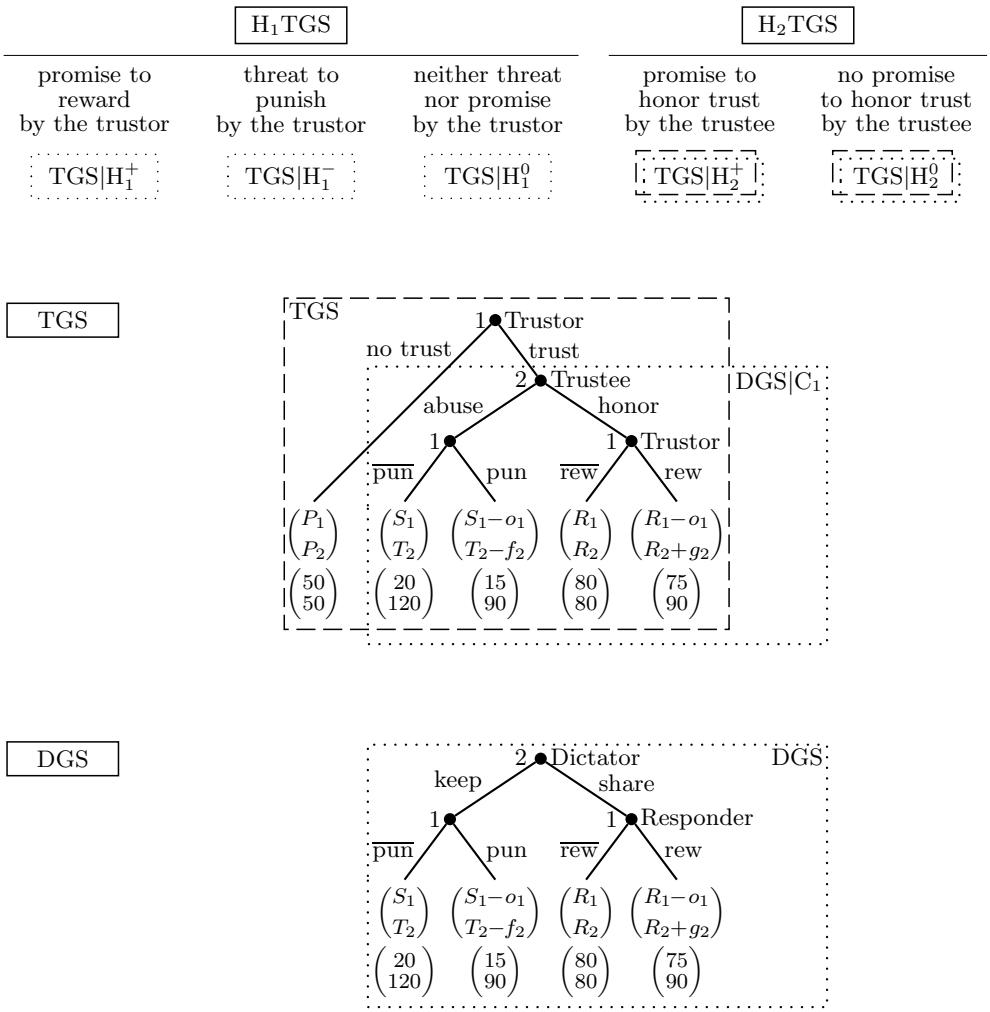
4.3 Design of the Experiment, Data, and Statistical Method

4.3.1 Experimental Design: Sets of (Sub)Games

The design of the experiment satisfies two main features. First, the behavior of each participant in decision situations with different behavioral contexts is recorded. Second, influences of outcome-based motivations can be controlled. For this purpose, participants made decisions in different games, for the analyses presented here, in TGSs, H_1TGSs, H_2TGSs, and DGSs (for details, see Vieth, 2008). Each game was a single encounter. Following the approach employed by Vieth and Weesie (2006; and see Chapter 2), the experiment was designed as within-subject sets of (sub)games that consist of the same behavioral options and the same outcomes for both actors (for related designs, see Snijders, 1996; Charness and Rabin, 2002, 2005; McCabe et al., 2003; Cox, 2004; and for further remarks, see Vieth and Weesie, 2006; and see Chapter 2). This design allows for the comparison of participants' behavior in decision situations with identical choice options and identical outcomes but in different behavioral contexts that result from preceding kind and unkind behavior (Figure 4.2). Intra-personal differences in behavior between the (sub)games should therefore reflect the impact of the behavioral contexts.

In the H_2TGS, the trustee decides whether or not to promise trustworthiness. Thus, an H_2TGS contains two TGS: one after the promise has been made ($TGS|H_2^+$) and one after the promise has been omitted ($TGS|H_2^0$). In the H_1TGS, the trustor chooses among three announcement options. Since the trustor's "announcement" decision implies that he placed trust, it does not constitute a behavioral context

Figure 4.2: Sets of games with identical subgames



In the H_1TGS, the trustor can only choose to announce sanctions (3 options) if he places trust. Full graphs of the H_1TGS and the H_2TGS are omitted because these graphs are complex without being more informative for cheap-talk announcements. The experimental design allows for the comparison of the trustor's trustfulness in (sub)games indicated by *dashed boxes* (3 contexts) and the trustee's trustworthiness in (sub)games indicated by *dotted boxes* (7 contexts). These sets of (sub)games constitute "subject-payoff response sets" used in statistical analyses. Numerical example: $S_1^{\text{low}} = 20$, $T_2^{\text{high}} = 120$, $R_1 = R_2 = 80$, $P_1 = P_2 = 50$, $o_1^{\text{low}} = 5$, $f_2^{\text{high}} = 30$, $g_2^{\text{low}} = 10$. Decision labels are abbreviated with "pun" for punishment and "rew" for reward; bar for "no".

Figure 4.3: Outcome parameters of the experimental design

DESIGN PARAMETERS:
$S_1(2) \times T_2(2) \times \pi_i(3) \times f_2(3) \times g_2(3)$

Symmetric payoff structure: $S_1(2) \times T_2(2)$

$S_1^{\text{low}} = 20 \quad T_2^{\text{low}} = 100$
$S_1^{\text{high}} = 40 \quad T_2^{\text{high}} = 120$

$R_1 = R_2 = 80$
$P_1 = P_2 = 50$

Sanctioning properties: $f_2(2) \times g_2(2)$

$f_2^{\text{low}} = g_2^{\text{low}} = \frac{1}{4}(T_2 - R_2) = \{5, 10\}$
$f_2^{\text{high}} = g_2^{\text{high}} = \frac{3}{4}(T_2 - R_2) = \{15, 30\}$

$o_1^{\text{low}} = \frac{1}{6}(R_1 - P_1) = 5$
$o_1^{\text{high}} = \frac{2}{6}(R_1 - P_1) = 10$

Asymmetric payoff structure: $\pi_i(3)$
$(\pi_1 \in \{R_1, P_1, S_1\};\ \pi_2 \in \{T_2, R_2, P_2\})$

Trustor advantage: $\pi_1 + 10$, $\pi_2 - 10$
Trustee advantage: $\pi_1 - 10$, $\pi_2 + 10$

The outlay was fixed per combination of T_2, S_1, f_2, g_2, varying for three of the four parameters.

for trustfulness, but only for the subsequent decisions. Note that these subsequent decisions are made in a behaviorally embedded TGS (not in a DGS) because of the trustor's preceding decision to place trust. Therefore, an H_1TGS contains three TGSs after placed trust: one in which reward has been promised (TGS$|H_1^+$), one in which punishment has been threatened (TGS$|H_1^-$), and one in which no announcement has been made (TGS$|H_1^0$). Each TGS consists of one DGS as a subgame. In addition to these behaviorally embedded subgames, the design also included TGSs and DGSs as separate games. The announcement options for trustors in the H_1TGS and for trustees in the H_2TGS were cheap-talk, i.e., the choice did not change objective outcomes. Therefore, the two TGSs of the same H_2TGS and the three TGSs of the same H_1TGS had identical payoffs. Sets of games were constructed consisting of one TGS, one H_1TGS, one H_2TGS, and one DGS. Within each set, (sub)games had identical payoffs.

Varying some outcome parameters yielded different sets of (sub)games (Figure 4.3) These variations were included in the design for methodological reasons (for details, see Vieth, 2008) and for further analyses. The payoffs resulting from abused trust (S_1 and T_2) were varied at two levels each (low, high). This yields four baseline payoff combinations with 20 and 40 for S_1 and 100 or 120 for T_2. The baseline payoffs after no trust (P_i) and after honored trust (R_i) were fixed at $P_1 = P_2 = 30$ and at $R_1 = R_2 = 60$. Thus, the baseline payoffs represent a symmetric payoff struc-

ture. Two asymmetric payoff structures were constructed by adding 10 to the payoffs for the advantaged position and simultaneously subtracting 10 from the payoffs for the disadvantaged position. This yielded four payoff combinations with a trustor advantage and four payoff combinations with a trustee advantage. The three sanctioning properties were likewise varied at two levels each (low, high) resulting in eight combinations. The fine (f_2) and the gratification (g_2) were varied at the scale of the trustee's temptation with $\frac{1}{4}(T_2 - R_2)$ as "low" values and $\frac{3}{4}(T_2 - R_2)$ as "high" values. Thus, fine and gratification vary with the trustee's payoff from abused trust (T_2). For instance, in the case of $T_2 = 120$, the trustee's temptation amounts to $T_2 - R_2 = 40$, such that $f_2^{\text{low}} = g_2^{\text{low}} = \frac{1}{4}(T_2 - R_2) = 10$ and $f_2^{\text{high}} = g_2^{\text{high}} = \frac{3}{4}(T_2 - R_2) = 30$. Note that sanctioning is effective if both fine and gratification have "high" values but is never credible because of $o_1 > 0$. The outlay (o_1) was defined at the scale of the difference between the trustor's sure payoff and the trustor's gain from honored trust with $o_1^{\text{low}} = \frac{1}{6}(R_1 - P_1) = 5$ and $o_1^{\text{high}} = \frac{2}{6}(R_1 - P_1) = 10$. The four combinations of sanctioning properties for player 2 and the four symmetric payoff combinations yielded 16 combinations. The outlay was fixed per combination in a way that it varied for three of the four parameters across these combinations (for details, see Vieth, 2008). Together with the three payoff structures, the parameter variations resulted in 48 different combinations of payoffs and sanctioning properties.

Each participant made decisions in two sets of games in the role of player 1 (trustor, responder, allocator) and in two game sets in the role of player 2 (trustee, dictator, receiver) and was for each encounter randomly and anonymously matched with another participant (stranger matching whereby the probability of re-matching was minimized within each type of game, see Vieth and Weesie, 2006). The sets of (sub)games were mixed by clustering types of games. For the results reported here, the clustering was as follows: first 10 TGSs, followed by 10 H_1TGSs, subsequently 10 H_2TGSs, and then 8 DGSs. The experiment also included some TGs after the H_2TGSs, some H_2TGs after the DGSs, and finally some distribution situations with a passive receiver such as Dictator Games (for details, see Vieth, 2008). The ordering of game clusters was fixed in a way that maximal differences between game clusters were assured concerning the presentation of decision situations (for details, see Vieth, 2008). Two TGSs, two H_1TGSs, and two H_2TGSs were constructed without objective incentive for trustees to abuse trust ($T_2 < R_2$). These decision situations are not used in the analyses but were included in the design in order to check for participants' attention. In fact, 93.0% trustfulness and 99.0% trustworthiness were observed in these decision situations (see Table 4.3 for averages per behavioral context when $T_2 > R_2$). This in-

dicates that participants strongly paid attention to the objective outcomes. Note that in the decision situations in which $T_2 < R_2$, neither full trustfulness nor full trustworthiness was expected, due to possible influences of other-regarding outcome-based motivations (e.g., aggressive or competitive tendencies). Two brief questionnaires concerning participants' socio-demographic characteristics (e.g., gender, age, education) separated TGSs from H_1TGSs and H_1TGSs from H_2TGSs. Other questions about personal attitudes and opinions followed at the end of the experiment (for details on questionnaires, see Vieth, 2008). Analyses of questionnaire items are not reported here. In each game cluster, player roles were changed after half of the periods. In addition to randomly changing interaction partners, payoffs and sanctioning properties changed from one period to the next. The combinations and sequences of payoffs and sanctioning properties were varied across experimental sessions employing a factorial design.

The experiment was computer-assisted, employing the software package "z-Tree" (Fischbacher, 2007) (for an example of the decision screens, see Appendix A.2). In addition to general information on paper, participants received on-screen instructions and a tutorial before each game cluster. Outcomes were displayed as points in decision trees and represented monetary gains (one British pence for four points). Participants were paid anonymously and immediately after the experiment. On average, participants earned approximately 14 GBP. The experiment was conducted in April 2008 at the CeDEx lab at the Nottingham School of Economics. Using "ORSEE" (Greiner, 2004), 166 students were recruited from the CeDEx participant pool and took part in nine groups of 16 to 20 participants. Participants were enrolled in various fields, primarily at Nottingham University.

4.3.2 Data and Statistical Method

The 166 subjects made 1992 decisions of whether or not to place trust in the role of the trustor and 1627 decisions of whether or not to share gains in the role of the trustee or in the role of the dictator (Table 4.2). Trustors' "placing trust" decisions involve 664 decisions made in the H_1TGS in which placing trust is combined with the decision about announcing sanctions. Each participant made 12 decisions in the trustor role, i.e., four in the TGS without behavioral context, four in the TGSs as a subgame of the H_2TGS, and four in the TGSs involved in the H_1TGS. Since trustees can only decide whether to honor trust if the trustor has placed trust, the number of decisions trustees made varies. Of 16 possible decisions in the role of the trustee or dictator, subjects actually made between 5 and 15 decisions, i.e., all four

Table 4.2: Number of cases and units of analyses

Number of ...	Placing trust	Honoring trust
subjects	166	166
total payoffs	48	48
subject-payoff response sets	664	664
decisions in total	1992	1627
decisions per subject	12	5–15
decisions per response set	3	1–4

Total payoffs are combinations of payoffs and promise properties. Decisions for placing trust without H_1TGS: 1328 in total, 8 per subject, 2 per response set.

"sure decisions" in the DGSs and at least one decision in one of the 12 TGSs in which the trustee can only make a decision if the trustor has chosen to place trust. Due to the "sure decisions" by the trustor in the TGSs and by the trustee in the DGSs, all 48 different combinations of payoffs and sanctioning properties were realized for trustors' decisions and for trustees' decisions. Each combination of payoffs and sanctioning properties determines the "total payoffs" that can be reached in a decision situation. Following the approach employed in Chapter 2, the data are grouped in a way that decisions made by each subject in identical (sub)games that differ with respect to the behavioral context constitute one group. This yields 664 "subject-payoff response sets" for trustors' decisions and also for trustee's decisions. Response sets for trustors always consist of 3 decisions, one made in the TGS, one in a TGS as a subgame of the H_2TGS, and one in the H_1TGS. Response sets for trustees involve between 1 and 4 decisions, with at least one decision in the DGS, because trustees are not always trusted. Since the focus is on analyses of trustfulness and trustworthiness, "sanctioning" decisions are not reported in this paper (for details, see Vieth, 2008; also see Chapter 5).

Table 4.3 provides an overview of the decisions made in subject-payoff response sets per (sub)game in order to show the composition of the response sets. For instance, of the 1992 "placing trust" decisions, 664 decisions have been made in response sets that involve the TGS. The same number holds for the H_2TGS ($575 + 89$) and the H_1TGS (see note below Table 4.3). Recall that each participant could face a (sub)game with certain total payoffs maximally once. Therefore, the number of decisions made per (sub)game equals the number of response sets in which the respective (sub)game is involved. However, since response sets involve decisions made in different (sub)games,

Table 4.3: Number of decisions within subject-payoff response sets per (sub)game

	Placing trust (x)					Honoring trust (z)				
	all $\bar{x}$	all x	mix	Σ	%x	all $\bar{z}$	all z	mix	Σ	%z
DGS						375	61	228	664	14.2
TGS	158	143	363	664	41.7	134	17	126	277	26.0
TGS$\vert H_2^+$	132	131	312	575	53.6	153	18	137	308	37.0
TGS$\vert H_2^0$	26	12	51	89	28.1	10	4	11	25	32.0
TGS$\vert H_1^+$						140	28	142	310	45.0
TGS$\vert H_1^-$						6	0	3	9	33.3
TGS$\vert H_1^0$						20	1	13	34	14.7
Σ	316	286	726	1328	45.9	838	129	660	1627	26.7

Decisions for "placing trust" in H_1TGS: 158 "all $\bar{x}$", 143 "all x", 363 "mix", and 664 decisions in total with 53.2% placed trust (%x = 48.3 across all (sub)games including the H_1TGS). Blank cells indicate situations that are logically impossible. "Placing trust" decisions are denoted by "$\bar{x}$" for withheld trust and by "x" for placed trust. Similarly, "honoring trust" decisions are denoted by "$\bar{z}$" for abused trust and by "z" for honored trust. The percentages of placed trust (%x) and honored trust (%z) are calculated for the respective sum of decisions (data in the analyses).

the sum of decisions across (sub)games does not equal the number of response sets. For instance, response sets for "placing trust" decisions consist of the TGS, the H_1TGS, and one of the two subgames of the H_2TGS, namely 575 response sets include the TGS$\vert H_2^+$ and 89 response sets include the TGS$\vert H_2^0$. As mentioned above (Table 4.2), response sets consisting of one single decision (singletons) therefore cannot occur for "placing trust" decisions. Similarly, 277 of the 664 response sets for "honoring trust" decisions in the DGS also include the TGS. Note that Table 4.3 does not provide information about what the other subgames that are involved.

Across (sub)games including the H_1TGS, but for given total payoffs, participants made 474 decisions (316 + 158) to always withhold trust (all $\bar{x}$) in the role of the trustor, and 429 decisions (286 + 143) to always place trust (all x). Thus, trustors have made 903 decisions that do not discriminate between the behavioral contexts for given total payoffs, whereas 1089 decisions (726 + 363) show some mixed pattern within response sets. Note again that any response set for "placing trust" decisions involves the TGS and the H_1TGS. Thus, the reported frequencies are identical in these two (sub)game because any difference in decision-making between these two behavioral contexts moves a response set into the category of mixed response sets.

Of course, the percentage of placed trust differs between these two (sub)games due to the differences in mixed response sets. In the role of the trustee, 838 decisions to always abuse trust (all $\bar{z}$) were made, and 129 decisions to always honor trust (all z). Of the 436 decisions $(375 + 61)$ made in response sets without variation that involve the DGS, 113 response sets are singletons (93 for all $\bar{z}$ and 20 for all z). In mixed response sets, in total 660 decisions were made by trustees.

On average, across (sub)games including the H_1TGS, trust was placed in 48.3% (963 cases) of the 1992 cases. Trust was honored in 26.7% (434 cases) of the 1627 cases in which trustees could decide whether to honor trust. The level of placed trust and of honored trust differs between behavioral contexts. The percentages of placed trust and of honored trust are averages of all cases per (sub)game without accounting for the grouping structure of the data. These percentages provide information about the actual decisions made in the (sub)games irrespective of the specific payoffs in the (sub)games and irrespective of the fact that each person made several decisions. For instance, it is possible that trustees more frequently receive a chance to honor trust, and then do so, in some behavioral contexts than in other behavioral contexts because the outcomes are perceived as favorable. Thus, influences of outcome-based motivations and of individual heterogeneity have to be controlled in order to test the hypotheses about effects of behavioral advances on subsequent decisions. In Chapter 2 logistic regression models with fixed effects for response sets have been used in order to analyze effects of behavioral contexts on people's decision-making. This method allows minimal assumptions to be made about subject-specific effects, outcome effects, and interaction between these effects. However, in a fixed effects approach, only the mixed response sets carry statistical information, and other response sets are thus excluded from the analyses.

The fixed effects approach could also be employed for the purpose of this study. However, the data in the experiment for the study presented here were much more skewed (see the discussion for further remarks). In particular, trustees promised trustworthiness in the H_2TGS in 575 of the 664 cases (86.5%). Thus, trustors made 89 $(664 - 575)$ "placing trust" decisions after the trustee did not promise trustworthiness. Since not all trustors placed trust, trustees made only 25 "honoring trust" decisions in the TGS$|H_2^0$. Similarly, trustors promised reward in the H_1TGS in 310 of the 353 cases (87.8%) of placed trust. This resulted in only 9 "honoring trust" decisions made after punishment was threatened (TGS$|H_1^-$). In the 3 cases in mixed response sets involving the TGS$|H_1^-$, trustees honored trust. The other mixed response sets for "honoring trust" decisions and for "placing trust" decisions have sufficient vari-

ation in decisions per (sub)game. Due to the specific response patterns in the data, the parameter indicating the difference in trustworthiness between the $\mathrm{TGS}|\mathrm{H}_1^-$ and the TGS is statistically not identified in a fixed effects approach. Therefore, logistic regression models with random effects for response sets are employed. Models are fitted by maximum marginal likelihood and have the following general form:

$$y_{ijk} = y^*_{ijk} > 0$$
$$y^*_{ijk} = \beta_0 + \eta'_{ijk}\beta + u_{0ij} + e_{0ijk}$$
$$\text{where } u_{0ij} \sim \text{Normal}(0, \sigma^2_u) \text{ and } e_{0ijk} \sim \text{Logistic}(0, \sigma^2_e)$$

The model is applied to describe the probability of trustfulness or trustworthiness of a subject i in the behavioral context of a (sub)game k that has a total payoff combination j. The intercept parameter (β_0) and the (sub)game effects ($\eta'_{ijk}\beta$), together with controls, constitute the fixed part of the model, i.e., effects are assumed to be the same across response sets. The random part consists of two random variables for the two levels of analysis. Errors at the level of decisions are assumed to have a standard logistic distribution with a mean of 0. The variance within response sets is fixed in logistic regression for identification purposes and serves as a scaling parameter ($\sigma^2_e = \frac{1}{3}\pi^2 \approx 3.29$, Long, 1997: 47–48). The intercept $\beta_0 + u_{0ij} + e_{0ijk}$ is allowed to vary randomly between response sets, reflecting that the average response probabilities differ between response sets. Models with random effects at the level of response sets require additional assumptions about the distribution of deviations that are specific to response sets. Specifically, it is assumed that the response sets are drawn from a population in which combinations of subjects and total payoffs are normally distributed with a mean of 0, i.e., $u_{0ij} \sim \text{Normal}(0, \sigma^2_u)$. The variance ($\sigma^2_u$) between response sets reflects the extent of unexplained deviations of the average probability per response set from the overall average probability. Note that the normality assumption might be more problematic for subject-payoff response sets than for subject response sets. However, subject-payoff response sets are preferred because this grouping of the data allows additive payoff effects to be controlled without further assumptions about specific representations of individually heterogeneous outcome-based motivations (for a discussion of this problem, see Chapter 3). In the study presented here, the reported results obtained with random intercept models are qualitatively similar to the results obtained with a fixed effects approach.

4.4 Results

4.4.1 Analyses for Trustworthiness

Recall that the trustee chooses whether or not to honor trust after the trustor has placed trust in a TGS. First, this can be in the TGS without behavioral context. Second, one of two TGSs as subgames of the H_2TGS results from the trustee's decision of whether or not to promise trustworthiness (TGS$|H_2^+$ and TGS$|H_2^0$). Third, three TGSs are involved as subgames of the H_1TGS after the trustor has chosen to place trust combined with a punishment threat (TGS$|H_1^-$), with a reward promise (TGS$|H_1^+$) or without an announcement of sanctioning (TGS$|H_1^0$). Moreover, sharing gains in the DGS represents the analogous decision of the trustee without a behavioral context. Thus, trustworthiness is observed in seven behavioral contexts (including the "empty" context).

The TGS serves as the reference context in the statistical analyses (Table 4.4). Pairwise comparisons of differences in trustworthiness between the behavioral contexts are reported in Table 4.5. The upper part of Table 4.4 shows the estimates for effects of the behavioral contexts (discussed below). The period in which a decision has been made (i.e., the number of past periods per type of game) is included as a control, but shows no significant effect on trustworthiness. Note again that influences of objective outcomes are controlled by grouping the data in subject-payoff response sets. Probabilities of trustworthiness per (sub)game are estimated at the mean of the number of periods. For computational convenience, the random effects are ignored, i.e., fixed to $u_{0ij} = 0$. The lower part of Table 4.4 summarizes the random part of the model, consisting of the non-estimated scaling parameter (SD of the error of decisions) and the (estimated) standard deviation of the random effect of response sets (SD of the error of response sets). The standard deviation of the random effect for response sets represents the deviation of average effects of response sets from the estimated overall mean of these effects. Approximately half of the total unexplained variance is due to influences of subjects and outcomes (rho = 0.51). Further analyses including a random effect for sessions in addition to the random effect for subject-payoff response sets show that approximately 10% of the total unexplained variance is at the session level. Thus, this part of the unexplained variance is due to influences of the experimental group and not only due to influences that are specific to subjects and outcomes. The fixed part of the model is basically not affected by incorporating the session level. Since the focus of this study is on the effects of behavioral contexts on subsequent behavior, the simpler two-level model is reported. At the bottom of

Table 4.4 (Panel B), the likelihood-ratio test against the model without (sub)game dummies is reported. This test shows that trustworthiness significantly differs between the behavioral contexts (LR $\chi^2_{6\,\mathrm{df}} = 166.95$ with $p < 0.0001$). In the following, the results for the specific behavioral contexts are described and discussed.

In line with Chapter 2, it has been argued that placed trust is a kind advance that creates an obligation to honor trust (Hypothesis 4.1). This idea receives strong support in the analyses. The estimated probability of trustworthiness in the TGS is 17.2%, but the probability of generosity in the DGS is only 5.9%. Thus, dictators are 11.2% less likely to share gains than trustees are. This difference in trustworthiness is due to the mere act of placing trust. Although less salient in previous studies (see Chapter 2; Gautschi, 2000; McCabe et al., 2003; Cox, 2004), the coefficient here is highly significant. This might be due to sanctioning possibilities because trustees might anticipate more punishment for abused trust than dictators for kept gains (Chapter 5).

In the H_2TGS, the trustee could choose whether or not to promise trustworthiness. Compared to the TGS without promise option, making the promise (TGS$|H_2^+$) significantly increases trustworthiness. The difference amounts to 15.5%, such that trust is expected to be honored every third time after the promise has been made (32.7% compared to 17.2% in the TGS). This supports the idea that self-consistency or obligation feelings drive the trustee to keep his promise (Hypothesis 4.2). It has been argued that the desire for self-consistency might also increase the feeling of obligation due to shared responsibility and induce trustees to honor trust. In Chapter 2, strong support has likewise been found for the promoting impact of promising to honor trust on trustworthiness in trust situations without sanctioning opportunity. Omitting the promise has been assumed to undermine obligation feelings and, thus, to reduce trustworthiness (Hypothesis 4.3). Although the probability in the TGS$|H_2^0$ (13.4%) compared to the TGS (17.2%) is indeed reduced by 3.8%, the difference is not significant. Previous studies likewise find no support for hampering influences of omitting (cheap-talk) promises to honor trust on trustworthiness (see Chapter 2; Snijders, 1996). Even the difference in trustworthiness between omitting (TGS$|H_2^0$ with 13.4%) and making (TGS$|H_2^+$ with 32.7%) the promise is only marginally significant (Table 4.5), although it amounts to 19.3%. Therefore, the lack of support for the idea that omitting the promise hampers trustworthiness might be due to the relatively small number of cases in this subgame (25 decisions, Tables 4.4 and 4.5). Note, however, that no support is found in Chapter 2 for a difference in trustworthiness between these two behavioral contexts in the cheap-talk case and that

Table 4.4: Logistic regression of trustworthiness with random intercepts for subject-payoff response sets

(A) Regression coefficients

	Hyp.	b	se	Pr(%)
Behavioral contexts				
DGS	H_1: –	-1.19^{***}	0.24	5.9
TGS		(ref.)		17.2
TGS|H_2^+	H_2: +	0.85^{***}	0.24	32.7
TGS|H_2^0	H_3: –	−0.30	0.64	13.4
TGS|H_1^+	H_6: +	1.22^{***}	0.24	41.3
TGS|H_1^-	H_7: –	1.10	1.05	38.3
TGS|H_1^0	H_8: +/–	−0.96	0.66	7.3
Past periods per game		−0.06	0.04	
Constant		-1.36^{***}	0.26	
SD(error decisions)		1.81	fixed	
SD(error response sets)		1.87	0.19	
rho		0.51	0.05	

(B) Likelihood-ratio tests

	χ^2	df
LR test (rho = 0)	108.78^{***}	1
LR test (control)	166.95^{***}	6

N(response sets) = 664, N(decisions) = 1627, N(subjects) = 166;
two-sided p-values: *** p = 0.001, ** p = 0.01, * p = 0.05, ° p = 0.1; (sub)games (0/1), past periods per game (0 . . . 7/9). The standard deviation (SD) for decision residuals (e_{0ijk}) is constant, and the SD of random intercepts (u_{0ij}) for response sets is estimated. The proportion of unexplained variance at the level of response sets is denoted by rho. The absolute probability of trustworthiness per (sub)game which is estimated for an average period ($\bar{t} = 3.78$) assuming $u_{0ij} = 0$. Likelihood-ratio tests are reported for the proportion of unexplained variance at the level of response sets (rho = 0) and for the presented model against null model with controls.

the coefficient for omitted promises is positive in the analyses. Therefore, it might also be that trustees' motivations are too diverse, i.e., obligation feelings might be undermined for some trustees, but increased for other trustees (see the discussion for further remarks on this issue; and for empirical evidence, see Chapter 2).

Now consider the H_1TGS in which the trustor decides about announcing sanctions in combination with placing trust. As previously argued, reward promises might serve as a kind advance and trigger an obligation feeling to return the favor (Hypothesis 4.6). Moreover, given that making a promise to honor trust increases trustworthiness (Hypothesis 4.2 and findings in Table 4.4), trustees might expect that trustors likewise feel bound to keep their promise and indeed reward honored trust (Chapter 5). The analyses provide support and show that promising reward strongly promotes trustworthiness. The probability of trustworthiness increases by 24.2% (from 17.2% in the TGS to 41.3% in the TGS$|H_1^+$). Trustees even tend to be significantly more inclined to honor trust after the trustor has promised a reward (TGS$|H_1^+$) than after trustees have promised trustworthiness (TGS$|H_2^+$) (Table 4.5). This might be due to trustees' pleasant anticipation of higher reward chances or to the particular kindness of the combination of trustfulness and reward promise.

In the case in which the trustor has chosen to threaten punishment, the trustee likewise might expect more punishment due to trustors' self-consistency. However, it has been argued that punishment threats indicate distrust and induce a feeling of indignation in trustees (Hypothesis 4.7). No support is found for a decrease in trustworthiness after punishment has been threatened (TGS$|H_1^-$) compared to the decision situation in which no announcements are possible. In fact, the coefficient is quite large and positive, such that the probability of trustworthiness increases by 21.1% due to the threat of punishment (from 17.2% in the TGS to 38.3% in the TGS$|H_1^-$). The reason that this change is not significant seems to be the low number of cases. As previously pointed out, trustors mostly chose to promise reward (87.8%) such that only 9 decisions by trustees about honoring trust are available after punishment has been threatened (Table 4.3). However, the positive sign suggest that threats might not be particularly hampering for trustworthiness (see also below), and that threats deserve further research (for positive influences of threats on cooperative behavior, see Voss and Vieth, 2006).

Next, recall the argument that the influence of neither threatening punishment nor promising reward might depend on the trustee's beliefs (Hypothesis 4.8). Since reward is promised in 87.8% of the cases, trustees might expect reward promises. If trustors omitted to promise a reward, trustees might be disappointed and might retal-

Table 4.5: Pairwise comparisons of behavioral contexts for trustworthiness

	Pr(%)	TGS	$TGS\|H_2^+$	$TGS\|H_2^0$	$TGS\|H_1^+$	$TGS\|H_1^-$	$TGS\|H_1^0$
$TGS\|H_2^+$	32.7	0.85^{***}					
(N = 308)		0.24					
$TGS\|H_2^0$	13.4	-0.30	-1.15°				
(N = 25)		0.65	0.65				
$TGS\|H_1^+$	41.3	1.22^{***}	0.37°	1.52^{*}			
(N = 310)		0.24	0.22	0.64			
$TGS\|H_1^-$	38.3	1.10	0.25	1.40	-0.12		
(N = 9)		1.05	1.02	1.21	0.85		
$TGS\|H_1^0$	7.3	-0.96	-1.81^{**}	-0.66	-2.18^{***}	-2.06°	
(N = 34)		0.65	0.65	0.87	0.66	1.21	
DGS	5.9	-1.19^{***}	-2.04^{***}	-0.89	-2.41^{***}	-2.29^{*}	-0.23
(N = 664)		0.24	0.24	0.62	0.24	1.05	0.66

The table presents differences between coefficients of (sub)games (row – column). Standard errors are reported underneath. The entries in the columns Pr(%) and TGS (with N = 277, Pr(%) = 17.2) are repeated from Table 4.4. Wald tests (two-sided p-values, not adjusted for multiple testing): *** p = 0.001, ** p = 0.01, * p = 0.05, ° p = 0.1.

iate for the unkindness by abusing trust. The probability of trustworthiness is indeed reduced by 9.8% after trust has been placed without an announcement of sanctions ($TGS|H_1^0$), to 7.3% compared to 17.2% in the TGS. However, this difference is not significant. However, the two differences in trustworthiness in the case of omitted announcement ($TGS|H_1^0$) compared to punishment threats ($TGS|H_1^-$) and compared to reward promises ($TGS|H_1^+$) are significantly negative (Table 4.5). Trustworthiness is decreased by 34%, from 41.3% in the $TGS|H_1^+$ to 7.3% in the $TGS|H_1^0$. This might support the idea that trustees punish for the omitted kindness of a reward promise, given the assumption that the frequency of reward promises shapes trustees' beliefs. The other decrease in trustworthiness from 38.3% in the $TGS|H_1^-$ to 7.3% in the $TGS|H_1^0$ amounts to 31%, but is only marginally significant. This difference might hint at a possible reason for the positive sign of the $TGS|H_1^-$ coefficient, given the strong decrease in trustworthiness after announcing sanctions has been omitted compared to in the decision situation in which the trustee has received a reward promise. Trustees might be more likely to honor trust after punishment has been threatened because the influence of anticipated punishment (due to trustors' self-consistency)

might be stronger than the trustee's feeling of indignation. Moreover, feelings of indignation might be undermined because trustees might have perceived the threat as legitimate. Previous studies suggest that only unfair threats have a detrimental effect on trustworthiness (Fehr and Rockenbach, 2003; Fehr and List, 2004; Houser et al., 2008). Nevertheless, the low number of cases involved in the TGS$|\mathrm{H}_1^-$ demands caution in interpreting this difference. Moreover, comparing trustworthiness in decision situations of made and of omitted threats (which has been done in the previous studies by Fehr and Rockenbach, 2003; Fehr and List, 2004; Houser et al., 2008) implies that the joint influence of different motivations in a given decision situation is tested, i.e., of motivation triggered by a received threat and of motivations triggered by an omitted threat (see the discussion for further remarks).

