UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN INFORMACIJSKE TEHNOLOGIJE

DOKTORSKA DISERTACIJA
(DOCTORAL THESIS)

VEDENJSKE LASTNOSTI SKOZI ČAS IN MED DRŽAVAMI: ESEJI O POSREDNI RECIPROČNOSTI, ZAVAJANJU TER NARODNOSTI
(BEHAVIORAL TRAITS ACROSS TIME AND COUNTRIES: ESSAYS ON INDIRECT RECIPROCITY, DECEPTION AND NATIONALITY)

ŽIGA VELKAVRH

KOPER, 2022

UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN INFORMACIJSKE TEHNOLOGIJE

DOKTORSKA DISERTACIJA
(DOCTORAL THESIS)

VEDENJSKE LASTNOSTI SKOZI ČAS IN MED DRŽAVAMI: ESEJI O POSREDNI RECIPROČNOSTI, ZAVAJANJU TER NARODNOSTI
(BEHAVIORAL TRAITS ACROSS TIME AND COUNTRIES: ESSAYS ON INDIRECT RECIPROCITY, DECEPTION AND NATIONALITY)

ŽIGA VELKAVRH

MENTORICA: PROF. DR. KLAVDIJA KUTNAR
SOMENTOR: IZR. PROF. DR. ALJAŽ ULE

KOPER, 2022

To my daughter, Eli.

Acknowledgements

I would like to express my gratitude to my mentor Prof. Klavdija Kutnar and my working mentor Prof. Aljaž Ule for their guidance and support. I could not have undertaken this journey without Prof. Aljaž Ule who provided knowledge and expertise and gave me the opportunity to run economic experiments in one of the major experimental laboratories in Europe. I would like to thank UP IAM and UP FAMNIT for accepting me as a young researcher, and their support staff. Many thanks also go to the Slovenian Research Agency (research program P1-0285, research projects J1-9110, J1-9186 and J1-2451 and Young Researchers Grant) and the University of Amsterdam (Speerpunt Behavioural Economics) for their financial support.

Special thanks go to Prof. Jernej Čopič for giving me the opportunity to work with him and to discuss the mechanism design topics. I am also grateful to Prof. Barbara Boldin, Prof. Ana Grdović Gnip, Margarita Leib, Prof. Arthur Schram and Ivan Soraperra for helping me understand various concepts and problems related to their areas of expertise.

Words cannot express my gratitude to Nataša and Aurora for their support and patience, and to Eli for the precious moments we spent together. I would also like to express my deepest gratitude to my brother Andraž for our endless conversations - mostly about sport, and to my parents Tanja and Rado for their support. Thanks should also go to Nataša's family for providing the excellent cuisine when I was working late. The family was always there for me.

Many thanks also go to my friend Mentor who went with me through many important phases of life, and my best colleagues from the faculty, Andrés, Blas and René.

Abstract

In the doctoral dissertation we first investigate behavioral strategies that subjects apply in a repeated helping game. Using the statistical mixture model-based method, we estimate that almost 90% of subjects consistently apply one of the strategies from our strategy set. In order to explain the nonstandard behaviors, we propose that previous estimations neglect an important strategy, motivated by personal experience. This strategy explains the behavior of more than half of the subjects in one of our experimental treatments. These “experientials” use strategies based on longer memory, suggesting that they are likely driven by learning and adaptation to social environment rather than emotions that trigger a strong immediate response. We also show that our subjects’ self-reports are not a reliable source of data about individuals' behavior.

Cooperation and altruism are important but when it comes to business and online trading honesty and trust become the key elements. They are investigated next, through a repeated sender-receiver type game (called deception game) where we vary the information that senders see about the past honesty of their current receiver. To study the influence of potential deception costs we compare the resulting dynamics to the dynamics of helping in helping game. Based on the evidence that reputation information increases helping, and that deception costs reduce selfishness, we expect that deception will be lower when senders can access receiver's reputation and that honesty will be higher than helping. We find support for our first expectation, but not for the second. We explore further our results by investigating strategies that subjects apply. The main finding is that rewarders who are modal in our helping game are rare in deception game.

We conclude the thesis by reporting the results of the experiment designed to detect differences in behavioral characteristics among Slovenian, Dutch and other international students. We investigate the differences using experimental measures of solidarity, trust, cooperation, positive and negative reciprocity, competition, honesty, and risk attitudes. We find that Slovenian and international students are similar whereas the Dutch students, when compared to Slovenian, are less solidary, generous and honest. They are also more often willing to take the dominant role. This points to differences in sociality between institutionally similar yet ideologically distant countries like Slovenia and the Netherlands.

Math. Subj. Class. (2020): 91-05, 91A05, 91A10, 91A90
Keywords: altruism, behavioral strategies, cross-national study, deception, deception game, experiential behavior, helping game, honesty, indirect reciprocity

Povzetek

V doktorski disertaciji najprej preučujemo vedenjske strategije, ki jih preiskovanci uporabljajo v ponavljani igri pomoči. Z uporabo statistične metode, ki temelji na modelu mešanic, ocenimo, da skoraj 90% preiskovancev dosledno uporablja eno od strategij iz naše množice strategij. Nestandardno vedenje skušamo pojasniti z dejstvom, da prejšnje ocene zanemarjajo pomembno strategijo, ki temelji na osebnih izkušnjah. Ta strategija pojasni vedenje več kot polovice preiskovancev v enem od naših eksperimentalnih okolij. Ti »izkustveniki« uporabljajo strategije, ki temeljijo na daljšem spominu, kar nakazuje, da sta njihovo vodilo verjetno učenje ter prilagajanje družbenemu okolju, ne pa čustva, ki sprožijo močan takojšen odziv. Pokažemo tudi, da samoporočila preiskovancev niso zanesljiv vir podatkov o vedenju posameznikov.

Sodelovanje in altruizem sta pomembna, vendar ko gre za poslovanje in spleto trgovanje, poštenost in zaupanje postaneta ključna elementa. Elementa sta preučevana s pomočjo ponavljane igre vrste pošiljatelj-prejemnik (imenovane igra zavajanja), v kateri spreminjamo informacijo, ki jo pošiljatelj ima o pretekli poštenosti njegovega prejemnika. Da bi preučili vpliv morebitnih stroškov zavajanja, primerjamo dinamiko poštenosti z dinamiko pomoči v igri pomoči. Na podlagi preteklih ugotovitev, da informacija o ugledu povečuje pomoč in da stroški zavajanja zmanjšujejo sebičnost, pričakujemo, da bo zavajanja manj, ko bodo pošiljatelji poznali ugled prejemnika, in da bo poštenost višja od pomoči. Naši rezultati podpirajo naše prvo pričakovanje, ne pa tudi drugega. Naše rezultate raziskujemo dalje s preučevanjem strategij posameznikov. Naša glavna ugotovitev je, da so v naši igri zavajanja nagrajevalci, ki so v naši igri pomoči najbolj pogosti, redki.

Doktorsko disertacijo zaključujemo s poročanjem o rezultatih poskusa, katerega namen je ugotavljanje razlik v vedenjskih značilnostih med slovenskimi, nizozemskimi in ostalimi mednarodnimi študenti. Razlike preučujemo s pomočjo eksperimentalnih meril solidarnosti, zaupanja, sodelovanja, pozitivne in negativne recipročnosti, tekmovanja, poštenosti in odnosa do tveganja. Ugotavljamo, da so slovenski in mednarodni študenti podobni, medtem ko so nizozemski študenti v primerjavi s slovenskimi manj solidarni, radodarni in pošteni. Prav tako so pogosteje pripravljeni prevzeti dominantno vlogo. To nakazuje na razlike v socialnosti med institucionalno podobnimi a ideološko različnimi državami, kot sta Slovenija in Nizozemska.

Math. Subj. Class. (2020): 91-05, 91A05, 91A10, 91A90
Ključne besede: altruizem, igra pomoči, igra zavajanja, izkustveno vedenje, meddržavna študija, poštenost, posredna recipročnost, vedenjske strategije, zavajanje

Contents

List of figures

List of tables

Chapter 1

Introduction and methodology

1.1 Introduction

In this doctoral dissertation we combine the methods of game theory and experimental economics to explore how people behave in interactions with unrelated anonymous individuals. Such experimental work is sometimes referred to as experimental or behavioral game theory. In game theory, a game refers to any interactive situation that involves at least two individuals called players. Any situation that involves only one individual (a decision-maker) is called a decision-making task.

After this introductory section, in which we provide motivation and outline of the doctoral dissertation, we present the methodology. We first present the experimental method in economics, with a special focus on laboratory experiments which were employed in all our research projects. Laboratory experiment is the most common type of experiments used by scientists studying economic environments. Next, we describe the experimental games along with their game-theoretic predictions. In the last part of Chapter 1 we present the experimental design and procedures of our main experiment on which the main part of the thesis (Chapters 2 and 3) is based.

In Chapter 2 we investigate altruism or, more precisely, helping behavior among unrelated anonymous individuals. Ever since human existence, strategic thinking and social interactions have played a prominent role in daily lives and helped societies to evolve and adapt to new situations and environments. An important aspect of human evolution has been prosocial behavior such as cooperation and altruism. These two types of behavior are similar in that both are costly for an individual (either monetary or psychologically) but beneficial for another individual or group, where the benefit is greater than the cost. The difference between them is that cooperation requires two or more decision-makers and is hence bilateral (or multilateral if an interaction involves more than two decision-makers), that is, the costly effort can be rewarded by other individuals in a group, whereas altruism involves one decision-maker (donor) and one or more recipients (passive individuals who do not make any decision) and is hence unilateral, that is, the costly effort cannot be rewarded.

In classical economic and biological models (e.g., Nowak & Sigmund, 1998a), altruism is efficient in that it involves a relatively small sacrifice for a donor but carries an important value for its recipient. An exchange of efficient altruistic acts between genetic relatives can be supported by kin selection (e.g., "*I help you because you are my brother*"), whereas altruism

between acquaintances can be supported by direct reciprocity ("*I help you because you have helped me*"), leading to a form of cooperation (Nowak, 2006). In large societies and particularly with the emergence of online trading and consumer-to-consumer sales, however, many interactions have moved online and occur between anonymous strangers, which lack opportunities for reciprocal exchange. In such conditions, achieving cooperation, trust and ultimately efficiency is more challenging, and new insights are needed to better understand human behavior and to create mechanisms that would improve efficiency.