4.4.2 Analyses for Trustfulness

The results for trustfulness are presented in a similar way as for trustworthiness (Tables 4.6 and 4.7). Trustors decide whether or not to place trust in the TGS and in the two TGSs as subgames of the H_2TGS after the trustee has decided whether or not to promise trustworthiness (TGS$|\mathrm{H}_2^+$ and TGS$|\mathrm{H}_2^0$). The TGS serves again as the reference category. The test of differences between the two coefficients for the subgames of the H_2TGS is presented in Table 4.7. In addition to the behavioral contexts resulting from preceding decisions, the H_1TGS is included, indicating the influence of whether or not the trustor had the opportunity to announce sanctions. The results show that trustfulness is significantly increased by 14.8% if the trustor has an opportunity to announce sanctions (H_1TGS), to 54% compared to 39.3% in the TGS. Note again that announcements are cheap-talk and therefore do not involve an objective barrier for lies, that mostly reward is promised (87.8%), and that sanctions are known to be costly for the trustor.

Roughly one-third of the unexplained variance is due to effects between response sets relative to the total unexplained variance (rho = 0.36). Further analyses including the session level show that 8% of the unexplained variance is due to influences of the experimental group. The effects of the behavioral contexts in the three-level model are again basically the same as obtained with the two-level model. Therefore, the simpler two-level model is reported. In contrast to the analyses for trustworthiness, the negative coefficient for the period in which a decision has been made indicates that trustfulness significantly decreases over time (Table 4.6). The likelihood-ratio test against the model without (sub)game dummies shows that trustfulness differs significantly between the behavioral contexts (LR $\chi^2_{2\,\mathrm{df}} = 41.15$ with $p < 0.0001$).

Table 4.6: Logistic regression of trustfulness with random intercepts for subject-payoff response sets

(A) Regression coefficients

	Hyp.	b	se	Pr(%)
Behavioral contexts				
TGS		(ref.)		39.3
TGS|H_2^+	H_4: +	0.64^{***}	0.14	55.2
TGS|H_2^0	H_5: –	-1.07^{***}	0.32	18.2
H_1TGS		0.60^{***}	0.13	54.0
Past periods per game		−0.20	0.03	
Constant		0.47^{**}	0.16	
SD(error decisions)		1.81	fixed	
SD(error response sets)		1.36	0.11	
rho		0.36	0.04	

(B) Likelihood-ratio tests

	χ^2	df
LR test (rho = 0)	111.38^{***}	1
LR test (control)	41.15^{***}	2

N(response sets) = 664, N(decisions) = 1992, N(subjects) = 166;
two-sided p-values: *** p = 0.001, ** p = 0.01, * p = 0.05, ° p = 0.1; (sub)games (0/1), past periods per game (0 . . . 9). The standard deviation (SD) for decision residuals (e_{0ijk}) is constant, and the SD of random intercepts (u_{0ij}) for response sets is estimated. The proportion of unexplained variance at the level of response sets is denoted by rho. The absolute probability of trustfulness per (sub)game which is estimated for an average period ($\bar{t} = 4.58$) assuming $u_{0ij} = 0$. Likelihood-ratio tests are reported for the proportion of unexplained variance at the level of response sets (rho = 0) and for the presented model against null model with controls.

Table 4.7: Pairwise comparisons of behavioral contexts for trustfulness

	Pr(%)	TGS	TGS|H_2^+	TGS|H_2^0
TGS|H_2^+	55.2	0.64^{***}		
(N = 575)		0.14		
TGS|H_2^0	18.2	-1.07^{***}	-1.71^{***}	
(N = 89)		0.32	0.33	
H_1TGS	54.0	0.60^{***}	−0.04	1.67^{***}
(N = 664)		0.13	0.12	0.32

The table presents differences between coefficients of (sub)games (row – column). Standard errors are reported underneath. The entries in the columns Pr(%) and TGS (with N = 664, Pr(%) = 39.3) are repeated from Table 4.6. Wald tests (two-sided p-values, not adjusted for multiple testing): *** p = 0.001, ** p = 0.01, * p = 0.05, ° p = 0.1.

Note again that the H_1TGS is not a behavioral context for the trustor's decision of whether or not to place trust. Again, results obtained using the fixed effects approach are qualitatively similar to the results presented here.

It has been argued that promises of trustworthiness motivate trustors to place trust (Hypothesis 4.4), because trustors feel an obligation to return the kindness and anticipate increased trustworthiness due to influences of the trustee's desire for self-consistency (see also Chapter 2). The analyses provide support for this reasoning and show that receiving a promise of trustworthiness (TGS|H_2^+) significantly increases trustfulness by 15.9%, such that trustors place trust with a probability of 55.2% in the TGS|H_2^+ compared to 39.3% in the TGS. As reported in Table 4.7, no significant difference is found between the impact of the trustee's promise (TGS|H_2^+) and the opportunity to announce sanctions (H_1TGS).

In contrast, trustfulness is strongly reduced when trustees omitted the promise (TGS|H_2^0). The probability that a trustor places trust drops by 21.1%, from 39.3% in the TGS to 18.2% in the TGS|H_2^0. It has been hypothesized that trustors are more reluctant to place trust because they anticipate reduced trustworthiness (Hypothesis 4.5). However, trustworthiness is not significantly reduced (Table 4.4). Given that player roles were exchanged, it is unlikely that beliefs are that inconsistent (for a similar argument, see Chapter 2). Rather, trustors punish for omitted kindness. In the absence of sanctioning opportunities, previous studies likewise find that trustfulness is increased by promises of trustworthiness (see Chapter 2) and strongly reduced by

omitted promises (see Chapter 2; Snijders, 1996; Gautschi, 2000). This suggests that the influences of making and omitting promises on trustfulness are quite stable.

4.5 Summary and Perspectives

4.5.1 Summary of Basic Ideas, Approach, and Contributions

This study has been based on the idea that people reciprocate kind and unkind behavior, even without influences of objective outcomes. Sociological and social-psychological research suggests that people feel an obligation to return favors, a need to retaliate against unkindness, and a desire to behave self-consistently (Gouldner, 1960; Coleman, 1990: ch. 12; Cialdini, 2001: chs. 2–3). These mechanisms drive people to respond to the mere act of kindness or unkindness. In Chapter 2, these insights have been applied to trust situations with and without an option for trustees to promise trustworthiness. It has been argued in Chapter 2 that trustfulness and promises of trustworthiness are kind advances that intrinsically demand a favor in return, whereas omitted promises are retaliated. Moreover, making or omitting a promise activates the desire for self-consistency. In addition to direct influences, self-consistency can also have indirect influences on subsequent behavior. Self-consistency can increase obligation feelings due to shared responsibility for others' decisions (e.g., after trustees have promised trustworthiness, they might feel responsible for the trustor's decision to place trust). However, self-consistency can also undermine perceptions of kindness (e.g., when trustees are confronted with placed trust after they have explicitly omitted to promise trustworthiness). The study presented here builds upon the study reported in Chapter 2 and extends it by incorporating opportunities for trustors to sanction trustees (punishment or reward) and to announce such sanctions (threat or promise). This allowed for contributions to previous research in three main respects. First, the study presented here investigated whether the previous findings can be replicated in the presence of sanctions, i.e., influences of placed trust on trustworthiness and influences of making or omitting promises to honor trust on trustworthiness and on trustfulness. Second, insights were provided on influences of punishment threats and reward promises. Third, this research examined whether perceptions of kindness and unkindness of preceding behavior depend on forgone outcomes of non-chosen options.

Following Vieth and Weesie (2006; and see Chapter 2), the experiment was designed as within-subject sets of structurally identical (sub)games resulting from kind and unkind actual behavior in single encounters. The baseline decision situation was the Trust Game with Sanctioning options for trustors (TGS). Sanctioning was costly

for trustors and not always effective for removing the trustee's temptation to abuse trust in objective terms. Employing a factorial design, payoffs and sanctioning properties were varied within and across sessions. Some decision situations involved an opportunity for trustees to promise their trustworthiness (H_2TGS). In other decision situations, trustors could announce punishment for abused trust or reward for honored trust (H_1TGS). Announcements of trustworthiness and of sanctions were entirely cheap-talk, i.e., did not affect objective outcomes. Trustors could choose whether or not to place trust in three behavioral contexts: in the Trust Game with sanctions, but without any announcement options (TGS), after the trustee promised trustworthiness (TGS$|H_2^+$), and after the trustee omitted the promise (TGS$|H_2^0$). In the H_1TGS, trustors combined their decision to place trust with their decision about announcing sanctions. For trustfulness, the availability of the announcement opportunity likewise constitutes a decision context that, however, does not result from previous behavior. Trustees decided whether to honor or to abuse trust (i.e., whether or not to share gains) in seven behavioral contexts: in the three contexts mentioned for the trustor's decision (TGS, TGS$|H_2^+$, TGS$|H_2^0$), in the three (sub)games after the trustor chose to place trust and decided about announcing sanctions (TGS$|H_1^+$, TGS$|H_1^-$, TGS$|H_1^0$), and in the Dictator Game with Sanctions (DGS) representing the trustee's decision of whether or not to share gains without behavioral context.

This design allows for the analysis of influences of behavioral contexts on decision-making while controlling for outcome-based motivations and individual heterogeneity. For this purpose, in Chapter 2, the data have been grouped in subject-payoff response sets and logistic regression models with fixed effects for response sets have been used for statistical analyses. In the study presented here, logistic regression models with random intercepts for response sets were employed. This is because trustors nearly always promised reward (87.8%), leaving only a few cases in the (sub)games after punishment was threatened (TGS$|H_1^-$) and after trust was placed with omitted sanctioning announcement (TGS$|H_1^0$). Similarly, trustees nearly always chose to make the promise of trustworthiness (86.5%), such that only a few cases are available after the promise was omitted (TGS$|H_2^0$).

The results show that the findings reported in Chapter 2 also hold for trust situations with sanctions. Trustworthiness is strongly increased by the mere act of placed trust, which is also found in previous studies, though not as strongly (see also McCabe et al., 2003; Cox, 2004). Promises of trustworthiness promote trustfulness and trustworthiness, while omitted promises are punished by withheld trust (see

also Snijders, 1996; Gautschi, 2000). This holds despite the fact that promises were entirely cheap-talk and trustworthiness was promised in 86.5% of the cases. Note the implications: trustees neither invested in making the promise nor were they objectively bound to keep the promise, and trustors had no indication of whether the trustee lied. Next, no support has been found for the hypothesis that omitting a promise of trustworthiness hampers decisions to honor trust (see also Snijders, 1996). Significant differences in trustworthiness after the trustee has omitted the promise to honor trust are only found in comparison with the decision situation after the trustee has made the promise. However, such comparisons test the joint influence of motivations induced by the made promise and those induced by the omitted promise.

Concerning the decision situation in which trustors can announce sanctions, trustworthiness is strongly increased by reward promises. This finding is in line with previous studies (e.g., Fehr and Schmidt, 2004, 2007; Fehr et al., 2007). It has also been found that trustors almost always promised reward (87.8%), indicating that they anticipate the promoting impact, and that the mere opportunity to announce sanctions increases trustfulness. No support has been provided for the idea that threatening punishment is unfriendly and would therefore be retaliated with abused trust. In fact, an increase in trustworthiness has been found when comparing the threat situation with the decision situation after the trustor has omitted any sanctioning announcement, whereas trustworthiness is decreased after the sanctioning announcement has been omitted compared to the decision situation that arises after a reward has been promised. This suggests that trustees might punish trustors for omitted reward promises by abusing trust because they might have expected a reward promise rather than a punishment threat (given the frequency). Moreover, a punishment threat might motivate trustees to honor trust because the threat emphasizes the prospect of sanctions. Indignation feelings might even be undermined if the threat is not perceived as unfair (for this argument, see also Fehr and Rockenbach, 2003; Fehr and List, 2004; Houser et al., 2008). However, as mentioned above, inferences from such comparisons are based on the joint influence of making and omitting the threat. Moreover, these results are based on a small number of cases. Further studies are necessary in order to investigate impacts of threats and associated influence factors.

4.5.2 Further Discussion and Perspectives

Some aspects discussed in Chapter 2 also apply to the study presented here. Further experiments are required, in which the ordering of game clusters is varied in order to investigate whether and to what extent the influences of behavioral contexts is

moderated by preceding experiences in other contexts. The ordering of games was fixed in order to minimize influences across types of games by maximizing differences concerning the presentation of decision situations. Statistical analyses are desirable that involve random coefficients for response sets (in order to relax the assumption that effects of behavioral contexts are the same for all response sets). In principle, this extension would be possible in the random effects approach taken here, but are overly demanding concerning the current data because the skewed responses reduce statistical power and the number of cases in certain subgames (see also remark below). Recall that accounting for the dependency of observations in sessions did not to make a difference for the effects of behavioral contexts that are of interest in this study. Next, beliefs, emotions, and perceived kindness should be elicited. However, this requires further research on how including such measures affects people's decision-making (e.g., Gächter and Renner, 2006). Some further aspects that are specific to this study deserve more detailed remarks: (1) cheap-talk announcements, (2) influences on and of sanctioning behavior, and (3) implications for theoretical modeling.

First, due to the cheap-talk character of announcements, trustees promised trustworthiness and trustors promised reward in most of the cases (86.5% and 87.8%). Since behavioral contexts are created by actual choices, only a small number of decisions is available in situations after no announcement has been made ($\text{TGS}|\text{H}_2^0$, $\text{TGS}|\text{H}_1^0$) and in situations after punishment has been threatened ($\text{TGS}|\text{H}_1^-$). As a result, findings (if any) that are based on behavior in these subgames have to be interpreted cautiously and provide only indications for further research (especially concerning the influence of threats). In order to balance responses, which would increase the number of observations in these subgames, making announcements could be associated with transaction costs. This might also accentuate differences in decision-making between behavioral contexts, because it limits incentives for making "non-serious" announcements (i.e., for lying).

Moreover, concerning the decision situation in which the trustor can announce sanctions, previous studies show that people threaten punishment far more frequently if no option to promise reward is available (Fehr and Rockenbach, 2003; Fehr and List, 2004; Voss and Vieth, 2006; Houser et al., 2008). Thus, decision situations with threat options could be separated from decision situations with promise options. This also has the advantage that omitting the announcement is less ambiguous, because omitting a threat is friendly, whereas omitting a reward it unfriendly. However, influences of made announcements might be reduced because promises might be perceived as

more kind if threats are possible and vice versa (for evidence on set-dependency of decision-making, see also Sandbu, 2007).

Another possibility to increase the number of threats is to assign different transaction costs for announcing punishment and for announcing reward. In cases in which transaction costs for punishment threats are lower than for reward promises, trustors have an objective incentive to make a threat rather than a reward. Given non-selfish motivations, it would be interesting to study whether and to what extent trustors follow such incentives and how trustees react upon such differently incentivized announcements. Note that it does not seem advisable to ask participants to make a choice for every possible part of the decision situation without knowing how others actually decide ("strategy method", Selten, 1967) because the emotional basis for other-regarding motivations might be undermined (for further remarks, see Chapter 2).

Second, it has been argued that people might react to preceding behavior by others (obligation or indignation) and to their own preceding behavior (self-consistency) because they anticipate increased or decreased sanctioning behavior. Previous studies suggest that sanctioning behavior likewise follows the principle of reciprocity. However, influences of outcome-based motivations have not been thoroughly controlled. In the analyses presented here, influences of kind and unkind behavior on trustfulness and trustworthiness are analyzed. Similarly, sanctioning behavior might likewise be motivated by an obligation to return favors, a thirst for revenge, and the desire for self-consistency. Therefore, the same approach should be employed in order to study how sanctioning behavior is influenced by the mere choice of a kind or unkind option, creating a behavioral context. This has been done in Chapter 5 and empirical support for such influences has been provided. The results also indicate influences of anticipated sanctions. For instance, it has been argued in Chapter 5 that the behavioral context influences punishing decisions differently than rewarding decisions, which is likewise supported in the analyses. However, it is unclear why people should focus, e.g., on a decrease in punishment in one behavioral context and on an increase in reward in another behavioral context. Moreover, concerning some behavioral contexts, the direction of the effect of behavioral advances on punishing behavior is the same as on rewarding behavior. In these cases, it is not obvious whether an anticipated increase in punishment, an increase in reward, or the combination of both motivates a certain decision.

Thus, concerning influences of anticipated sanctions on trustworthiness and trustfulness, further research is desirable. Further experiments should be conducted in which the decision context differs with respect to future options and properties of

the options. Some decision situations should include punishment options, others reward options, others punishment and reward, and yet others no sanctioning options. This setup would also allow the question to be investigated whether and to what extent available sanctioning options support or hamper process-based motivations. Previous research on sanctions shows that intrinsic motivations can be crowded-out by extrinsic incentives. Therefore, it is important to vary the ordering of different decision situations in a within-subject design. Note that actual behavior does not necessary match people's anticipations. Therefore, beliefs should be elicited which, however, influences peoples decision-making and requires further research (see remarks above).

Third, this study provides evidence that people evaluate others' behavior in terms of kindness, even without actual changes in objective outcomes. In theoretical models, perceived kindness is determined by forgone outcomes of non-chosen options (e.g., Falk and Fischbacher, 2006). Such models can account for the promoting impact of placed trust on trustworthiness because due to placing trust, the outcome from withheld trust is ruled out, such that the trustees' (expected) objective outcome is increased. However, the strong influences of making and omitting cheap-talk announcements are not captured (for similar findings, see studies on communication mentioned above). Thus, the question is what it is that makes a certain behavior kind and another behavior unkind. Promises and threats have been defined as expressed intentions to perform a certain action. Consider the arguments that promises are friendly advances, because they provide the perspective of gains, whereas threats involve unkindness, because they shift the focus on potential losses. In other words, promises are indications of decision paths that yield favorable objective outcomes, whereas threats indicate decision paths that result in unfavorable objective outcomes. Thus, in theoretical models, expected outcomes from indicated decision paths could be the basis for determining the kindness of an action. This "baseline kindness" can then be moderated by influences of forgone outcomes. Note that such a theoretical model would account for the empirical indications that a threat does not necessarily trigger retaliation if the behavior associated with the threat is sufficiently kind. This idea to take indicated decision paths into account reflects the intuition that preceding decisions indicate "focal points" (Schelling, 1960) that help people to solve the coordination problem that arises in the presence of other-regarding motivations (for similar intuitions, see also Falk et al., 2002). Moreover, the desire for self-consistency likewise influences decision-making, both directly and indirectly by moderating influences of intention-based motivations. Note that social-psychological research indicates that

the desire for self-consistency can be increased if people make investments. Considering the various influences, developing a suitable theoretical model would be a fruitful effort that would aid in deriving hypotheses.

Chapter 5

Revenge and Gratitude in Trust Situations Involving Promises and Threats

Experimental Evidence on Reciprocity by Intention-Based Sanctioning

This study follows an approach developed together with Jeroen Weesie. The experiment was prepared and conducted while visiting Simon Gächter at Nottingham University. I am grateful for all the support both of them provided. I thank members of CeDEx at the Nottingham School of Economics for support and comments, in particular, Michail Drouvelis, Maria Montero, and Ping Zhang for help during pre-tests, Ruslan Kabalin for technical support, and Jo Morgan for checking language of instruction texts. Comments are acknowledged which have been made by members of the CREED/CeDEx/UEA meeting 2008 in Amsterdam.

Abstract

People are inclined to reward others' kindness and to retaliate for others' unkindness. Based on obligation feelings, indignation feelings, and self-consistency, the mere choice of an action without any change of objective outcomes can influence subsequent decisions. These influences have been studied in trust situations with sanctioning options for trustors. Some trust situations also involve announcement options for sanctions by trustors or for trustworthiness by trustees. Announcements are cheap-talk without a reply option. Sanctions are costly and not always effective in objective terms. The experiment is designed as within-subject sets of structurally identical (sub)games resulting from kind and unkind actual behavior in single encounters. This design allows effects of objective outcomes and of individual heterogeneity to be controlled. Sanctioning behavior is found to be strongly influenced by preceding behavior rather than by outcome-based motivations. Even cheap-talk announcements strongly increase actual rewarding and punishing decisions, except for kept promises of trustworthiness, which tend to be less rewarded.

5.1 Introduction

In social and economic interactions, people seek various opportunities for expressing their gratefulness and for taking revenge. For instance, people send a postcard, flowers, chocolates, or a bottle of good wine in order to express their gratitude for received help, kept promises, or a particularly satisfying deal, especially if the dealer kept promises that were not contractually fixed. In turn, people also seek ways to vent their anger about someone's betrayal or to retaliate for losses. Thereby, people are often ready to incur substantial costs in order to confront the other with an outburst of rage or to take revenge by playing a dirty trick on the other. The prospect of sanctions can limit incentives for opportunistic behavior (Williamson, 1985). For example, in trust situations, people are tempted to take advantage of those who have trusted them (e.g., Coleman, 1990). However, if people expect to be punished, the temptation diminishes or even completely vanishes. The same holds for expected rewards that might outweigh the additional gain from exploitation. The crucial point is that people have to be sufficiently convinced that they will be punished if they misbehave or that they will receive a reward for good conduct. The mere possibility of punishment or reward is not always a credible prospect. Especially if substantial effort is required for performing punishment or reward, people have to feel very angry or very grateful about something. For instance, people might get particularly angry about being misled by lies. Someone who asked for a favor and explicitly promised to behave well, but then took advantage of the other whom he owes the favor, might be penalized harshly. In other interaction situations, promising reward or threatening with punishment can enhance beliefs about the credibility of sanctions. Thus, misbehavior and good conduct might be punished and rewarded differently after such announcements have been made. *How does preceding behavior affect subsequent sanctioning decisions?*

In previous research, powerful social-psychological forces have been identified that drive people to respond to kind and unkind behavior and to keep their word: feelings of obligation to return favors, feelings of indignation that induce people to take revenge, and the desire for self-consistency (Gouldner, 1960; Coleman, 1990: ch. 12; Cialdini, 2001: chs. 2–3). In this study, these ideas are employed in order to investigate revengefulness and gratefulness in trust situations with and without opportunities to promise trustworthiness or to announce sanctions. Data from a game-theoretical lab experiment are used in order to explore the influences of trustfulness, promises of trustworthiness, promises of reward, and threats of punishment on sanctioning decisions. All announcements are cheap-talk (i.e., objectively costless and non-binding), while sanctions are costly and not always effective in objective terms. Following Vieth

and Weesie (2006; see also Chapter 2), the experiment is designed as within-subject sets of structurally identical (sub)games resulting from kind and unkind behavior in single encounters.

This study is based on the idea that people bear costs in order to punish for unkindness and to reward kindness, even if preferences concerning the distribution of objective outcomes do not play a role. Punishment is motivated by revengefulness and reward by gratefulness, not only by objective outcomes. Previous research indicates support for this intuition (Falk et al., 2003, 2005; Masclet et al., 2003; Walker and Halloran, 2004; Sefton et al., 2007; Masclet and Villeval, 2008; Vyrastekova and van Soest, 2008). However, in these studies, influences of motivations that are based on objective outcomes have not been ruled out. Moreover, perceived kindness has been determined by forgone objective outcomes, whereas in the study presented here, cheap-talk announcements are also assumed to shape sanctioning behavior. Some previous studies provide evidence that lies are punished particularly strongly (Brandts and Charness, 2003; for repeated interactions Bochet and Putterman, 2007), but do not find support for the idea that cooperation would be rewarded differently if promised beforehand (Brandts and Charness, 2003). Few studies address influences of sanctioning announcements on sanctioning decisions. Voss and Vieth (2006) find that threatening with punishment increases actual punishment. No support has been found that a promised reward is actually paid (Fehr and Schmidt, 2004, 2007; Fehr et al., 2007). Moreover, the design of these studies involves several confounding factors, e.g., the opportunity to reward is only available if a reward has been promised. Furthermore, none of these studies accounts for influences of various outcome-based motivations and employs a within-subject design for studying individual motivations (for a discussion of these two issues, see also Chapters 2 and 3).

5.2 Reciprocity of Sanctioning in Trust Situations with Announcements

5.2.1 Reciprocal Behavior and Other-Regarding Motivations behind Informal Sanctions

People reward kindness and retaliate for unkindness, even if this requires incurring costs or forgoing gains. This behavioral pattern is called reciprocity (for reviews, see Fehr and Schmidt, 2006; Hann, 2006; Kolm, 2006; Lévy-Garboua et al., 2006). Sociological and social-psychological research suggests that *feelings of obligation* drive people to return favors, in order to escape the "shadow of indebtedness" (Gouldner, 1960: 174; Coleman, 1990: ch. 12; Cialdini, 2001: ch. 2). Intrinsic distress and

emotional tension arise from delaying or not fulfilling an outstanding obligation. In contrast to positive reciprocity, experienced harm invokes *feelings of indignation* (or "sentiments of retaliation", Gouldner, 1960: 172) that create a thirst for revenge.

The idea of reciprocity has received support in numerous experimental studies on social dilemmas (for reviews see, e.g., Camerer, 2003: ch. 2; Ostrom and Walker, 2003; Kopelman et al., 2002; Kollock, 1998; Komorita and Parks, 1996; Ledyard, 1995; van Lange et al., 1992; Messick and Brewer, 1983; Pruitt and Kimmel, 1977). Particularly, experimental studies on sanctioning behavior in social dilemmas have directed attention to the principle of reciprocity. Most of these studies are devoted to *informal punishment*, i.e., to voluntary punishment decisions that are not enforced exogenously by contracts or by third parties. In bargaining problems, people reject offers they perceive as unfairly low, and in public good problems, people punish free-riders (for reviews see, e.g., Roth, 1995; Camerer, 2003: ch. 2; Shinada and Yamagishi, 2008). Some studies also address *informal reward*, either as the only sanctioning option (on contribution to public goods, see Vyrastekova and van Soest, 2008) or in combination with punishment opportunities (on sharing benefits, see Offerman, 2002; Andreoni et al., 2003; on risky investments, see Abbink et al., 2000; Falk et al., 2003; Rigdon, 2009; and on contribution to public goods, see Walker and Halloran, 2004; Sefton et al., 2007). These studies show that people also reward others' cooperation or generosity. In all these studies, people are found to invest in sanctioning others, even in single encounters with strangers. The findings provide strong evidence for *other-regarding motivations* (i.e., people are not only concerned with their own objective outcomes) as a basis for reciprocal behavior in general and for revengefulness and gratefulness in particular.

Many studies propose that fairness motivations induce people to punish others in order to reinstall equality in outcomes (for a review, see Camerer, 2003: ch. 2). This idea receives support in experimental studies in which performing punishment is less costly than the inflicted fine (e.g., Fehr and Gächter, 2000, 2002; Gächter et al., 2008) and in studies exploring the influence of various cost-effect ratios on sanctioning behavior (e.g., Falk et al., 2005; Nikiforakis and Normann, 2008; Anderson and Putterman, 2006; Carpenter, 2007; Egas and Riedl, 2008; Masclet and Villeval, 2008). Similar evidence has been provided for reward (e.g., Vyrastekova and van Soest, 2008). Theoretical models of fairness in terms of inequality aversion incorporate emotional utility resulting from guilt or envy about outcome differences that complements utility an actor derives from his own objective outcomes (e.g., Kelley and Thibaut, 1978; MacCrimmon and Messick, 1976; Weesie, 1994a; Ledyard, 1995; van Lange, 1999;

Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). The studies involving various cost-effect ratios also include a one-to-one ratio, i.e., sanctioning does not change the difference between people's own and others' outcomes. Nevertheless, people do invest in sanctioning, although it is often not sufficient to maintain cooperation (see also Walker and Halloran, 2004; Sefton et al., 2007; and on non-monetary punishment, Masclet et al., 2003). This finding has been interpreted as evidence that motivations other than outcome-based motivations drive people to invest in sanctioning others. However, "equalitarian orientations" (MacCrimmon and Messick, 1976) are only one example of social (value) orientations (Messick and McClintock, 1968; McClintock, 1972; Liebrand, 1984). Various preferences concerning the distribution of people's own and others' outcomes have been identified (for reviews, see Au and Kwong, 2004; McClintock and van Avermaet, 1982). For instance, models of altruism account for benevolence and spite (e.g., Brew, 1973; Taylor, 1987/1976; Weesie, 1993, 1994b; Snijders, 1996). Even in the case of a one-to-one ratio of costs and effects of sanctions, altruistic inclinations can motivate people to reward others, while aggressive tendencies can drive people to punish others. Thus, not all influences of outcome-based motivations on sanctioning behavior are ruled out in these studies.

In addition to outcome-based motivations, it has been suggested that people also respond to others' kind and unkind intentions (for a review, see Fehr and Schmidt, 2006). The basic idea is that people take into account the behavioral processes of how certain outcomes are obtained (for experimental evidence see, e.g., Snijders, 1996; Gallucci and Perugini, 2000; Gautschi, 2000; Brandts and Solà, 2001; Falk et al., 2003, 2008; McCabe et al., 2003; Cox, 2004; Charness and Rabin, 2005; also see Chapters 2–4). Preceding behavior is evaluated in terms of kindness and unkindness. In a theoretical model, Falk and Fischbacher (2006) assume that intentionality and the size of outcome changes caused by others' behavior determines perceived kindness (for other models of intentions see, e.g., Rabin, 1993; Levine, 1998; Charness and Rabin, 2002; Dufwenberg and Kirchsteiger, 2004). Some studies employ this idea for investigating whether kindness and unkindness of preceding decisions motivate people to punish or to reward others (e.g., Falk et al., 2003, 2008). However, in these experiments participants are asked to make a decision for every possible choice of the other person without knowing the other person's actual decision ("strategy method", Selten, 1967). This design feature gives rise to several concerns, e.g., undermined influences of emotions and artificial consistency in people's decisions (for more details, see Chapter 2).

5.2.2 Informal Sanctions and Announced Intentions

Theoretical and experimental research on intention-based motivations is mostly based on the assumption that the evaluation of others' decisions in terms of kindness is entirely based on actually forgone objective outcomes (for further remarks, see Chapter 4). However, previous studies have shown that cheap-talk communication strongly promotes cooperative behavior (for reviews see, e.g., Sally, 1995; Crawford, 1998; Kopelman et al., 2002; Bicchieri, 2002; Shankar and Pavitt, 2002; Ostrom and Walker, 2003; Brosig, 2006). The experimental evidence indicates that people explicitly promise to behave cooperatively and tend to keep their promise, despite the fact that these promises are entirely cheap-talk on objective grounds (Dawes et al., 1977; Snijders, 1996; Brandts and Charness, 2003; Bochet and Putterman, 2007; also see Chapters 2 and 4).

Promises are expressed intentions to perform a certain action that yield a gain to the other person. Due to the prospect of gains, promises have an inherent kindness which creates an obligation to return the favor in the person who receives the promise (Cialdini, 2001; also see Chapters 2 and 4). Moreover, in expressing an intention to perform a certain action, the *desire for self-consistency* induces people to "keep their word" (for reviews see, e.g., Webster, 1975; Cialdini, 2001: ch. 3; Kunda, 2002; Gass and Seiter, 2007). People do so, in order to avoid or to reduce cognitive dissonance (Heider, 1944, 1958; Festinger, 1957; Aronson, 1992; Akerlof and Dickens, 1982). In this sense, self-consistency creates intrinsic bonds, such that promises serve as commitments. A commitment is a "voluntary strategic action", costly or not, with the purpose of "reducing one's freedom of choice" or of changing the outcomes, i.e., a "strategic move" by which an actor voluntarily offers a "hostage" in the sense of a bond (Schelling, 1960). In contrast to cheap-talk promises, announcements have been associated with objective incentives (binding values, compensating values, transaction costs) in research on commitments. Such extrinsic commitments have been theoretically studied in trust situations and cooperation problems (Weesie and Raub, 1996; Voss, 1998a; Raub and Weesie, 2000; Raub, 2004; and including other-regarding motivations Snijders, 1996). Experimental research provides evidence that imperfectly binding commitments also promote cooperative behavior and that even small transaction costs hamper commitment posting (Raub and Keren, 1993; Snijders, 1996; Mlicki, 1996; also see Chapters 2 and 4; and for negotiation problems, also see Prosch, 2006).

In addition to intrinsic bonds that are not based on objective incentives, sanctions can provide additional incentives (also extrinsic ones) to keep one's word. For instance, Brandts and Charness (2003) provide evidence that lies are punished more

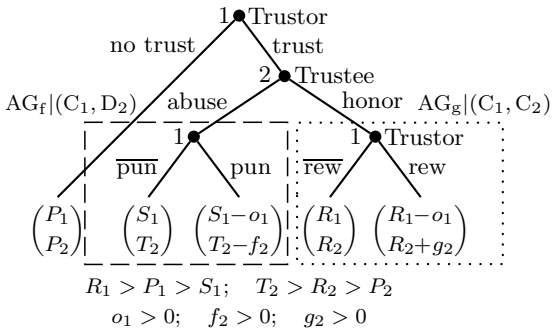
strongly than unkindness that does not involve a reneged promise (for repeated interactions, also see Bochet and Putterman, 2007). They do not find evidence that keeping cooperation promises makes a difference for rewarding. However, the possible reward was only minimal in their study. In fact, their experimental design involves an asymmetry between reward and punishment, because punishing is five times more effective than rewarding and half as costly as rewarding. Next, in the presence of sanctions, people can not only promise cooperation but also reward for cooperative behavior. Previous studies on bonus contracts do not provide support for the idea that the promised reward is actually paid (Fehr and Schmidt, 2004, 2007; Fehr et al., 2007). However, the design of these experiments inhibits the study of influences that making the reward promise as such has on subsequent behavior, because participants only have reward options after they have made a promise. Thus, changes in expectations about objective outcomes are induced because of the suddenly available reward options, such that influences of the mere choice of promising a reward on subsequent behavior cannot be studied. In addition to promising a reward, people can also threaten with punishment. Threats are expressed intentions to perform a certain action that inflicts a loss upon the other person. Voss and Vieth (2006) find evidence that actual punishment increases with the amount of punishment announced, indicating support for the idea of self-consistency.

5.2.3 Revenge and Gratitude in Trust Situations

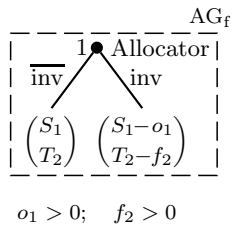
Intention-Based Sanctioning in Situations of Trust and Sharing

In order to represent sanctions in trust situations, the standard Trust Game (Dasgupta, 1988; Kreps, 1990) is supplemented with sanctioning options for the trustor (TGS, Figure 5.1a). First, the trustor decides whether or not to place trust. In the case in which trust has been placed, the trustee chooses between abusing and honoring trust, i.e., between keeping and sharing gains. Depending on the trustee's choice, the trustor decides either whether or not to punish for abused trust or whether or not to reward for honored trust. The indicated payoffs represent objective outcomes, e.g., in terms of money. Honored trust is beneficial for both actors ($R_i > P_i$, with $i = 1, 2$). This creates an incentive for the trustor to place trust. However, the trustee is tempted to abuse trust ($T_2 > R_2$), which inflicts a loss upon the trustor ($S_1 < P_1$). Sanctioning options limit the trustee's temptation by way of a fine (f_2) and a gratification (g_2), given that the trustor incurs the outlay (o_1). Sanctioning is *effective* in objective terms if it removes the trustee's temptation to abuse trust ($f_2 + g_2 > T_2 - R_2$) and *credible* if the trustor has no costs to carry ($o_1 \leq 0$). How-

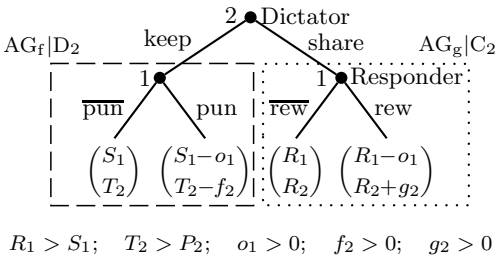
Figure 5.1: Sanctioning situations with different behavioral contexts



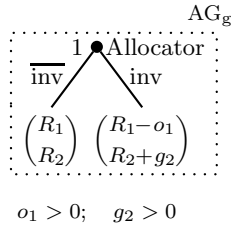
a) Trust Game with Sanctions (TGS)



b) Allocation Game for investment in a fine (AG$_f$)



c) Dictator Game with Sanctions (DGS)



d) Allocation Game for investment in gratification (AG$_g$)

Decisions labels are abbreviated with "pun" for punishment, "rew" for reward, and "inv" for investment; bar for "no".

ever, if sanctions are *costly* ($o_1 > 0$) and actors are primarily concerned with their own objective outcomes, trustors neither reward nor punish the trustee. Under these conditions, trustors withhold trust because it would be abused.

Outcome-based motivations can induce trustors to perform costly sanctions. For instance, inequality aversion can create an incentive to punish abused trust, and collectivist preferences for maximizing joint outcomes can motivate trustors to reward honored trust. However, if people cared only about the distribution of outcomes, their behavior would not be different in a decision situation in which the trustee decides about sharing his original property. Such a decision situation is described by a dichotomous Dictator Game with Sanctions (DGS, Figure 5.1c). The trustee then is in the position of a dictator but restricted by an active responder (formerly, the trustor) who can reward shared gains and punish if the dictator has kept the gains. Similarly, the subgames that start with the trustor's sanctioning decision represent two Allocation Games (AG), one for punishment (AG_f, Figure 5.1b) and one for reward (AG_g, Figure 5.1d). The trustor then is in the position of an allocator deciding whether to retain a certain distribution of objective outcomes or to invest in decreasing (AG_f) or in increasing (AG_g) the receiver's outcome (formerly, the trustee's outcome). The part of the decision situation starting with the sanctioning decision is exactly the same, regardless of whether it is the trustor's decision in the TGS, the receiver's decision in the DGS, or the allocator's decision in the respective AG. In the case of punishment, the decision is always an investment in order to reduce the other's outcome by a fine, and in the case of reward, it is always an investment in order to increase the other's outcome by a gratification. The only difference is the behavioral context, i.e., the preceding choice options and actual decisions made. In contrast to outcome-based motivations, intention-based motivations induce actors to respond to previous decisions and, thus, to discriminate between the decision situations.

In the TGS, a trustee who abuses trust refuses to return the favor of placed trust and inflicts harm on the trustor to whom he owes the favor. Therefore, abused trust demands retaliation more strongly for than a lack of generosity that does not fail to fulfill an outstanding obligation. At first sight, a similar reasoning applies to rewarding trustworthiness. By sharing the gains that result from the trustor's trustfulness, the trustee behaves in a friendly manner. Trustors might therefore feel an obligation to return the favor. However, it seems reasonable to distinguish between rewarding favors and rewarding returned favors. The feeling of obligation to return a favor might be stronger than the desire to reward the fact that another person's has been so kind as to return a favor. Fulfilling an obligation might even be perceived as something

that can be expected and not as something that deserves to be particularly rewarded (see also Coleman, 1990: ch. 12). In contrast to the TGS, the dictator's generosity in the DGS is an original favor. Moreover, it might also seem more legitimate in the DGS not to share the gains generously, because they are the dictator's property anyway. This reduces motivations to punish greediness but can increase motivations to reward for a surprising gift. Note that the trustor's gratefulness in the TGS might well increase with the sure gain given up by the trustor, but that gratefulness should also be hampered by the improved outcomes provided to the trustee.

Hypothesis 5.1: Unkindness of abused trust and kindness of unexpected generosity
Compared to punishing and rewarding in the TGS, greediness is *less* likely to be punished and generosity is *more* likely to be rewarded in the DGS.

Concerning the allocator's decision, intention-based motivations cannot play a role. However, the allocator might well be concerned with the distribution of outcomes. For instance, an aggressive or competitive social orientation can motivate the allocator to reduce the receiver's outcome in the AG_f, while altruistic or collectivistic orientations can induce the allocator to increase the receiver's outcome in the AG_g. Inequality aversion can motivate investments only if this reduces inequality between an actor's own and others' outcomes. In the DGS and in the TGS, the other's kind or unkind behavior precedes the decision of whether or not to invest in changing the other's outcome and activates intention-based motivations. The dictator's generosity is a favor that demands a favor in return (Hypothesis 5.1). Although keeping gains might be perceived as more legitimate in the DGS than in the TGS (Hypothesis 5.1), omitting a friendly option is unkind and triggers revenge. Responders in the DGS should therefore invest more in both increasing and decreasing the other's outcome than allocators in the AG. Similarly, abused trust is particularly unkind, because the trustee has refused to fulfill an obligation and thereby inflicted a loss upon the trustor (Hypothesis 5.1), whereas relief and gratitude over trustworthiness motivate reward.

Hypothesis 5.2: Kindness of generosity and unkindness of greediness
Compared to rewarding and punishing in the DGS as well as in the TGS, investments in order to change the other person's outcome are *less* likely to be made in the AG.

Promises of Trustworthiness in Trust Situations with Sanctions

The TGS can be extended with an option for the trustee to promise his trustworthiness (H_2TGS). Specifically, the trustee decides whether or not to make the promise prior to the trustor's decision of whether to place or to withhold trust. The trustor is informed about the trustee's decision. This creates two additional behavioral contexts for the trustor's sanctioning decisions: one as the TGS after the trustee has promised trustworthiness (TGS$|H_2^+$) and one as the TGS after the promise has been omitted (TGS$|H_2^0$). Sanctioning decisions in each of these two behaviorally embedded TGSs are compared to decisions made in the TGS without promise option. Since promises are cheap-talk, the outcomes in the subgames of the H_2TGS are identical to those in the TGS, regardless of whether or not the promise has been made.

If the trustee abuses trust after he has promised trustworthiness (TGS$|H_2^+$), he not only omits to return the favor of placed trust (Hypothesis 5.1) but also reneges on his promise. The trustee misled the trustor in order to exploit his trustfulness. Therefore, abusing trust is more unfriendly after having promised trustworthiness than in the decision situation in which no promise is possible. The trustor, realizing that the trustee's promise is in fact a lie, can be expected to feel particularly angry. This increases the trustor's thirst for revenge and, thus, his motivation to incur the costs for punishing the trustee. Also previous experiments provide support for the idea that inflicted harm is punished more strongly if the other reneged on a promise than if the other abused trust without having promised trustworthiness beforehand (Brandts and Charness, 2003). Keeping a promise is at first sight something kind and deserves to be rewarded. However, recall the distinction made between rewards in terms of returning favors and rewards for returned favors (Hypothesis 5.1). Trustfulness rewards the trustee's promise, but in the expectation that the promise will be kept. If the trustee then indeed keeps the promise, he rewards the trustfulness he asked for and for which he thus shares some responsibility (see Chapters 2 and 4). This reduces the trustor's feeling of obligation to return the favor of honored trust. Kept promises that required a preceding favor can even more be perceived as a matter of course. Therefore, trustors might feel less inclined to reward honored trust, if the trustee promised to do so in advance in order to receive the favor of placed trust. In this sense, not rewarding a second time for the kept promise might be perceived as perfectly legitimate.

Hypothesis 5.3: Influence of promised trustworthiness on sanctioning
Compared to punishing and rewarding in the TGS (i.e., without promise opportunity), abused trust is *more* likely to be punished and honored trust is *less* likely to be rewarded after a promise has been made (TGS|H_2^+).

The opposite holds, if the trustee did not make the promise (TGS|H_2^0). The trustor might have nevertheless placed trust in order to induce the trustee to feel obliged to return the favor. If trust then gets abused, the trustor should realize that thit is his own mistake, given that the trustee explicitly omitted to promise his trustworthiness. Therefore, the trustor should be less motivated to punish the trustee for abused trust than in the decision situation in which the trustee has no opportunity to make or omit a promise. Omitting the promise can even be perceived as a kind hint that the gain from abusing trust is overly tempting and that the trustor therefore should not place trust unless he would be so altruistic as to allow the trustee to abuse trust. If the trustee gave in to the induced obligation to return the unwanted favor and honored trust, it is a particularly kind act. Trustworthiness despite the omitted promise is a favor provided to the trustor. Therefore, the trustor can be expected to be especially grateful that the trustee nevertheless decided for the jointly improved outcomes. This gratitude creates a strong obligation to return the favor, thus motivating the trustor to incur even large costs in order to reward the trustee. Note, however, that the trustor worries more about abused trust after the promise has been omitted. Stress associated with favors reduces gratitude, especially, if the stress is unnecessary and avoidable. Considering that no transaction costs are involved, the trustor might even be annoyed that the trustee did not make the promise, given the fact that the trustee honored trust. However, it seems reasonable to assume that trustors hope for trustworthiness and do not place trust while anticipating irritation about honored trust.

Hypothesis 5.4: Influence of not promised trustworthiness on sanctioning
Compared to punishing and rewarding in the TGS (i.e., without promise opportunity), abused trust is *less* likely to be punished and honored trust is *more* likely to be rewarded after a possible promise of trustworthiness has not been made (TGS|H_2^0)

Reward Promises and Punishment Threats in Trust Situations with Sanctions

Now consider the decision situation in which the trustor can announce sanctions (H_1TGS), while the trustee has no opportunity to promise trustworthiness. Recall that the trustee has no decision to make if the trustor withheld trust. Therefore, announcing sanctions is only meaningful in combination with placing trust. Thus, the trustor chooses one of four options: withholding trust, placing trust with promising a reward (TGS|H_1^+), placing trust with threatening punishment (TGS|H_1^-), or placing trust without announcing sanctions (TGS|H_1^0). Thereafter, the trustee decides whether or not to honor trust, followed by the trustor's decision of either whether or not to punish or whether or not to reward. The combinations of placed trust and announcement decision create three behavioral contexts for the trustor's sanctioning decisions. The sanctioning decisions in these three contexts are compared to those in the TGS without announcement option.

Promising a reward (TGS|H_1^+) is a favor that requires repayment, especially given its combination with placed trust (Chapter 4). If the trustee abuses trust, he omits to fulfill two obligations. This induces a particularly strong thirst for revenge that increases the trustor's motivation to punish the trustee. Next, it has been argued that trustworthiness is a returned favor (Hypothesis 5.1). Moreover, trustors might think that the trustee has been motivated by the prospect of a reward rather than by the desire to fulfill the obligation to return the favor of placed trust. Trustors might therefore feel less obliged to reward returned favors and perceive omitting the reward as legitimate. However, the desire for self-consistency induces trustors to keep their promise. In addition, trustors share some responsibility for the trustee's decision to forgo the gain from abused trust (see also the arguments for Hypothesis 5.3). This increases the feeling of obligation to reward the returned favor.

> **Hypothesis 5.5: Influence of a reward promise on sanctioning**
> Compared to punishing and rewarding in the TGS (i.e., without promise opportunity), abused trust is *more* likely to be punished and honored trust is *more* likely to be rewarded after a reward for trustworthiness has been promised (TGS|H_1^+).

The desire for self-consistency promotes not only promise-keeping but also actually performing the threatened punishment (TGS|H_1^-). It might also be perceived as more legitimate to inflict harm on someone if this consequence has been indicated in advance (see also Chapter 4). Moreover, trustors who employ a threat although

they could instead have chosen to promise a reward might be more inclined to punish rather than to reward. Therefore, trustors should be much more motivated to carry the outlay for taking revenge after they have explicitly threatened to do so. If the trustee honored trust, trustors might perceive it as more kind, if they are aware of the hampering influence of threats. However, given that the trustor chose to threaten punishment, it seems reasonable to assume that he expects that the threat promotes trustworthiness. The trustor might then have the impression that the trustee did not voluntarily return the favor of placed trust but honored trust because of the expected punishment. This undermines feelings of obligation to reward for trustworthiness. Moreover, the trustor might perceive omitting a reward as more legitimate, given that he chose to threaten punishment. Therefore, the trustor might have little motivation to reward trustworthiness after he has chosen to threaten with punishment.

Hypothesis 5.6: Influence of a punishment threat on sanctioning
Compared to punishing and rewarding in the TGS (i.e., without promise opportunity), abused trust is *more* likely to be punished and honored trust is *less* likely to be rewarded after a punishment for abused trust has been threatened ($\text{TGS}|\text{H}_1^-$).

Finally, consider the decision situation that arises after the trustor has neither promised a reward nor threatened punishment ($\text{TGS}|\text{H}_1^0$). In the case of abused trust, the desire for self-consistency then limits the trustor's motivation to take revenge. The trustor might even perceive it as his own mistake that he placed trust but omitted to provide an incentive to the trustee (see also Hypothesis 5.4). Moreover, by explicitly omitting to announce sanctions, the trustor might indicate that he will neither punish nor reward (see also Hypothesis 5.4). This would even be an honest signal in case a high outlay would be required. Therefore, the trustor might also perceive it as more legitimate to omit the reward. In addition, recall that trustworthiness is a returned favor that might create a weaker feeling of obligation than having received a generous share in the DGS (Hypothesis 5.1). Therefore, the trustor might hardly be motivated to reward trustworthiness in this decision situation.

Hypothesis 5.7: Influence of an omitted sanctioning announcement on sanctioning
Compared to punishing and rewarding in the TGS (i.e., without promise opportunity), abused trust is *less* likely to be punished and honored trust is *less* likely to be rewarded after neither reward has been promised nor punishment has been threatened ($\text{TGS}|\text{H}_1^0$).

Table 5.1: Overview of hypotheses and notation

	Punishing	Rewarding	
AG	– –	– –	Allocation Game
DGS	–	+	Dictator Game with Sanctions (no placed trust)
TGS	(ref.)	(ref.)	Trust Game with Sanctions
TGS$\|H_2^+$	+	–	TGS after a made promise to honor trust
TGS$\|H_2^0$	–	+	TGS after an omitted promise to honor trust
TGS$\|H_1^+$	+	+	TGS after placed trust with a reward promise
TGS$\|H_1^-$	+	–	TGS after placed trust with a punishment threat
TGS$\|H_1^0$	–	–	TGS after placed trust with an omitted announcement of sanctions

The hypotheses are formulated in terms of differences toward the TGS (i.e., the behavioral context without anouncement options). In the Allocation Game the allocator decides whether or not to investment in changing the other's outcome, which represents the trustor's "sanctioning decision" without behavioral context.

Summarizing the hypotheses highlights the differences between influences of behavioral contexts on revengefulness and on gratefulness (Table 5.1). Due to feelings of indignation, revengefulness should be increased after the trustee has refused to return a favor (in the TGS vs. DGS, as well as in the TGS$|H_1^+$, and in the TGS$|H_2^+$; see Hypotheses 5.1, 5.3, and 5.5). Gratefulness might be increased, if the favor is less likely to be expected, thus inflicting stronger feelings of obligation (in the DGS and in the TGS$|H_2^0$; see Hypotheses 5.1 and 5.4), and decreased if the preceding decision is already a returned favor (in the TGS vs. DGS, and in the TGS$|H_2^+$; see Hypotheses 5.1 and 5.3). Self-consistency should increase the motivation to punish after punishment has been threatened (TGS$|H_1^-$) and the motivation to reward after reward has been promised (TGS$|H_1^+$; see Hypotheses 5.5 and 5.6), while reduced sanctioning is expected after sanctioning announcements have been explicitly omitted (TGS$|H_1^0$; see Hypothesis 5.7). Furthermore, since neither obligation nor indignation nor self-consistency are activated in the AG, investments in changing the other's outcomes are expected to be much less likely than in the TGS or in the DGS (Hypothesis 5.2).

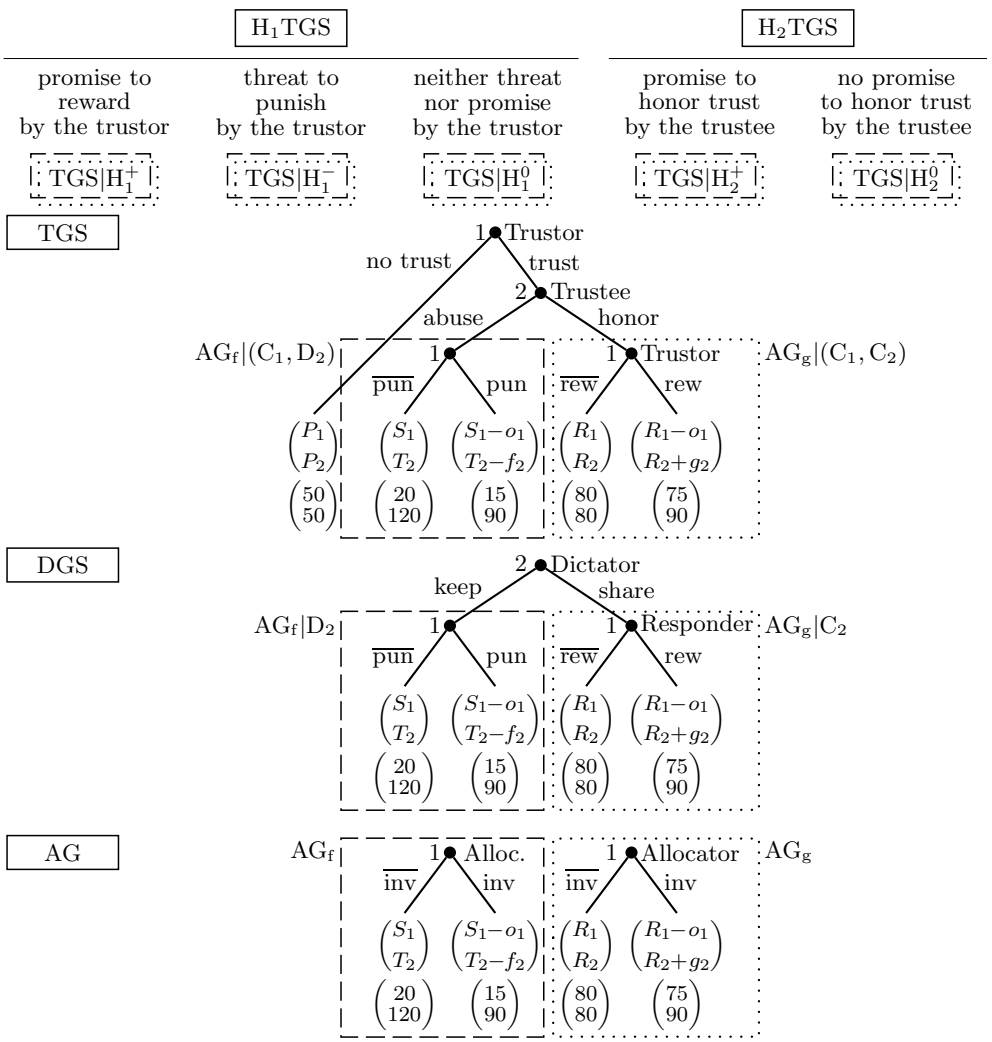
5.3 Design of the Experiment, Data, and Statistical Method

5.3.1 Experimental Design: Sets of (Sub)Games

The design of the experiment satisfies two main features. First, the behavior of each participant in decision situations with different behavioral contexts is recorded. Second, influences of outcome-based motivations can be controlled. For this purpose, participants made decisions in different games, for the analyses presented here, in TGSs, H_1TGSs, H_2TGSs, DGSs, and AGs (for details, see Vieth, 2008). Each game was a single encounter. Following the approach employed by Vieth and Weesie (2006; and see Chapter 2), the experiment was designed as within-subject sets of (sub)games that consist of the same behavioral options and the same outcomes for both actors (for related designs, see Snijders, 1996; Charness and Rabin, 2002, 2005; McCabe et al., 2003; Cox, 2004; and for further remarks, see Vieth and Weesie, 2006; and see Chapter 2). This design allows for the comparison of participants' behavior in decision situations with identical choice options and identical outcomes but in different behavioral contexts that result from preceding kind and unkind behavior (Figure 5.2). Intra-personal differences in behavior between the (sub)games should therefore reflect the impact of the behavioral contexts.

In the H_2TGS, the trustee decides whether or not to promise trustworthiness. Thus, an H_2TGS contains two TGS: one after the promise has been made (TGS$|H_2^+$) and one after the promise has been omitted (TGS$|H_2^0$). In the H_1TGS, the trustor chooses among three announcement options. Since the trustor's "announcement" decision implies that he placed trust, it does not constitute a behavioral context for trustfulness, but only for the subsequent decisions. Note that these subsequent decisions are made in a behaviorally embedded TGS (not in a DGS) because of the trustor's preceding decision to place trust. Therefore, an H_1TGS contains three TGSs after placed trust: one in which reward has been promised (TGS$|H_1^+$), one in which punishment has been threatened (TGS$|H_1^-$), and one in which no announcement has been made (TGS$|H_1^0$). Each TGS contains one DGS as a subgame, and each DGS contains two AGs as subgames. One of the two AGs describes the trustor's decision of whether or not to invest in reducing the trustee's outcome (AG_f for punishment by a fee). The other AG describes the trustor's decision of whether or not to invest in increasing the trustee's outcome (AG_g for reward by a gratification). In addition to these behaviorally embedded subgames, the design also included TGSs, DGSs, and AGs as separate games. The announcement options for trustors in the H_1TGS and for trustees in the H_2TGS were cheap-talk, i.e., the choice did not change objective

Figure 5.2: Sets of games with identical subgames



In the H_1TGS, the trustor can only choose to announce sanctions (3 options) if he places trust. Full graphs of the H_1TGS and the H_2TGS are omitted because these graphs are complex without being more informative for cheap-talk announcements. The experimental design allows for the comparison of the trustor's revengefulness in (sub)games indicated by *dashed boxes* (8 contexts) and the trustor's gratefulness in (sub)games indicated by *dotted boxes* (8 contexts). These sets of (sub)games constitute "subject-payoff response sets" used in statistical analyses. Numerical example: $S_1^{\text{low}} = 20$, $T_2^{\text{high}} = 120$, $R_1 = R_2 = 80$, $P_1 = P_2 = 50$, $o_1^{\text{low}} = 5$, $f_2^{\text{high}} = 30$, $g_2^{\text{low}} = 10$. Decision labels are abbreviated with "pun" for punishment, "rew" for reward, and "inv" for investment; bar for "no".

Figure 5.3: Outcome parameters of the experimental design

DESIGN PARAMETERS:
$S_1(2) \times T_2(2) \times \pi_i(3) \times f_2(3) \times g_2(3)$

Symmetric payoff structure: $S_1(2) \times T_2(2)$

$S_1^{\text{low}} = 20 \quad T_2^{\text{low}} = 100$
$S_1^{\text{high}} = 40 \quad T_2^{\text{high}} = 120$

$R_1 = R_2 = 80$
$P_1 = P_2 = 50$

Sanctioning properties: $f_2(2) \times g_2(2)$

$f_2^{\text{low}} = g_2^{\text{low}} = \frac{1}{4}(T_2 - R_2) = \{5, 10\}$
$f_2^{\text{high}} = g_2^{\text{high}} = \frac{3}{4}(T_2 - R_2) = \{15, 30\}$

$o_1^{\text{low}} = \frac{1}{6}(R_1 - P_1) = 5$
$o_1^{\text{high}} = \frac{2}{6}(R_1 - P_1) = 10$

Asymmetric payoff structure: $\pi_i(3)$
$(\pi_1 \in \{R_1, P_1, S_1\};\ \pi_2 \in \{T_2, R_2, P_2\})$

Trustor advantage: $\pi_1 + 10$, $\pi_2 - 10$
Trustee advantage: $\pi_1 - 10$, $\pi_2 + 10$

The outlay was fixed per combination of T_2, S_1, f_2, g_2, varying for three of the four parameters. Figure repeated from Chapter 4.

TGS, one H_1TGS, one H_2TGS, one DGS, and two AGs. Within each set, (sub)games had identical payoffs.

Varying some outcome parameters yielded different sets of (sub)games (Figure 5.3) These variations were included in the design for methodological reasons (for details, see Vieth, 2008) and for further analyses. The payoffs resulting from abused trust (S_1 and T_2) were varied at two levels each (low, high). This yields four baseline payoff combinations with 20 and 40 for S_1 and 100 or 120 for T_2. The baseline payoffs after no trust (P_i) and after honored trust (R_i) were fixed at $P_1 = P_2 = 30$ and at $R_1 = R_2 = 60$. Thus, the baseline payoffs represent a symmetric payoff structure. Two asymmetric payoff structures were constructed by adding 10 to the payoffs for the advantaged position and simultaneously subtracting 10 from the payoffs for the disadvantaged position. This yielded four payoff combinations with a trustor advantage and four payoff combinations with a trustee advantage. The three sanctioning properties were likewise varied at two levels each (low, high) resulting in eight combinations. The fine (f_2) and the gratification (g_2) were varied at the scale of the trustee's temptation with $\frac{1}{4}(T_2 - R_2)$ as "low" values and $\frac{3}{4}(T_2 - R_2)$ as "high" values. Thus, fine and gratification vary with the trustee's payoff from abused trust (T_2). For instance, in the case of $T_2 = 120$, the trustee's temptation amounts to $T_2 - R_2 = 40$, such that $f_2^{\text{low}} = g_2^{\text{low}} = \frac{1}{4}(T_2 - R_2) = 10$ and $f_2^{\text{high}} = g_2^{\text{high}} = \frac{3}{4}(T_2 - R_2) = 30$.

Note that sanctioning is effective if both fine and gratification have "high" values but is never credible because of $o_1 > 0$. The outlay (o_1) was defined at the scale of the difference between the trustor's sure payoff and the trustor's gain from honored trust with $o_1^{\text{low}} = \frac{1}{6}(R_1 - P_1) = 5$ and $o_1^{\text{high}} = \frac{2}{6}(R_1 - P_1) = 10$. The four combinations of sanctioning properties for player 2 and the four symmetric payoff combinations yielded 16 combinations. The outlay was fixed per combination in a way that it varied for three of the four parameters across these combinations (for details, see Vieth, 2008). Together with the three payoff structures, the parameter variations resulted in 48 different combinations of payoffs and sanctioning properties.

Each participant made decisions in two sets of games in the role of player 1 (trustor, responder, allocator) and in two game sets in the role of player 2 (trustee, dictator, receiver) and was for each encounter randomly and anonymously matched with another participant (stranger matching whereby the probability of re-matching was minimized within each type of game, see Vieth and Weesie, 2006). The sets of (sub)games were mixed by clustering types of games. For the results reported here, the clustering was as follows: first 10 TGSs, followed by 10 H_1TGSs, subsequently 10 H_2TGSs, then 8 DGSs, and finally 16 AGs. The experiment also included some TGs after the H_2TGSs, some H_2TGs after the DGSs, and some DGs between the AGs (for details, see Vieth, 2008). The ordering of game clusters was fixed in a way that maximal differences between game clusters were assured concerning the presentation of decision situations (for details, see Vieth, 2008). Two TGSs, two H_1TGSs, and two H_2TGSs were constructed without objective incentive for trustees to abuse trust ($T_2 < R_2$). These decision situations are not used in the analyses but were included in the design in order to check for participants' attention. In fact, 93.0% trustfulness and 99.0% trustworthiness were observed in these decision situations. This indicates that participants strongly paid attention to the objective outcomes (compared to the highest averages of 53.6% trustfulness and 45.0% trustworthiness per behavioral context when $T_2 > R_2$, as reported in Chapter 4). Note that in the decision situations in which $T_2 < R_2$, neither full trustfulness nor full trustworthiness was expected , due to possible influences of other-regarding outcome-based motivations (e.g., aggressive or competitive tendencies). Two brief questionnaires concerning participants' socio-demographic characteristics (e.g., gender, age, education) separated TGSs from H_1TGSs and H_1TGSs from H_2TGSs. Other questions about personal attitudes and opinions followed the AGs (for details on questionnaires, see Vieth, 2008). Analyses of questionnaire items are not reported here. In each game cluster, player roles were changed after half of the periods. In addition to randomly changing interaction part-

ners, payoffs and sanctioning properties changed from one period to the next. The combinations and sequences of payoffs and sanctioning properties were varied across experimental sessions employing a factorial design.

The experiment was computer-assisted, employing the software package "z-Tree" (Fischbacher, 2007) (for an example of the decision screens, see Appendix A.2). In addition to general information on paper, participants received on-screen instructions and a tutorial before each game cluster. Outcomes were displayed as points in decision trees and represented monetary gains (one British pence for four points). Participants were paid anonymously and immediately after the experiment. On average, participants earned approximately 14 GBP. The experiment was conducted in April 2008 at the CeDEx lab at the Nottingham School of Economics. Using "ORSEE" (Greiner, 2004), 166 students were recruited from the CeDEx participant pool and took part in nine groups of 16 to 20 participants. Participants were enrolled in various fields, primarily at Nottingham University.

5.3.2 Data and Statistical Method

The 166 subjects made 2955 "sanctioning" decisions about investing in changing the other's outcome, i.e., 1857 "punishing" decisions (reducing the other's outcome) and 1098 "rewarding" decisions (increasing the other's outcome) (Table 5.2). Since the focus in this study is on analyses of sanctioning behavior, "placing trust" and "honoring trust" decisions are not reported (for details, see Vieth, 2008; and Chapter 4). Recall that participants made "sanctioning" decisions in the role of the trustor (in the differently embedded TGSs), in the role of the recipient (in the DGSs), and in the role of the allocator (in the AGs). Since participants always reacted to decisions actually made, the number of decisions per subject varies. Each participant could maximally make 20 decisions of either punishment or reward, i.e., up to 12 decisions in the three differently embedded TGSs (four decisions per game), up to 4 decisions in the DGSs, and 4 "sure decisions" in the AGs. Participants actually made between 8 and 16 "punishing" decisions and between 4 and 13 "rewarding" decisions. The "sure decisions" in the AG helped realize all 48 different combinations of payoffs and sanctioning properties for both kinds of "sanctioning" decisions. Each of these combinations generates the "total payoffs" of a decision situation. Each subject made "punishing" decisions and "rewarding" decisions in identical (sub)games that only differ with respect to the behavioral context. For the analyses, decisions are grouped per subject and per combination of total payoffs and constitute 664 "subject-payoff

Table 5.2: Number of cases and units of analyses

Number of ...	Punishing	Rewarding
subjects	166	166
total payoffs	48	48
subject-payoff response sets	664	664
decisions in total	1857	1098
decisions per subject	6–18	4–13
decisions per response set	1–5	1–5

Total payoffs are combinations of payoffs and promise properties. Decisions for rewarding without TGS|H_2^0: 1090 in total, 4–13 per subject, 1–5 per response set.

response sets". Response sets involve between 1 and 5 "sanctioning" decisions for "punishing" and also for "rewarding".

In order to describe the composition of response sets and the available data, an overview of the decisions made in subject-payoff response sets per (sub)game is given in Table 5.3. For instance, in response sets that involve the TGS, 205 of the 1857 "punishing" decisions and 72 of the 1098 "rewarding" decisions have been made. All decisions made in a (sub)game belong to different response sets, because each participant could face a (sub)game with certain total payoffs at most once. Across (sub)games, the number of decisions does not correspond to the number of response sets, because several decisions made in different (sub)games constitute one response set. For instance, of the 664 response sets involving the AG, 205 "punishing" decisions and 72 "rewarding" decisions also include the TGS. Note that some of these response sets in addition involve other subgames, if the trustor placed trust in the respective decision situation.