In principle, the level of altruism in a society may depend on unconditional altruistic individuals who always help. They can be easily exploited, however, fare poorly and are rare (Nowak & Sigmund, 1998a; Ule et al., 2009). An alternative path to social cooperation is offered by conditional altruism which directs generosity towards strangers that are generous themselves. This behavior is again reciprocal, but in a different, indirect way ("*I help you because you helped someone else*"). This type of reciprocity is called indirect reciprocity. In particular, altruistic behavior can emerge in anonymous interactions between strangers if individual reputations can be shared, for instance through gossip. Such reputation mechanism assigns a label to each individual thus giving everyone a potentially useful information about their otherwise anonymous partners.

The type of indirect reciprocity presented above is based on reputation. This reputation-based reciprocity and especially behavioral rules that induce indirect reciprocal behavior are well studied experimentally (Seinen & Schram, 2006; Ule et al., 2009; Swakman et al., 2016). Altruism, however, can also be promoted through another type of indirect reciprocity driven by personal experience and adaptation to a social environment. This experience-based indirect reciprocity and corresponding experiential behavior are not systematically studied in laboratory economic experiments that investigate indirect reciprocity, despite the evidence that experience matters (Bolton et al., 2005; Swakman et al., 2016).

In general, the success of generous types of individuals and spread of generosity depend on the behavioral rules used by members of a population, as well as on the scope of reputation and potentially experience. In order to be able to apply the analytical results to human societies, we need to understand which behavioral heuristics people will consider, whether their behavior can indeed be captured by the specified strategies, how popular these are, and when will helping be profitable.

Although cooperation and altruism play an important role in our lives, when it comes to personal relationships, business and online trading two other determinants become more essential, namely honesty and trust. Honesty and trust among unrelated anonymous individuals are investigated in Chapter 3. Trust in the advice of strangers is an increasingly important element of market and daily interactions. This matters when incentives of interactive parties are aligned but may be hard to achieve when they are in conflict, especially if there is

information asymmetry where one party in transaction is more informed than the other. A typical real-life example is a negotiation between an informed used-car seller, who knows the exact value of her car, and an uninformed buyer who only knows some general market statistics about this type of a car, for instance, the average value of such a car. A private information (e.g., the value of a car) that is only know to one party in transaction (e.g., the seller) may give rise to profitable dishonesty at the expense of the other party (e.g., the buyer).

Given its prominence in everyday life, economics and business, dishonest behavior has been extensively studied in laboratory and field experiments over the past two decades. In experimental literature, dishonesty is a very broad term. In fact, it can be seen as an umbrella term for deception, lying and other related behaviors. In the game-theoretic context, a clear distinction between deception and lying is made by Sobel (2020), for whom a lie is simply a statement (about the private information) that a liar believes is false, whereas deception is a statement that induces in others incorrect beliefs about private information that a deceiver has. In the thesis we focus on deception because it is present in many daily situations that involve two anonymous parties, where one is more informed that the other. Examples include used-car sales, online trading, and consumer-to-consumer sales. To date, little is known how honesty, and any psychological costs associated with deception, develop with time, especially in the presence of reputation mechanisms which are very important as they mimic the real-life mechanisms such as gossiping and information sharing and spreading. On the one hand, a truth-telling norm may emerge with social sanctions imposed on deceivers. For example, individuals may sanction the deceivers either by not trusting them or by deceiving them when the roles are reversed. On the other hand, reservations against deception may disappear after substantial experience of dishonesty, unravelling any previous trust in strangers in a society. Related phenomena have been predicted and observed in the theoretical and experimental research on indirect reciprocity in helping games (Nowak & Sigmund, 1998a; 1998b; Seinen & Schram, 2006). In those, reciprocity may promote the development of generosity between strangers, but also lead to a vicious cycle of retaliation.

A large part of Chapter 3 is devoted to the identification and exploration of behavioral strategies that subjects apply in deception game, especially when access to reputation information is provided. To date, literature is agnostic about that, and we are curious whether indirect reciprocal behavior, detected in environments with altruistic opportunities, translates to environments that offer opportunities for honesty. We investigate both the reputation-based and experience-based indirect reciprocity.

Honesty is certainly not identical to generosity, but with the right game structures and experimental design it is possible to draw certain parallels between them and compare them. However, honesty may be driven by more than just social preferences, which Gneezy (2005) demonstrated comparing deception games and dictator games with identical payoffs. He

proposed that while generosity is motivated by social preferences, honesty is driven by both social preferences and aversion to deception, and this has been confirmed by a number of studies on deception (e.g., Sutter, 2009; Erat & Gneezy, 2012; Vranceanu & Dubart, 2019). The long run interaction of these two distinct behavioral motivations is still unclear, and one aim of this study is to investigate whether any aversion to deception in one-shot interactions persists over time as groups learn and adapt. Our design will therefore allow comparison between the long run helping behavior in the repeated helping game and the long run honesty in the repeated deception game with identical payoffs. As honesty of senders may depend on the past honesty of their current receiver, we will also investigate the role of information that senders have about past actions of their receivers. In both the repeated helping and the repeated deception games we will implement the stranger matching protocol, where interacting pairs randomly rematch in every period, but provide information to the senders about the past honesty or generosity of their receivers. With this matching and such reputation information the helping game becomes identical to the games used in the literature on indirect reciprocity and we will provide its detailed investigation in Chapter 2. The deception game with such matching and reputation has not yet been investigated and we focus on it in Chapter 3.

Chapter 3 concludes the main part of the doctoral dissertation devoted to the investigation of altruism, honesty and trust among unrelated anonymous individuals who are given repeated opportunities to engage in helping or honest and trusting behavior. The last part of the doctoral dissertation, Chapter 4, takes us on completely different route, as it focuses on cross-national study where experimental subjects engage in a series of one-shot economic tasks. In particular, in the last part of the doctoral dissertation, we investigate the role of nationality in eight standard economic tasks used to measure solidarity, trust, cooperation, positive and negative reciprocity, competition, honesty and risk attitudes. These eight types of sociality measures play an important role in our lives, as they dictate many of our economic and personal decisions. They are also one of the reasons why our behavior often deviates from the behavior of the *homo economicus* considered in the classical economic and game-theoretic theory.

Cross-national and cross-cultural studies are attractive but challenging to execute, because researchers must control for potential currency, experimenter, and language effects (Roth et al., 1991; Thöni, 2019). There are many cross-country experimental studies comparing sociality measures, but most of them focus on just one or few of the above eight sociality measures. To date, the literature provides mixed evidence as to whether a nationality or culture component is significant. Furthermore, at least to our knowledge, only one of them focuses on the differences in behavioral characteristics between Slovenian and other nationality groups. In particular, Roth et al. (1991) compared the bargaining behavior in Israel, Japan, Slovenia and United States and found that Slovenians proposed more generous offers than Japanese and Israelis. Our main objective is to conduct a comprehensive study with students in order to detect differences in behavioral characteristics between Slovenian and other nationality student

groups. In general this is a difficult and costly task, as it requires international cooperation of researchers. Luckily, we have established a cooperation with University of Amsterdam, giving us an opportunity to run experiments in both countries and to directly compare the behavior of Slovenian students with the behavior of Dutch and other international students. Our main purpose is to see whether significant differences between Slovenian and other groups exist, and if yes, on which dimensions. Our broader objective, however, is to bring experimental economics closer to general audience in Slovenia and present its role in science. Finally, our results can also be used in meta-analyses which often miss the data about Slovenia.

To summarize, the doctoral dissertation is organized as follows. In Chapter 2 we investigate altruism and helping behavior among strangers. Chapter 3 is devoted to honesty (deception) and trust among strangers. Chapter 4 presents a cross-national study on sociality measures. Chapter 5 concludes the dissertation.

1.2 Generosity and honesty in economics: Methodology

In our daily life, we often read about, hear about, or observe acts of generosity and honesty towards complete strangers. Unfortunately, the real life often offers neither enough opportunities nor information to systematically analyze, for example, individuals' behavior, its dynamics over time, or the effect of communication. An alternative approach that solves the problem of missing or duplicate data, while ensuring complete control over all variables, is experimental approach that is now a standard practice among scientists who study economic interactions. All three studies in this thesis obtain their empirical data via this approach. Paragraphs below, up to the subsubsection *Laboratory experiments: Strengths and weaknesses*, are a minor modification of the introductory section in Velkavrh and Ule's (2022) paper entitled "*Indicators of human sociality in Slovenia and the Netherlands: Evidence from experiments with students*" that was published in Teorija in praksa journal (doi: 10.51936/tip.59.2.487-508).

1.2.1 The experimental method in economics

Experiments are the original way of doing research in the natural sciences. In contrast, since the mid-20th century in the social sciences experiments have only been the key method of research in psychology. The Asch conformity experiments and Milgram experiments about hierarchical submission had a strong impact on both the expert and general public in the 1960s (M. Ule, 2004). In economics, the experiments initially focused on testing the standard assumptions about the efficiency of free markets (V. L. Smith, 1962), but eventually covered the general area of human decision-making, overlapping with fields such as psychology, social psychology, and evolutionary biology.

Experiments, the laboratory in particular, are valuable because they offer an important tool for both basic and auxiliary research that is able to yield important, systematic, controlled and highly replicable insights into social human behavior. They are carefully designed, carried out and analyzed so that usually several years pass from the initial idea to the publication of the results of an experiment. In standard experiments, including ours, subjects (volunteers) from the same population (i.e., subjects who register for an experiment) are randomly assigned to different experimental conditions/environments called treatments, whereby the treatments that are compared differ in only one factor that is being studied (e.g., reputation information, punishment, pre-play communication). The random assignment assures that groups of volunteers exposed to different treatments are ex-ante similar in every respect and that the studied factor (i.e., the effect of experimental condition) is random and uncorrelated with everything that might be omitted, meaning that there is no omitted variable bias problem which is often present when a researcher is dealing with observational data. So, given that all subjects come from the same population and are randomly distributed across treatments, the distribution of demographic and unobserved variables should be similar across treatments too. In addition, different sessions for the same treatment are usually ran at different times and days, to further minimize the possibility that the results are caused by some uncontrolled environmental factors. For example, by running sessions with different treatments in random order rather than one after another, researchers assure that subjects who registered for the experiment first are not all exposed to the same treatment. If we instead ran sessions with different treatments sequentially, one after another, then the difference in behavior could occur because those "early birds" had experimental experience, were more profit-oriented, etc.