Table 5.3 also provides information about the frequency of abused trust ("punishing" decisions), honored trust ("rewarding" decisions), and placed trust (sum of "punishing" decisions and "rewarding" decisions) per subgame and in total. Moreover, it becomes visible that 308 $(194 + 114)$ of the 333 cases $(194 + 114 + 17 + 8)$ in the H_2TGS were observed after the trustee had promised trustworthiness (TGS|H_2^+), which amounts to 92.5%. Decision situations in which trust has been withheld are not included in the analyses reported here because no subsequent sanctioning decision is involved. For the whole data set (i.e., including withheld trust), trustees promised trustworthiness in 575 of the 664 cases (86.5%) in the H_2TGS (see Chapter 4). Similarly, in 87.8% of the decision situations, trustors combined placing trust

Table 5.3: Number of decisions within subject-payoff response sets per (sub)game

	Punishing (x)					Rewarding (z)				
	all $\bar{x}$	all x	mix	Σ	%x	all $\bar{z}$	all z	mix	Σ	%z
AG	441	16	207	664	5.3	514	28	122	664	4.7
DGS	373	11	186	570	15.4	59	1	34	94	26.6
TGS	111	4	90	205	22.9	39	2	31	72	26.4
TGS$\|H_2^+$	102	6	86	194	40.2	71	5	38	114	16.7
TGS$\|H_2^0$	12	0	5	17	11.8	5	0	3	8	0.0
TGS$\|H_1^+$	81	4	87	172	45.3	49	4	85	138	64.5
TGS$\|H_1^-$	1	0	5	6	66.7	1	0	2	3	33.3
TGS$\|H_1^0$	23	1	5	29	13.8	4	1	0	5	20.0
Σ	1144	42	671	1857	18.1	742	41	315	1098	16.8

"Punishing" decisions are denoted by "$\bar{x}$" for no punishment and by "x" for punishment. Similarly, "rewarding" decisions are denoted by "$\bar{z}$" for no reward and by "z" for reward. The percentages of punishment (%x) and reward (%z) are calculated for the respective sum of decisions (data in the analyses).

with a reward promise in the H_1TGS, namely in 310 $(172 + 138)$ of the 353 cases $(172 + 138 + 6 + 3 + 29 + 5)$ of placed trust (Table 5.3). Since behavioral contexts endogenously depend on actual decisions made, only a few observations of decisions are available in some subgames.

Concerning all (sub)games with identical total payoffs in which a participant decided whether to punish the other, participants always punished in 42 cases (all x) and always decided to refrain from punishing (all $\bar{x}$) in 1144 cases (Table 5.3). Of these 1186 "punishing" decisions, 457 decisions $(441 + 16)$ have been made in the AG, of which 36 decisions (34 for all $\bar{x}$ and 2 for all x) constitute singletons (i.e., response sets consisting of one single decision). Similarly, in 41 cases (all z), participants always chose to reward and always omitted a reward in 742 cases (all $\bar{z}$). Of these 783 "rewarding" decisions, 542 decisions have been made in response sets that involve the AG, with 344 singletons (327 for all $\bar{z}$ and 17 for all z). Participants always making the same decision across (sub)games do not allow for the discrimination between behavioral contexts, for given total payoffs. In response sets that show some mixed pattern, 671 "punishing" decisions and 315 "rewarding" decisions have been made.

Across (sub)games, "punishing" decisions were made in 18.1% (336 cases) of all 1857 cases and "rewarding" decisions in 16.8% (185 cases) of all 1098 cases (Table 5.3). The level of costly sanctioning is typically found to be rather low in other

studies (e.g., on sanctioning for non-cooperation, see Voss and Vieth, 2006; and on rejection rates of low offers, see Camerer, 2003: ch. 2). Nevertheless, the levels of punishing and rewarding differ considerably across behavioral contexts, ranging from approximately 5% to over 60%. These average percentages of punishing and rewarding provide information about the actual decisions made in the (sub)games irrespective of the specific payoffs in the (sub)games and irrespective of the fact that each person made several decisions. For instance, it is possible that trust was abused and punished more frequently in some behavioral contexts than in others just because of the specific outcomes. Thus, testing the hypotheses about effects of behavioral advances on subsequent decisions requires accounting for the grouping structure of the data in order to control for influences of outcome-based motivations and of individual heterogeneity. For analyzing effects of behavioral contexts on people's decision-making, in Chapter 2, logistic regression models with fixed effects for response sets have been used. This method allows minimal assumptions to be made about subject-specific effects, about effects of outcomes, and about interactions between them. However, only the mixed response sets carry statistical information in a fixed effects approach, while other response sets are excluded from the analyses.

In contrast to the fixed effects approach employed in Chapter 2, the analyses reported in the study presented here treat the effects of response sets as random effects. The reason is mainly technical: Due to specific response patterns in the data, some of the model parameters became unidentified using the fixed effects estimator (see the discussion for further remarks). This concerns the $\mathrm{TGS}|\mathrm{H}_2^0$, the $\mathrm{TGS}|\mathrm{H}_1^+$, and the $\mathrm{TGS}|\mathrm{H}_1^0$ in analyses for gratefulness. Without these subgames, the fixed effects analysis of gratefulness would only be based on 52 of the 664 response sets, 40 of the 166 subjects, and 29 of the 48 total payoffs. At the cost of stronger assumptions (see below), the random effects estimators allow all but one subgame effect to be estimated. The exception is the $\mathrm{TGS}|\mathrm{H}_2^0$ in which a reward has been omitted in all 8 cases (0% gratefulness, see Table 5.3). The logistic regression models with random effects for response sets are fitted by maximum marginal likelihood and have the following general form:

$$
\begin{gathered}
y_{ijk} = y^*_{ijk} > 0 \\
y^*_{ijk} = \beta_0 + \eta'_{ijk}\beta + u_{0ij} + e_{0ijk} \\
\text{where } u_{0ij} \sim \text{Normal}(0, \sigma^2_u) \text{ and } e_{0ijk} \sim \text{Logistic}(0, \sigma^2_e)
\end{gathered}
$$

The model is applied to describe the probability of revengefulness or gratefulness of a subject i in the behavioral context of a (sub)game k that has a total payoff com-

bination j. The intercept parameter (β_0) and the (sub)game effects ($\eta'_{ijk}\beta$), together with controls, constitute the fixed part of the model, i.e., effects are assumed to be the same across response sets. The random part consists of two random variables for the two levels of analysis. Errors at the level of decisions are assumed to have a standard logistic distribution with a mean of 0. The variance within response sets is fixed in logistic regression for identification purposes and serves as a scaling parameter ($\sigma_e^2 = \frac{1}{3}\pi^2 \approx 3.29$, Long, 1997: 47–48). The intercept $\beta_0 + u_{0ij} + e_{0ijk}$ is allowed to vary randomly between response sets, reflecting that the average response probabilities differ between response sets. Models with random effects at the level of response sets require additional assumptions about the distribution of deviations that are specific to response sets. Specifically, it is assumed that the response sets are drawn from a population in which combinations of subjects and total payoffs are normally distributed with a mean of 0, i.e., $u_{0ij} \sim \text{Normal}(0, \sigma_u^2)$. The variance ($\sigma_u^2$) between response sets reflects the extent of unexplained deviations of the average probability per response set from the overall average probability. Note that the normality assumption might be more problematic for subject-payoff response sets than for subject response sets. However, subject-payoff response sets are preferred because this grouping of the data allows additive payoff effects to be controlled without further assumptions about specific representations of individually heterogeneous outcome-based motivations (for a discussion of this problem, see Chapter 3). In the study presented here, the reported results obtained with random intercept models are qualitatively similar to the results obtained with a fixed effects approach.

5.4 Results

5.4.1 Analyses for Trustworthiness

First, consider the situation in which the trustor is confronted with abused trust and decides whether or not to punish the trustee. This decision situation can occur in the following behavioral contexts: in the TGS ("empty context"), in the two TGSs as subgames of the H_2TGS after the trustee has decided whether or not to promise trustworthiness (TGS$|H_2^+$ and TGS$|H_2^0$), and in the three TGSs as subgames of the H_1TGS after the trustor has chosen to place trust combined with a punishment threat (TGS$|H_1^-$), with a reward promise (TGS$|H_1^+$), or without sanctioning announcement (TGS$|H_1^0$). Moreover, in the DGS the responder decides about punishing the dictator for having kept the gains without the preceding favor of placed trust. And, finally, in the AG_f the trustor is in the role of an allocator choosing whether or not to invest in reducing the other's outcome, which represents the trustor's aggressive tendencies.

Table 5.4: Logistic regression of revengefulness with random intercepts for subject-payoff response sets

(A) Regression coefficients

	Hyp.	b	se	Pr(%)
Behavioral contexts				
AG_f	H_2: – –	-2.01^{***}	0.34	1.5
DGS	H_1: –	-0.63^{*}	0.27	5.6
TGS		(ref.)		10.1
TGS$\|H_2^+$	H_3: +	1.46^{***}	0.32	32.7
TGS$\|H_2^0$	H_4: –	-0.35	0.97	7.3
TGS$\|H_1^+$	H_5: +	1.78^{***}	0.33	39.9
TGS$\|H_1^-$	H_6: +	3.17^{*}	1.29	72.7
TGS$\|H_1^0$	H_7: –	-0.62	0.78	5.7
Past periods per game		-0.05	0.03	
Constant		-1.89^{***}	0.30	
SD(error decisions)		1.81	fixed	
SD(error response sets)		2.01	0.21	
rho		0.55	0.05	

(B) Likelihood-ratio tests

	χ^2	df
LR test (rho = 0)	108.88^{***}	1
LR test (control)	221.35^{***}	7

N(response sets) = 664, N(decisions) = 1627, N(subjects) = 166;
two-sided p-values: *** p = 0.001, ** p = 0.01, * p = 0.05, ° p = 0.1; (sub)games (0/1), past periods per game (0 . . . 7/9/19). The standard deviation (SD) for decision residuals (e_{0ijk}) is constant, and the SD of random intercepts (u_{0ij}) for response sets is estimated. The proportion of unexplained variance at the level of response sets is denoted by rho. The absolute probability of revengefulness per (sub)game which is estimated for an average period ($\bar{t} = 5.85$) assuming $u_{0ij} = 0$. Likelihood-ratio tests are reported for the proportion of unexplained variance at the level of response sets (rho = 0) and for the presented model against null model with controls.

This yields eight behavioral contexts in which the decision about investing in reducing the other's outcome can be analyzed.

In the statistical analyses, revengefulness in the various behavioral contexts is compared to revengefulness in the TGS (Table 5.4). Pairwise comparisons of differences in revengefulness between the behavioral contexts are reported in Table 5.5. The upper part of Table 5.4 shows the estimates for effects of the behavioral contexts (discussed below). The period in which a decision has been made (i.e., the number of past periods per kind game) is included as a control but shows no significant influence on revengefulness. Note again that influences of objective outcomes are involved in subject-payoff response sets and are therefore controlled. For each behavioral context the probability of revengefulness is estimated at the mean of the number of periods. The random effects are ignored for reasons of computational convenience, i.e., fixed to $u_{0ij} = 0$. The random part of the model summarized in the lower part of Table 5.4 consists of the non-estimated scaling parameter (SD of the error of decisions) and of the (estimated) standard deviation of the random effects of response sets (SD of the error of response sets). The standard deviation of the random effect for response sets represents the deviation of average effects of response sets from the estimated overall mean of these effects. These deviations indicate unexplained influences that are specific to subjects and outcomes. About half of the total unexplained variance is at the level of response sets (rho = 0.55). Further analyses including a random effect for subject-payoff response sets and a random effect for sessions show that there is basically no unexplained variance at the session level. Therefore, the simpler two-level model is reported here. The likelihood-ratio test against the model without (sub)game dummies reported at the bottom of Table 5.4 (Panel B) shows that revengefulness significantly differs between the behavioral contexts (LR $\chi^2_{7\,\mathrm{df}} = 221.35$ with $p < 0.0001$). In the following, the results for the specific behavioral contexts are described and discussed.

Recall that the dictator in the DGS decides about sharing gains that are his own property and do not result from the favor of placed trust. Therefore, keeping gains as a dictator is less unfriendly than keeping gains as a trustee (Hypothesis 5.1). The results show that revengefulness is indeed significantly lower in the DGS (5.6%) than in the TGS (10.1%). The decrease of 4.5% seems relatively small. However, the probability of punishment anyway is somewhat low in the TGS, which leaves little room for hampering influences. Since no unkindness is preceding the allocator's decision in the $\mathrm{AG_f}$, aggressive behavior should be even more reduced (Hypothesis 5.2). The probability of aggressive investments is only 1.5% in the $\mathrm{AG_f}$. The differences

Table 5.5: Pairwise comparisons of behavioral contexts for revengefulness

	Pr(%)	TGS	AG_f	DGS	$TGS\|H_2^+$	$TGS\|H_2^0$	$TGS\|H_1^+$	$TGS\|H_1^-$
AG_f	1.5	-2.01^{***}						
(N = 664)		0.34						
DGS	5.6	-0.63^{*}	1.38^{***}					
(N = 570)		0.27	0.30					
$TGS\|H_2^+$	32.7	1.46^{***}	3.47^{***}	2.09^{***}				
(N = 194)		0.32	0.37	0.28				
$TGS\|H_2^0$	7.3	-0.35	1.66°	0.28	-1.81°			
(N = 17)		0.97	0.99	0.93	0.97			
$TGS\|H_1^+$	39.9	1.78^{***}	3.79^{***}	2.41^{***}	0.32	2.13^{*}		
(N = 172)		0.33	0.38	0.30	0.31	0.98		
$TGS\|H_1^-$	72.7	3.17^{*}	5.18^{***}	3.80^{**}	1.71	3.52^{*}	1.39	
(N = 6)		1.29	1.30	1.28	1.28	1.58	1.28	
$TGS\|H_1^0$	5.7	-0.62	1.39°	0.01	-2.08^{**}	-0.27	-2.40^{**}	-3.79^{**}
(N = 29)		0.78	0.78	0.79	0.78	1.21	0.79	1.47

The table presents differences between coefficients of (sub)games (row – column). Standard errors are reported underneath. The entries in the columns Pr(%) and TGS (with N = 205, Pr(%) = 10.1) are repeated from Table 5.4. Wald tests (two-sided p-values, not adjusted for multiple testing): *** p = 0.001, ** p = 0.01, * p = 0.05, ° p = 0.1.

of −8.6% between the AG_f and the TGS (Table 5.4) and of −4.1% between the AG_f and the DGS (Table 5.5) are highly significant, despite the decreased revengefulness in the DGS.

In the H_2TGS, in which the trustee decides about promising his trustworthiness, the trustor is assumed to be particularly annoyed if the trustee reneged on his promise (Hypothesis 5.3). Lies are indeed punished significantly more than abused trust as such. Revengefulness increases by 22.6%, from 10.1% in the TGS to 32.7% in the $TGS|H_2^+$ (Table 5.4). In fact, lies are punished every third time. Note again that this is the "pure" influence of indignation and does not result from outcome-based motivations. In contrast to reneged promises, trustors might perceive abused trust as their own mistake, if the promise of trustworthiness has been explicitly omitted (Hypothesis 5.4). Revengefulness is indeed reduced by 2.8% (from 10.1% in the TGS to 7.3% in the $TGS|H_2^0$), but this difference is not significant. In fact, some remaining

indignation plays a role, since trustors still tend to be more revengeful in the TGS|H_2^0 than in the AG_f (Table 5.5).

After the trustor has promised reward (TGS|H_1^+) but is nevertheless confronted with abused trust, revengefulness is 39.9%, significantly higher than the 10.1% in the TGS (Table 5.4). This supports the argument that the trustor's indignation rises, because the trustee omits fulfilling two obligations, one resulting from placed trust and one resulting from the reward promise (Hypothesis 5.5). The increase in revengefulness by 29.8% in the TGS|H_1^+ is even larger than after the trustor has discovered a lie in the TGS|H_2^+ (22.6%), but the difference of 7.2% is not significant (Table 5.5). If the trustor threatened punishment (TGS|H_1^-) and the trustee dared to abuse trust, revengefulness jumps up by 62.6% and amounts to 72.7%. This difference is highly significant. Differences in revengefulness in the threat situation (TGS|H_1^-) compared to the promise situation (TGS|H_1^+) and to the situation in which no announcement of sanctions has been made (TGS|H_1^0) are not significant (Table 5.5). The reason seems to be that only 6 decisions have been made in this (sub)game (Table 5.3). Therefore, the huge increase in revengefulness should be interpreted carefully. Nevertheless, this finding indicates some support for the argument that self-consistency is also relevant for performing unfriendly behavior, not only for friendly behavior (Hypothesis 5.6). In previous studies, no support can be provided for unfriendly self-consistency in terms of reduced trustworthiness after omitted cheap-talk promises of trustworthiness (see Chapters 2 and 4; also see Snijders, 1996). Note, however, that it has been found in Chapter 2 that the influence of omitting a promise to honor trust on trustworthiness depends on the properties of the omitted promise. If the trustor neither threatened punishment nor promised reward (TGS|H_1^0), it has been assumed that the desire for self-consistency limits feelings of indignation (Hypothesis 5.7). Revengefulness is indeed reduced by 4.4% (from 10.1% in the TGS to 5.7% in the TGS|H_1^0), but this difference is not significant (Table 5.4).

5.4.2 Analyses for Gratefulness

Concerning the trustor's decision of whether or not to reward the trustee, the six behavioral contexts of the TGS are distinguished again (TGS, TGS|H_2^+, TGS|H_2^0, TGS|H_1^+, TGS|H_1^-, TGS|H_1^0). Note that the difference now is that the trustee honored trust. In addition, the responder decides about rewarding the dictator for shared gains in the DGS, and the allocator in the AG_g chooses whether or not to invest in increasing the other's outcome, representing the trustor's altruistic tendencies. The results are presented in the same format as for revengefulness (Tables Table 5.6

Table 5.6: Logistic regression of gratefulness with random intercepts for subject-payoff response sets

(A) Regression coefficients

	Hyp.	b	se	Pr(%)
Behavioral contexts				
AG_g	H_2: $--$	-3.46^{***}	0.80	0.2
DGS	H_1: $+$	0.39	0.64	9.5
TGS		(ref.)		6.7
$TGS|H_2^+$	H_3: $-$	-1.33°	0.68	1.9
$TGS|H_1^+$	H_5: $+$	3.87^{***}	0.92	77.5
$TGS|H_1^-$	H_6: $-$	-1.88	2.23	1.1
$TGS|H_1^0$	H_7: $-$	-1.98	2.58	1.0
Past periods per game		-0.04	0.05	
Constant		-2.35^{***}	0.68	
SD(error decisions)		1.81	fixed	
SD(error response sets)		3.22	0.75	
rho		0.76	0.08	

(B) Likelihood-ratio tests

	χ^2	df
LR test (rho = 0)	42.01^{***}	1
LR test (control)	236.14^{***}	6

N(response sets) = 664, N(decisions) = 1627, N(subjects) = 166;
two-sided p-values: *** p = 0.001, ** p = 0.01, * p = 0.05, ° p = 0.1; (sub)games (0/1), past periods per game ($0 \ldots 7/9/19$). The $TGS|H_2^0$ is excluded because all 8 cases are "no reward" decisions (versus H_4: $+$). The standard deviation (SD) for decision residuals (e_{0ijk}) is constant, and the SD of random intercepts (u_{0ij}) for response sets is estimated. The proportion of unexplained variance at the level of response sets is denoted by rho. The absolute probability of gratefulness per (sub)game which is estimated for an average period ($\bar{t} = 7.21$) assuming $u_{0ij} = 0$. Likelihood-ratio tests are reported for the proportion of unexplained variance at the level of response sets (rho = 0) and for the presented model against null model with controls.

and Table 5.7). The proportion of unexplained variance at the level of response sets amounts to 76%. Due to numerical and convergence problems, models including a random effect for sessions could not be fitted for gratefulness. As for revengefulness, the negative coefficient for the number of past periods per game is also not significant for gratefulness (Table 5.6). The likelihood-ratio test against the model without (sub)game dummies shows that gratefulness differs significantly between the behavioral contexts (LR $\chi^2_{6\,\mathrm{df}} = 236.14$ with $p < 0.0001$). The results presented here are again qualitatively similar to the results obtained with the fixed effects approach.

Generally, the level of gratefulness is very low: lower than 10%, disregarding one outstanding exception, and only 6.7% in the TGS. This leaves little room for a further decrease motivated by preceding behavior. However, it also strengthens the evidence for reductions that are found to be significant. Moreover, the direction of the effects is in line with the hypotheses. In the DGS, kept gains are less unfriendly than in the TGS (Hypothesis 5.1 and Table 5.4). It has been argued that shared gains in the DGS are even more kind, because they are an original favor and not a returned favor (Hypothesis 5.1). However, no support could be found for the implication of this argument: Although gratefulness is increased from 6.7% in the TGS to 9.5% in the DGS (Table 5.6), this 2.9% increase is not sufficient to become significant. As previously addressed, this might be due to the low overall reward rate. Next, similar to aggressive investments discussed with respect to revengefulness, altruistic investments should also be much lower in the absence of preceding decisions, as represented by the $\mathrm{AG_g}$ (Hypothesis 5.2). The results indeed show that altruism is significantly decreased in the $\mathrm{AG_g}$ compared to the TGS (Table 5.6) and, thus, also to the DGS (Table 5.7). Altruistic inclinations are reduced by 6.5% compared to the TGS and by 9.3% compared to the DGS. In the $\mathrm{AG_g}$, allocators invest in increasing the other's outcome with only a probability of 0.2%, compared to 6.7% in the TGS and 9.5% in the DGS.

In the $\mathrm{H_2TGS}$, the trustee decides whether or not to promise his trustworthiness. It has been argued that keeping a promise can be seen as a matter of course, especially if it required a preceding favor (Hypothesis 5.3). Gratefulness indeed tends to be decreased by 4.8%, from 6.7% in the TGS to 1.9% after the trustor has received a promise of trustworthiness ($\mathrm{TGS}|\mathrm{H}_2^+$). Gratefulness is even approximately 7.6% lower in the $\mathrm{TGS}|\mathrm{H}_2^+$ than in the DGS without preceding placed trust. Concerning the decision situation that arises after the trustee has omitted the promise of trustworthiness ($\mathrm{TGS}|\mathrm{H}_2^0$) the trustor is assumed to be particularly grateful if the trustee has fulfilled the obligation induced by the trustor's decision to place trust despite the promise

Table 5.7: Pairwise comparisons of behavioral contexts for gratefulness

	Pr(%)	TGS	AG_g	DGS	$TGS\|H_2^+$	$TGS\|H_1^+$	$TGS\|H_1^-$
AG_g	0.2	-3.46^{***}					
(N = 644)		0.80					
DGS	9.5	0.39	3.85^{***}				
(N = 94)		0.64	0.86				
$TGS\|H_2^+$	1.9	-1.33°	2.13^{***}	-1.72^{*}			
(N = 114)		0.68	0.62	0.71			
$TGS\|H_1^+$	77.5	3.87^{***}	7.33^{***}	3.48^{***}	5.20^{***}		
(N = 138)		0.92	1.33	0.81	1.11		
$TGS\|H_1^-$	1.1	-1.88	1.58	-2.27	-0.55	-5.75^{*}	
(N = 3)		2.23	2.09	2.25	2.08	2.52	
$TGS\|H_1^0$	1.0	-1.98	1.48	-2.37	-0.65	-5.85^{*}	-0.10
(N = 5)		2.58	2.50	3.84	2.46	2.79	3.24

The table presents differences between coefficients of (sub)games (row – column). Standard errors are reported underneath. The entries in the columns Pr(%) and TGS (with N = 72, Pr(%) = 6.7) are repeated from Table 5.6. The $TGS|H_2^0$ is excluded because "no reward" was chosen in all 8 cases Wald tests (two-sided p-values, not adjusted for multiple testing): *** p = 0.001, ** p = 0.01, * p = 0.05, $^{\circ}$ p = 0.1.

has been omitted (Hypothesis 5.4). However, this behavioral context ($TGS|H_2^0$) could not be included in the analyses, because variation in responses is lacking. In fact, all trustors refrained from rewarding trustworthiness (Table 5.3). Although only 8 decisions had been made in this subgame, the complete lack of rewarding indicates that trustors are not grateful. This suggests that trustors might omit the reward, because the trustee made the decision to place trust more stressful than necessary. This argument has been discussed as an alternative for Hypothesis 5.4.

Now consider the situation in which the trustor decides about accompanying placed trust with a sanctioning announcement. The obligation due to the shared responsibility for the trustee's decision and the desire for self consistency should induce the trustor to indeed reward, if he promised to do so (Hypothesis 5.5). The results show that gratefulness jumps by 70.8%, from 6.7% in the TGS to 77.5% in the $TGS|H_1^+$. Thus, in more than three of four cases the reward promise is actually fulfilled. This high level of gratefulness sharply contrasts with the low level in the other behavioral contexts ($< 10\%$). The increase is highly significant and provides strong evidence for the power of self-consistency and of shared responsibility. In contrast,

trustors who have threatened punishment might perceive it as legitimate to refrain from rewarding trustworthiness (Hypothesis 5.6). In the analyses presented here, no support could be found for such a hampering impact of punishment threats on gratefulness ($TGS|H_1^-$). In models with fixed effects for response sets, the coefficient for the $TGS|H_1^-$ is strongly negative and highly significant. However, the estimates are only based on 2 decisions in mixed response sets (of 3 decisions in total, see Table 5.3). Therefore, the decrease in gratefulness by 5.6% in the model presented here can only provide indications for further research. Explicitly omitting any sanctioning announcement has likewise been assumed to hamper rewarding behavior (Hypothesis 5.7). The results show that gratefulness is reduced by 5.7% (from 6.7% in the TGS to 1.0% in the $TGS|H_1^0$), but this difference is not significant.

5.5 Summary and Perspectives

5.5.1 Summary of Basic Ideas, Approach, and Contributions

People take revenge for unkindness and express their gratitude for others' kindness, and they are ready to expend their resources in doing so. Sociological and social-psychological research (Gouldner, 1960; Coleman, 1990: ch. 12; Cialdini, 2001: chs. 2–3) suggests that people reciprocate others' kind and unkind behavior because of obligation feelings to return favors, indignation feelings that create a thirst for revenge, and the desire for self-consistency (e.g., promoting that people are inclined to keep their word). Based on these ideas, the study presented here investigated revengefulness and gratefulness in trust situations in which trustors could punish abused trust and could reward honored trust. The aim was to study how the mere choice of kind or unkind actions influences people's subsequent decisions to punish or to reward. For this purpose, different behavioral contexts have been distinguished that were created by kind and unkind behavior while outcomes were identical. In addition to the impact of trustfulness as such, the focus was on influences of made and omitted cheap-talk announcements. In some decision situations, the trustee could promise his trustworthiness, while in other situations, the trustor could combine placing trust with a punishment threat or a reward promise. In addition to influences of abused and of honored trust, this allowed for the study of revenge for lies, reward for kept promises, and influences of announced sanctions on actual sanctioning decisions. Previous studies provide evidence for punishment of uncooperative behavior (for reviews see, e.g., Roth, 1995; Camerer, 2003: ch. 2; Shinada and Yamagishi, 2008). Some studies also explore rewarding cooperative behavior (e.g., Vyrastekova and van Soest, 2008). Concerning announcements, previous research only addressed the punishment of lies in

a way comparable to the study presented here, and showed that lies are punished more strongly than uncooperative behavior without a broken promise (Brandts and Charness, 2003). However, previous studies neither employ a within-subject design nor rule out influences of outcome-based motivations. Moreover, some studies involve confounding factors in the experimental design such that the influence of preceding behavior on sanctioning decisions cannot be assessed.

In order to control for outcome-based motivations, this study follows an approach employed by Vieth and Weesie (2006; see also Chapter 2). In doing so, the experiment is designed as within-subject sets of single encounters in structurally identical (sub)games generated by kind and unkind actual behavior. Eight behavioral contexts were distinguished, in which participants decided whether or not to invest in changing the other's outcome. In the Trust Game with Sanctions (TGS), participants decided either whether or not to punish for abused trust or whether or not to reward for honored trust. The same holds for the two TGSs as subgames after the trustee has decided whether or not to make the promise of trustworthiness and for the three TGSs as subgames after the trustor has placed trust and has decided whether or not to combine it with an announcement of sanctions, i.e., with a punishment threat or with a reward promise. Moreover, the Dictator Game with Sanctions (DGS) describes the trustee's decision to share gains without preceding trustfulness. Two kinds of Allocation Games (AGs) completed the design, one for the decision to invest in reducing the other's outcome (AG_f) and one for the decision to invest in increasing the other's outcome (AG_g). Decisions made in the AGs indicate the trustor's aggressive and altruistic motivations that are not induced by preceding kind and unkind behavior. In addition to the costly nature of sanctions, sanctioning was not always effective on objective grounds for removing the trustee's temptation to abuse trust, whereas announcements were always cheap-talk in objective terms. Sanctioning properties and outcome parameters were varied by a factorial design. In order to control for outcome-based variations, the data were grouped in subject-payoff response sets. Due to the decisions participants made, variation of gratefulness within response sets is lacking in some subgames. Therefore, instead of the fixed effects approach employed in Chapter 2, logistic regression models with random intercepts for response sets were used. This allows more parameters to be identified, at the cost of stronger assumptions. The results that could be estimated in a fixed effects approach were very similar to those obtained from the presented models.

Summarizing the results yields three main findings. First, this study provides evidence that both punishing behavior and rewarding behavior are strongly motivated

by revengefulness against others' unkindness and by gratefulness for others' kindness, respectively. Evidence has been provided that both revengefulness and gratefulness are strongly decreased in the decision situation without behavioral context, i.e., the situation in which the trustor has made a decision in the role of the allocator (AG) motivated by outcome concerns. This decrease has also been found in comparison to the situation after the trustor has placed trust (AG vs. TGS) and to the situation after the trustee has decided about sharing gains that do not arise from trustfulness (AG vs. DGS). Similarly, abused trust increases revengefulness (DGS vs. TGS). Moreover, lies are punished particularly strongly (TGS$|\mathrm{H}_2^+$ vs. TGS) (see also Brandts and Charness, 2003; Bochet and Putterman, 2007). Trustees are punished similarly strongly if they have omitted to fulfill two obligations, as it is in the case of placed trust combined with a reward promise (TGS$|\mathrm{H}_1^+$ vs. TGS). Second, the results support the idea that self-consistency strongly motivates people to keep their word concerning both reward promises and punishment threats. Gratefulness is strongly increased if the trustor promised reward (TGS$|\mathrm{H}_1^+$ vs. TGS). This contrasts with previous studies on bonus contracts, in which no support has been found for the idea that promises of higher rewards would actually result in payments of higher rewards (Fehr and Schmidt, 2004, 2007; Fehr et al., 2007). The evidence in the study presented here might also indicate the promoting impact of shared responsibility for the trustee's decision to honor trust. This argument has also been suggested in previous studies that address keeping promises of trustworthiness (see Chapters 2 and 4). Moreover, the findings suggest that trustors are particularly revengeful after trust has been abused despite punishment has been explicitly threatened (TGS$|\mathrm{H}_1^-$ vs. TGS). However, this should only be interpreted as an indication that requires to be studied in further experiments, because the number of decisions in this subgame is very small and responses are skewed. These limitations in the data might also have caused the lack of support for the arguments that punishment threats should legitimize omitting a reward and that both revengefulness and gratefulness should be reduced after omitted promises of both trustworthiness and reward. Third, some support could be found for the argument that returned favors induce weaker obligation feelings than original favors. While shared gains without the favor of placed trust do not significantly increase the trustor's motivation to reward, gratefulness indeed tends to be reduced after the trustee has kept his promise to honor trust (TGS$|\mathrm{H}_2^+$ vs. TGS).

5.5.2 Further Discussion and Perspectives

Some aspects concerning the approach employed in the presented study have been discussed in Chapters 2 and 4. Further experiments investigating sequence effects by varying the ordering of game clusters would be desirable. In the study presented here, the ordering of games was fixed in order to minimize influences across types of games by maximizing differences concerning the presentation. However, it is possible that the strong decrease in revengefulness and in gratefulness in the AGs is due to presenting them as the last decision situations. In further experiments, in which the number of decisions is sufficiently large in each behavioral context, statistical models could (a) account for the dependency of observations within subjects and within sessions and (b) test for random coefficients in order to allow for differences in influences of behavioral contexts between response sets. Eliciting beliefs, emotions, and perceived kindness would be desirable to (a) investigate whether the actual feelings and perceptions support the interpretations suggested here and (b) disentangle opposing effects that can inhibit finding support for hypothesized effects in some behavioral contexts (for remarks on drawbacks and dangers of including standard measures, see Chapter 2; and for empirical evidence, see Gächter and Renner, 2006). The small number of decisions in some subgames might be increased by incorporating transaction costs for making announcements and by separating decision situations with threat options from decision situations with promise options. Some further aspects require more detailed remarks: (1) sanctioning options and (2) properties of sanctions and of announcements.

First, in this study, the trustor can reward honored trust and can punish abused trust. Reward options are not available after abused trust, and punishment options are not available after honored trust. This could have prompted the socially desirable option. However, since the decision situations are relatively simple, it seems unlikely that participants would not be aware of the social desirability of their decisions, anyway. In a sense, they should realize it in order to make decisions that are meaningful for inferring people's other-regarding motivations. The data also show rates of cooperative behavior that are comparable to those in previous studies. Another concern is that the restriction of available options limits people's freedom of choice and, thereby, people's opportunities to express their motivations. The question then is why people might prefer punishing honored trust or rewarding abused trust.

Outcome-based motivations are a possible reason. For instance, strong altruistic motivations can induce people to sacrifice remaining resources in order to increase the other's outcome even after abused trust, while strong spiteful or even aggressive

motivations can drive people to punish after honored trust (for evidence on anti-social punishment, see Herrmann et al., 2008; Falk et al., 2005; Nikiforakis, 2008). In fact, allowing for both punishment and reward options in each case could provide interesting insights if outcome asymmetries are studied. From this perspective, the motivation for such behavior is to change the distribution of outcomes, i.e., such behavior is not caused by feelings of indignation or obligation. Of course, influences of the desire for self-consistency might also play a role. However, since reward promises and punishment threats are conditional on preceding behavior, it is unlikely that self-consistency would drive people to punish honored trust after they have threatened to punish abused trust. Therefore, not providing punishment and reward options irrespective of the trustee's decision, limits the influence of distributive concerns. Reward and punishment can still be to some extent due to outcome-based motivations. However, despite the assumption that such outcome-based motivations might influence sanctioning decisions, the presented results show the influences of preceding behavior.