Experiments can position subjects – recruited using standard protocols (e.g., an online recruitment system) – in real social or economic situations where each decision holds real social or economic consequences for all involved. When these situations mimic real-life conflicts and trade-offs, they raise real moral dilemmas, which offers an insight into non-hypothetical values and actual human decision processes. A typical example of such an approach is experimental economics using game theory to design simple versions of actual economic dilemmas and offering performance-based monetary incentives for the realism of decisions. Subjects therefore volunteer for experiments in order to earn money. A design of this nature can increase both the internal and external validities of laboratory experiments for the social sciences (Hertwig & Ortmann, 2001; Schram, 2005).

When subjects participate in only one treatment (session) we say that an experiment has a "between-subjects" design. In contrast, when subjects participate in different treatments, we say that an experiment has a "within-subjects" design. Both have their advantages and disadvantages (see, e.g., Greenwald, 1976; Charness et al., 2012). For example, a between-subject design is better if one wants to avoid the learning or carry-over effect (according to which the exposure to one treatment affects subjects' behavior in subsequent treatments) or

reduce demand effect (according to which subjects form beliefs about experimenters' expectations and then behave accordingly to satisfy these expectations). It, however, requires more subjects (new subjects for each treatment) and individual differences may confound the results. In general, the decision about the type of design depends largely on the research interests and questions.

The main advantage of the experimental method is a clear, efficient, transparent, aggregate and reliable tool for detecting causal relationships (e.g., Ule & Živoder, 2018). In the social sciences, it facilitates exact analyses of phenomena up to a medium scope such as interpersonal relations, conformism, biased judgement, and social exchange. Controlled experimentation has in recent decades thus led to substantial conceptual revolutions in several social disciplines. Economists have developed theories of prosocial motives that are not driven by individual market success, political scientists have developed and then criticized the theory of rational electoral choice, while communication scientists have engaged in a systematic exploration of the influence process (Webster & Sell, 2014). One additional advantage offered by experiments is replicability of the decision environment across different locations like cities, countries and cohorts. This facilitates cross-cultural research that is low on noise and confounds.

Laboratory experiments: Strengths and weaknesses

In this section we take a closer look at laboratory experiments which are the most common type of experiments used by experimental economists and discuss their main advantages and limitations. The main advantage of running an experiment in laboratory is control over the environment where subjects interact and make decisions. For example, in laboratory experiment researchers have control over individuals' incentives and payoffs (by assigning values to outcomes or objects) and experimental environments (e.g., by setting market rules, determining institutions, choosing reputation mechanisms). Since researchers have complete control over experimental environments, they can directly examine how changing a single factor (i.e., reputation information, the type of matching protocol) affects the behavior of subjects. So, experiments provide efficient tests for causality. One particular strength of the laboratory experiment is that it offers the opportunity to gather data that is otherwise not available in the field, for example the data about individuals' beliefs or risk, intertemporal and social preferences, which helps researchers measure individuals' rationality and generosity, or risk and lying/deception aversion, i.e., the tendency of people to prefer honesty to lying/deception. Another advantage of laboratory experiments is that they are replicable. Replications are essential for science because they either confirm the previous results making them robust or dispute them resulting in further replications of the experiment and further study of the problem. Laboratory experiments can also be used to test the effect and efficiency of various reputation systems, types of auctions or policy proposals before they are actually implemented in practice, thus preventing the potential inefficiencies due to ill-designed

systems, mechanisms or policies. Last but not the least, laboratory experiments provide control over matching protocols and allow the implementation of random events. For example, laboratory experiments allow researchers to make the matching between the subjects random if the theory they are testing assumes so, and allows the implementation of risky lotteries or die roll simulations based on which risk preferences and lying aversion can be measured, respectively (Schram & Ule, 2019).

By having a complete control over the experiment, the researchers can make clear causal inferences from the experimental results, which means that laboratory experiments (if properly designed) have high *internal validity*. High internal validity, however, usually requires the experiment to be abstract and simple, as otherwise the research may not be tractable. This may negatively affect the *external validity* of the experiment which corresponds to the extent to which the experimental results are generalizable outside the laboratory, i.e., to real-world situations. Ideally, a researcher wants to design an experiment with high internal and external validity. In practice there is usually tension between the internal and external validity. In general, laboratory experiments seem to provide higher internal validity and lower external validity, field studies higher external validity and lower internal validity, whereas the internal and external validity of field experiments falls somewhere in-between (Schram, 2005).

There are several potential reasons for a relatively low external validity, which seems to be the major limitation of laboratory experiments. For example, they usually have lower number of observations than field studies, so that researchers must resort to non-parametric statistics. Then, it is also possible that due to the artificiality and "coldness" (e.g., the rigor and formality of experimental procedures, seriousness of experimenters) of experimental design, experimental subjects might not behave in the same way as they would in the analogous real-life situation (Schram & Ule, 2019). Furthermore, using students as experimental subjects, despite being standard in experimental economics, may not be ideal for external validity, as their behavior may deviate from the behavior of a typical (randomly chosen) member of a community (e.g., Carpenter et al., 2008; Anderson et al., 2013; Falk et al., 2013). Also, student subject pool is not the best choice if one wants to explain and understand the general behavior of certain groups of people like CEOs, traders, children or pensioners. Anyway, often we do not predict real world behavior from one experiment but study causality and effects of environmental variables using treatment comparison. This is where it is important that subjects in all treatments come from the same population, more than which population they come from. We do not predict the magnitude of real-world effects from experiments, but instead investigate its direction and significance.

Measuring dishonesty

Given that a substantial part of the thesis is devoted to dishonesty, which is in the experimental literature a very broad term, we dedicate this short subsubsection to dishonesty where we

clarify the term and describe how it is typically measured in experiments. Dishonesty is in the experimental literature an umbrella term for deception, lying and other related behaviors. To date, researchers have proposed many different paradigms that can be used to measure dishonest behavior, among which the following four are the most standard and all discussed in the recent meta-analysis by Gerlach et al. (2019): sender-receiver or deception games à la Gneezy (2005), die-roll tasks à la Fischbacher and Föllmi-Heusi (2013), coin flip tasks à la Bucciol and Piovesan (2011), and matrix tasks à la Mazar et al. (2008).

A *sender-receiver* or *deception game* involves two players, an informed sender who knows the payoff scheme of the game (i.e., the available options and the corresponding payoffs, e.g., option A earns 2 EUR to a sender and 3 to a receiver, options B, C and D all earn 3 EUR to a sender and 2 to a receiver) and an uninformed receiver who only knows the available options (e.g., there are options A, B, C and D). In this game a sender sends a message to a receiver informing him which option will earn him the most money. The sender can send an honest message by providing true information (e.g., informing the receiver that option A will earn him the most money when option A indeed earns him the most money) or a deceptive message by providing false information (e.g., informing the receiver that option B, C or D will earn him the most money when option A in fact earns him the most money). The receiver, upon observing the sender's message, chooses one option which determines the payoffs of both the sender and the receiver. In this game the sender who believes that the receiver will choose the recommended option faces a dilemma between sending an honest message and a more profitable deceptive message.[1] In a *die-roll* task subjects are instructed to report a number they roll on a die. Since subjects are instructed to roll a die in private, they are the only one who observe the number, and hence may lie by reporting another number. As their payoff depends on the reported number and not on the observed number, they are incentivized to report the most profitable number even if they roll another one on their die. The degree of (dis)honesty is then determined at the aggregate level by comparing the distribution (or just the average) of the numbers reported with the expected uniform distribution of the numbers observed. A *coin-flip* task is like a die-roll task, except that in a coin-flip task subjects are instructed to report the outcome of a coin toss, so there are only two possibilities (e.g., heads and tails, 0 or 1) and not multiple as in a die-roll task (e.g., 1 to 6). In a *matrix task* subjects are provided with several matrices, each containing twelve three-digit numbers such as 2.74. In each matrix, subjects are then instructed to find the unique two numbers whose sum equals 10.00 (e.g., 2.74+7.26=10.00). If they find such pair, the matrix is "solved". There is however the time limit (e.g., 4 minutes for 20 matrices), making it extremely difficult to solve all matrices in a

[1]The sender who believes that the receiver will *not* choose the recommended option also faces a dilemma between sending a deceptive message and a more profitable honest message. This type of sender is however less commonly expected, especially if there are many options available.

given time. The more matrices they solve, the higher the payoff. In a standard experiment with the matrix task, subjects are randomly divided in two groups: control and experimental group. In the control group the number of solved matrices is counted by the experimenter, so the results of this control group generate the distribution of solved matrices under honest reporting. In the experimental group, when the time runs out, the subjects are instructed to report the number of matrices they solved which determines their payoffs. Since in the experimental group the subjects are the only one who know the true number of solved matrices, they may lie by reporting another number. So, the subjects have an incentive to claim that they solved all matrices even if they did not solve them all. The degree of (dis)honesty is then determined at the aggregate level by comparing the distribution (or just the average) of matrices solved across two groups.

In all four experimental paradigms there is information asymmetry and temptation to be dishonest. Unlike the deception game, the other three paradigms involve no interaction between the subjects. So, they are merely decision-making tasks. As such, the deception game is used to measure a different type of dishonest behavior, namely deception, whereas the other three are used to measure lying. In the literature there are different definitions of deception and lying and many overlap. Deception can be defined as "…an act that is *intended* to foster in *another person* a belief or understanding which the deceiver *considers* false" (Zuckerman et al., 1981, p.3, emphasis in original). Lying refers to "making a statement believed to be false, with the intention of getting another to accept it as true" (Primoratz, 1984, p. 54n2). In the context of experimental games we consider deception as an act that requires (at least) two interacting experimental subjects with one (a deceiver) seeking to manipulate the *beliefs* of the other (a victim of the deceit), whereas we consider lying simply as misreporting of private information as in a die-roll, coin-flip and matrix tasks. While deception directly affects another experimental subject (by manipulating his beliefs which may affect his decision), lying affects only the liar and none of the other experimental subjects. Similar distinction between deception and lying is made by Sobel (2020), for whom a lie is simply a statement (about private information) that a liar believes is false, whereas deception is a statement that induces in others incorrect beliefs about private information that a deceiver has, e.g., about a realized (privately observed) event. In the standard versions of the paradigms, the dishonest behavior can be detected at the individual level only in deception game, as in other paradigms, the die-rolls, coin-flips and matrices solved are observed in private. If an experiment is computerized, though, and a program stores the actual die-roll, coin-flip or the number of matrices solved, the dishonesty can be measured at the individual level too. The advantage of the die-roll and matrix task over the other two is that they can measure the degree of dishonesty and not just whether there is a significant dishonesty or not. In general, subjects might not be dishonest to full extent. For example, by rolling number 1 in the die-roll task a subject can lie to the full extent by repoting number 6 or lie to a lesser extent by reporting some other number which also increases

her payoff but potentially raises less suspicion. Similarly, in the matrix task if a subject solves 5 matrices, she can lie to the full extent by stating that she solved all 20 matrices, which is extremely unbelievable but maximizes her payoffs, or she can lie to a lesser extent by stating that she solved 10 matrices which still increases her payoffs yet seems plausible. In deception games and coin-flip tasks a subject can only be honest or dishonest to the full extent. The disadvantage of the matrix task is that misreporting of the number of solved matrices is not necessarily a sign of dishonest behavior, like in other paradigms, as misreporting in the matrix task might happen because subjects truly believe that the numbers they circled sum to 10.00, or they made a counting error while counting the number of solved matrices. Moreover, in the matrix task, for some subjects the primary reason for lying may not be higher payoff but the desire to appear competent (Gerlach et al., 2019).

1.2.2 Game theoretic models of rational economic interaction

Many real-life situations and dilemmas can be translated into simplified decision-making tasks or games. In this section we formally introduce two games used in our main experiment whose results are reported in the main part of the thesis (Chapters 2 and 3). Chapter 4 of the thesis will describe the experiment based on six classic economic games and two economic tasks, and we will describe that experiment and the corresponding games in that chapter.

In our two games experimental subjects were earning *francs* which was our experimental unit. At the end of the experiment francs were translated into money the subjects earned at a constant exchange rate, so that the private interest of subjects who volunteered for experiments to earn money is to have as many francs as possible.

Experimental games

Helping game

In game-theoretic language individuals are usually called players or agents. A helping game is a simple strategic game involving two players, a *sender* and a *receiver*. In the literature (e.g., Nowak & Sigmund, 1998a; Seinen & Schram, 2006), players of a helping game are usually called a donor and a recipient, respectively, but we use terms “sender” and “receiver”, to be consistent with the terminology of our second (deception) game described below. Consistency will make the text in subsequent chapters, when the behavior of subjects across games will be compared, easier to read and understand. Throughout the thesis, for the purpose of making the text more readable, the sender will be referred to as “she” or “her” and the receiver to as “he” or “him”.

In a helping game only the sender makes a decision, whereas the receiver takes a passive role. The sender has two available actions: either helps the receiver or passes. If she helps, she incurs the cost of 150 francs and the receiver earns 250 francs. If she passes, neither player earns nor loses francs. Helping is therefore socially efficient, but costly for the sender (see Figure 1).

Sender's decision	Payoffs Sender, Receiver
Help	-150, 250
Pass	0, 0

In each cell with numbers, the first number corresponds to the sender's payoff and the second to the receiver's payoff. Payoffs are expressed in francs which was our experimental unit. Francs earned were at the end of the experiment translated into money at a constant exchange rate.

Figure 1: Helping game.

If the game is played only once, the standard game-theoretic prediction is that sender will pass, as this is costless. The prediction remains the same even if the game is repeated finitely many times where the players in every round meet a different randomly chosen player from a population of players. This holds because in this case the finitely repeated game is nothing but the collection of finitely many distinct one-shot helping games. Then, by backward induction, all senders in the last round have an incentive to pass, regardless of the outcomes in previous rounds; knowing that all senders will pass in the last round, in the next-to-last round senders have also an incentive to pass, regardless of the outcomes in previous rounds; and so on the reasoning proceeds to the first round where all senders also have an incentive to pass.

The prediction may change towards more socially desirable behavior, however, if one assumes that players interact for an unknown number of rounds and discount the future (Dilmé, 2016; Camera & Gioffré, 2022), or if the environment allows for reputation building, which helps individuals distinguish between generous and selfish individuals (Nowak & Sigmund, 1998a; Leimar & Hammerstein, 2001).

A helping game was developed to study indirect reciprocal behavior (generosity and selfishness, in particular) in repeated interactions where helping is individually costly but socially beneficial. In large societies and particularly with the emergence of online trading, especially after Covid-19 crisis, many interactions have moved online, occur between anonymous strangers and are asymmetric (e.g., one buys a product, the other sells it; one needs technical help, the other provides it). Given the high number of online users and their daily-changing demands, the probability of the same two individuals meeting again in the future is low, so the opportunities for direct reciprocal exchange are rare. In such conditions, achieving cooperation, trust and ultimately efficiency is more challenging. One way to promote the exchange of goods is via indirect reciprocity which gained interest among economists, biologists, and other scholars because, to function properly, it only requires that people know the reputation of a person they meet. Reputation describes the general perception of actions this person took recently. This mechanism does not require i) that interacting individuals are genetically related as kin selection does, ii) repeated interactions between the same individuals as direct reciprocity does or iii) special network topologies as network reciprocity and group

selection do. Nowadays, reputation systems are widely used in electronic commerce and, in the form of customer reviews, for instance on websites dedicated to lodging and tourism activities.

Deception game

The deception game is a dynamic game with incomplete information used to study information transmission between two players, a privately informed *sender* and an uninformed *receiver* (Crawford & Sobel, 1982; Gneezy, 2005). In the standard setting, a sender privately observes the state of the world and informs about that, via costless message, a receiver who knows nothing about the state of the world except the prior probability of its realization. Upon receiving a message, the receiver takes an action which determines the payoffs for both players. Players' incentives are often misaligned, i.e., their utilities are different and action that is ex-post best for the receiver is not ex-post optimal for the sender. The sender is therefore motivated to misrepresent the true state with the goal to deceive the receiver into choosing the action that is optimal for her but not for the receiver.

Our particular game has eight possible states, $\theta \in \{\theta_A, \theta_B, \dots, \theta_H\}$, each of which is realized with equal prior probability 1/8. Each state is associated with a unique payoff scheme that consists of eight payoff allocations (options), of which one is better (worse) than the other seven payoff-equivalent allocations for the receiver (sender). One such payoff scheme is illustrated in Table 1 and corresponds to the state Θ_F. If state Θ_i, for $i \in \{A, B, \dots, H\}$, is realized, then option *i* brings 250 francs to the receiver and a 150 francs loss to the sender while all other options bring no gains or losses to both the sender and the receiver. The receiver therefore benefits if option *i* is chosen but the sender prefers any other option, inducing a conflict of interest between them.

This game has several steps. In step 1, nature (e.g., a computer) randomly choses one state and reveals that information to the sender only. In step 2, the sender sends a message $m \in \{m_A, m_B, \dots, m_H\}$, informing the receiver about the realized state, where m_i has a natural meaning that "*the realized state is state* Θ_i*"*. This is essentially the same as saying "option *i* will bring you, the receiver, 250 francs", so the message can be interpreted as a recommendation to the receiver which option he should choose. Note that in her message the sender can either reveal the true state, in which case the message is *honest*, or provide false information, in which case the message is *deceptive*. In step 3, the receiver receives sender's message and chooses an option (action) $a \in \{A, B, \dots, H\}$ that determines the payoffs for both the sender and the receiver. Note that the receiver can either *trust* the message and choose the option recommended in the sender's message, or he can *doubt* that the message is honest and choose one of the options not recommended in the sender's message. If the receiver chooses option *i* when the state is Θ_i, the sender incurs the cost of 150 francs and the receiver gains 250 francs. Otherwise, both receive 0 francs. So, if the realized state is Θ_F, and the receiver chooses

Option	Payoffs (sender, receiver)
A	(0, 0)
B	(0, 0)
C	(0, 0)
D	(0, 0)
E	(0, 0)
F	(-150, 250)
G	(0, 0)
H	(0, 0)

Table 1: Payoff scheme of the game if state Θ_F is realized.

option *F*, the sender incurs the cost of 150 francs whereas the receiver gains 250 francs. In contrast, if the receiver choses any other option than option *F*, both get 0 francs (see Table 1). In this game the sender's message is a "cheap-talk" message, meaning that it is costless, and that it does not directly determine the payoffs, although it may influence the receiver's beliefs and his subsequent action.

In this game, guessing the true state (i.e., choosing option *i* when the realized state is Θ_i) leads to a socially efficient outcome – in social value terms, since a gain of 250 minus cost of 150 (i.e., 100) is better than 0. This is, however, rather difficult to achieve because the sender has no incentives to reveal the true state as that would result in correct guess by the receiver, in which case she (the sender) would incur the cost of 150 francs. Since the sender's message will not reveal the true state and will be thus uninformative, the receiver will ignore her message and try to guess the true state based on his prior information that any state can be realized with probability 1/8. Such games are typically examined through the sequential (perfect Bayesian) equilibrium analysis. On the one hand, our game does not have an informative (separating) equilibrium in which the sender chooses a different message for each state, and the receiver chooses the option that maximizes his payoff for each message, resulting in correct guess. To see this, note that if the sender knew that the receiver would trust her message, she would rather send a deceptive message and thus increase her payoffs from -150 to 0 francs. On the other hand, our game has several uninformative (pooling) equilibria, in which the sender chooses the same message (or randomizes between all of them) regardless of the realized state, and the receiver chooses the same option (or randomizes between all of them) regardless of the message sent. If the sender always chooses the same message or randomizes between them, the receiver cannot update his prior beliefs and hence believes that any of the states was realized with probability 1/8. Under such beliefs, each of his options generates the same (maximal) expected utility of 250/8 (so, he plays a best response to his beliefs and sender's choice). Knowing that the receiver will ignore sender's message, the sender will get the expected utility of -150/8, regardless of her message, so she can choose either of the messages or randomize over them (so, she plays a best response to receiver's choice). Moreover, if the game is repeatedly played among strangers, as in our experiment, the equilibria remain the same, because each round can

be treated as one-shot game, in which the sender does not have an incentive to send an informative message and the receiver does not have an incentive to condition his action on sender's message.