Moreover, the question studied here is whether trustors invest more, e.g., in reducing the trustee's outcome after the trustee has behaved in an unfriendly manner compared to the decision situation in which the trustee did not have a choice (i.e., behaved in neither a friendly nor unfriendly manner). In contrast, comparing sanctioning behavior in the decision situation after the other has behaved in a friendly manner with the decision situation after the other has behaved in an unfriendly manner, would lump the influences of two behavioral contexts together such that joint influences would be tested.

Nevertheless, available sanctioning options can influence preceding behavior due to beliefs about anticipated sanctioning behavior. In this sense, further experiments in which decision situations differ with respect to subsequent options (i.e., a different future instead of a different past) would provide further insights (for a more detailed discussion, see Chapter 4).

Second, sanctioning properties (i.e., outlay, fine, and gratification) can moderate the impact of preceding behavior. For instance, reward promises might be more kind the higher the required outlay. Abused trust could then be punished even more strongly. Sanctioning properties are varied in the experimental design of the study presented here and can be analyzed as far as possible given the small number of decisions in some behavioral contexts. Similarly, outcome-based influences can also interact with preceding behavior (e.g., see Chapter 3). This would also allow further questions to be investigated, such as whether fairness orientations in terms of inequality aversion limit revenge-taking or rewarding. Interactions with sanctioning

properties and outcome-based motivations can be analyzed with the current data for subgames in which a sufficiently large number of decisions was made.

Moreover, announcement properties can likewise shape influences of behavioral contexts. For instance, keeping a promise of trustworthiness might be more kind if making the promise was associated with high transaction costs. Trustors might then be more grateful for the kind decision support and inclined to reward trustworthiness in order to compensate for the incurred transaction costs. Since announcements in this study are entirely cheap-talk in terms of objective outcomes, investigating such influences of announcement properties on sanctioning decisions requires new experiments in which announcement properties are varied.

Chapter 6

Summary, Discussion, and Perspectives

6.1 Summary

6.1.1 Theoretical Foundation

Many social and economic interactions involve interdependencies between people accompanied with incentives for "opportunistic behavior" (Williamson, 1985), i.e., for taking advantage of situations at the cost of others. This causes two types of problems known as "social dilemmas": cooperation problems, resulting in inefficient and suboptimal outcomes, and distribution problems, creating problematic inequalities. Both cooperation and distribution problems threaten the social order in societies and within groups of people as well as between societies and between groups by reducing cohesion among people and giving rise to conflicts (see also Voss, 1982, 1985). Social norms of cooperation and of fair distributions are desirable that help mitigate incentive problems and maintain social order. However, the enforcement of social norms requires suitable sanctions, i.e., punishment for norm deviations or reward for norm conformity. Sanctions can be based on objective incentives and, in the case of internalized social norms, on intrinsic motivations rooted in emotions. One fundamental behavioral pattern in social interactions that arises from sanctions is reciprocity. People reward kind behavior and retaliate for unkind behavior, even if it is against their objective self-interest.

Reciprocity can be implied by other-regarding motivations, such as outcome-based motivations and intention-based motivations (for a review, see Fehr and Schmidt, 2006). Outcome-based motivations arise from people's concern about the distribution of objective outcomes between themselves and others. This idea is based on social comparisons in the sense that people's own well-being (utility) also depends on others' outcomes, to some positive or negative degree. Thereby, people's concern with their own objective outcomes (selfish motivation) is complemented by some other-regarding motivation. In social-psychological research, various types and degrees of outcome concerns have been distinguished and empirically identified, known as social (value) orientations (for reviews, see McClintock and van Avermaet, 1982; Au and Kwong, 2004). Examples are cooperative orientations (maximizing the joint outcome) and competitive orientations (maximizing the advantageous difference between one's own and others' outcomes). Fairness orientations based on inequality aversion (minimizing differences between one's own and others' outcomes) have likewise received much attention (e.g., MacCrimmon and Messick, 1976; Weesie, 1994a; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). Social orientations are based on emotions that can induce people to reciprocate. For instance, inequality-averse people are driven to

reward others' (expected) cooperative behavior through their own cooperation if this reduces the difference between their own and others' outcomes. A similar reasoning applies to punishment of others' non-cooperation.

Now, consider a situation in which people decide about sharing gains with another person. If these gains are initially not people's own property, but are the result of the other person's preceding decision, the principle of reciprocity requires that people share the gains more generously. Outcome-based motivations cannot explain differences in behavior between decision situations that differ with respect to preceding decision, but in which objective outcomes are identical. Intention-based motivations, however, induce people to take into account the behavioral process of how certain outcomes are obtained and to evaluate others' kindness. Sociological and social-psychological research suggests that the principle of reciprocity is rooted in fundamental social-psychological forces that are based on obligation and indignation. Others' kindness induces feelings of obligation that urge people to return the favor (Gouldner, 1960; Coleman, 1990: ch. 12; Cialdini, 2001: ch. 2), whereas others' unkind behavior inflicts feelings of indignation that drive people to seek revenge (Gouldner, 1960).

In addition to intention-based motivations induced by others' preceding behavior, people's own preceding behavior gives rise to intra-personal process-based motivations, such as the desire for self-consistency. People are inclined to behave consistently in order to avoid cognitive dissonance (Festinger, 1957; Webster, 1975; Cialdini, 2001: ch. 3; Kunda, 2002). For instance, promises and threats are expressed intentions to perform a certain behavior. Self-consistency demands that people keep their word and creates intrinsic bonds, such that promises and threats serve as commitments (in the sense of a "strategic move", Schelling, 1960). People can make a promise to someone in order to induce the other person to provide a favor in advance. Reciprocal behavior can then also result from peoples' desire to behave consistently with their promise, which induces people to indeed return the favor, and not because of obligation feelings. Moreover, people share some responsibility for the other person's decisions because the other person might only have agreed to provide the favor in advance because of the promise. Due to such shared responsibility for the other's decision, the desire for self-consistency can also increase feelings of obligation to return the favor. In turn, received promises involve a prospect of a gain that likewise creates an obligation for repayment. In contrast, threats involve unkindness due to the prospect of a loss. Therefore, threats can inflict feelings of indignation and induce the threatened person to retaliate for the unkindness. Similarly, omitted promises

are likewise unkind which gives rise to indignation feelings and, thus, revenge. In turn, for those who explicitly omit to make a promise, mechanisms of cognitive dissonance reduction (Festinger, 1957) can legitimate behaving unkindly. Even if people receive a favor despite an explicitly omitted promise, the desire for self-consistency competes with feelings of obligation and can undermine the influence of obligation feelings. Note that promises and threats can be "cheap-talk" such that they do not change objective outcomes. Nevertheless, reciprocity can result due to feelings of obligation or indignation and due to the desire for self-consistency. This contrasts with contemporary theoretical models that account for intention-based motivations in which perceived kindness is assumed to be determined by forgone outcomes of intentionally non-chosen options (e.g., Falk and Fischbacher, 2006). Empirical evidence for the promoting influence that cheap-talk promises exert on cooperative behavior is also found by numerous studies on communication (for reviews see, e.g., Sally, 1995; Shankar and Pavitt, 2002). To conclude, obligation, indignation, and self-consistency are powerful mechanisms that drive people to respond to the mere act of kindness or unkindness.

6.1.2 Four Studies: Basic Ideas, Approach, and Contributions

Theoretical Ideas and Contributions

The studies presented in this book investigated how process-based motivations affect people's behavior in social dilemmas. For this purpose, people's decision-making was observed in decision situations in which objective outcomes were identical, but that differed with respect to the behavioral context generated endogenously by preceding kind or unkind behavior. The focus was on interpersonal trust in decision situations between two persons in single encounters. For all four studies, the insights of sociological and social-psychological research on feelings of obligation or indignation and on self-consistency provided the theoretical foundation. These insights seem to be largely neglected in previous research on social dilemmas. The following list provides an overview of the four studies and summarizes the substantive contributions to previous research.

Study 1: Trust and Promises as Friendly Advances. Experimental Evidence on Reciprocated Kindness

Study 1 (Chapter 2) examined how trustfulness affects trustworthiness and how making and omitting promises to honor trust influences trustworthiness on trustfulness. This allowed for the study of positive reciprocity, which some other researchers claimed

to be largely non-existent (e.g., see the discussion by Falk et al., 2003). Moreover, influences of promises of trustworthiness that were combined with objective bonds were investigated. In other studies, the content of communication was less controlled and promises were cheap-talk (e.g., Brandts and Charness, 2003) or influences of only some specific outcome-based motivations were controlled (e.g., Snijders, 1996).

Study 2: Temptation, Loss, and Promises of Trustworthiness. Experimental Evidence on Context-Dependency of Outcome-Based Motivations

Study 2 (Chapter 3) focused on how behavioral contexts resulting from preceding behavior moderate effects of outcome-based motivations on trustfulness and on trustworthiness. For this purpose, a classical altruism model (Brew, 1973; Weesie, 1993, 1994b) was informally (not analytically) applied which consists of a selfish utility component and an individually weighted other-regarding utility component. This allowed influences of people's own outcomes to be separated from influences of the other person's outcomes. The idea of context-dependency of outcome-based motivations questioned the assumption made in theoretical models that outcome-based motivations (social orientations) are individually stable across decision situations. Some previous research has investigated social orientations in decision situations (a) that differ with respect to outcomes, (b) that are simultaneous, rarely sequential, and (c) that have different choice options (e.g., McClintock and Liebrand, 1988; Blanco et al., 2006). Such a setup does not allow for the study of how procedural motivations moderate outcome effects.

Study 3: Influences of Promises and Threats on Trust and Trustworthiness. Experimental Evidence on Reciprocated Behavioral Advances

Study 3 (Chapter 4) analyzed how trustfulness and trustworthiness are influenced by promises and threats. In this study, the basic approach taken in Study 1 (Chapter 2) was applied to trust situations with sanctioning options for trustors. First, this allowed the question to be investigated whether the main findings of Study 1 are also found when trustors have explicit sanctioning options. This was questionable because previous studies found that sanctions can have adverse effects and thereby undermine cooperative behavior (e.g., Gürerk et al., 2004; Voss and Vieth, 2006). In contrast, in repeated interactions, a particularly promoting influence on cooperative behavior was found when both communication and sanctioning were possible (e.g., Ostrom et al., 1992; Bochet and Putterman, 2007). Second, a further aim was to investigate influences of reward promises and punishment threats on trustworthiness. Previous

studies did not assess the mere influence of announcing sanctions on subsequent decision-making independent of other factors (e.g., Fehr and Rockenbach, 2003; Voss and Vieth, 2006; Fehr et al., 2007).

Study 4: Revenge and Gratitude in Trust Situations Involving Promises and Threats. Experimental Evidence on Reciprocity by Intention-Based Sanctioning

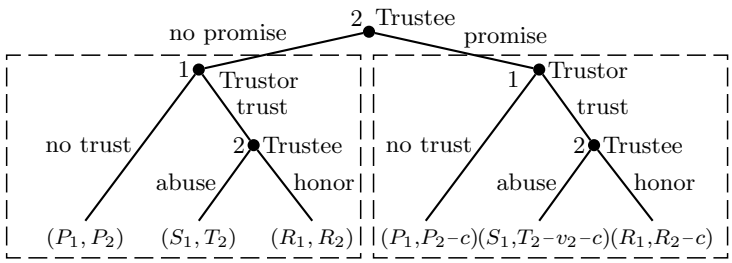
Study 4 (Chapter 5) shed light on how preceding behavior affects revengefulness and gratefulness. Thus, the focus was shifted from trustworthiness and trustfulness to sanctioning behavior. In previous research on sanctioning only influences of some specific outcome-based motivations were controlled. This also holds for studies in which the costs of sanctioning others were equal to the effect of sanctions (e.g., Falk et al., 2005; Vyrastekova and van Soest, 2008; Sefton et al., 2007). These studies overlook that inequality aversion is not the only outcome-based motivation that might be activated. Moreover, other studies on punishment of lies (Brandts and Charness, 2003), reward of promise-keeping (e.g., Fehr et al., 2007), and influences of sanctioning announcements on actual sanctioning decisions (e.g., Fehr and Rockenbach, 2003; Voss and Vieth, 2006; Fehr et al., 2007) include confounding factors in their experimental design or lack control for influences of outcome-based motivations.

Methodological Approach and Contributions

Two lab experiments were designed in which the behavior of a person in different behavioral contexts was observed (see Table 6.1). The decision situations were structurally identical, i.e., they consisted of the same choice options and of the same objective outcomes for both the trustor and the trustee (for similar designs, see Snijders, 1996; McCabe et al., 2003; Cox, 2004). The only difference was the behavioral context, i.e., the preceding behavior that generated the specific decision situation. For instance, consider the decision situation in which the trustee chooses whether or not to promise his trustworthiness (Figure 6.1) (see also Raub, 1992; Weesie and Raub, 1996; Raub, 2004). The behavior in each of the two subsequent trust situations resulting from the trustee's decision of whether or not to make the promise was compared to the behavior in the trust situation with identical outcomes in which the trustee did not have an opportunity to promise his trustworthiness.

In the second experiment, the trustor had an option to reward honored trust or to punish for abused trust. In some of these trust situations, the trustor also had an option to accompany placed trust with a promise of reward or with a threat of punishment. Again, behavior in situations resulting from the trustor's decision to

Figure 6.1: Trust Game with promises of trustworthiness



Trust Game with initial payoffs Trust Game with modified payoffs

$R_1 > P_1 > S_1; \quad T_2 > R_2 > P_2; \quad v_2 \geq 0; \quad c \geq 0$

announce sanctions was compared to behavior in the trust situation without this announcement option. Similarly, the influence of placing trust on trustworthiness was investigated by comparing the trustee's decision to share gains that resulted from trustfulness with the trustee's decision in the position of a dictator about sharing gains that are obtained exogenously (i.e., that are their own property). This procedure was also applied to the trustor's sanctioning decisions. Punishment is an investment in order to reduce the other's outcome by a fine, and reward is an investment in order to increase the other's outcomes by gratification. Each of these decisions of whether or not to invest in changing the other's outcome in situations after the other has chosen between keeping and sharing gains was compared to the two respective situations in which the other had no preceding choice options. This allowed the influence of the other's mere act of sharing or keeping gains to be assessed, i.e., whether people punish or reward due to feelings of obligation or indignation without influences of outcome-based motivations.

In order to study influences of the behavioral contexts on decision-making, various outcome-based motivations and individual heterogeneity were controlled in the statistical analyses. This was achieved by grouping the data into "subject-payoff response sets" consisting of the decisions that a subject made in different behavioral contexts with identical objective outcomes. In Study 2, subject response sets were created in order to investigate the interaction of outcome-based motivations (represented as benevolence and spitefulness by informally applying a well-established altruism model) with the behavioral context (Chapter 3). Data of the first experiment were analyzed using logistic regression models with fixed effects for response sets (Chapters 2 and 3). Due to the actual decisions that participants made in the

Table 6.1: Overview of main differences between the two experiments

	Experiment 1	Experiment 2
	Chapters 2 and 3	Chapters 4 and 5
Time and location	Nov. 2006 at ELSE lab, ICS/Sociology at Utrecht University	Apr. 2008 at CeDEx lab, School of Economics at Nottingham University
Sanctioning options for trustors (punish, reward)	not included	in all situations (costly, mostly ineffective)
Promise of trustworthiness by trustees	in some situations	in some situations
Sanctioning announcement by trustors (punishment threat, reward promise)	not included	in some situations
Announcement properties	either with binding value and/or transaction costs, or cheap-talk	always cheap-talk
Codebook	Vieth and Weesie (2006)	Vieth (2008)

Effective sanctions remove the trustee's temptation to abuse trust in objective terms.

second experiment, logistic regression models with random effects for response sets were employed to analyze the data of the second experiment (Chapters 4 and 5).

All four studies involved cheap-talk announcements (see also studies on communication, for reviews see, e.g., Sally, 1995; Shankar and Pavitt, 2002). This allowed the question to be investigated whether perceived kindness is determined by forgone objective outcomes of non-chosen options, as assumed in theoretical models (e.g., Falk and Fischbacher, 2006), or by process-based motivations (obligation, indignation, and self-consistency). Moreover, the four studies in this book contributed in methodological respects to previous research (for more detailed discussions, see Chapter 2).

First, designing sets of *structurally identical (sub)games* allows the "pure" influences of behavioral contexts to be analyzed while controlling for various outcome-based motivations without making any assumptions about such motivations. If reasonable assumptions could be made, modeling these assumptions in statistical analyses would allow for more efficient tests. However, given the current state of research, outcome-based motivations cannot adequately be modeled and measured (for more details, see Chapter 3; Aksoy and Weesie, 2008). Previous experiments only control

for some specific representations of outcome-based motivations (e.g., linear influences of inequality aversion).

Second, a *within-subject design* was employed which allows for the analysis of intra-personal differences in decision-making between behavioral contexts. With few exceptions, previous studies used a between-subjects design, which is less suitable for studying individual motivations. Within-subject designs have advantages but also disadvantages (Keren, 1993; Putt, 2005). As discussed in Chapter 2, one major disadvantage is that practice effects and carryover effects are involved. However, a within-subject design appears to be more suitable for the type of studies presented in this book because influences of motivations can be studied on the individual level, whereas between-subjects designs only allow for the comparison of the average behavior at an aggregate level (on the ecological fallacy, see Robinson, 1950). Moreover, by employing a within-subject design it is possible to control for (additive) individual heterogeneity and for influences of various objective outcomes without making assumptions about specific outcome-based motivations.

Third, the *behavioral contexts were generated endogenously* by kind and unkind behavior of participants. In many previous experiments the "strategy method" (Selten, 1967) is used, especially in the rare studies using a within-subject design. Concerning the strategy method, one main problem is that decisions remain hypothetical. This undermines influences of emotions, which are the underlying forces of other-regarding motivations. In addition, the strategy method implies simultaneous decision-making (see also McCabe et al., 2003), which undermines studying influences of process-based motivations. Moreover, biases due to artificial consistency in responses are more likely if the strategy method is employed because participants then seek to answer in a consistent manner.

Fourth, *binary-choice situations* assure that decisions are relatively unambiguous for both participants and researchers, concerning the interpretation of decisions in terms of kindness and unkindness. Moreover, the number of possible (sub)games is conveniently small. An exception to the binary-choice decisions was the trustor's announcement decision, which involved three options (punishment threat, reward promise, no announcement). Concerning perceived kindness, the three options are clearly interpretable.

6.1.3 Summary of Results

The results of the four studies provide evidence for reciprocal behavior and support for the idea that reciprocity is based on obligation feelings, indignation feelings, and

the desire for self-consistency. People's decision-making and underlying motivations are found to be influenced by the behavioral context, even without any changes in objective outcomes.

> **Result 1:** Trustfulness begets trustworthiness.

The mere act of placing trust increases trustworthiness (Chapters 2 and 3; also see McCabe et al., 2003; Cox, 2004). This influence has been found especially in decision situations in which sanctioning options are available (Chapter 4). Due to trustfulness, the hampering impact that the trustee's temptation exerts on trustworthiness is reduced (Chapter 3). These findings support the idea that trustees feel an obligation to return the favor of placed trust by behaving trustworthily.

> **Result 2:** Promising trustworthiness promotes trustfulness and actual trustworthiness.

The mere act of promising trustworthiness increases trustfulness and trustworthiness, even if promises are objectively cheap-talk (Chapters 2 and 3). This effect is particularly strong in decision situations in which sanctioning options are available (Chapter 4). The promoting impact of promising to honor trust on actual trustworthiness indicates the influence of obligation feelings or self-consistency. As argued, self-consistency can also foster the influence of obligation feelings due to shared responsibility for the trustor's decision to place trust. Transaction costs associated with making the promise of trustworthiness promote the increase in trustworthiness (Chapter 2). Due to making the promise of trustworthiness, the positive effect of the trustee's concern about the trustor's loss on trustworthiness is reduced (Chapter 3). The promoting impact of promised trustworthiness on trustfulness provides evidence for the influence of obligation feelings and of trustors' beliefs about the increased trustworthiness. No support has been found for influences of promise properties on trustfulness when controlling for various outcome-based motivations (Chapter 2). The hypothesis that receiving a promise of trustworthiness would change the hampering influences that loss and temptation exert on trustfulness has likewise not been supported (Chapter 3).

> **Result 3:** Lies trigger a thirst for revenge, but promise-keeping receives less reward compared to honored trust without preceding option to promise trustworthiness.

Reneged promises of trustworthiness increase revengefulness (Chapter 5; also see Brandts and Charness, 2003; and for repeated public good situations, Bochet and Putterman, 2007). This supports the idea that a feeling of indignation drives people to retaliate for suffered losses. If trust is honored after trustworthiness has been promised, gratefulness tends to be reduced (Chapter 5). This indicates that returned favors induce weaker obligation feelings than original favors and that trustors might expect the trustee to share the responsibility for requested trustfulness. Note that no support has been found for the idea that gratefulness about gains shared in return for trustfulness is lower than gratefulness about gains shared as an original favor (Chapter 5). Thus, feelings of obligation to return the favor of honored trust are particularly reduced after trustworthiness has been promised.

> **Result 4:** Omitting a promise of trustworthiness is retaliated against by withheld trust, while the impact on trustworthiness depends on the properties of the omitted promise.

Omitting a possible promise of trustworthiness hampers trustfulness, even if promises are objectively cheap-talk (Chapters 2 and 3; also see Snijders, 1996; Gautschi, 2000). This especially holds in decision situations in which sanctioning options are available (Chapter 4). The finding supports the idea that indignation feelings drive trustors to withhold trust in order to punish trustees for omitted promises. No support has been found for the idea that promise properties would moderate the influence of the omitted promise on trustfulness (Chapter 2). However, trustworthiness increases with transaction costs that would have been associated with making the promise, and trustworthiness decreases with the binding value of the omitted promise (Chapter 2). This indicates two implications. First, trustees feel an obligation to reward the trustor's trustfulness, which is stronger the more trustfulness is a sign of understanding that the promise has been omitted because of high transaction costs. Second, self-consistency undermines feelings of obligation by legitimating to abuse trust if trust has been placed despite a high binding value of the omitted promise. It has also been found that the hampering impacts of the trustee's temptation on trustworthiness and of the trustor's loss on trustfulness are increased after the promise of trustworthiness has been omitted (Chapter 3). This increase in the hampering effect of the trustee's temptation supports the idea that feelings of obligation are undermined after omitted promises. Note that no support has been found for the reasoning that the influence of the trustor's loss on trustworthiness would be more promoting after the trustee has omitted the promise to honor trust due to obligation feelings induced by placed trust despite the omitted promise (Chapter 3). Moreover, no support has been found

for generally decreased trustworthiness after the promise has been omitted (Chapter 2; also see Snijders, 1996) or in the case in which the omitted promise has been cheap-talk (Chapters 2 and 3).

> **Result 5:** Reward promises increase trustworthiness, revengefulness, and gratefulness.

The mere act of promising a reward for honored trust promotes trustworthiness, despite the promise is cheap-talk (Chapter 4). This provides evidence for the influence of increased obligation feelings that arise from the combination of two favors (i.e., placed trust and a reward promise) and for the influence of anticipated sanctioning behavior. Abused trust after a reward has been promised is punished particularly strongly (Chapter 5). This supports the idea that indignation feelings trigger a thirst for revenge. Moreover, promising a reward increases gratefulness (Chapter 5). Recall that the trustor's gratitude is lower in case the trustor decides whether or not to reward the trustee for having kept his promise of trustworthiness (Result 3). It has been mentioned that this can be understood by considering that obligation feelings for rewarding returned favors are weaker than for rewarding original favors. Given this, the positive effect of promising a reward on actually performing the reward indicates the strong influence of self-consistency.

> **Result 6:** Punishment threats seem to promote revengefulness.

If trustors can combine placing trust with announcing sanctions, they mostly (87%) promise a reward (Chapters 4 and 5). Therefore, hardly any observations were available for analyzing influences of punishment threats. It can thus only be mentioned with caution that the mere act of threatening with punishment has been found to increase revengefulness (Chapter 5; also see Voss and Vieth, 2006). This indicates some support for the idea that self-consistency drives trustor's to actually perform the threatened punishment, whereby self-consistency can also increase the trustor's feeling of indignation. No support has been found for the idea that threatening with punishment would be retaliated by abused trust (Chapter 4). In fact, the indications suggest a positive influence (see also Voss and Vieth, 2006). Further research is required on the impact of threats on both cooperative behavior and sanctioning behavior. Thereby, the moderating influences of perceived unfairness of threats should also be investigated (for this argument, also see, e.g., Fehr and Rockenbach, 2003).

> **Result 7:** Punishing behavior and rewarding behavior are only marginally motivated by purely distributional concerns about objective outcomes.

The mere act of keeping gains (while sharing would have been possible) increases revengefulness (Chapter 5). Similarly, the mere act of sharing gains (while keeping gains would have been possible) increases gratefulness, even if sharing gains is a returned favor for placed trust (Chapter 5). Only few people voluntarily incur costs for reducing or increasing the other's outcome in decision situations without preceding kind or unkind decision of the other person. This supports the idea that punishing behavior is motivated by indignation feelings that drive people to take revenge and that feelings of obligation to return favors motivate rewarding behavior.

> **Result 8:** Outcome-based motivations interact with behavioral contexts, and the influences depend on the person's decision position.

Study 2 also provides evidence that outcome-based motivations interact with behavioral contexts, i.e., they are not individually stable across decision situations with identical outcomes. Moreover, it has been found that influences of outcome-based motivations depend on the decision position a person takes in a decision situation (i.e., influences are role-dependent). Altruistic inclinations in the role of the trustee, who feels benevolence concerning the trustor's outcome, seem to turn into aggressive tendencies in the role of the trustor, who is rather spiteful about the trustee's gains. This interpretation receives support by the indication that hampering impacts of the other's selfish motivations seem to be mirrored such that the behavioral context affects the influences of people's own selfish motivations rather than the influence of people's other-regarding motivations. This also implies that the influence of beliefs is not mediated by the same outcome components.

The results of the four studies presented in this book provide strong evidence that behavioral contexts resulting from preceding behavior influence trustfulness and trustworthiness, as well as revengefulness and gratefulness. Due to self-consistency, obligation feelings, or both, people are driven to keep their promises and to perform a threat. Obligation feelings motivate people to return the favor of placed trust and the favor of received promises. However, people who return favors cannot expect to be particularly rewarded because obligation feelings are stronger for original favors. Explicitly omitted promises of trustworthiness trigger indignation feelings that drive trustors to withhold trust. For the person who omitted the promise, self-consistency conflicts with obligation feelings. Whether self-consistency has a stronger influence than the feeling of obligation depends on the promise properties. Next, irrespective of influences of objective outcomes, people are strongly motivated by indignation feelings

to punish others for unkindness and by obligation feelings to reward others' kindness. In fact, the influence of outcome-based motivations, such as spite and benevolence, tends to differ between decision roles and to depend on preceding behavioral processes. The findings demonstrate the power of obligation feelings, indignation feelings, and the desire for self-consistency. Based on these motivations, behavioral patterns of reciprocity emerge that reveal the strong influence of internalized social norms on people's behavior and thereby constitute a basis for maintaining social order.

6.2 Discussion and Perspectives

6.2.1 Summary of Selected Main Discussion Points

Since a study cannot be perfect and insights typically generate new questions, further research is desirable that improves on identified drawbacks and contributes further insights. Various aspects have been discussed in one or more reports of the four studies presented in this book. In particular, the following main aspects have been addressed that concern all four studies: (1) sequence effects, (2) eliciting beliefs and emotions, (3) self-consistency and feelings of obligation or indignation, (4) statistical analyses, and (5) theoretical models.

First, whenever participants make several decisions, practice effects and carryover effects can bias subsequent decisions. This is an inherent feature of within-subject designs, although such a design is particularly usful for the type of studies collected in this book. Participants' experiences in previous encounters can alter the participants' mood or beliefs and thus influence decision-making in the actual encounter (see also indirect reciprocity, e.g., Nowak, 2006). For instance, positive experiences typically bring participants into a positive mood, which might increase their generosity and their belief that the interaction partner in the current encounter is a kind person. Thus, in further analyses, perceived kindness of previous encounters could be assessed, and influences on actual decision-making could be studied. Moreover, influences of such experiences could also be controlled in the analyses presented in this book.

Next, fixing the ordering of the types of decision situations (games) allowed for the optimization of the sequence. Specifically, differences in parameters and visualization between decision situations were maximized in order to hide the underlying sequence of identical (sub)games. Similarly, parameters, interaction partners, and decision roles were varied from one decision situation to the next. Moreover, subject-payoff response sets consist of decisions made in different decision periods. Nevertheless, the fixed ordering of games could have caused biases. For instance, participants could have become increasingly selfish in the course of the experiment. In this case, the reduced

generosity of trustees in the role of dictators at the end of the experiment would not necessarily be due to the absence of placed trust as a favor that induces obligation feelings. In order to investigate sequence effects, further experiments are required in which the ordering of games is varied between experimental groups (sessions).

Second, further experiments should include measures that elicit beliefs, emotions, and perceived kindness. This would allow the question to be investigated whether the actual feelings toward the other person and perceptions of the other's kindness support the proposed interpretations. Moreover, it would help disentangle opposing effects that could have inhibited finding support in some behavioral contexts. However, standard measures used for eliciting beliefs are found to affect participants' decision-making towards other-regarding behavior (Gächter and Renner, 2006; Hoffman et al., 2008) or increase selfish behavior (Croson, 2000). Similarly, influences on participants' decision-making can also be expected from asking questions about their feelings toward the other person. Therefore, methodological research is required that explores the conditions under which certain measures have certain influences on decision-making. This might also allow for the development of new measures and procedures in order to remove or to control for certain biases. Considering the context-effects on outcome-based motivations reported in Chapter 3, such influences of measures might also differ between decision situations.

Third, the arguments for the hypotheses and the interpretation of results show that it is difficult to disentangle influences of self-consistency from influences of obligation or indignation. For instance, consider the trustee's decision of whether or not to honor trust after the trustee has omitted the promise of trustworthiness but the trustor nevertheless has placed trust. The desire for self-consistency then competes with the feeling of obligation to return the favor, and both self-consistency and obligation feelings appear to cancel each other out. The analyses in Chapter 2 provide evidence that the influences of self-consistency and obligation can be separated by accounting for the properties of the promise. However, in other behavioral contexts, the direction of the influences of self-consistency and of obligation is the same and disentangling the effects of the two motivations is not possible in the studies reported here. For instance, consider the trustee's decision to keep his promise of trustworthiness. It has been argued that the trustee then shares some responsibility for the trustor's decision to place trust. Thus, the trustee's decision to keep his promise can be due to self-consistency (while obligation feelings are undermined), obligation feelings (while self-consistency might play no role), or a combination of these two motivations (self-consistency might boost the impact of obligation feelings) (Chapter 2).

Further experiments are required to investigate the driving motivations behind such decisions. For instance, using a (sequential) Prisoner's Dilemma or a Chicken Game and unconditional cooperation promises or non-cooperation threats would yield further decision situations in which the influence of self-consistency conflicts with feelings of obligation or indignation.

Fourth, statistical analyses could be extended in several respects, partly requiring further experiments. As mentioned above, the influence of positive and negative experiences in previous encounters on decision-making in the actual encounter could be analyzed (see discussion point 1). Next, statistical models could account for the dependency of observations within subjects and within sessions. In particular, accounting for group dynamics within sessions might yield valuable insights into the formation and updating of beliefs between and within the interactions with strangers. Moreover, the statistical models employed in the four studies reported in this book made strong homogeneity assumptions: Effects of behavioral contexts and of outcomes were assumed to be the same for all individuals. Differences in influences of behavioral contexts on decision-making between response sets should be explored by allowing coefficients to vary randomly. Similarly, the moderating influences of outcomes and of outcome asymmetry could be studied (see discussion point 3 in the next section). Models with random coefficients could not be fitted due to restricted sample size and thus require further or new experiments to be conducted. Furthermore, the experiments include questionnaires on participants' personal characteristics. Using the information from questionnaire items, would thus allow for the exploration of individual heterogeneity.

Fifth, in the arguments for the hypotheses, influences of obligation or indignation, self-consistency, and beliefs have been discussed. At times, these influences appear to oppose one another. For instance, concerning trustworthiness after trust has been placed despite an omitted promise to honor trust, self-consistency conflicts with obligation feelings. Therefore, theoretical models should be developed and employed to derive hypotheses. For instance, suitable models would allow effects for certain parameter spaces to be assessed. Moreover, perceived kindness could be quantified. For instance, the influence of cheap-talk announcements could be incorporated by taking expected outcomes from indicated decision paths (not actually realized ones) as a basis in order to determine the kindness of an action. This "baseline kindness" can then be moderated by influences of forgone outcomes (e.g., as proposed by Falk and Fischbacher, 2006). Formalization could help derive hypotheses, e.g., by employing a random utility approach (McFadden, 1973) and by calculating quantal response equi-

libriums (McKelvey and Palfrey, 1998). Using a series of computer simulations, results could be obtained for various values of individual utility weights that shape the influence of various social orientations, intention-based motivation, and self-consistency.

6.2.2 Selected Examples of Further Research Perspectives

In addition to the summarized aspects discussed in more detail in the respective studies, some further perspectives deserve to be addressed for future research: (1) requests and coercion (2) "real-life" decision situations, (3) asymmetries between actors, and (4) social networks.