Our deception game is a variant of games used in Crawford and Sobel (1982) and Gneezy (2005), but is more closely related to Gneezy's which is simpler in terms of the number of monetary allocations (there are only two) and designed to be as close as possible to the dictator game, which enabled a direct comparison of behavior between the games and the estimation of deception aversion, i.e., the tendency of people to prefer honesty to deception. Our game differs from that of Gneezy (2005) in two aspects. First, our state set has eight states instead of two. This is done to avoid the sophisticated deception through honesty that occurs with higher probability when the state space is small. A sophisticated deception through honesty is a situation where senders maximize their expected payoffs by telling the truth because they believe their receiver will doubt the message and choose another option (Sutter, 2009). Second, in our game both players know the payoff scheme, whereas in Gneezy (2005) the receivers do not know the payoff scheme, that is, do not know that the incentives are misaligned. This is done because in our experiment the game is played for 100 rounds, not just once as in Gneezy's (2005) experiment, and players would figure out the payoff scheme anyway as soon as they become senders for the first time (on average, this occurs in the second round). Our game differs from that of Crawford and Sobel (1982) in many aspects, for example 1) the state, message and action sets are finite in our model, whereas in their model the state and message sets are [0, 1]-intervals with the Lebesgue measure and the action set is $\mathbb{R}$; 2) in our model, the state is a discrete random variable with probability $P(\Theta=\Theta_i)=1/8$, whereas in their model the state is a random variable with density on the interval [0, 1]; and 3) in our model, the utility functions are not continuous with respect to state and action chosen, whereas in their model they are continuous and twice continuously differentiable.

Our deception game can be simplified to a 2x2 simultaneous game, presented in Figure 2, assuming that i) senders do not use a different strategy for different states but decide only whether to send the honest message (by revealing the true state) or a deceptive message (by providing false information), whatever the state, and ii) receivers do not use a different strategy for different messages but decide only whether to trust or doubt a message, whatever its content, where trusting the veracity of the message means choosing according to the sender's recommendation and doubting the veracity of the message means choosing one of the options not recommended in the message. If the sender sends a deceptive message, she does not care which false option she recommends in her message as along as her behavior is independent of the realized state. Given that all states are equally likely the rational receiver only needs to consider his beliefs about the honesty of the sender. He should choose the option recommended in the message if this belief is sufficiently high and otherwise choose any other option. In this interpretation of the deception game, both players have two available actions. The sender can

	Trust	Doubt
Honest	-150, 250	0, 0
Deceptive	0, 0	-150/7, 250/7

Each cell in the table corresponds to unique combination of actions chosen by a sender and a receiver. Each cell shows the resulting (sender's, receiver's) payoff.

Figure 2: Deception game – a simultaneous move variant.

send the *honest* or a *deceptive* message while the receiver can *trust* or *doubt* the veracity of the message.

If the sender sends the honest message and the receiver trusts the message, the sender incurs the cost of 150 francs and the receiver earns 250 francs. If the sender sends a deceptive message and the receiver doubts the message, the receiver guesses the true state with probability 1/7, because he chooses one of the seven not recommended states of which one is the true state. In this scenario, the sender incurs the expected loss of 150/7 francs and the receiver receives the expected gain of 250/7 francs. In the remaining two cases both earn 0 francs, as trusting a deceptive message or doubting the honest message results with probability 1 in the receiver choosing an option yielding 0 francs to both players. This game can now be analyzed using the standard Nash equilibrium analysis. The Nash equilibrium prediction for this game is that the sender will send the honest message with probability 1/8 and that the receiver will trust the message with the same probability, 1/8. Their equilibrium expected earnings are -150/8 francs (sender) and 250/8 francs (receiver).

This game has only one Nash equilibrium in which the sender is honest with probability 1/8 and the receiver trusts with probability 1/8. Since the finitely repeated game between strangers can be seen as the collection of finitely many distinct one-shot games, the prediction for such repeated game remains the same, namely that after each outcome (history) all players will play equilibrium strategies, that is, will be honest and trust with probability 1/8. The prediction may change if we assume that some senders are deception averse (or have social preferences) in which case honesty and trust are likely to be seen throughout the repeated game. Repeated deception game can be used to study the dynamics of both honesty and trust between strangers.

Game comparison

In this research project we are mainly interested in generosity, honesty and deception aversion which are measured by analyzing the behavior of senders. Our research on generosity and honesty addresses several research questions by examining either the helping game data or deception game data alone. However, certain research questions require a direct comparison of senders' behavior between the games, which requires that the two games, or more precisely the experimental designs of the two games, are comparable. As described in later sections, our

experiment was indeed designed to study both games with as similar interface as possible. We do not compare the behavior of receivers between our deception and helping game, since in the helping game receivers make no decisions.

Consider first the differences between the helping game and our original "non-simplified" deception game. One major difference is that in the helping game the sender determines the payoffs (i.e., monetary allocation) and the receiver is passive, whereas in the deception game the receiver determines the payoffs, while the sender just sends a "cheap-talk" message that at most influences receiver's beliefs about which option brings him 250 francs. Therefore, the receiver is active in deception game causing the sender to think strategically. In contrast, the receiver is passive in the helping game, so the sender needs not consider any beliefs about the receiver's action. In particular, while the profit-maximizing sender should always pass in the one-shot helping game, the profit-maximizing sender in deception game might choose any message, depending on her beliefs about the receiver behavior. First, if she believes that the receiver will trust her message, she should send a deceptive message. Second, if she believes that the receiver will surely doubt her message, she should send the honest message. Third, if she believes that the receiver will completely ignore the message and always choose the same option or randomize between them, then she is indifferent and may always send the same message or a random message.

The sender-receiver experiments showed that receivers trust the recommended options more often than predicted by the theory (e.g., Cai & Wang, 2006; Sánchez-Pagés & Vorsatz, 2007). If the senders anticipate that the receivers are overly trusting, i.e., that the chance of trust is over 1/8, then they should always send a deceptive message if they want to maximize their own payoff and the honest message if they want to maximize the receiver's payoff (or the sum of payoffs, i.e., social welfare). This implies that sending a deceptive message is the same as passing in the helping game as both maximize sender's payoff, and sending the honest message is the same as helping in the helping game as both maximize receiver's payoff or social welfare. So, under the assumption that senders believe that receivers are overly trusting, the senders in both games are confronted with the payoff-equivalent decisions.

It is easier to see the analogy between the decisions of senders in the helping and deception game by looking at the helping game (Figure 1) and at the simplified deception game (Figure 2). The only difference between these two games (besides labelling) is that in the deception game the receiver can react to the sender's decision (by doubting the message), whereas in the helping game he can not – he can only accept the sender's decision. However, if senders in the deception game believe that the chance of trust is over 1/8 (i.e., the equilibrium probability), then their best response to such a belief is equivalent to their best response to a belief that the chance of trust is 1. This means that the sender when making her decision essentially considers only the left column (labelled "trust") of the deception game in Figure 2. Such assumption

regarding the senders' beliefs is reasonable and can be justified. First, because of the past experimental evidence cited above. Second, as will be revealed in Results section of Chapter 3, because in our DREP and DBASE the receivers in almost every group and in almost all ten-round blocks on average trusted more than 1/8 of the messages – which senders could learn through personal experience.

1.2.3 Generosity and honesty in laboratory

In Chapters 2 and 3, i.e., the main part of the doctoral dissertation, we investigate generosity/selfishness and honesty/deception among strangers. These two chapters are based on the data gathered from the same experiment and we describe the experimental design and procedures of our main experiment in this section rather than in those chapters. Before describing them, we briefly present the matching protocol that we employed in our experiment.

Matching protocols

In our experiment subjects play a repeated game. In general, a subject can play a repeated game either with the same fixed subject (i.e., a partner) or with different randomly chosen subjects (i.e., strangers). The first matching protocol is usually called a fixed or partners' matching protocol, whereas the second is usually referred to as a random or strangers' matching protocol. In the following we briefly describe the strangers' matching protocol, since it was employed in our experiment.

Under a strangers' matching protocol, a repeated game can be interpreted as a sequence of one-shot games. This protocol prevents reputation building through long-term relationships and is therefore ideal for researchers who want to study the dynamics and convergence of one-shot decisions in large populations. It allows for experiential learning about the social environment and social norms but prevents learning about a specific individual. Ideally, under a strangers' matching protocol a subject meets a new unknown subject every round. In practice, this is difficult to implement in laboratory for more than several rounds due to laboratory size restrictions. For example, sessions typically have no more than 40 subjects while repeated experimental games can last even for 100 rounds, so some pairs of subjects will surely meet more than once. In fact, subjects are typically divided into smaller independent matching groups sized from 4 to 10, to increase the number of independent observations, i.e., data points. If subjects were not divided into smaller matching groups, then the choices of all individuals would be correlated because subjects would interact and impact everyone in the session with their choices. As a result, we would end up with only one independent data point. Independent observations are necessary for statistical analyses and with only one independent data point the statistical analysis would practically be meaningless. By having several matching groups per session or treatment, the chance that the same pair meets several times becomes much higher. A way to handle this problem is to randomly rematch subjects between rounds and not reveal

subjects' identities or tags to their counterparties, so that no one knows for sure with whom they are playing.

Experimental design

In our experiment, all subjects in a session either play a deception or a helping game (but not both) over 100 identical rounds. Games are designed to be as similar as possible, for instance, the screen design, the monetary allocations, information, and terminology. This creates an opportunity to directly compare honesty with helping which is one of our goals. Similar approach, albeit in one-shot experiment, has been taken by Gneezy (2005).

In both our games there are two players, a sender and a receiver. Both games are played in a matching group of six anonymous subjects, with subjects being unaware of the matching group sizes. In each round the subjects are first randomly divided into three pairs, and then randomly assigned the sender/receiver roles. We thus implement the strangers' matching in all our experimental treatments. Our experiment could therefore be viewed as a sequence of one-shot interactions almost à la Gneezy (2005). We will first describe the deception game because it needs more detailed explanation.

Deception game design

In our deception game the sender observes eight colored options and the corresponding payoff scheme (see left part of Figure 3). Options are labelled from A to H, of which one (randomly chosen) is colored blue (and called Blue) and the rest are colored green (and called Green). The sender could see, with equal probability, any of the eight different payoff schemes which correspond to eight different states: one where A is colored blue, one where B is colored blue, …., one where H is colored blue. With Blue option the sender incurs the cost of 150 francs, and the receiver earns 250 francs, whereas with any Green option both earn 0 francs. We used color codes instead of words like "honest" or "deceptive" to ensure neutral framing.

The possible payoff allocations are known to both players, but only the sender knows which option corresponds to allocation (-150, 250), i.e., which option is Blue, and which to allocation (0, 0), i.e., which options are Green. Upon seeing the options, the sender sends a message to the receiver, claiming which option will earn the receiver 250 francs (see the right part of Figure 3). Note that this is the same as claiming which option is Blue. Therefore, she has eight possible actions: claim that option A will earn the receiver 250 francs, claim that option B will earn the receiver 250 francs, …, claim that option H will earn the receiver 250 francs. Importantly, the message sent does not have to reveal the truth, i.e., the sender can deceive in her message. In

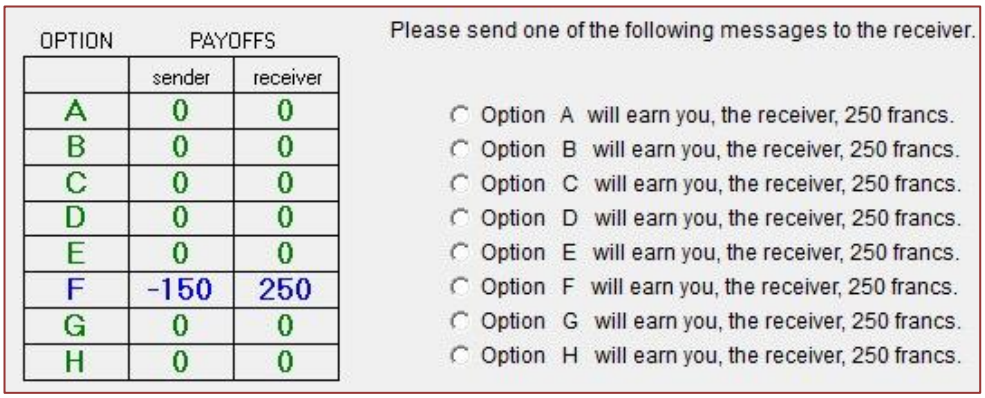
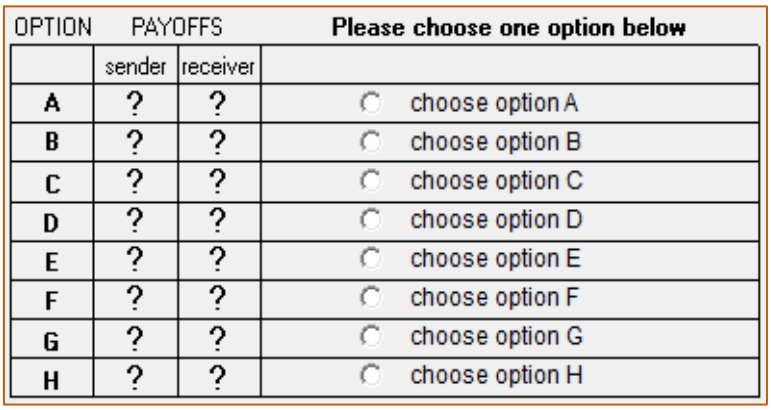


Figure 3: Deception game - insert from sender's screen.

particular, the sender can indicate any option, including Green ones, to be Blue. We say that the sender sends the honest message if she recommends the actual Blue option in her message to the receiver. We say that the sender sends a deceptive message if she recommends one of the Green options in her message to the receiver. In this game the sender only sends a message, and hence does not directly determine the payoffs (as in the helping game), although her message may influence receiver's beliefs about which option is Blue and his subsequent decision. After the sender sends her message, the receiver observes the message (which might be deceptive) before making his choice. The receiver knows that there are eight options and that one of them is Blue and the rest are Green. He also knows what payoffs correspond to Blue and Green options but does not know which option is Blue (see Figure 4). Finally, the receiver makes his choice by choosing one of the eight feasible options A-H. If the receiver chooses the option recommended in the sender's message, we say that he trusts the veracity of the message. If the receiver chooses an option that is not recommended in the sender's message, we say that he doubts the veracity of the message. After both players made their decision, they both learned which option was Blue, what the sender recommended (at this point both players already knew this), and what the receiver chose. From that information the sender could learn whether the receiver trusted her, and the receiver could learn whether the sender was honest. The round ends with the information about the incurred costs or francs earned in current round in this pair.



OPTION	PAYOFFS		Please choose one option below
	sender	receiver	
A	?	?	choose option A
B	?	?	choose option B
C	?	?	choose option C
D	?	?	choose option D
E	?	?	choose option E
F	?	?	choose option F
G	?	?	choose option G
H	?	?	choose option H

Figure 4: Deception game - insert from receiver's screen.

Helping game design

The design of our helping game mimics the design of our deception game as close as possible. Namely, the sender also observes eight colored options and the corresponding payoff scheme (see left part of Figure 5). Options are labelled from A to H, of which one (randomly chosen) is colored blue (and called Blue) and the rest are colored green (and called Green). The sender could see, with equal probability, any of the eight different payoff schemes which correspond to eight different states: one where A is colored blue, one where B is colored blue, …., one where H is colored blue. With Blue option the sender incurs the cost of 150 francs, and the receiver earns 250 francs, whereas with any Green option both earn 0 francs. We used color codes instead of words like “generous” or “selfish” to ensure neutral framing.

The possible payoff allocations are known to both players, but only the sender knows which option corresponds to allocation (-150, 250), i.e., which option is Blue, and which to allocation (0, 0), i.e., which options are Green. Upon seeing the options, the sender makes her choice by choosing one of the eight feasible options A-H. We say that the sender chooses the generous option if she chooses the Blue option. We say that the sender chooses a selfish option if she chooses one of the Green options. In this game only the sender makes a decision (the receiver is passive), which directly determines the payoffs of both the sender and the receiver. After the sender made her decision, both players learned which option was Blue and what sender chose. From that information the receiver could learn whether the sender was generous. The round ends with the information about the incurred costs or francs earned in current round in this pair.

As a final remark, in Gneezy (2005), the dictator choices (analogous to our helping game choices) were implemented with probability 0.8 to make dictator games more comparable to his deception games where receivers followed the advice with such probability. In his experiment, however, deception games were run prior to dictator games, so he knew how likely the sender's chosen option is implemented. In our experiment, sessions of our treatments were

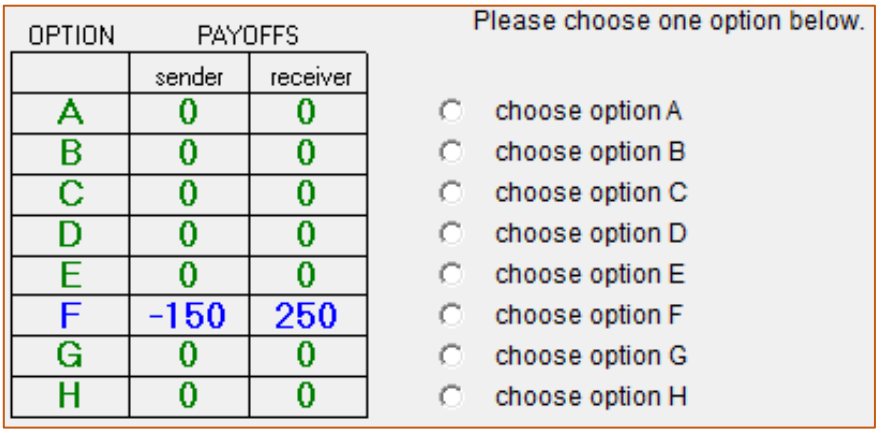


Figure 5: Helping game - insert from sender's screen.

run in a mixed sequence, meaning that at the time of the first helping game session we did not have the results of deception game, because not all deception game sessions had been conducted yet. Without this data, we decided to implement all helping game choices with probability 1, as has also been done by Hurkens and Kartik (2009) who had the same problem but for a different reason, namely they had a within-subject design.

Reputation mechanism

For both games we consider two experimental conditions that vary the information that senders receive about their receiver's past sending actions. That is, the sender may observe the colors of the three most recent options that her current receiver has chosen (for the helping game) or recommended by a message (for the deception game) when he had a sender role in the previous rounds, e.g., "Blue: 2; Green: 1". This reputation information was displayed on the sender's screen (see Figure 6).

This information reveals to the sender the reputation of her receiver. The reputation always contains information about past sending behavior of the sender's current receiver. In all our treatments (described below) the receivers receive no information about their sender's reputation. While it may be interesting to investigate the effect of the sender's reputation on receiver's trust in future experiments, we kept it hidden in this experiment in order to maximize the similarity between the helping and deception game designs. Also, the behavior of receivers cannot be compared between deception and helping game, since in the helping game receivers make no decisions.

In our games, direct reciprocity is not possible, because subjects are anonymous, and pairs change between the rounds. Indirect reciprocity is possible, however, if a sender can observe the reputation of her current receiver. Like previous theoretical and experimental studies on indirect reciprocity (see, e.g., Nowak & Sigmund, 1998a; Leimar & Hammerstein, 2001; Seinen & Schram, 2006; Ule et al., 2009), we limit reputation to the last few actions (three in

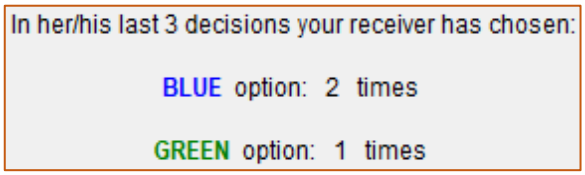


Figure 6: Reputation information.

particular). In addition, subject's own reputation was displayed to herself at the end of every round as an ordered sequence of colors, e.g., "Green Blue Green", showing the colors of the last three options she has recommended when she was the sender. As a remark, in the following chapters of the doctoral dissertation we often use the terms good and bad reputation to motivate the strategies and facilitate their descriptions. In our experiment, however, these words were not used to avoid any framing effect.