First, in the studies reported in this book, promises and rewards were voluntary favors to the other person. Received favors induce feelings of obligation that demand repayment, whereas self-consistency drives people to keep their promises. What would happen if the other person would have requested a promise or a reward? Trustors could request promises of trustworthiness (Bruins and Weesie, 1996) and also directly request trustworthiness. Similarly, trustees could request a reward and reward promises. Considering the weakened obligation feeling to repay returned favors, requested favors might likewise have a less promoting or even a hampering influence on subsequent decisions. Moreover, requests and demands can involve an element of coercion, i.e., reducing the other's freedom of choice. For instance, if the trustor requested a promise that inflicted high transaction costs on the trustee, the trustee might seek revenge rather than be induced by self-consistency to behave in accordance with the promise he has made. Following the theoretical and methodological approach employed in the studies reported in this book, the influence of requests and omitted requests could be compared to the decision situation in which the request option is not available. Note that this can be extended such that people select a decision environment with properties of announcement or sanctioning options, or even with or without certain options at all. For instance, if the trustee could make such decisions and would deprive the trustor of the punishment option, the trustor might perceive this as an unkind intervention and withhold trust as a revenge. Moreover, the trustor's willingness to reward honored trust might also be reduced. The influence of beliefs could be reduced by comparing the situation created by the trustee's decision with a situation in which the punishment option is randomly removed (e.g., Houser et al., 2008).

Second, abstract decision situations were presented to participants in the two experiments conducted for the studies reported in this book. Rather than abstract decision situations, a cover story could be used. Further research could then also

explore whether the influences of process-based motivations differ between situational contexts (cover stories). For instance, previous studies show that framing a decision situation as a "community project" increased cooperative behavior compared to framing the same decision situation as a "market situation" (Rege and Telle, 2004; see the discussion in Chapter 3). It is possible that such priming of cooperative or competitive settings also moderates the influences of process-based motivations. In addition to "fixed cover stories", certain elements in cover stories could also be varied. Such "vignette lab experiments" would combine the advantages of the two approaches: non-hypothetical decision situations of lab experiments (i.e., real interaction partners and real incentives due to payment) and "real-life" associations of factorial surveys. Experiments could be designed as within-subject sequences of games with identical outcomes but different cover stories. One difficulty might be in assigning objective payoffs to the outcomes in the described "real-life" decision situations. This is straightforward as long as outcomes are quantifiable (e.g., money, time, grades). However, in many social interactions, outcomes can often only be rank ordered. Therefore, research using "vignette lab experiments" would also have to explore how people value certain social outcomes.

Third, the influences of outcome-based motivations and of intention-based motivations might be shaped by people's own and others' positions in a decision situation, e.g., positions in terms of status, power, and dependency. As previously mentioned, influences of outcome-based motivations and process-based motivations might differ between asymmetry structures (see discussion point 3 in the previous section). For instance, asymmetric social dilemmas involve advantaged and disadvantaged positions in terms of objective outcomes. Unkindness of people in advantaged positions might induce more indignation, thus increasing revengefulness, than unkindness of people in disadvantaged positions. The opposite might be the case if the disadvantaged person is in some ways dependent on the advantaged person. Similarly, kindness of people in disadvantaged positions might be particularly rewarded. Asymmetric outcome structures are involved in the two experiments presented here, and influences of such asymmetries can be addressed in further analyses as far as the data allow.

Fourth, influences of process-based motivations can also be studied in social networks. For example, according to Coleman (1990: ch. 12) created obligations serve as "credit slips". People help community members who are in need and receive help in return when they are in need. Groups or societies in which people engage in creating dependencies by exchanging obligations should therefore be better off than more individualistic groups or societies. However, Coleman (1990) also addresses examples

in which obligations can be created inflationary and unwanted or repaid when it costs little, even if the other person is not in need. Lab experiments would be fruitful in order to study the dynamics of obligations in social networks, the conditions for certain outcomes that arise, and the moderating influences of changes in the network structure.

Studying process-based motivations in social networks would also allow for the study of structural asymmetries that arise from people's position in a social network (see also the discussion point 3 in this section). People in less powerful positions (e.g., in terms of centrality or in combination with outcome asymmetries) might behave in a more friendly manner due to the dependency structure. In an experiment, a person's decision-making could be compared in various behavioral contexts as well as in various social network structures and network positions. An extension to dynamic networks would allow for the study of partner-selection and group formation. In addition to the addressed moderating influences of network structures and network positions, decisions directly affecting the network can the interpreted in terms of providing favors (initiating or maintaining a relationship) and of retaliation (terminating a relationship).

Moreover, mere group membership can shape the influence of people's motivations. Social-psychological research suggests that people treat in-group members in a friendlier manner than out-group members (e.g., for minimal-group experiments, see Tajfel et al., 1971). Given the same preceding behavior, obligation feelings might thus be stronger toward in-group members than toward out-group members. However, if an in-group member behaved in an unkind manner, it might inflict stronger indignation feelings and thus increase revengefulness than unkindness by an out-group member.

Extending this line of research to these fields would provide valuable insights into how feelings of obligation, feelings of indignation, and the desire for self-consistency shape people's behavior under different conditions. Commitments and reciprocity are fundamental elements in human interactions. The four studies presented in this book have investigated some basic principles of the interplay of commitments and reciprocity and, thereby, may inspire further research.

Appendix A

Decision Screens in the Experiments

A.1 Example of a Decision Screen in Experiment 1

This example is a decision screen of the HTG in Experiment 1 with $P_1 = P_2 = 30$, $R_1 = R_2 = 60$, $S_1^{\text{low}} = 0$, $T_2^{\text{high}} = 100$, $v_2^{\text{low}} = 10$, $c^{\text{low}} = 5$. It shows the decision situation of the trustee (role B) in which he chooses whether or not to honor trust after he has sent the message "I will choose up" and the trustor (role A) placed trust. The trustee can choose either "up" ("omhoog") representing his decision to honor trust or "down" ("omlaag") representing his decision to abuse trust. The last two columns of the table show the outcomes for the trustee after he "did send" ("wel gestuurd") the message and after he "did not send" ("niet gestuurd") the message. The participant's own choice options, outcomes, and labels were displayed in red color and those of the other person in blue color. Parts of the table that were no longer accessible due to the previous decisions made were changed into grey color. The codebook provides details on experimental design, instructions and screen setup (Vieth and Weesie, 2006). The experiment was programmed using the software package "z-Tree" version 2 (Fischbacher, 2007). The data were collected in November 2006 at ELSE lab of the Sociology Department (ICS-Utrecht) at Utrecht University.

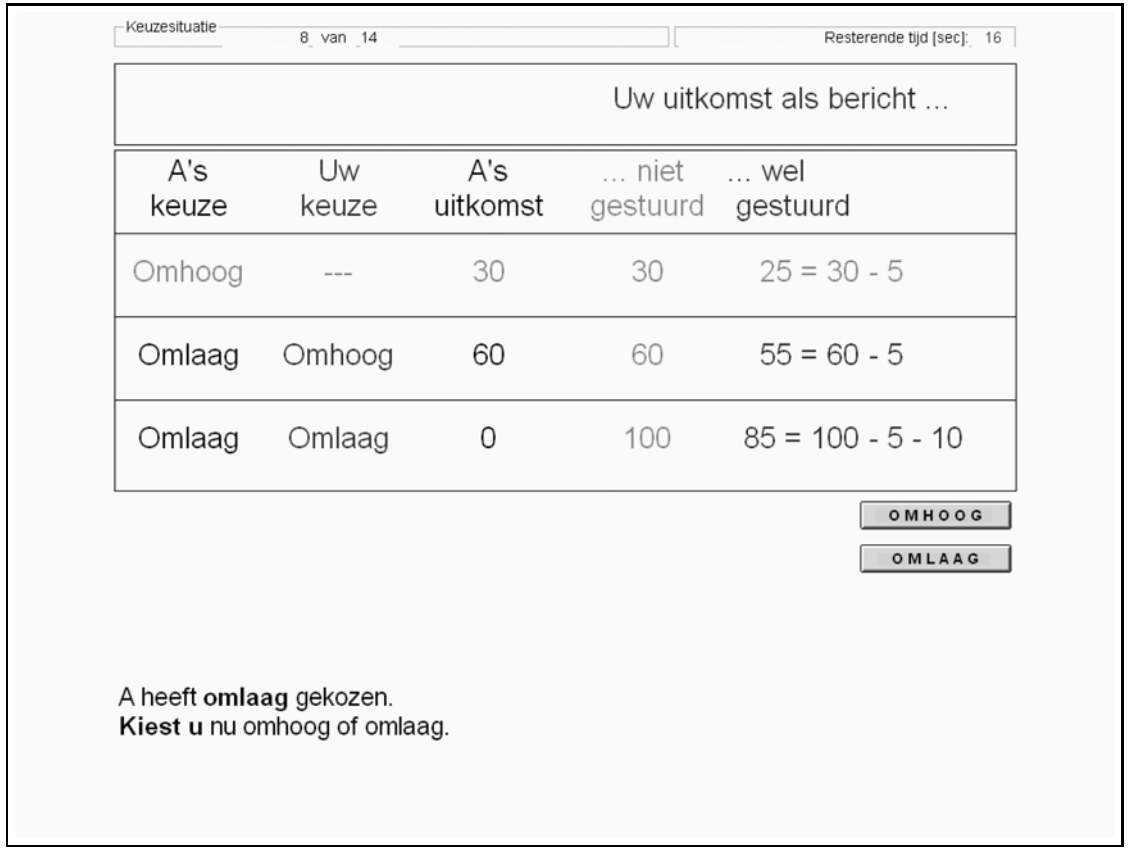


Figure A.1: Example of a decision screen in Experiment 1 for the trustee's choice of whether or not to honor trust after having promised trustworthiness in the HTG

A.2 Example of a Decision Screen in Experiment 2

This example is a decision screen of the H_1TGS in Experiment 2 with $P_1 = P_2 = 50$, $R_1 = R_2 = 80$, $S_1^{\text{low}} = 20$, $T_2^{\text{low}} = 100$, $o_1^{\text{high}} = 10$, $f_2^{\text{high}} = 15$, $g_2^{\text{low}} = 5$. It shows the decision situation of the trustee (role B) in which he chooses whether or not to honor trust after the trustor (role A) has placed trust combined with the message "If you choose up, I will choose down" (reward promise). The trustee can choose either "up" representing his decision to honor trust or "down" representing his decision to abuse trust. The active decision position (i.e., node in the game tree) was market with a frame in yellow color. The participant's own choice options, outcomes, and labels were displayed in red color and those of the other person in blue color. Parts of the table that were not reachable anymore due to the previous decisions made were changed into grey color. Moreover, the participant's own outcomes were displayed in the first column and those of the other person in the second column. The codebook provides details on experimental design, instructions and screen setup (Vieth, 2008). The experiment was programmed using the software package "z-Tree" version 3 (Fischbacher, 2007). The data were collected in April 2008 at CeDEx lab of the Nottingham School of Economics at Nottingham University.

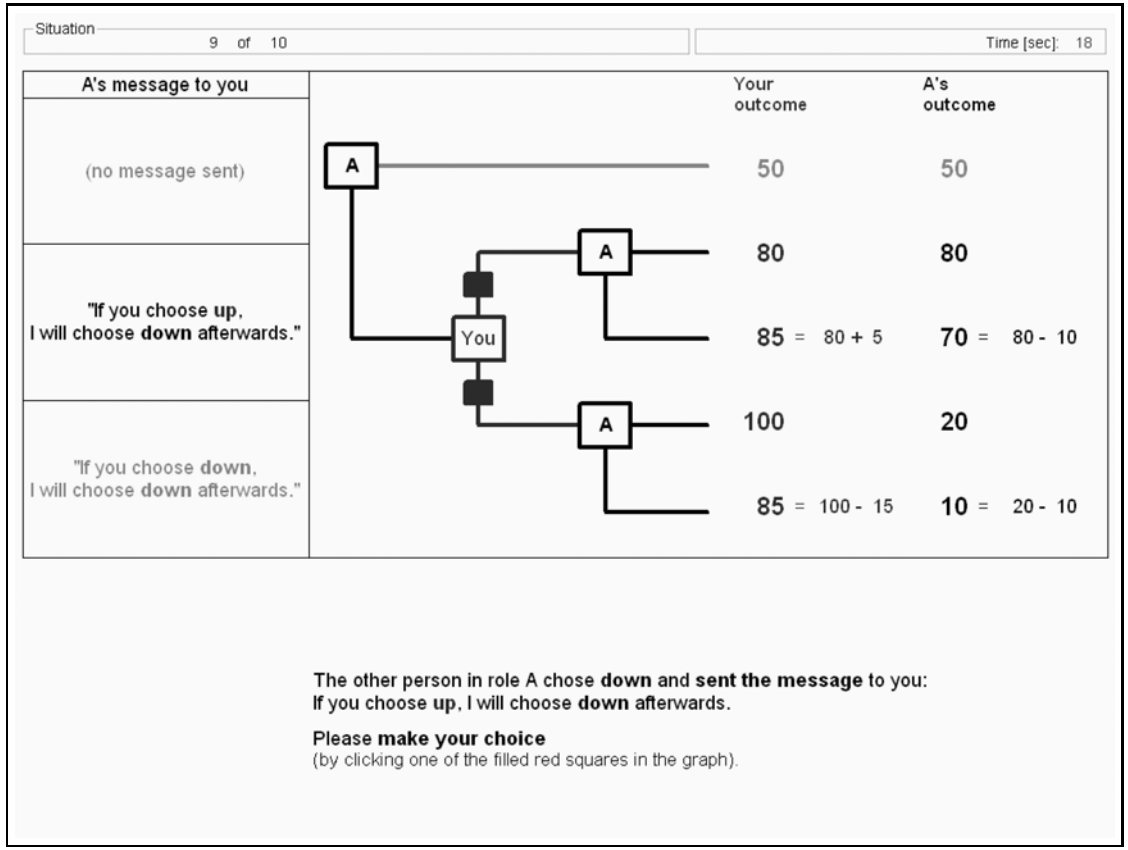


Figure A.2: Example of a decision screen in Experiment 2 for the trustee's choice of whether or not to honor trust after having received a reward promise in the H_1TGS

Appendix B

Glossary

The collection of studies presented in this book combines approaches and ideas from economics, psychology, and sociology. Some key terms are not used consistently within disciplines and are given different meanings across disciplines. This glossary summarizes the definitions used in this book.

Commitment

A commitment is a voluntary strategic action with the purpose of "reducing one's freedom of choice" or of changing the → outcomes by incurring or offering a "hostage" in the sense of a bond (based on Schelling, 1960; also see Williamson, 1985; Raub, 1992; Snijders, 1996; Weesie and Raub, 1996; Raub, 2004). Commitments involve intrinsic costs and bonds (see also → process-based motivations, → self-consistency) and can also be combined with objective incentives (binding values, compensating values, transaction costs). Binding values and compensating values typically modify the objective → outcomes in the case in which the committed actor chooses another option than the one the actor committed himself to choose. A commitment is "perfectly binding" if it removes the objective temptation to choose another option than the one an actor committed himself to choose. A commitment is "perfectly compensating" if it outweighs the objective loss that others incur if the committed actor chooses another option than the one the actor committed himself to choose. Transaction costs can be irreversible investments or can be returned depending on others' decisions. A commitment is "affordable" if the transaction costs do not outweigh the possible objective gain that the committed actor receives if he indeed chooses the option he committed himself to choose.

Context, behavioral

The behavioral context of a decision is generated by preceding behavior in a decision situation. If no decisions precede an actual decision, the behavioral context is "empty". The behavioral context comprises information about preceding decisions actually made, non-chosen options, forgone → outcomes, and outcome prospects. The behavioral processes generating a behavioral context give rise to intra-personal and inter-personal → process-based motivations (see also → self-consistency, → intention-based motivations, → perceived kindness, → obligation, → indignation).

Indignation, feeling of

A feeling of indignation is a felt need to retaliate for a loss (see also → perceived kindness, → intention-based motivations, → process-based motivations). Indignation feelings induce a thirst for revenge ("sentiments of retaliation", Gouldner, 1960: 172) (see also → punishing, → reciprocity, → social norms).

Intention-based motivations

Intention-based motivations are procedural motivations that are invoked by others' intentional behavior that is perceived as kind or unkind. Kind advances induce → obligation feelings, whereas others' unkind preceding behavior inflicts → indignation feelings (see also → other-regarding motivations, → perceived kindness, → process-based motivations).

Obligation, feeling of

A feeling of obligation is a felt need to return a favor (Gouldner, 1960; Coleman, 1990: ch. 12; Cialdini, 2001: ch. 2) (see also → perceived kindness, → intention-based motivations, → process-based motivations). Omitting or delaying to fulfill this obligation inflicts intrinsic distress and emotional tension ("shadow of indebtedness", Gouldner, 1960: 174) (see also → rewarding, → reciprocity, → social norms).

Other-regarding motivations

Other-regarding motivations are rooted in emotions and complement utility derived from selfishness by inducing people to take into account others' objective → outcomes or intentions. Two types of other-regarding motivations are typically distinguished: → outcome-based other-regarding motivations (see also → selfishness) and → intention-based motivations (see also → process-based motivations).

Outcome-based motivations

Outcome-based motivations, known as social (value) orientations (Messick and McClintock, 1968), are preferences concerning distributions of an actor's own and others' objective → outcomes. These distributive preferences shape the influence of objective outcomes on people's decision-making by transforming objective outcomes into subjective utilities. Outcome-based motivations consist of a selfish utility component (see also → selfishness) and one or more other-regarding components (see also → other-regarding motivations).

Outcomes, objective

Outcomes are the results of one's own and others' choices in a decision situation. The term "objective outcomes" refers to objectively measurable → payoffs, such as money or time, without subjective or intrinsic evaluations that are, e.g., based on emotions.

Payoffs, total

Payoffs are the numerical results in a game used for describing the structure of a decision situation (see also → outcomes). The term "total payoffs" refers to the sum of payoffs and modifying payoff components that depend on the decisions that can be made (e.g., properties of → commitments or properties of → sanctions).

Perceived kindness

Perceived kindness refers to the extent to which others' (preceding) behavior is evaluated as kind or unkind (see also → intention-based motivations). The evaluation is based on the direction and extent to which an actor's own expected outcomes and expected others' outcomes are shaped due to others' intentional decisions (see also Falk and Fischbacher, 2006).

Process-based motivations

Process-based motivations induce actors to respond to processes of how specific outcomes are obtained. These processes can be exogenous or endogenous. Exogenous processes determine the number of actors and their decision positions, options, and outcomes (e.g., random devices creating uncertain contingencies or specific situations, third-party influences etc.). Endogenous processes are created by people's intentional decisions. For instance, outcome inequalities can arise from technical rules of how resources are assigned or from people's decisions. Endogenous processes can result from people's own decisions or from decisions of interaction partners. People's own decisions activate intra-personal process-based motivations (see → self-consistency), whereas others' decisions activate inter-personal process-based motivations (see → intention-based motivations, → indignation, → obligation, → perceived kindness).

Promise

A promise is an expressed intention to perform a certain action that yields a gain to the other person (see also → commitment, → obligation, → rewarding, → self-consistency, → threat). The promised action can be conditional on others' behavior. Examples are cooperation promises and reward promises. Reward promises are always conditional on previous behavior, whereas cooperation can also be promised unconditionally (e.g., in simultaneous decision situations). Promises can be accompanied by objective incentives (see also → commitment). A promise is "cheap-talk" if the promise does not change objective → outcomes.

Punishing

Punishing (negative sanctioning) is an action that reduces the other's outcome by a fine after the other has acted in an unfriendly manner (see also → sanctions, → social norms, → indignation, → threat). The punishment is "effective" if it removes others' objective temptation to behave in an unfriendly manner. Costly punishment requires an investment (outlay) of one's own resources in order to punish others. Thus, costly punishment is not "credible" in objective terms. Note that non-cooperation can be interpreted as a punishment in some decision situations.

Reciprocity

Reciprocity is a behavioral pattern of returning favors and retaliating for unkind actions. Returning favors is positive reciprocity (see also → rewarding, → obligation, → promise) and retaliating unkindness is negative reciprocity (see also → punishing, → indignation, → threat). In single encounters, reciprocity can be an implication of → other-regarding motivations (see also → outcome-based motivations, → intention-based motivations) and can also arise from → self-consistency.

Rewarding

Rewarding (positive sanctioning) is an action that increases the other's outcome by a gratification after the other has acted in a friendly manner (see also → sanctions, → social norms, → obligation, → promise). The reward is "effective" if it removes others' objective temptation to behave in an unfriendly manner. Costly reward requires an investment (outlay) of one's own resources in order to reward others. Thus, costly reward is not "credible" in objective terms. Note that cooperation can be interpreted as a reward in some decision situations.

Sanctions, informal

Sanctions are behavioral options that allow pleasure or disapproval about others' preceding behavior to be expressed. Sanctioning actions often also change others' → outcomes depending on previous actions and thereby create objective incentives for good conduct (see also → social norms). Sanctioning can be negative by inflicting a fine on others for their unkindness (→ punishing) or positive by providing others with gratification for their kindness (→ rewarding). The term "informal sanctions" refers to sanctions that are not codified (laws or contracts) and not enforced by third parties, but are voluntary decisions in a given situation. Informal sanctions typically have an emotional basis (see also → other-regarding motivations, → self-consistency).

Self-consistency, desire for

The desire for self-consistency is an intra-personal → process-based motivation to behave consistently with one's beliefs, attitudes, and previous choices (Cialdini, 2001: ch. 3; Gass and Seiter, 2007: ch. 3; Kunda, 2002) in order to avoid or to reduce cognitive dissonance (Festinger, 1957). The desire for self-consistency can moderate the influences of → indignation feelings and → obligation feelings. This is based on shared responsibility for others' subsequent decisions and on various mechanisms to reduce cognitive dissonance (Cialdini, 2001; Gass and Seiter, 2007).

Selfishness

Selfishness is an outcome-based motivation that induces people to exclusively care about their own objective → outcome (see also → other-regarding motivations, → outcome-based motivations).

Social dilemma

A social dilemma (also known as a "social trap", Platt, 1973; or "mixed motive game", Schelling, 1960) is a problematic decision situation of strategic interdependence that involves a conflict of interests between or within people that can result in sub-optimal outcomes. The conflict of interests arises from incentives for "opportunistic behavior" (Williamson, 1985), i.e., incentives to take advantage of a situation at the costs of others (see also → selfishness, → other-regarding motivations). Social dilemmas involve cooperation problems or distribution problems (also labeled bargaining problems or negotiation problems) (Harsanyi, 1977). In cooperation problems, all people involved share a common interest to overcome individual incentives for opportunistic behavior in

order to improve the joint outcome. In distribution problems, some of the people involved would have to sacrifice resources in order to improve the outcome of other people involved (opposed interests). Solving or mitigating social dilemmas demands → social norms. Note that the definition of social dilemmas is based on objective → outcomes and on the assumption of → selfishness. Note further that pure coordination problems and pure conflict problems ("zero-sum games", i.e., the sum of objective outcomes per decision combination equals zero, or more generally "constant-sum games") are classified as "pure motive games" that do not involve a conflict of interests, whereas "mixed motive games" describe a mixture of coordination and conflict strategies (Schelling, 1960). In contrast to social dilemmas, coordination problems are solved or mitigated by conventions that are self-enforcing (e.g., Voss, 2001). However, for the definition given here, pure conflict problems are social dilemmas without sub-optimal outcomes and thus demand disjoint social norms regulating compromises of sharing gains and losses, just as other distribution problems.

Social norms

Social norms are behavioral regularities in recurrent interactions in a population of actors who expect that deviant behavior will be punished (Voss, 2001: 108) or that conformity will be rewarded. Punishment and reward can be based on extrinsic incentives (see also → punishing, → rewarding, → sanctions) and intrinsic incentives (see also → other-regarding motivations, → self-consistency). Internalized social norms (e.g., Coleman, 1990) are supported by intrinsic incentives on the basis of emotions (e.g., Frank, 1988). Social norms solve or mitigate → social dilemmas. In cooperation problems, "conjoint social norms" require all parties to improve joint outcomes (common interests), and in distribution problems, "disjoint social norms" require some parties to improve the outcomes of others (opposed interests) (Coleman, 1990).

Threat

A threat is an expressed intention to perform a certain action that inflicts a loss upon the other person (see also → commitment, → indignation, → punishing, → self-consistency, → promise). The threatened action can be conditional on others' behavior. Examples are threats of non-cooperation and punishment threats. Punishment threats are always conditional on previous behavior, whereas non-cooperation can also be threatened unconditionally (e.g., in simultaneous decision situations). Threats can be accompanied by objective

incentives (see also → commitment). A threat is "cheap-talk" if the threat does not change objective → outcomes.

Trust, inter-personal

Inter-personal trust is "initiating an exchange" with someone who has opportunities that involve "a possibility of exit, betrayal, defection" (based on Snijders, 1996; Gambetta, 1988a) (see also → social dilemmas, → social norms). Separate definitions: "Trust is initiating an exchange before you know how the other person will reciprocate" (Snijders, 1996: 10). "For trust to be relevant, there must be a possibility of exit, betrayal, defection" (Gambetta, 1988a: 218–219).

Samenvatting in het Nederlands

Vincent Buskens and Nynke van Miltenburg translated the English summary into Dutch language. I am grateful for all the effort they put into it. In addition, I also thank Mariëlle Bedaux-de Jonge, Werner Raub, and Jeroen Weesie for their careful reading and editing of the Dutch translation.

Vincent Buskens en Nynke van Miltenburg hebben de Engelstalige samenvatting naar het Nederlands vertaald. Ik ben hun dankbaar voor de moeite die ze hierin hebben gestoken. Daarnaast dank ik Mariëlle Bedaux-de Jonge, Werner Raub en Jeroen Weesie voor het zorgvuldig lezen en corrigeren van de Nederlandse vertaling.

Een aantal Engelstalige begrippen is niet goed in het Nederlands te vertalen. Voor deze begrippen zullen de Engelse woorden gebruikt worden. Deze worden dan tussen aanhalingstekens gezet. Waar wenselijk voegen we een korte uitleg toe.

Korte Samenvatting

Het verklaren van sociale orde is een van de hoofdproblemen van sociologische theorieën en dit vraagt om de bestudering van hoe sancties kunnen helpen zodat sociale normen worden nageleefd. Het bestraffen van slecht gedrag en het belonen van goed gedrag zijn gedragpatronen die bekend staan als reciprociteit. Reciprociteit is geworteld in gevoelens die de basis vormen van geïnternaliseerde sociale normen. Mensen laten zich niet alleen leiden door hun eigen opbrengsten en de opbrengsten van anderen, maar hun motieven hangen ook af van eerder gedrag van henzelf en van anderen. Vriendelijk gedrag van anderen zorgt voor gevoelens van een verplichting om iets terug te doen. Onvriendelijk gedrag van anderen leidt tot gevoelens van verontwaardiging en een neiging tot vergelding. Bovendien zorgt consistentie ervoor dat mensen hun woord houden omdat beloften en bedreigingen op een intrinsieke manier dienen als 'commitments'. Dit boek bestaat uit vier studies over de invloed van deze op het proces gebaseerde motieven op vertrouwen, betrouwbaarheid en sanctioneren en ook over effecten van op uitkomsten gebaseerde motieven. De nadruk ligt op vertrouwenssituaties en soortgelijke interacties tussen mensen die elkaar niet kennen. In sommige keuzesituaties kan betrouwbaarheid beloofd worden en in andere kunnen beloningen beloofd worden of kan er met straffen gedreigd worden. Twee laboratoriumexperimenten zijn gedaan waarmee de zuivere invloeden van eerder gedrag bestudeerd kunnen worden zonder dat er aannames hoeven gemaakt te worden over de preferenties van de proefpersonen ten aanzien van de opbrengsten. De resultaten tonen aan dat eerder gedrag latere beslissingen beïnvloedt en bovendien een invloed heeft op de effecten van op uitkomsten gebaseerde motieven.

1. Theoretische achtergrond

In veel sociale en economische interacties bestaat er naast een wederzijdse afhankelijkheid tussen mensen ook een prikkel om 'opportunistisch gedrag' te vertonen (Williamson, 1985). Dat betekent dat mensen kunnen profiteren van de situatie ten koste van anderen. Dit veroorzaakt twee soort problemen, die bekend staan als sociale dilemma's, namelijk coöperatieproblemen, met mogelijk inefficiënte en suboptimale uitkomsten, en distributieproblemen met problematische ongelijkheid. Coöperatie- en distributieproblemen kunnen de sociale orde bedreigen. Deze problemen verminderen de cohesie tussen mensen en ze geven aanleiding tot conflicten zowel binnen als tussen samenlevingen en groepen mensen (zie ook Voss, 1982, 1985). Vandaar dat sociale normen over coöperatie en eerlijke verdelingen gewenst zijn. Ze helpen de problemen met een prikkel tot opportunistisch gedrag te verminderen en de sociale orde te bewaren. Voor de handhaving van sociale normen zijn er echter passende sancties vereist. Dat wil zeggen dat er straffen moeten zijn voor afwijkingen van de norm en beloningen voor het volgen van de norm. Sancties kunnen gebaseerd zijn op objectieve prikkels en, in het geval van geïnternaliseerde sociale normen, op intrinsieke motieven gebaseerd op emoties. Reciprociteit is een fundamenteel gedragspatroon in sociale interacties dat voortkomt uit sancties. Mensen belonen vriendelijk gedrag en vergelden onvriendelijk gedrag, zelfs als deze sancties tegen hun eigenbelang indruisen.

Reciprociteit kan veroorzaakt worden door sociale motieven, bijvoorbeeld door motieven die op uitkomsten gebaseerd zijn en door motieven die op intenties gebaseerd zijn (voor een overzicht zie Fehr en Schmidt, 2006). Op uitkomsten gebaseerde motieven komen voort uit het belang dat mensen hechten aan de verdeling van objectieve uitkomsten tussen zichzelf en anderen. Dit idee is gebaseerd op sociale vergelijkingen in de zin dat het eigen welzijn (nut) op een positieve of negatieve manier mede-afhankelijk is van uitkomsten voor anderen. Hierbij wordt het belang dat mensen hechten aan eigen objectieve uitkomsten (zelfzuchtige motieven) aangevuld met bepaalde sociale motieven. In sociaalpsychologisch onderzoek zijn verschillende typen en gradaties van hoe mensen belang hechten aan de verdeling van uitkomsten onderscheiden en empirisch vastgesteld. Deze staan bekend als sociale (waarde)oriëntaties (voor een overzicht zie McClintock en van Avermaet, 1982; Au en Kwong, 2004). Voorbeelden zijn coöperatieve oriëntaties (het maximaliseren van de gezamenlijke uitkomst) en competitieve oriëntaties (het maximaliseren van het eigen voordeel ten opzichte van anderen). Oriëntaties gebaseerd op een weerzin tegen ongelijkheid (het minimaliseren van de verschillen tussen de eigen uitkomsten en uitkomsten voor anderen) hebben evenzo veel aandacht gekregen (bijvoorbeeld MacCrimmon en Messick, 1976;

Weesie, 1994b; Fehr en Schmidt, 1999; Bolton en Ockenfels, 2000). Sociale oriëntaties zijn gebaseerd op emoties die mensen aanzetten tot reciprociteit. Mensen met een weerzin tegen ongelijkheid willen bijvoorbeeld (verwachte) coöperatie van anderen belonen met hun eigen coöperatie als dit het verschil tussen hun eigen uitkomst en die van anderen kleiner maakt. Een vergelijkbare redenering kan worden gehouden met betrekking tot het bestraffen van oncoöperatief gedrag van anderen.

Denk aan een situatie waarin mensen beslissen opbrengsten al dan niet te delen met een ander. Als deze opbrengsten in eerste instantie niet het eigendom van mensen zelf zijn, maar het resultaat van een voorgaand besluit van een ander, vereist het principe van reciprociteit dat mensen deze opbrengsten vrijgeviger delen. Op uitkomsten gebaseerde motieven kunnen echter verschillen in gedrag niet verklaren tussen keuzesituaties die verschillen met betrekking tot eerdere keuzes, maar waarin objectieve uitkomsten identiek zijn. Op intenties gebaseerde motieven onderkennen echter wel dat mensen rekening houden met het proces van eerdere keuzes waardoor bepaalde uitkomsten verkregen worden. Op intenties gebaseerde motieven kunnen ook de vriendelijkheid van anderen in een beslissing betrekken. Sociologisch en sociaalpsychologisch onderzoek suggereert dat het principe van reciprociteit diep geworteld is in fundamentele sociaalpsychologische krachten die gebaseerd zijn op verplichting en verontwaardiging. Vriendelijkheid van anderen leidt tot gevoelens van verplichting die mensen ertoe aanzetten om voor deze gunst iets terug te doen (Gouldner, 1960; Coleman, 1990: hoofdstuk 12; Cialdini, 2001: hoofdstuk 2), terwijl onvriendelijk gedrag van anderen gevoelens van verontwaardiging oproept die mensen ertoe drijven iets onvriendelijks terug te doen (Gouldner, 1960).

Naast op intenties gebaseerde motieven met betrekking tot eerder gedrag van anderen, geeft ook eerder gedrag van mensen zelf aanleiding tot op het proces gebaseerde motieven binnen de persoon. Een voorbeeld is het verlangen om consistent te zijn. Mensen zijn geneigd zich consistent te gedragen om cognitieve dissonantie te vermijden (Festinger, 1957; Webster, 1957; Caldini, 2001: hoofdstuk 3; Kunda, 2002). Beloften en bedreigingen zijn voorbeelden van geuite intenties om een bepaald gedrag te vertonen. Consistentie vereist dat mensen hun woord houden en zorgt op een intrinsieke manier dat beloften en bedreigingen dienen als 'commitments' (in de zin van een 'strategische zet', Schelling, 1960). Mensen kunnen een belofte doen aan iemand om deze ertoe aan te zetten ze een gunst te verlenen voordat de belofte wordt ingelost. Reciprociteit kan dan ook het resultaat zijn van de behoefte van mensen om zich consistent met hun beloften te gedragen. Ze doen dan iets vriendelijks terug niet vanwege gevoelens van verplichting maar omdat ze zich consistent willen gedragen.