Treatment design

Our experiment consisted of four treatments in a two-by-two design. One treatment dimension was the *game* the subjects played for 100 rounds with strangers matching: deception game (DG) or helping game (HG). The other treatment dimension was the *reputation information* that senders received about their receiver: either they received it in every round, or never. DG and HG treatments without reputation information are labelled BASE, while DG and HG treatments with reputation information are labelled REP, so the abbreviations used for our treatments were DBASE, HBASE, DREP and HREP (see Table 2 and Figure 7). Except for reputation information, REP and BASE treatments are identical, facilitating a direct comparison.

		Rep. Info.	
		BASE	REP
Game	DG	DBASE N=10	DREP N=10
	HG	HBASE N=9	HREP N=9

DG and HG denote deception and helping game, respectively. BASE and REP denote the treatment without or with reputation information, respectively. DBASE, HBASE, DREP and HREP are abbreviations for our four treatments.

Table 2: Experimental treatments and the number of independent observational units.

Each treatment session consisted of 100 rounds. For each treatment we ran three sessions, each with multiple matching groups of six subjects. Because subjects in different matching groups never interacted, we consider each matching group of six subjects as one independent observational unit for our statistical tests. Table 2 shows the number of independent matching groups per treatment.

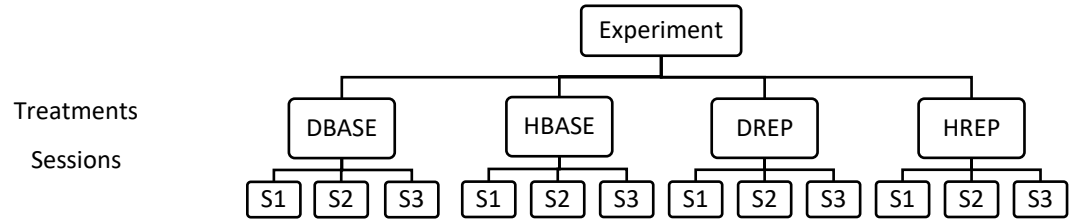


Figure 7: Experimental design. The experiment consisted of four treatments, each of which had three sessions.

With the first dimension (game) we investigate the different dynamics of honesty and helping in the long run, extending the investigation in Gneezy (2005) from a single round to multiple rounds. With the second dimension (reputation information) we investigate if the reputation system promotes honesty in an analogous way that it promotes generosity.

Complete instructions of our treatments are provided in Appendix A3 (A3.1 and A3.4).

Experimental procedures

The experiment was conducted in early 2020 at the laboratory of the Center for Research in Experimental Economics and Political Decision Making (CREED) at the University of Amsterdam. It was programmed and conducted with the experimental software z-Tree (Fischbacher, 2007). Subjects were recruited from the CREED subject pool. In total 228 subjects participated in the experiment, 120 in the deception game treatments and 108 in the helping game treatments, each in only one treatment and session.

In our experiment, subjects were students from different study disciplines (i.e., economists and non-economists) and either had or did not have previous experimental experience. These two factors might affect the results, as evidenced, for example, by Benndorf et al. (2017), López-Pérez and Spiegelman (2019) and us in Chapter 4. Namely, Benndorf et al. (2017) found that subjects with experimental experience are less trustworthy and trust less often than inexperienced subjects, while López-Pérez and Spiegelman (2019) found that business and economics students lie more than students from other disciplines. We tested whether the distributions of disciplines and experimental experience were similar across our four treatments and found that they are ($p>0.1$, confirmed by Chi-square and Fisher exact test). This assured us that subjects in all four treatments are sampled from the same population and that we can assign the differences in behavior between treatments to the treatment conditions (i.e., reputation information, honesty/generosity) rather than the demographic structure of the subjects. We also controlled for gender of our subjects but were not able to collect the gender data in all our sessions due to a glitch in our experimental software, so we omit gender from

all our analyses. We expect a similar gender structure in all our treatments due to subject randomization, though.

Each session started with instructions about laboratory conduct where subjects learned that communication between them is not permitted throughout the experiment. After that, subjects were provided printed detailed description of the repeated deception or helping game. After the experiment the subjects completed a questionnaire about the background information and demographics. They were also asked to describe how they were making their decisions in the experiment and to advise a new subject in such experiment. Throughout the experiment subjects sat in their private cubicles which protected anonymity and obstructed communication and observation of each other's decisions.

To make sure that subjects do not leave the experiment with a loss, we gave them 3000 francs at the beginning of round 1. The francs a subject earned and lost during the 100 rounds were then added to this initial endowment. At the end of the experiment the total was converted to euros at the rate of 1 EUR per 250 francs and privately paid to the subject when she left the laboratory, together with a show-up fee of 7 EUR. In our four treatments, the average and median earnings ranged from around 22 EUR to around 28 EUR, and the average duration of a session ranged from 60 minutes in HBASE to 140 minutes in DREP. We had such variation in duration because 1) in our deception game both the senders and receivers were making decisions, whereas in our helping game only senders were making decisions; so, the deception game has twice as many decisions as the helping game, and 2) in REP treatments senders were endowed with additional reputation information increasing the complexity of decision-making.

Chapter 2

Helping among strangers

2.1 Introduction

Generosity is efficient when it carries an important value for receiver (recipient) at a relatively small cost for sender (donor). An exchange of efficient generous acts between acquaintances can be supported by direct reciprocity, leading to a form of cooperation. In large societies, however, many interactions involve anonymous strangers, which lack opportunities for reciprocal exchange. Generosity may then depend on altruistic individuals who help indiscriminately, but they can be exploited, fare poorly and are rare (Nowak & Sigmund, 1998a; Ule et al., 2009). An alternative path to social cooperation is offered by conditional altruism which directs generosity towards strangers that are generous themselves. This behavior is again reciprocal, but in an indirect way.

In particular, generosity can emerge in short-term interactions if individual reputations can be shared, for instance through gossip, so that encounters are not entirely anonymous. For illustration consider the theoretical setup from Nowak and Sigmund (1998a) where matches in a population of individuals are random and in each pair one individual can help another. Help is costly but socially efficient, because the benefit to the receiver outweighs the cost for the sender. Help is not in the immediate private interest of the sender, however, so there is a dilemma. This calculation changes when such helping game (formally presented in Chapter 1) is played repeatedly in a population. Now let the matching and the helping decision be made repeatedly and let each sender observe the past generosity of its current receiver. This reputation may for instance be shared by gossip. How individuals will behave towards receivers with different reputations is described by their strategies. In the simple setup every individual is either an altruist, a defector, or a rewarder. Altruists always help, defectors never help, and rewarders help those with good reputation. Individuals will occasionally replace their type if they find that another performs better. Rewarders can flourish in this system because they may help each other but cannot be exploited by defectors if their share is small. In general, rewarding is not sufficient for widespread social cooperation. Whether indirect reciprocity will promote generosity depends on the normative definition of good reputation in this society, and on the initial distribution of strategies and matching (Nowak & Sigmund, 1998a; 1998b).

Rewarding can be exploited by more strategic behavior, however. Cautious subjects, for instance, help only when this protects their own reputation. They receive help from rewarders at a smaller helping cost and will invade a population of rewarders, destroying cooperation. This invasion can in turn be prevented with a richer reputation information, containing

motivations behind past choices (Leimar & Hammerstein, 2001). In general, the success of helpful types and spread of generosity depend on strategies used by members of a population, as well as on the scope of reputation. In order to apply these analytical results to human societies, we need to understand which behavioral heuristics people will consider, whether their behavior can indeed be captured by unique strategies, how popular these are, and when will helpful behavior be profitable.

The empirical studies of indirect reciprocity initially tested only the general predictions about its role for the spread of generosity. Wedekind and Milinski (2000) confirmed that reputation can sustain generosity, Bolton et al. (2005) found that many subjects consider the motivational element in reputation, and Engelmann and Fischbacher (2009) found evidence of cautious behavior. Three experimental studies investigated the behavioral strategies that individuals apply in various helping games, each applying a different statistical classification method, tailored to a specific empirical question. Seinen and Schram (2006) classified experimental subjects into six key strategy classes based on the best fit with individual data, but did not correct for random, noisy or inconsistent behavior. As people make mistakes or experiment, an attempt to fit their actions into a deterministic model with this method may lead to misclassification. Ule et al. (2009) therefore allowed for noise in their classification but biased it towards rewarder strategy and did not consider sophisticated subjects who condition on both the sender's and the receiver's reputation. Swakman et al. (2016) investigated correlations between information and generosity but did not tie the resulting classification to a theoretical strategy set. Each method classifies subjects whose behavior is sufficiently regular and consistent with theoretically plausible behavioral types. This does not imply a conscious application of a behavioral rule, or an awareness of regularity in behavior. The methods are agnostic on whether regularity emerges from rational utility maximization, adaptation, habit, or another mechanism such as desire for simplicity. It is unclear to what extent these three classification procedures are compatible.

It is also unclear how these procedures relate to a more general strategy classification approach such as the mixture model estimation. This method relies on standard statistical techniques such as the maximum likelihood estimation and has several advantages. It classifies behavioral types according to the best fit analyses, based on how well the predicted sequence of choices fits with the sequence of actual observed choices. The method allows for noise, complex and stochastic strategies, as well as non-strategic random choice. It is independent of the order of classification and does not advantage any set of types over another. It also provides a standardized way to include new strategies into the set of considered types. Its large number of parameters may be considered a disadvantage, but this is controlled by the iterative elimination of irrelevant strategies from the initial set. Its accuracy is ultimately an empirical question, although it is natural to expect that it will provide a better fit to the data than the above mentioned three methods because it estimates more parameters. This approach was applied by

Dal Bó and Fréchette (2011) to estimate the distribution of strategies in a population playing a repeated prisoner's dilemma game, treating past choices as strategic determinants for future choices. Stahl (2013) and Aoyagi et al. (2021) then used mixture-model estimates to determine the individual strategies used by experimental subjects. While the mixture model estimation is now the standard technique for repeated prisoners' dilemma games (see for example the literature in Dvorak, 2020b), it has not been applied to the conceptually different repeated helping game.