Bovendien delen mensen dan ook een verantwoordelijkheid voor het besluit van de ander omdat deze misschien alleen heeft toegezegd om de gunst te verlenen als gevolg van de belofte die gedaan is. Vanwege de gedeelde verantwoordelijkheid voor het besluit van de ander, kan het verlangen consistent te zijn ook gevoelens van een verplichting vergroten om iets voor de gunst terug te doen. Ontvangen beloften houden op hun beurt het vooruitzicht in van een winst die op dezelfde wijze een verplichting creëert tot terugbetaling. Bedreigingen daarentegen hebben betrekking op onvriendelijkheid door het vooruitzicht van een verlies. Hierdoor kunnen bedreigingen gevoelens van verontwaardiging oproepen en de bedreigde persoon ertoe aanzetten zich ook onvriendelijk te gedragen. Ook het niet doen van een belofte is vaak onaardig en zet op eenzelfde manier aan tot gevoelens van verontwaardiging en dus vergelding. Personen die expliciet nalaten een belofte te doen, kunnen op hun beurt onvriendelijk gedrag legitimeren via mechanismen tot vermindering van cognitieve dissonantie (Festinger, 1957). Zelfs als mensen een gunst krijgen ondanks het expliciet nalaten van een belofte, strijdt het verlangen naar consistentie met gevoelens van verplichting en dit kan de invloed van de gevoelens van verplichting ondermijnen. Merk op dat beloften en bedreigingen 'cheap-talk' (goedkope praatjes) kunnen zijn, dat wil zeggen dat ze de objectieve uitkomsten niet veranderen. Toch kunnen ze reciprociteit veroorzaken dankzij gevoelens van verplichting of verontwaardiging en door het verlangen consistent te zijn. Dit is tegengesteld aan hedendaagse theoretische modellen die rekening houden met op intenties gebaseerde motieven, waarin waargenomen vriendelijkheid geacht wordt bepaald te worden door verloren gegane uitkomsten van moedwillig niet gekozen opties (bijvoorbeeld Falk en Fischbacher, 2006). Empirisch bewijs voor de stimulerende invloed die 'cheap-talk' beloften uitoefenen op coöperatief gedrag is ook gevonden door talrijke studies op het gebied van communicatie (voor een overzicht zie bijvoorbeeld Sally, 1995; Shankar en Pavitt, 2002). Samenvattend is het uitgangspunt van deze studie dan ook dat verplichting, verontwaardiging en consistentie krachtige mechanismen zijn die mensen ertoe drijven, daden van louter vriendelijkheid of onvriendelijkheid 'gepast' te beantwoorden.

2. Vier studies: Uitgangspunten, benadering en vernieuwingen

2.1 Theoretische ideeën en bijdragen

De studies die in dit boek gepresenteerd worden onderzochten hoe op het proces gebaseerde motieven het gedrag van mensen beïnvloeden in sociale dilemma's. Voor dit doel zijn beslissingen van mensen geobserveerd in keuzesituaties waarin de objectieve uitkomsten identiek waren maar die verschilden met betrekking tot de gedragscontext.

Deze gedragscontext kwam op een endogene manier tot stand doordat mensen eerder vriendelijke of onvriendelijke keuzes hadden gemaakt. De focus lag op vertrouwen tussen personen in keuzesituaties met twee mensen in eenmalige ontmoetingen. De theoretische fundering voor alle vier de studies kwam uit inzichten van sociologisch en sociaalpsychologisch onderzoek naar gevoelens van verplichting of verontwaardiging en naar consistentie. Deze inzichten lijken grotendeels genegeerd te zijn in eerder onderzoek naar sociale dilemma's. Het vervolg van deze samenvatting geeft een overzicht van de vier studies en van de substantiële aanvullingen op voorgaand onderzoek.

Study 1: Vertrouwen en beloften als vriendelijke tegemoetkomingen. Experimentele bevindingen over reciprociteit na eerdere vriendelijkheid

In studie 1 (hoofdstuk 2) werd onderzocht hoe vertrouwen betrouwbaarheid beïnvloedt en hoe het doen en nalaten van beloften om vertrouwen te honoreren een invloed heeft op betrouwbaarheid en vertrouwen. Dit leidde tot een studie naar positieve reciprociteit. Sommige onderzoekers claimden dat deze vorm van reciprociteit niet of nauwelijks bestaat (zie bijvoorbeeld de discussie van Falk e.a., 2003). Bovendien werden de invloeden van beloften van betrouwbaarheid onderzocht in combinatie met objectieve onderpanden. In andere studies werd de inhoud van communicatie minder goed gecontroleerd. In eerdere studies waren beloften 'cheap-talk' (bijvoorbeeld Brandts en Charness, 2003) of werd er alleen gecontroleerd op invloeden van een aantal specifieke op uitkomsten gebaseerde motieven (bijvoorbeeld Snijders, 1996).

Study 2: Verleiding, verlies en beloften van betrouwbaarheid. Experimentele bevindingen over contextafhankelijkheid van op uitkomsten gebaseerde motieven

In studie 2 (hoofdstuk 3) lag de focus op hoe gedragscontexten die het gevolg zijn van eerder gedrag, de effecten van op uitkomsten gebaseerde motieven op vertrouwen en betrouwbaarheid beïnvloeden. Voor dit doel is een klassiek model van altruïsme (Brew, 1973; Weesie, 1993, 1994b) op informele wijze toegepast. Dit model bestaat uit een zelfzuchtige nutscomponent en een individueel gewogen sociale nutscomponent. Het model maakt het mogelijk om invloeden van eigen uitkomsten te scheiden van invloeden van de uitkomsten voor de andere persoon. Het idee van contextafhankelijkheid van op uitkomsten gebaseerde motieven trok een belangrijke assumptie in twijfel die gemaakt wordt in eerdere theoretische modellen, namelijk dat op uitkomsten gebaseerde motieven (sociale oriëntaties) individueel stabiel zijn tussen keuzesituaties. Sommige eerdere onderzoeken hebben sociale oriëntaties bestudeerd in keuzesituaties die (a) verschillen met betrekking tot uitkomsten, (b) gelijktijdig of na

elkaar plaatsvinden en (c) verschillende keuzeopties hebben (bijvoorbeeld McClintock en Liebrand, 1988; Blanco e.a., 2006). Met zo'n design is het echter niet mogelijk om te onderzoeken hoe op het proces gebaseerde motieven uitkomsteffecten beïnvloeden.

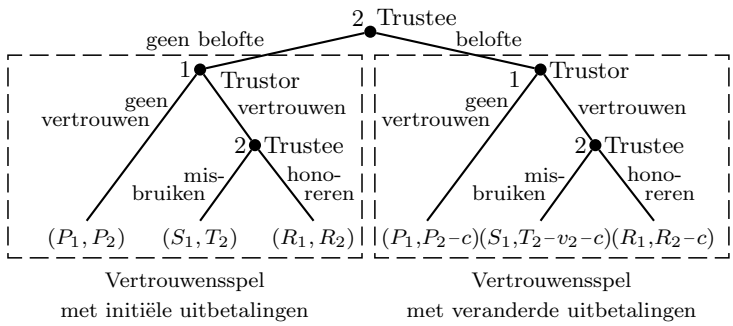
Study 3: De invloed van beloften en bedreigingen op vertrouwen en betrouwbaarheid. Experimentele bevindingen over reciprociteit na beloften en bedreigingen

In studie 3 (hoofdstuk 4) werd geanalyseerd hoe vertrouwen en betrouwbaarheid beïnvloed worden door beloften en bedreigingen. In deze studie werd de benadering die was gekozen in studie 1 (hoofdstuk 2) toegepast op vertrouwenssituaties waarbij 'trustors' (degenen die beslissen om vertrouwen al dan niet te geven) de mogelijkheid hebben om sancties toe te passen. Hiermee kon ten eerste de vraag onderzocht worden of de belangrijkste bevindingen van studie 1 ook gevonden worden wanneer 'trustors' expliciete opties voor sancties hebben. Dit was twijfelachtig omdat eerdere studies lieten zien dat sancties averechtse effecten kunnen hebben en dat daarmee coöperatief gedrag juist ondermijnd wordt (bijvoorbeeld Gürerk e.a., 2004; Voss en Vieth, 2006). In herhaalde interacties werd juist een stimulerende invloed op coöperatief gedrag gevonden wanneer zowel communicatie als het plaatsen van sancties mogelijk waren (bijvoorbeeld Ostrom e.a., 1992; Bochet en Putterman, 2007). Een tweede doel was om invloeden van een belofte tot beloning en een dreiging met bestraffing op betrouwbaarheid te onderzoeken. Eerdere studies bekeken alleen de invloed van de aankondiging van sancties op de daaropvolgende besluitneming onafhankelijk van andere factoren (bijvoorbeeld Fehr en Rockenbach, 2003; Voss en Vieth, 2006; Fehr e.a., 2007).

Study 4: Vergelding en dankbaarheid in vertrouwenssituaties met beloften en bedreigingen. Experimentele bevindingen over reciprociteit bij op intenties gebaseerde sancties

In studie 4 (hoofdstuk 5) werd aandacht besteed aan de vraag hoe eerder gedrag vergelding en dankbaarheid beïnvloedt. De focus verschoof dus van betrouwbaarheid en vertrouwen naar het sanctioneren. In voorgaand onderzoek naar sanctioneren werd alleen gecontroleerd op invloeden van sommige specifieke, op uitkomsten gebaseerde motieven. Dit geldt ook voor studies waarin de kosten van sanctioneren voor anderen gelijk waren aan het effect van de sancties (zie bijvoorbeeld Falk e.a., 2005; Vyrastekova en van Soest, 2008; Sefton e.a., 2007). Deze studies zien over het hoofd dat een weerzin tegen ongelijkheid niet het enige op uitkomsten gebaseerde motief is dat geactiveerd kan worden. Er zijn nog een aantal studies naar de bestraffing van leugens

Figuur 1: Vertrouwensspel met een belofte om betrouwbaar te zijn



$$R_1 > P_1 > S_1; \quad T_2 > R_2 > P_2; \quad v_2 \geq 0; \quad c \geq 0$$

(Brandts en Charness, 2003), de beloning van het houden van beloftes (bijvoorbeeld Fehr e.a., 2007) en de invloeden van aankondigingen van sancties op daadwerkelijke beslissingen tot het uitdelen van sancties (bijvoorbeeld Fehr en Rockenbach, 2003; Voss en Vieth, 2006; Fehr e.a., 2007). Maar deze laatste studies hadden zogenaamde 'confounding' factoren in het experimentele design of er werd niet voldoende gecontroleerd op invloeden van op uitkomsten gebaseerde motieven.

2.2 Methodologische benadering en bijdragen

Er werden twee laboratoriumexperimenten ontworpen waarin het gedrag van een persoon in verschillende gedragscontexten werd geobserveerd (zie Tabel 1). De keuzesituaties hadden een identieke structuur. Dat wil zeggen dat ze bestonden uit dezelfde keuzeopties en uit dezelfde objectieve uitkomsten voor zowel de 'trustor', degene die al dan niet vertrouwen geeft, als de 'trustee', degene die besluit om eventueel geplaatst vertrouwen al dan niet te honoreren (voor gelijksoortige designs zie Snijders, 1996; McCabe e.a., 2003; Cox, 2004). Het enige verschil was het eerdere gedrag waardoor de specifieke keuzesituatie gegenereerd werd. Denk bijvoorbeeld aan een keuzesituatie waarin de 'trustee' al dan niet belooft betrouwbaar te zijn (Figuur 1) (zie ook Raub, 1992; Weesie en Raub, 1996; Raub, 2004). Het gedrag in beide daaropvolgende vertrouwenssituaties werd vergeleken met het gedrag in de vertrouwenssituatie waarin de 'trustee' geen gelegenheid had zijn betrouwbaarheid te beloven.

In het tweede experiment had de 'trustor' een optie om gehonoreerd vertrouwen te belonen of om misbruik van vertrouwen te bestraffen. In sommige van deze situaties kon de 'trustor' naast het plaatsen van vertrouwen ook vooraf aankondigen zo'n beloning of straf daadwerkelijk te willen uitvoeren. Opnieuw werd gedrag in situaties

Tabel 1: Overzicht van de belangrijkste verschillen tussen de twee experimenten

	Experiment 1	Experiment 2
	Hoofdstuk 2 en 3	Hoofdstuk 4 en 5
Tijd en locatie	Nov. 2006 in het ELSE laboratorium, ICS/Sociology, Universiteit Utrecht	Apr. 2008 in het CeDEx laboratorium, School of Economics, Nottingham University
Opties voor het uitdelen van sancties voor de "trustor" (bestraffing of beloning)	geen opties	opties in alle situaties (kostbaar en meestal niet effectief)
Belofte van betrouwbaarheid door "trustees"	opties in sommige situaties	opties in sommige situaties
Aankondiging van sancties door "trustors" (dreiging met bestraffing, belofte van beloning)	geen opties	opties in sommige situaties
Eigenschappen van de aankondiging	met een onderpand en/of transactiekosten, of "cheap-talk"	altijd "cheap-talk"
Codebook	Vieth en Weesie (2006)	Vieth (2008)

Sancties zijn effectief als de "trustee" geen prikkel heeft om onbetrouwbaar te zijn.

resulterend uit de beslissing van de 'trustor' om sancties aan te kondigen, vergeleken met situaties zonder de mogelijkheid voor deze aankondiging. Op een soortgelijke manier werd de invloed van het plaatsen van vertrouwen op betrouwbaarheid onderzocht. Hierbij werd de beslissing van de 'trustee' om betrouwbaar te zijn na geplaatst vertrouwen vergeleken met een beslissing van de 'trustee' waarbij deze opbrengsten op een soortgelijke manier kan verdelen over zichzelf en ander, maar nu zijn deze opbrengsten exogeen verkregen en niet het gevolg van het vertrouwen van de ander. Deze procedure werd ook toegepast op de sanctiebeslissingen van de 'trustor'. Bestraffing is een investering die de uitkomsten van de ander vermindert via een boete. Een beloning is een investering die de uitkomsten van de ander verhoogt via een vergoeding. Elk van deze beslissingen om wel of niet te investeren in het veranderen van de uitkomsten van de ander in situaties nadat de ander gekozen heeft tussen opbrengsten houden of delen werd vergeleken met de twee situaties in welke de ander

geen voorgaande keuzeoptie had. Hiermee kon de zuivere invloed van de keuze van de ander om opbrengsten te houden of verdelen, bepaald en nagegaan worden of mensen straffen of belonen vanwege gevoelens van verplichting of verontwaardiging zonder dat er invloeden waren van op uitkomsten gebaseerde motieven.

Om de invloeden van de gedragscontext op het nemen van beslissingen te kunnen bestuderen is in de statistische analyses gecontroleerd voor de op uitkomsten gebaseerde motieven en voor individuele heterogeniteit. Dit is gedaan door de data te groeperen in 'subject-payoff reponse sets' bestaande uit de beslissingen die een proefpersoon maakte in verschillende gedragscontexten met identieke objectieve uitkomsten. In studie 2 werden 'subject response sets' gecreëerd om interactie-effecten te onderzoeken tussen op uitkomsten gebaseerde motieven (gerepresenteerd door welwillendheid en vijandigheid op informele wijze toegepast binnen een gevestigd altruïstisch model) en de gedragscontext (hoofdstuk 3). Bij het analyseren van de data van het eerste experiment werd gebruik gemaakt van logistische regressiemodellen met 'fixed effects' voor 'response sets' (hoofdstuk 2 en 3). Voor het analyseren van de beslissingen die deelnemers maakten in het tweede experiment werd gebruik gemaakt van logistische regressiemodellen met 'random effects' voor 'response sets' (hoofdstuk 4 en 5).

Bij alle vier de studies werden beloftes in de vorm van 'cheap-talk' toegepast (zie ook studies op het gebied van communicatie, voor een overzicht zie bijvoorbeeld Sally, 1995; Shankar en Pavitt, 2002). Dit maakte het mogelijk om systematisch de vraag te onderzoeken of gepercipieerde vriendelijkheid bepaald wordt door niet bereikte objectieve uitkomsten van niet gekozen opties, zoals wordt aangenomen in andere theoretische modellen (bijvoorbeeld Falk en Fischbacher, 2006) of door op het proces gebaseerde motieven (verplichting, verontwaardiging en consistentie). Bovendien dragen de vier studies in dit boek in methodologische zin bij aan eerder onderzoek (zie hoofdstuk 2 voor een uitgebreide discussie).

Ten eerste staat het construeren van een set van *structureel identieke (sub)spellen* toe om de zuivere invloed van gedragscontexten te analyseren terwijl gecontroleerd wordt voor op uitkomsten gebaseerde motieven zonder assumpties te maken over zulke motieven. Als aannemelijke assumpties gemaakt zouden kunnen worden over op uitkomsten gebaseerde motieven, zou het modelleren van deze assumpties efficiëntere toetsen toestaan in de statistische analyse. Op uitkomsten gebaseerde motieven kunnen echter niet voldoende precies gemodelleerd en gemeten worden, gegeven de huidige staat van het onderzoek (voor meer details zie hoofdstuk 3; Aksoy en Weesie, 2008). Eerdere experimenten controleren alleen voor een aantal specifieke op uitkomsten gebaseerde motieven (bijvoorbeeld lineaire invloeden van weerzin tegen ongelijkheid).

Ten tweede is gebruik gemaakt van een *'within-subject' experimenteel design*. Zo'n design staat toe verschillen te analyseren in het nemen van beslissingen van een en dezelfde persoon in verschillende gedragscontexten. Op enkele uitzonderingen na gebruikten eerdere studies een 'within-subject' design. Zo'n design is minder geschikt om individuele motieven te bestuderen. 'Within-subject' designs hebben voordelen maar ook nadelen (Keren, 1993; Putt, 2005). Zoals behandeld in hoofdstuk 2, is een belangrijk nadeel dat er ervaringseffecten en volgorde-effecten kunnen optreden. Een 'within-subject' design lijkt echter beter geschikt voor het type studies dat gepresenteerd wordt in dit boek, omdat invloeden van motieven bestudeerd kunnen worden op individueel niveau, terwijl 'between-subjects' designs alleen een vergelijking tussen gemiddeld gedrag op groepsniveau toestaat (voor meer informatie over de ecologische valkuil, zie Robinson, 1950). Door gebruik te maken van een 'within-subject' design is het bovendien mogelijk, te controleren voor additieve individuele heterogeniteit en voor invloeden van verschillende objectieve uitkomsten, en wel zonder assumpties te maken over specifieke op uitkomsten gebaseerde motieven.

Ten derde werden *gedragscontexten endogeen gegenereerd* door vriendelijk en onvriendelijk gedrag van deelnemers. In veel voorgaande experimenten wordt de 'strategiemethode' (Selten, 1967) gebruikt, vooral in de weinige studies die gebruik maken van een 'within-subject' design. Een belangrijk probleem met de strategiemethode is echter dat beslissingen hypothetisch blijven. Dit ondermijnt invloeden van emoties, die zorgen voor de onderliggende krachten van sociale motieven. Bovendien zijn onzuiverheden in geschatte effecten waarschijnlijker door de artificiële consistentie in antwoorden als de strategiemethode gebruikt wordt.

Ten vierde verzekert het gebruik van *binaire keuzesituaties* dat beslissingen relatief ondubbelzinnig zijn met betrekking tot de interpretatie of keuzes vriendelijk of onvriendelijk zijn. Bovendien maakt het gebruik van binaire keuzes het aantal (sub)spellen klein. Een uitzondering op de binaire keuzesituaties was de keuze van de 'trustor' om een aankondiging te doen, waarin drie opties mogelijk waren (dreiging met een straf, beloven van een beloning of geen aankondiging). De drie opties zijn duidelijk interpreteerbaar met betrekking tot gepercipieerde vriendelijkheid.

3. Samenvatting van de resultaten

De resultaten van de vier studies leveren ondersteuning voor de aanname van reciprociteit en sterken het idee dat reciprociteit gebaseerd is op gevoelens van verplichting, gevoelens van verontwaardiging en het verlangen, consistent te zijn. De besluitvor-

ming van mensen en de onderliggende motieven bleken beïnvloed te worden door de gedragscontext, zelfs zonder enige verandering in de objectieve uitkomsten.

Result 1: Vertrouwen leidt tot betrouwbaarheid.

Alleen al het plaatsen van vertrouwen verhoogt betrouwbaarheid (hoofdstuk 2 en 3; zie ook McCabe, 2003; Cox, 2004). Dit is vooral gevonden in keuzesituaties waarin opties beschikbaar zijn om sancties op te leggen (hoofdstuk 4). Dankzij het geven van vertrouwen wordt de negatieve invloed verkleind voor onbetrouwbaar gedrag van de 'trustee' (hoofdstuk 3). Deze bevindingen steunen het idee dat 'trustees' een verplichting voelen zich betrouwbaar te gedragen om iets terug te doen voor de gunst van het gekregen vertrouwen.

Result 2: Het beloven van betrouwbaarheid bevordert vertrouwen en daadwerkelijke betrouwbaarheid.

Enkel het beloven van betrouwbaarheid verhoogt vertrouwen en betrouwbaarheid, zelfs als deze belofte objectief gezien 'cheap-talk' is (hoofdstuk 2 en 3). Dit effect is vooral sterk in keuzesituaties waarin de mogelijkheid tot sanctioneren bestaat (hoofdstuk 4). De bevorderende impact van de belofte om vertrouwen te honoreren wijst op de invloed van gevoelens van verplichting en consistentie. Zoals beargumenteerd kan consistentie ook de invloed bevorderen van gevoelens van verplichting, dankzij een gedeelde verantwoordelijkheid voor het besluit van de 'trustor' om vertrouwen te plaatsen. Transactiekosten die gemaakt moeten worden voor het doen van de belofte om betrouwbaar te zijn, bevorderen de toename in betrouwbaarheid (hoofdstuk 2). Als gevolg van het maken van de belofte om betrouwbaar te zijn, wordt het positieve effect van de zorg van de 'trustee' over het verlies voor de 'trustor' mocht de 'trustee' vertrouwen misbruiken, verminderd (hoofdstuk 3). Deze twee bevindingen steunen het idee van consistentie. De bevorderende invloed van beloofde betrouwbaarheid op vertrouwen ondersteunt de ideeën over de invloed van gevoelens van verplichting en het geloof van de 'trustor' in verhoogde betrouwbaarheid. Er zijn geen aanwijzingen gevonden dat eigenschappen van de belofte een invloed hebben op vertrouwen wanneer gecontroleerd wordt voor op uitkomsten gebaseerde motieven (hoofdstuk 2). De hypothese dat het ontvangen van een belofte van betrouwbaarheid de belemmerende invloed van verlies en verleiding op vertrouwen zou veranderen, is evenmin bevestigd (hoofdstuk 3).

> **Result 3:** Leugens brengen vergelding teweeg maar het zich houden aan beloften om betrouwbaar te zijn, leidt tot minder beloning dan betrouwbaar zijn zonder de eerdere optie om betrouwbaarheid te beloven.

Het onvermogen zich te houden aan een belofte van betrouwbaarheid verhoogt de neiging tot vergelding (hoofdstuk 5; zie ook Brandts en Charness, 2003; en voor herhaalde interacties met betrekking tot publieke goederen, Bochet en Putterman, 2007). Dit bevestigt het idee dat een gevoel van verontwaardiging mensen er toe aanzet geleden verliezen te vergelden. Als vertrouwen gehonoreerd wordt nadat het is beloofd, heeft dankbaarheid de neiging minder te worden (hoofdstuk 5). Dit wijst erop dat vriendelijk gedrag na verleende gunsten leidt tot zwakkere gevoelens van verplichting dan oorspronkelijke gunsten en dat 'trustors' mogelijk verwachten dat de 'trustee' de verantwoordelijkheid deelt voor het gevraagde vertrouwen als betrouwbaarheid beloofd is. Merk op dat er geen ondersteuning is gevonden voor het idee dat dankbaarheid over gedeelde winsten in ruil voor vertrouwen kleiner is dan dankbaarheid over opbrengsten die gedeeld worden als een oorspronkelijke gunst (hoofdstuk 5). Gevoelens van verplichting om gehonoreerd vertrouwen te belonen worden dus in het bijzonder verminderd nadat betrouwbaarheid beloofd is.

> **Result 4:** Het nalaten van het doen van een belofte van betrouwbaarheid wordt vergolden door het inhouden van vertrouwen terwijl de invloed op betrouwbaarheid afhangt van de eigenschappen van de nagelaten belofte.

Nalaten van het doen van een belofte van betrouwbaarheid belemmert vertrouwen, zelfs als deze belofte objectief gezien 'cheap-talk' is (hoofdstuk 2 en 3; zie ook Snijders, 1996; Gautschi, 2000). Dit is vooral het geval in keuzesituaties waarin de optie tot sanctioneren bestaat (hoofdstuk 4). Deze bevinding ondersteunt het idee dat gevoelens van verontwaardiging 'trustors' ertoe aanzetten geen vertrouwen te plaatsen met als doel de 'trustee' te bestraffen voor het nalaten van een belofte. Er is geen bewijs gevonden voor het idee dat eigenschappen van de belofte het effect van een nagelaten belofte op vertrouwen beïnvloeden (hoofdstuk 2). Betrouwbaarheid stijgt wel met transactiekosten die gemaakt moeten worden voor het doen van een belofte. Betrouwbaarheid vermindert met de waarde die een nagelaten belofte zou hebben (hoofdstuk 2). Dit heeft twee implicaties. Ten eerste voelen 'trustees' een verplichting het vertrouwen van de 'trustor' te belonen. Dit gevoel is sterker naarmate vertrouwen meer een teken is dat de 'trustor' begrepen heeft dat de belofte is nagelaten vanwege de hoge transactiekosten. Ten tweede ondermijnt consistentie de gevoelens van verplichting om gegeven vertrouwen te honoreren wanneer vertrouwen is geplaatst ondanks

een hoge waarde van een onderpand van een nagelaten belofte. Nog een bevinding is dat de belemmerende invloeden van de verleiding van de 'trustee' op betrouwbaarheid en van het verlies als gevolg van misbruikt vertrouwen voor de 'trustor' groter zijn nadat de belofte betrouwbaar te zijn is nagelaten (hoofdstuk 3). Deze toename van het belemmerende effect van de verleiding van de 'trustee' bevestigt het idee dat gevoelens van verplichting ondermijnd worden na nagelaten beloften. Merk op dat er geen bevestiging is gevonden voor de redenering dat de invloed van de het verlies van de 'trustor' op betrouwbaarheid positiever zou zijn na een nagelaten belofte dankzij gevoelens van verplichting veroorzaakt door geplaatst vertrouwen ondanks dat een belofte was nagelaten (hoofdstuk 3). Bovendien is er geen bewijs gevonden voor een algemene vermindering in betrouwbaarheid nadat is nagelaten een belofte te doen (hoofdstuk 2; zie ook Snijders, 1996) noch in het geval dat de nagelaten belofte 'cheap-talk' is geweest (hoofdstuk 2 en 3).

> **Result 5:** Beloften leiden tot een toename in betrouwbaarheid, neiging tot vergelding en dankbaarheid.

Enkel de belofte gehonoreerd vertrouwen te belonen verhoogt betrouwbaarheid, ondanks het feit dat de belofte 'cheap-talk' is (hoofdstuk 4). Dit levert ondersteuning voor de invloed van verhoogde gevoelens van verplichting die voortkomen uit de combinatie van twee gunsten (dat wil zeggen geplaatst vertrouwen en de belofte van een beloning) en voor de invloed van geanticipeerd sanctiegedrag. Beschaamd vertrouwen nadat een beloning is beloofd, wordt bijzonder streng bestraft (hoofdstuk 5). Dit ondersteunt het idee dat gevoelens van verontwaardiging een neiging tot vergelding met zich meebrengen. Bovendien verhoogt het beloven van een beloning dankbaarheid (hoofdstuk 5). Eerder was te zien dat de dankbaarheid van de 'trustor' lager was wanneer de 'trustor' besloot de 'trustee' al dan niet te belonen voor het houden aan zijn belofte van betrouwbaarheid (resultaat 3). Zoals eerder aangegeven kan dit begrepen worden door in acht te nemen dat gevoelens van verplichting om beantwoorde gunsten te belonen zwakker zijn dan om oorspronkelijke gunsten te belonen. Daarom duidt het positieve effect van het beloven van een beloning op een daadwerkelijke beloning op een sterke invloed van consistentie.

> **Result 6:** Dreigen met bestraffing lijkt de neiging tot vergelding te bevorderen.

Wanneer 'trustors' het plaatsen van vertrouwen kunnen combineren met het aankondigen van sancties beloven ze meestal (87%) een beloning (hoofdstuk 4 en 5). Daarom

waren er nauwelijks observaties beschikbaar voor het analyseren van de invloeden van het dreigen met bestraffing. Daarom kan alleen met enige voorzichtigheid geconcludeerd worden dat enkel het dreigen met bestraffing de neiging tot vergelding vergroot (hoofdstuk 5; zie ook Voss en Vieth, 2006). Dit wijst op enige ondersteuning voor het idee dat consistentie de 'trustors' ertoe aanzet om daadwerkelijk de bestraffing waarmee gedreigd is uit te voeren. Dit kan ook het gevoel van verontwaardiging van de 'trustor' versterken. Er is geen ondersteuning gevonden voor het idee dat dreigen met bestraffing leidt tot vergelding via het misbruiken van vertrouwen (hoofdstuk 4). In feite suggereren de empirische bevindingen eerder een positieve invloed van dreigen met bestraffen op het honoreren van vertrouwen (zie ook Voss en Vieth, 2006). Verder onderzoek naar de invloed van bedreigingen op zowel coöperatief gedrag als sanctiegedrag is vereist. Daarbij zouden de invloeden van waargenomen oneerlijkheid ook onderzocht moeten worden (voor dit standpunt zie bijvoorbeeld ook Fehr en Rockenbach, 2003).

> **Result 7:** Het feit dat mensen zich druk maken om de verdeling van objectieve uitkomsten bepaalt maar voor een klein deel de keuzes met betrekking tot bestraffen en belonen.

Enkel het voor zichzelf houden van opbrengsten, wanneer delen mogelijk was, verhoogt al de neiging tot vergelding (hoofdstuk 5). Op eenzelfde manier verhoogt louter het delen van opbrengsten, wanneer zelf houden mogelijk was, dankbaarheid, zelfs als delen een antwoord is op de gunst van geplaatst vertrouwen (hoofdstuk 5). Slechts een enkeling aanvaardt vrijwillig kosten voor het verminderen of vermeerderen van de uitkomsten van de ander in situaties waaraan geen onvriendelijke of vriendelijke beslissing van de ander vooraf gaat. Dit versterkt het idee dat bestraffingsgedrag gemotiveerd wordt door gevoelens van verontwaardiging die mensen aanzetten tot vergelding en dat gevoelens van verplichtingen om iets terug te doen voor ontvangen gunsten, beloningsgedrag motiveren.

> **Result 8:** Op uitkomst gebaseerde motieven hangen samen met gedragscontexten en de invloeden hangen af van de beslissingsposities van de persoon in kwestie.

In studie 2 worden ook aanwijzingen gevonden voor het idee dat op uitkomsten gebaseerde motieven samenhangen met gedragscontexten. Dat wil zeggen dat deze motieven niet individueel stabiel zijn tussen keuzesituaties met identieke uitkomsten.

Bovendien is gevonden dat de invloeden van op uitkomsten gebaseerde motieven afhangen van de beslissingspositie die een persoon inneemt in een keuzesituatie. Altruïstische tendensen in de rol van de 'trustee', die welwillend is richting de 'trustor', lijken te veranderen in agressieve neigingen in de rol van de 'trustor', die kwaadwillend is richting de 'trustee'. Deze interpretatie wordt gesteund door indicaties over de belemmerende invloed van de egoïstische motieven van de ander: de gedragscontext heeft eerder een effect op de zelfzuchtige motieven dan op de sociale motieven. Dit impliceert ook dat de invloed van overtuigingen niet gemedieerd wordt door dezelfde componenten van de uitkomst.

De resultaten van de vier studies die in dit boek gepresenteerd worden, bieden sterke ondersteuning voor de gedachte dat de gedragscontext die het resultaat is van eerder gedrag, vertrouwen en betrouwbaarheid beïnvloedt via gevoelens van vergelding en dankbaarheid. Dankzij consistentie, gevoelens van verplichting of beide worden mensen gedreven hun beloften te houden en een bedreiging uit te voeren. Gevoelens van verplichting motiveren mensen ertoe iets terug te doen voor de gunst van geplaatst vertrouwen en ontvangen beloften. Mensen die iets terugdoen voor ontvangen gunsten, kunnen niet verwachten om rijkelijk beloond te worden, omdat gevoelens van verplichting groter zijn voor oorspronkelijke gunsten. Expliciet nagelaten beloften van betrouwbaarheid veroorzaken gevoelens van verontwaardiging die 'trustors' ertoe aanzetten geen vertrouwen te plaatsen. Voor de persoon die de belofte naliet, conflicteert consistentie met gevoelens van verplichting. Of consistentie een sterkere invloed heeft dan gevoelens van verplichting hangt dan af van de eigenschappen van de belofte. Daarnaast worden mensen, ongeacht de invloeden van objectieve uitkomsten, sterk gedreven door gevoelens van verontwaardiging om anderen te straffen voor onvriendelijkheid en door gevoelens van verplichting om vriendelijkheid van anderen te belonen. De invloed van op uitkomsten gebaseerde motieven, zoals vijandigheid en welwillendheid, verschilt tussen beslissingsrollen en hangt af van eerdere keuzes. De bevindingen tonen de kracht aan van gevoelens van verplichting, gevoelens van verontwaardiging en het verlangen consistent te zijn. Gebaseerd op deze motieven ontstaan patronen van reciprociteit die een sterke invloed onthullen van geïnternaliseerde sociale normen op het gedrag van mensen en ze vormen daarmee een basis voor het behouden van de sociale orde. 'Commitments' en reciprociteit zijn fundamentele onderdelen van interacties tussen mensen. De vier studies in dit boek hebben een aantal basisprincipes van de samenhang tussen 'commitments' en reciprociteit onderzocht. Hopelijk vormen de bevindingen een inspiratiebron voor verder onderzoek.

References

Abbink, Klaus, Bernd Irlenbusch, and Elke Renner. 2000. "The Moonlighting Game. An Experimental Study of Reciprocity and Retribution." *Journal of Economic Behavior and Organization* 42:265–277.

Akerlof, George A. 1970. "The Market for "Lemons": Qualitative Uncertainty and the Market Mechanism." *Quarterly Journal of Economics* 84:488–500.

Akerlof, George A. and William T. Dickens. 1982. "The Economic Consequences of Cognitive Dissonance." *American Economic Review* 72:307–319.

Aksoy, Ozan and Jeroen Weesie. 2008. "Social Motives and Expectations in One-Shot Asymmetric Prisoner's Dilemmas." Mimeo, Utrecht University.

Anderson, Christopher and Louis Putterman. 2006. "Do Non-Strategic Sanctions Obey the Law of Demand? The Demand for Punishment in the Voluntary Contribution Mechanism." *Games and Economic Behavior* 54:1–24.

Andreoni, James, William Harbaugh, and Lise Vesterlund. 2003. "The Carrot or Stick: Reward, Punishment, and Cooperation." *American Economic Review* 93:893–902.

Andreoni, James and Hal Varian. 1999. "Preplay Contracting in the Prisoner's Dilemma." *Proceedings of the National Academy of Science* 96:10933–10938.

Aronson, Elliot. 1992. "The Return of the Repressed: Dissonance Theory Makes a Comeback." *Psychological Inquiry* 3:303–311.

Au, Wing Tung and Jessica Y. Y. Kwong. 2004. "Measurements and Effects of Social Orientation in Social Dilemmas: A Review." In *Contemporary Psychological Research on Social Dilemmas*, edited by Ramzi Suleiman, David V. Budescu, Ilan Fischer, and David M. Messick, pp. 71–98. Cambridge (MA): Cambridge University Press.

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior* 10:122–142.

Bicchieri, Cristina. 2002. "Covenants Without Sword: Group Identity, Norms, and Communication in Social Dilemmas." *Rationality and Society* 14:192–228.

Biddle, Bruce J. 1986. "Recent Development in Role Theory." *Annual Review of Sociology* 12:67–92.

Binmore, Ken. 1994. *Game Theory and the Social Contract: Playing Fair*, volume 1. Cambridge (MA): MIT Press.

Binmore, Ken. 1998. *Game Theory and the Social Contract: Just Playing*, volume 2. Cambridge (MA): MIT Press.

Binmore, Ken. 2005. *Natural Justice*. Oxford: Oxford University Press.

Binmore, Ken and Larry Samuelson. 1994. "An Economist's Perspective on the Evolution of Norms." *Journal of Institutional and Theoretical Economics* 150:45–63.

Blanco, Mariana, Dirk Engelmann, and Hans-Theo Normann. 2006. "A Within-Subject Analysis of Other-Regarding Preferences." Mimeo, University of London.

Blau, Peter M. 1964. *Exchange and Power in Social Life*. New York: Wiley.

Bochet, Olivier, Talbot Page, and Louis Putterman. 2006. "Communication and Punishment in Voluntary Contribution Experiments." *Journal of Economic Behavior and Organization* 60:11–26.

Bochet, Olivier and Louis Putterman. 2007. "Not Just Babble: Opening the Black Box of Communication in a Voluntary Contribution Experiment." Mimeo, Brown University.

Bolton, Gary E. and Axel Ockenfels. 2000. "A Theory of Equity, Reciprocity, and Competition." *American Economic Review* 100:166–193.

Boyd, Robert and Peter J. Richerson. 1985. *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Brandts, Jordi and Gary Charness. 2000. "Hot vs. Cold: Sequential Responses and Preference Stability in Experimental Games." *Experimental Economics* 2:227–238.

Brandts, Jordi and Gary Charness. 2003. "Truth or Consequences: An Experiment." *Management Science* 49:116–130.

Brandts, Jordi and Carles Solà. 2001. "Reference Points and Negative Reciprocity in Simple Sequential Games." *Games and Economic Behavior* 36:138–157.

Brehm, Jack W. 1956. "Postdecision Changes in Desirability of Alternatives." *Journal of Abnormal and Social Psychology* 52:384–389.

Brehm, Jack W. 1966. *A Theory of Psychological Reactance*. New York: Academic Press.

Brew, J. S. 1973. "An Altruism Parameter for Prisoner's Dilemma." *Journal of Conflict Resolution* 17:351–367.

Brosig, Jeannette. 2006. "Communication Channels and Induced Behavior." Mimeo, University of Magdeburg.

Brosig, Jeannette, Weimann Joachim, and Chun-Lei Yang. 2003. "The Hot versus Cold Effect in a Simple Bargaining Experiment." *Experimental Economics* 6:75–90.

Brosig, Jeannette, Thomas Riechmann, and Weimann Joachim. 2007. "Selfish in the End? An Investigation of Consistency and Stability of Individual Behavior." Mimeo, University of Magdeburg.

Bruins, Joost and Jeroen Weesie. 1996. "HIN95Exp. A Set of Experiments Conducted in Connection with HIN95. Codebook of HIN95Exp." ISCORE paper 77, ICS/Sociology, Utrecht University.

Burks, Stephen V., Jeffrey P. Carpenter, and Eric Verhoogen. 2003. "Playing Both Roles in the Trust Game." *Journal of Economic Behavior and Organization* 51:195–216.

Buskens, Vincent. 2002. *Social Networks and Trust*. Boston (MA): Kluwer.

Buskens, Vincent and Werner Raub. 2002. "Embedded Trust: Control and Learning." *Advances in Group Processes* 19:167–202.

Buskens, Vincent and Jeroen Weesie. 2000. "An Experiment on the Effects of Embeddedness in Trust Situations: Buying a Used Car." *Rationality and Society* 12:227–253.

Camerer, Colin F. 1995. "Individual Decision Making." In Kagel and Roth (1995), pp. 587–683.

Camerer, Colin F. 2003. *Behavioral Game Theory*. Princeton (NJ): Princeton University Press.

Carpenter, Jeffrey P. 2007. "The Demand for Punishment." *Journal of Economic Behavior and Organization* 62:522–542.

Casari, Marco and Timothy N. Cason. 2009. "The Strategy Method Lowers Trustworthy Behavior." Mimeo, Purdue University.

Chamberlain, Gary. 1980. "Analysis of Covariance with Qualitative Data." *Review of Economic Studies* 47:225–238.

Charness, Gary and Matthew Rabin. 2002. "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics* 117:817–869.

Charness, Gary and Matthew Rabin. 2005. "Expressed Preferences and Behavior in Experimental Games." *Games and Economic Behavior* 53:151–169.

Cialdini, Robert B. 2001. *Influence: Science and Practice*. Boston (MA): Allyn & Bacon.

Coleman, James S. 1990. *Foundations of Social Theory*. Cambridge (MA): Belknap Press.

Cook, Karen S. and Robin M. Cooper. 2003. "Experimental Studies of Cooperation, Trust, and Social Exchange." In Ostrom and Walker (2003), pp. 209–244.

Cox, James C. 2004. "How to Identify Trust and Reciprocity." *Games and Economic Behavior* 46:260–281.

Crawford, Vincent. 1998. "A Survey of Experiments on Communication via Cheap Talk." *Journal of Economic Theory* 78:286–298.

Croson, Rachel T. A. 2000. "Thinking like a Game Theorist: Factors Affecting the Frequency of Equilibrium Play." *Journal of Economic Behavior and Organization* 41:299–314.

Dasgupta, Partha. 1988. "Trust as a Commodity." In Gambetta (1988b), pp. 49–72.

Dawes, Robyn M., Jeanne McTavish, and Harriet Shaklee. 1977. "Behavior, Communication, and Assumptions about Other People's Behavior in a Commons Dilemma Situation." *Journal of Personality and Social Psychology* 35:1–11.

Dufwenberg, Martin and Georg Kirchsteiger. 2004. "A Theory of Sequential Reciprocity." *Games and Economic Behavior* 47:268–290.

Egas, Martijn and Arno Riedl. 2008. "The Economics of Altruistic Punishment and the Maintenance of Cooperation." *Proceedings of the Royal Society B, Biological Sciences* 275:871–878.

Ellickson, Robert C. 1991. *Order Without Law: How Neighbors Settle Disputes*. Cambridge (MA): Harvard University Press.

Emerson, Richard M. 1976. "Social Exchange Theory." *Annual Review of Sociology* 2:335–362.

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2002. "Appropriating the Commons: A Theoretical Explanation." In Ostrom et al. (2002), pp. 157–191.

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2003. "On the Nature of Fair Behavior." *Economic Inquiry* 41:20–26.

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2005. "Driving Forces behind Informal Sanctions." *Econometrica* 73:2017–2030.

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2008. "Testing Theories of Fairness—Intentions Matter." *Games and Economic Behavior* 62:287–303.

Falk, Armin and Urs Fischbacher. 2006. "A Theory of Reciprocity." *Games and Economic Behavior* 54:293–315.

Fehr, Ernst and Armin Falk. 2002. "Psychological Foundation of Incentives." *European Economic Review* 46:687–724.

Fehr, Ernst and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90:980–994.

Fehr, Ernst and Simon Gächter. 2002. "Altruistic Punishment in Humans." *Nature* 415:137–140.

Fehr, Ernst, Alexander Klein, and Klaus M. Schmidt. 2007. "Fairness and Contract Design." *Econometrica* 75:121–154.

Fehr, Ernst and John A. List. 2004. "The Hidden Costs and Rewards of Incentives." *Journal of the European Economic Association* 2:743–771.

Fehr, Ernst and Bettina Rockenbach. 2003. "The Detrimental Effect of Sanctions on Human Altruism." *Nature* 422:137–140.

Fehr, Ernst and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114:817–868.

Fehr, Ernst and Klaus M. Schmidt. 2004. "Fairness and Incentives in a Multi-Task Principal-Agent Model." *Scandinavian Journal of Economics* 106:453–474.

Fehr, Ernst and Klaus M. Schmidt. 2006. "The Economics of Fairness, Reciprocity and Altruism—Experimental Evidence and New Theories." In Kolm and Ythier (2006), pp. 615–691 (ch. 8).

Fehr, Ernst and Klaus M. Schmidt. 2007. "Adding a Stick to the Carrot? The Interaction of Bonuses and Fines." *American Economic Review, Papers and Proceedings* 97:177–181.

Festinger, Leon. 1957. *A Theory of Cognitive Dissonance*. Stanford (CA): Stanford University Press.

Fischbacher, Urs. 2007. "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics* 10:171–178.

Fischer, Gerhard H. and Ivo W. Molenaar (eds.). 1995. *Rasch Models. Foundations, Recent Developments and Applications*. New York: Springer.

Frank, Robert H. 1988. *Passions Within Reason: The Strategic Role of Emotions*. New York: Norton.

Fudenberg, Drew and David K. Levine. 1998. *The Theory of Learning in Games*. Cambridge (MA): MIT Press.

Gächter, Simon and Elke Renner. 2006. "Effects of (Incentivized) Belief Elicitation in Public Goods Experiments." CeDEx Discussion Paper 16, School of Economics, Nottingham University.

Gächter, Simon, Elke Renner, and Martin Sefton. 2008. "The Long-Run Benefits of Punishment." *Science* 322:1510.

Gallucci, Marcello and Marco Perugini. 2000. "An Experimental Test of a Game-Theoretical Model of Reciprocity." *Journal of Behavioral Decision Making* 13:367–389.

Gambetta, Diego. 1988a. "Can We Trust Trust?" In Gambetta (1988b), pp. 213–237.

Gambetta, Diego (ed.). 1988b. *Trust: Making and Breaking Cooperative Relations*. Oxford: Blackwell.

Gass, Robert H. and John S. Seiter. 2007. *Persuasion, Social Influence, and Compliance Gaining*. Boston (MA): Pearson, 3rd edition.

Gautschi, Thomas. 2000. "History Effects in Social Dilemma Situations." *Rationality and Society* 12:131–162.

Gigerenzer, Gerd and Reinhard Selten (eds.). 2001. *Bounded Rationality: The Adaptive Toolbox*. Cambridge (MA): MIT Press.

Glaeser, Edward L., David I. Laibson, José A. Scheinkman, and Christine L. Soutter. 2000. "Measuring Trust." *Quarterly Journal of Economics* 115:811–846.

Gouldner, Alvin W. 1960. "The Norm of Reciprocity." *American Sociological Review* 25:161–178.

Greiner, Ben. 2004. "An Online Recruitment System for Economic Experiments." In *Forschung und wissenschaftliches Rechnen. GWDG Bericht 63*, edited by Kurt Kremer and Volker Macho, pp. 79–93. Göttingen: Gesellschaft für Wissenschaftliche Datenverarbeitung.

Gürerk, Özgür, Bernd Irlenbusch, and Bettina Rockenbach. 2004. "On the Evolvement of Institutions in Social Dilemmas." Mimeo, University of Erfurt.

Hann, Chris. 2006. "The Gift and Reciprocity: Perspectives from Economic Anthropology." In Kolm and Ythier (2006), pp. 207–223 (ch. 4).

Harsanyi, John C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.

Heider, Fritz. 1944. "Social Perception and Phenomenal Causality." *Psychological Review* 51:358–374.

Heider, Fritz. 1958. *The Psychology of Interpersonal Relations*. New York: Wiley.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter. 2008. "Antisocial Punishment across Societies." *Science* 319:1362–1367.

Hobbes, Thomas. 1651/1966. *Leviathan, or the Matter, Forme, and Power of Commonwealth Ecclesiastical and Civil*. Oxford: Blackwell.

Hoffman, Elizabeth, Kevin McCabe, and Veron L. Smith. 2008. "Prompting Strategic Reasoning Increases Other-Regarding Behavior." In *Handbook of Experimental Economics Results. Handbooks in Economics 28*, edited by Charles R. Plott and Veron L. Smith, volume 1, pp. 423–428. Amsterdam: Elsevier (North-Holland).

Hoijtink, Herbert and Anne Boomsma. 1995. "On Person Parameter Estimation in Dichotomous Rasch Models." In Fischer and Molenaar (1995), pp. 53–68.

Homans, George C. 1974. *Social Behavior*. New York: HBJ, rev. edition.

Houser, Daniel, Erte Xiao, Kevin McCabe, and Veron Smith. 2008. "When Punishment Fails: Research on Sanctions, Intentions, and Non-Cooperation." *Games and Economic Behavior* 62:509–532.

Hume, David. 1739/1978. *A Treatise of Human Nature*. Oxford: Clarendon, 2nd edition.

Iedema, Jurjen. 1993. *The Perceived Consensus of One's Social Value Orientation*. Ph.D. thesis, Katholieke Universiteit Brabant.

Kagel, John H. and Alvin E. Roth (eds.). 1995. *The Handbook of Experimental Economics*. Princeton (NJ): Princeton University Press.

Kelley, Harold H. and John W. Thibaut. 1978. *Interpersonal Relations: A Theory of Interdependence*. New York: Wiley.

Keren, Gideon. 1993. "Between or Within Subjects Design: A Methodological Dilemma." In *A Handbook for Data Analysis in the Behavioral Sciences*, edited by Gideon Keren and Charles Lewis, volume 1, pp. 257–272. Hillsdale (NJ): Lawrence Erlbaum.

Knight, George P. and Alan F. Dubro. 1984. "Cooperative, Competitive, and Individualistic Social Values: An Individualized Regression and Clustering Approach." *Journal of Personality and Social Psychology* 46:98–105.

Kollock, Peter. 1998. "Social Dilemmas: The Anatomy of Cooperation." *Annual Review of Sociology* 24:183–214.

Kolm, Serge-Christophe. 2006. "Reciprocity: Its Scope, Rationales, and Consequences." In Kolm and Ythier (2006), pp. 371–541 (ch. 6).

Kolm, Serge-Christophe and Jean Mercier Ythier (eds.). 2006. *Handbook of the Economics of Giving, Altruism and Reciprocity. Handbooks in Economics 23*. Amsterdam: Elsevier (North-Holland).

Komorita, Samuel S. and Craig D. Parks. 1996. *Social Dilemmas*. Oxford: Westview.

Kopelman, Shirli, J. Mark Weber, and David M. Messick. 2002. "Factors Influencing Cooperation in Commons Dilemmas: A Review of Experimental Psychological Research." In Ostrom et al. (2002), pp. 113–156.

Kreps, David M. 1990. "Corporate Culture and Economic Theory." In *Perspectives on Positive Political Economy*, edited by James E. Alt and Kenneth A. Shepsle, pp. 90–143. Cambridge: Cambridge University Press.

Kunda, Ziva. 2002. *Social Cognition. Making Sense of People*. Cambridge (MA): MIT Press.

Ledyard, John O. 1995. "Public Goods: A Survey of Experimental Research." In Kagel and Roth (1995), pp. 111–194.

Levine, David K. 1998. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics* 1:593–622.

Lévy-Garboua, L., C. Meidinger, and B. Rapoport. 2006. "The Formation of Social Preferences: Some Lessons from Psychology and Biology." In Kolm and Ythier (2006), pp. 545–613 (ch. 7).

Liebrand, Wim B. G. 1984. "The Effects of Social Motives, Communication and Group Size on Behavior in an N-Person, Multi-Stage, Mixed-Motive Game." *European Journal of Social Psychology* 14:239–264.

Lindenberg, Siegwart. 1998. "Solidarity: Its Microfoundations and Macro-Dependence. A Framing Approach." In *The problem of Solidarity: Theories and Models*, edited by Patrick Doreian and Thomas J. Fararo, pp. 61–112. Amsterdam: Gordon and Breach.

Lindenberg, Siegwart. 2001. "Social Rationality versus Rational Egoism." In *Handbook of Sociological Theory*, edited by Jonathan H. Turner, pp. 635–668. New York: Kluwer.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks (CA): Sage.

MacCrimmon, Kenneth R. and David M. Messick. 1976. "A Framework for Social Motives." *Behavioral Science* 21:86–100.

Malinowski, Bronislaw. 1922. *Argonauts of the Western Pacific*. London: Routledge.

Masclet, David, Charles Noussair, Steven Tucker, and Marie-Claire Villeval. 2003. "Monetary and Non-Monetary Punishment in the Voluntary Contribution Mechanism." *American Economic Review* 93:366–380.

Masclet, David and Marie-Claire Villeval. 2008. "Punishment, Inequality, and Welfare: A Public Good Experiment." *Social Choice and Welfare* 31:475–502.

Mauss, Marcel. 1950. *The Gift*. New York: Free Press.

McCabe, Kevin A., Mary L. Rigdon, and Veron L. Smith. 2003. "Positive Reciprocity and Intentions in Trust Games." *Journal of Economic Behavior and Organization* 52:267–275.

McCabe, Kevin A., Veron L. Smith, and Michael LePore. 2000. "Intentionality Detection and "Mindreading": Why Does Game Form Matter?" *Proceedings of the National Academy of Sciences* 97:4404–4409.

McClintock, Charles G. 1972. "Social Motivation—A Set of Propositions." *Behavioral Science* 17:438–454.

McClintock, Charles G. and Wim B. G. Liebrand. 1988. "Role of Interdependence Structure, Individual Value Orientation, and Another's Strategy in Social Decision Making: A Transformational Analysis." *Journal of Personality and Social Psychology* 55:396–409.

McClintock, Charles G. and Eddy van Avermaet. 1982. "Social Values and Rules of Fairness: A Theoretical Perspective." In *Cooperation and Helping Behavior*, edited by Valerian J. Derlega and Janusz L. Grzelak, pp. 43–71 (ch. 3). New York: Academic Press.

McFadden, Daniel. 1973. "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers of Econometrics*, edited by Paul Zarembka, pp. 105–142. New York: Academic Press.

McKelvey, Richard D. and Thomas R. Palfrey. 1998. "Quantal Response Equilibria for Extensive Form Games." *Experimental Economics* 1:9–41.

Messick, David M. and Marilynn B. Brewer. 1983. "Solving Social Dilemmas: A Review." In *Review of Personality and Social Psychology*, edited by Ladd Wheeler and Philip Shaver, volume 4, pp. 11–44. Beverly Hills: Sage.

Messick, David M. and Charles G. McClintock. 1968. "Motivational Bases of Choice in Experimental Games." *Journal of Experimental Social Psychology* 4:1–25.

Mlicki, Pawel P. 1996. "Hostage Posting as a Mechanism for Co-Operation in the Prisoner's Dilemma Game." In *Frontiers in Social Dilemmas Research*, edited by Wim B. G. Liebrand and David M. Messick, pp. 165–183. Berlin: Springer.

Nikiforakis, Nikos. 2008. "Punishment and Counter-Punishment in Public Good Games: Can We Really Govern Ourselves?" *Journal of Public Economics* 92:91–112.

Nikiforakis, Nikos and Hans-Theo Normann. 2008. "A Comparative Statics Analysis of Punishment in Public-Good Experiments." *Experimental Economics* 11:358–369.

Nowak, Martin A. 2006. "Five Rules for the Evolution of Cooperation." *Science* 314:1560–1563.

Offerman, Theo. 2002. "Hurting Hurts More than Helping Helps." *European Economic Review* 46:1423–1437.

Ostrom, Elinor, Thomas Dietz, Nives Dolšak, Paul C. Stern, Susan Stonich, and Elke U. Weber (eds.). 2002. *The Drama of the Commons*. Washington (DC): National Academy Press.

Ostrom, Elinor and James Walker (eds.). 2003. *Trust and Reciprocity. Interdisciplinary Lessons from Experimental Research*. New York: Russell Sage.

Ostrom, Elinor, James Walker, and Roy Gardner. 1992. "Covenants With and Without a Sword: Self-Governance is Possible." *American Political Science Review* 86:404–417.

Oxoby, Robert J. and Kendra N. McLeish. 2004. "Sequential Decision and Strategy Vector Methods in Ultimatum Bargaining: Evidence on the Strength of Other-Regarding Behavior." *Economics Letters* 84:399–405.

Parsons, Talcott. 1937. *The Structure of Social Action*. New York: Free Press.

Platt, John. 1973. "Social Traps." *American Psychologist* 28:641–651.

Prosch, Bernhard. 2006. "Kooperation durch soziale Einbettung und Strukturveränderung." Mimeo (Habilitationsschrift), University of Erlangen-Nürnberg.

Pruitt, Dean G. and Melvin J. Kimmel. 1977. "Twenty Years of Experimental Gaming: Critique, Synthesis, and Suggestions for the Future." *Annual Review of Psychology* 28:363–392.

Putt, Mary E. 2005. "Carryover and Sequence Effects." In *Encyclopedia of Statistics in Behavioral Science*, edited by Brian S. Everitt and David C. Howell, pp. 197–201. New York: Wiley.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83:1281–1302.

Rasch, Georg. 1960/1980. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.

Raub, Werner. 1992. "Eine Notiz über die Stabilisierung von Vertrauen durch eine Mischung von wiederholten Interaktionen und glaubwürdigen Festlegungen." *Analyse und Kritik* 14:187–194.

Raub, Werner. 2004. "Hostage Posting as a Mechanism of Trust." *Rationality and Society* 16:319–366.

Raub, Werner and Gideon Keren. 1993. "Hostages as a Commitment Device." *Journal of Economic Behavior and Organization* 21:43–67.

Raub, Werner and Jeroen Weesie. 1990. "Reputation and Efficiency in Social Interactions: An Example of Network Effects." *American Journal of Sociology* 96:626–654.

Raub, Werner and Jeroen Weesie. 2000. "Cooperation via Hostages." *Analyse und Kritik* 22:19–43.

Rege, Mari and Kjetil Telle. 2004. "The Impact of Social Approval and Framing on Cooperation in Public Good Situations." *Journal of Public Economics* 88:1625–1644.

Rigdon, Mary. 2009. "Trust and Reciprocity in Incentive Contracting." *Journal of Economic Behavior and Organization* 70:93–105.

Robinson, W. S. 1950. "Ecological Correlations and the Behavior of Individuals." *American Sociological Review* 15:351–357.

Ross, Lee and Richard E. Nisbett. 1991. *The Person and the Situation: Perspectives of Social Psychology*. New York: McGraw Hill.

Roth, Alvin E. 1995. "Bargaining Experiments." In Kagel and Roth (1995), pp. 253–348.

Sally, David. 1995. "Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992." *Rationality and Society* 7:58–92.

Samuelson, Larry. 1997. *Evolutionary Games and Equilibrium Selection*. Cambridge (MA): MIT Press.

Sandbu, Martin E. 2007. "Fairness and the Roads Not Taken: An Experimental Test of Non-Reciprocal Set-Dependence in Distributive Preferences." *Games and Economic Behavior* 61:113–130.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Cambridge (MA): Harvard University Press.

Sefton, Martin, Robert Shupp, and James M. Walker. 2007. "The Effect of Rewards and Sanctions in Provision of Public Goods." *Economic Inquiry* 45:671–690.

Selten, Reinhard. 1967. "Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments." In *Beiträge zur experimentellen Wirtschaftsforschung*, edited by Heinz Sauermann, pp. 136–168. Tübingen: Mohr-Siebeck.

Shankar, Anisha and Charles Pavitt. 2002. "Resource and Public Goods Dilemmas: A New Issue for Communication Research." *The Review of Communication* 2:251–272.

Shinada, Mizuho and Toshio Yamagishi. 2008. "Bringing Back Leviathan into Social Dilemmas." In *New Issues and Paradigms in Research on Social Dilemmas*, edited by Andreas Biel, Daniel Eek, Tommy Gärling, and Mathias Gustafsson, pp. 93–123. New York: Springer.

Simon, Herbert A. 1957. *Models of Man*. New York: MIT Press.

Smeesters, Dirk, Luk Warlop, and Eddy van Avermaet. 2002. "Exploring the Role of Consistency of Social Value Orientations: Temporal Stability, Reciprocal Cooperation, and Forgiveness." DTEW Research Report 0238, Catholic University of Leuven.

Smith, Adam. 1759/1976. *The Theory of Moral Sentiments*. Oxford: Clarendon.

Snijders, Chris. 1996. *Trust and Commitments*. Amsterdam: Thela Thesis.

Tajfel, Henri, M. G. Billig, R. P. Bundy, and Claude Flament. 1971. "Social Categorization and Intergroup Behaviour." *European Journal of Social Psychology* 1:149–178.

Taylor, Michael. 1987/1976. *The Possibility of Cooperation*. Cambridge: Cambridge University Press. Revised edition of Anarchy and Cooperation.

Thaler, Richard H. 1992. *The Winner's Curse: Paradoxes and Anomalies of Economic Life*. New York: Basic Books.

Thibaut, John W. and Harold H. Kelley. 1959. *The Social Psychology of Groups*. New York: Wiley.

Trivers, Robert L. 1971. "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology* 46:35–57.

Tversky, Amos and Daniel Kahneman. 1981. "The Framing of Decisions and the Psychology of Choice." *Science* 211:453–458.

van Lange, Paul A. M. 1999. "The Pursuit of Joint Outcomes and Equality in Outcomes: An Integrative Model of Social Value Orientation." *Journal of Personality and Social Psychology* 77:337–349.

van Lange, Paul A. M., Wim B. G. Liebrand, David M. Messick, and Henk A. M. Wilke. 1992. "Introduction and Literature Review." In *Social Dilemmas. Theoretical Issues and Research Findings*, edited by Wim B. G. Liebrand, David M. Messick, and Henk A. M. Wilke, pp. 3–28. Oxford: Pergamon Press.

Vieth, Manuela. 2003. "Die Evolution von Fairnessnormen im Ultimatumspiel. Eine spieltheoretische Modellierung." *Zeitschrift für Soziologie* 32:346–367.

Vieth, Manuela. 2008. "Codebook of CoRe Experiments 2: Announcements and Sanctions in Trust Games. Sets of Identical (Sub)Games." Mimeo, Utrecht University.

Vieth, Manuela and Jeroen Weesie. 2006. "Codebook of CoRe Experiments 1: Promises of Trustworthiness in Trust Games. Sets of Identical (Sub)Games." Mimeo, Utrecht University.

Vieth, Manuela and Jeroen Weesie. 2007. "Trust and Promises as Friendly Advances. Experimental Evidence on Reciprocated Kindness." Mimeo, Utrecht University.

Voss, Thomas. 1982. "Rational Actors and Social Institutions: The Case of the Organic Emergence of Norms." In *Theoretical Models and Empirical Analyses. Contributions to the Explanation of Individual Actions and Collective Phenomena*, edited by Werner Raub, pp. 76–100. Utrecht: ESP.

Voss, Thomas. 1985. *Rationale Akteure und soziale Institutionen. Beitrag zu einer endogenen Theorie des sozialen Tauschs*. München: Oldenbourg.

Voss, Thomas. 1998a. "Strategische Rationalität und die Realisierung sozialer Normen." In *Norm, Herrschaft und Vertrauen: Beiträge zu James S. Colemans Grundlagen der Sozialtheorie*, edited by Hans-Peter Müller and Michael Schmid, pp. 117–135. Opladen: Westdeutscher Verlag.

Voss, Thomas. 1998b. "Vertrauen in modernen Gesellschaften." In *Der Transformationsprozess*, edited by Regina Metze, Kurt Mühler, and Karl-Dieter Opp, pp. 91–129. Leipzig: Universitätsverlag.

Voss, Thomas. 2001. "Game Theoretical Perspectives on the Emergence of Social Norms." In *Social Norms*, edited by Michael Hechter and Karl-Dieter Opp, pp. 105–136. New York: Russell Sage.

Voss, Thomas and Manuela Vieth. 2006. "Kooperationsnormen und vergeltende Sanktionen. Experimentelle Untersuchungen." Working paper 50, Institute of Sociology, Leipzig University.

Vyrastekova, Jana and Daan van Soest. 2008. "On the (In)Effectiveness of Rewards in Sustaining Cooperation." *Experimental Economics* 11:53–65.

Walker, James M. and Matthew A. Halloran. 2004. "Rewards and Sanctions and the Provision of Public Goods in One Shot Settings." *Experimental Economics* 7:235–247.

Webster, Murray. 1975. *Actions and Actors. Principles of Social Psychology*. Cambridge (MA): Winthrop.

Weesie, Jeroen. 1988. *Mathematical Models for Competition, Cooperation, and Social Networks*. Ph.D. thesis, ICS/Sociology, Utrecht University.

Weesie, Jeroen. 1993. "Social Orientations in the Prisoner's Dilemma." ISCORE paper 14, ICS/Sociology, Utrecht University.

Weesie, Jeroen. 1994a. "Fairness Orientations in Symmetric 2x2 Games." Mimeo, Utrecht University.

Weesie, Jeroen. 1994b. "Social Orientations in Symmetric 2x2 Games." ISCORE paper 17, ICS/Sociology, Utrecht University.

Weesie, Jeroen and Werner Raub. 1996. "Private Ordering." *Journal of Mathematical Sociology* 21:201–240.

Williamson, Oliver E. 1985. *The Economic Institutions of Capitalism*. New York: Free Press.

Yamagishi, Toshio. 1986. "The Provision of a Sanctioning System as a Public Good." *Journal of Personality and Social Psychology* 51:110–116.

Young, Peyton H. 1998. *Individual Strategy and Social Structure*. Princeton (NJ): Princeton University Press.

Further Acknowledgements

In addition to the acknowledgements noted for each chapter, I thank Vincent Buskens, Werner Raub, and, in particular, Jeroen Weesie for their comments on my manuscript.

Moreover, I thank Michał Bojanowski for exchanging thoughts about scientific and other issues in the past years and for his tireless help with LaTeX, to which he introduced me some weeks ago. I have gladly learned from him during each of these joint moments. For providing LaTeX files in which I could look at examples, I thank Ozan Aksoy, Roger Berger, Michał Bojanowski, Vincent Buskens, Rense Corten, Thomas Gautschi, Ben Jann, Marco Vieth, and Jeroen Weesie. I also thank Rense Corten for vivid exchanges while preparing the manuscript for printing.

Furthermore, I am grateful to Mariëlle Bedaux-de Jonge for her support with and the discussions about all kinds of issues other than the content of my research.

About the Author

Manuela Vieth studied Sociology at the Universities of Leipzig (Germany) and Bern (Switzerland), with minors in Journalism and in German Language and Literature Studies. Her previous research projects include computer simulations in the field of evolutionary game theory, a factorial online survey on sanctions in the field of social dilemma research, a study of risk perceptions of nuclear power as part of a larger postal survey, and game-theoretical lab experiments in the field of behavioral game theory. In 2004, Manuela Vieth joined the Ph.D. program of the Interuniversity Center for Social Science Theory and Methodology (ICS) in the Netherlands and developed her project "Commitments and Reciprocity" at Utrecht University. During her period as a Ph.D. student she was visiting scholar in the Sociology teams at ETH Zurich (Switzerland) in 2005 and 2007 and visited Nottingham School of Economics (UK) in 2008. Currently, she continues her research as a postdoctoral researcher at the ICS at Utrecht University.

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.

Explaining social order is of primary concern for social theories. It requires the study of sanctioning mechanisms that help enforce social norms. Punishment for misbehavior and reward for good conduct are forms of reciprocity. Reciprocal behavior can be rooted in emotions that constitute the basis of internalized social norms. Not only are motivations generated by people's concern with their own and others' outcomes but also by people's own and others' preceding behavior. Others' kind behavior induces feelings of obligation to return the favors and others' unkind behavior inflicts feelings of indignation that trigger a thirst for revenge. Furthermore, due to people's desire for self-consistency, promises and threats can intrinsically serve as a commitment. This book comprises four studies that investigate influences of these process-based motivations on trustfulness, trustworthiness, and sanctioning behavior, as well as on effects of outcome-based motivations. The focus is on trust situations and related sharing situations among two strangers. Some decision situations involve promises of trustworthiness and others reward promises or punishment threats. Two lab experiments have been conducted in order to analyze "pure" effects of preceding decisions without making specific assumptions about people's outcome preferences. The results provide evidence that preceding behavior affects subsequent decision-making and also shapes the influence of outcome-based motivations on people's behavior.

Manuela Vieth studied sociology at the Universities of Leipzig and Bern. She conducted the research presented in this book as a member of the Interuniversity Center for Social Science Theory and Methodology (ICS), where she has been located at the Department of Sociology at Utrecht University.

Universiteit Utrecht

ISBN 978-90-393-50836