We begin this chapter by comparing the four classification procedures using new data from our experimental helping game. We consider the three classification procedures from the helping game literature, and a procedure that we base on finite mixture model estimation. We apply each method to the same data, obtaining four classification results. We consolidate them into a consensus classification, giving us a comparison benchmark. We found that the mixture model estimation yields an almost perfect match with the consensus, while the other three classifications perform less well. They nevertheless provide quite reliable classifications and are more user-friendly for researchers not familiar with maximum likelihood techniques and their implementation in computer software.

We next use the flexibility of the mixture model-based approach to expand our set of feasible strategies in order to study whether our experimental subjects adapt to their personal experience. While it is generally assumed that reputation mechanism will prevail over experience, Baker & Bulkley (2014) found that it is stronger only in the short-term, while private experience becomes more important in the long-term, perhaps because it is less cognitively demanding. Such 'experiential' heuristic is often neglected in classifications of behavioral types, however. In particular, it is not considered in the experimental studies of indirect reciprocity, despite the evidence that experience matters (Bolton et al., 2005; Swakman et al., 2016). We account for this behavior in the mixture model estimation and indeed find that 7% of subjects consistently react to private experience. Even more, experiential behavior is modal when reputation is private.

We finally investigate whether post experimental (retrospective) self-reports, in which subjects summarized how they were making decisions during the experiment, are a reliable source of data for strategy classification. Such reports are not normally used for detection of strategic behavior, perhaps because of economists' reservations about external validity of unincentivized decisions (V. L. Smith, 1976). They can also suffer from inaccuracy due to people's cognitive limitations for self-reflection and other biases (Nisbett & Wilson, 1977). Experimental economists sometimes use them to supplement choice data (Seinen & Schram, 2006), but usually consider them inferior to choice data (McCloskey, 1983). On the other hand, Capra (2019) found that self-reports, collected while subjects are making decisions, can yield an insight into subjects' levels of reasoning in short games. There has been no evaluation of their

utility for strategy classification in longer games, however. Here we show that retrospective reports, collected after the experiment, are not very reliable. While around 80% of subjects provided feasible strategy descriptions, more than 30% of these in HBASE and almost 50% in HREP were not consistent with the choice-based statistical classification.

To summarize, this chapter has three main contributions. First, our analysis shows that finite mixture models can be successfully used in repeated helping games for classification of individuals' strategies. Second, with this method we detect experiential behavior, an important type that was overlooked by previous classifications. This also substantially reduces the number of unclassified behaviors. Finally, we compare statistical estimates of strategies to the verbal explanations written by the subjects themselves after the experiment and show that they do not lead to a reliable classification.

In the following section we formally introduce the strategies and the four choice-based classifications. Section 2.3 contains the results of our comparison and the complete classification including experiential behavior. This section also describes the classification based on the retrospective self-reports and discusses its veridicality. The complete list and definitions of behavioral (sub)strategies and detailed descriptions of the four methods are in Appendix A (A1 and A2, in particular).

2.2 Behavioral strategies and classification methods

We compare four different methods for classifying subjects' behavior in our experimental treatments with repeated helping game. Three methods were previously used by Seinen and Schram (2006), Ule et al. (2009), and Swakman et al. (2016). We label them as "DFIT", "SFIT" and "TREND" to indicate that they consider deterministic strategies, stochastic strategies and conditional behavior, respectively. The fourth method is based on the statistical approach used to identify strategies in repeated prisoner's dilemma game. It was introduced by Dal Bó and Fréchette (2011) and extended by Breitmoser (2015) and Dvorak (2020b). Since this method is based on finite mixture models and maximum likelihood estimates, we label it as "MLFIT''.

We classify subjects' behavior in HREP four times, according to each method, obtaining four classifications based on the same choice and information data. For classifications we remain consistent with the previous literature (Seinen & Schram, 2006; Ule et al., 2009) and skip rounds 91-100 to avoid the end-game effect, and those earliest rounds when subject's reputation may have less information or be even empty. In particular, we consider only rounds where the sender has already made at least two decisions, has already been receiver at least three times and her receiver has already made at least three decisions in the past.

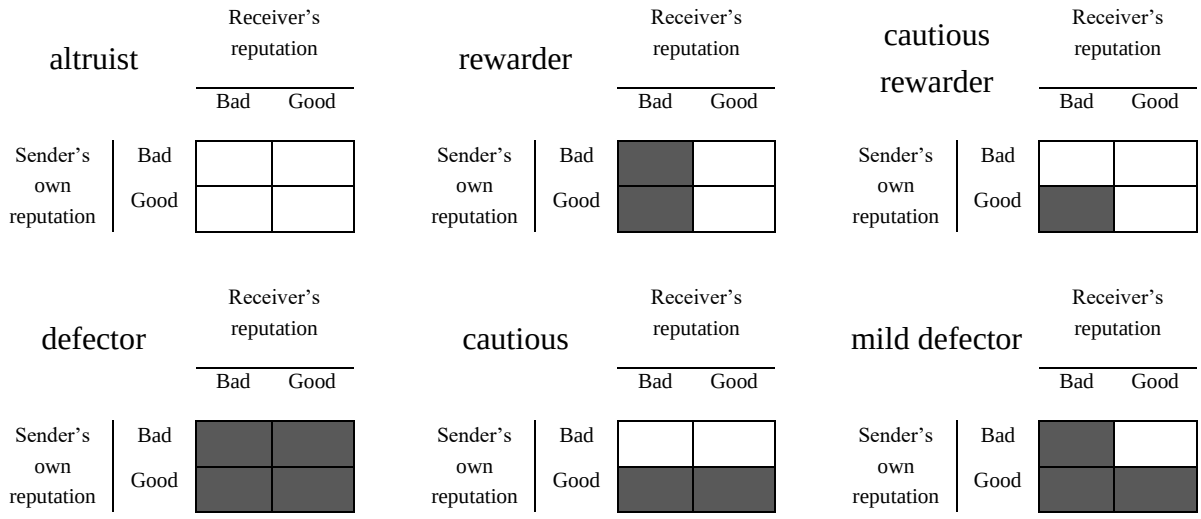


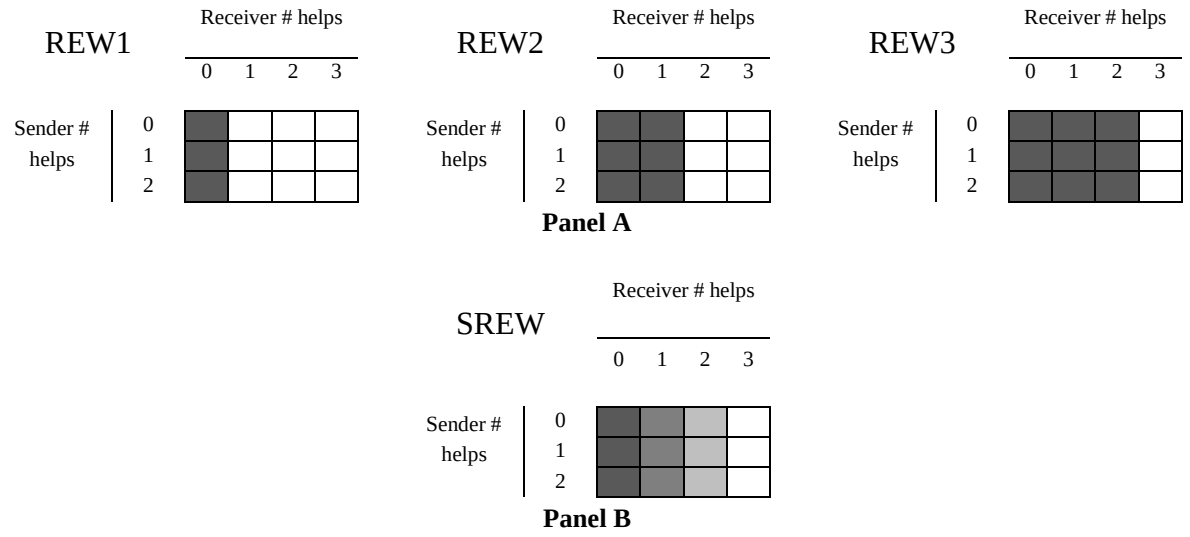
Figure 8: The six main behavioral strategies for the repeated helping game. For each strategy, a gray cell indicates when a sender will pass, and a white cell indicates when a sender will help.

The literature on the theory of indirect reciprocity (e.g., Nowak & Sigmund, 1998a; Leimar & Hammerstein, 2001; Seinen & Schram, 2006) identified six key types of behavioral strategies. The simplest two behavioral types are a*ltruist* and *defector*. An altruist always helps while a defector never helps. Both use unconditional strategies which always dictate the same action. Next, we have two behavioral types that condition the sender's action on reputation. A *rewarder* conditions her help only on receiver's reputation (the image score), while a *cautious* subject conditions help on her own reputation. More precisely, a rewarder only helps receivers with good reputation. This is a quintessential kind of indirect reciprocity that rewards past good behavior towards third parties.[2] In contrast, someone who is cautious helps only when her own reputation is bad, irrespectively of her receiver's reputation. This may be a rational selfish response to widespread reciprocity and can destroy social generosity. Leimar & Hammerstein (2001) show that cautious subjects can invade a population of rewarders before themselves being invaded by defectors.

The last two behavioral types use strategies that condition on the reputations of both paired subjects. A *cautious rewarder* helps either when her receiver's reputation is good or when her own reputation is bad. A *mild defector* helps only when both her receiver's reputation is good and her own reputation is bad. We call these two strategies *sophisticated.* All six strategies can be illustrated by two-by-two tables shown in Figure 8. The columns correspond to receiver's reputation and the rows correspond to sender's own reputation. The colors of table cells

[2]Technically, this is called downstream reciprocity, to contrast it with upstream reciprocity where a sender helps a stranger in gratitude for help she had received in the past (Nowak & Sigmund, 2005). Upstream reciprocity has not been considered in the literature that studies behavioral strategies in repeated helping games, so we skip it here, but we will consider it in the following section.

indicate helping (white) and passing (gray). For simplicity we assume in the illustrations that reputations can only be bad or good.



The top panel A shows the three possible deterministic substrategies used by a rewarder. The bottom panel B shows a stochastic rewarder strategy. For each substrategy, the vertical dimension counts the number of times the receiver has helped in her last 3 decisions, and the horizontal dimension counts the number of times the sender has helped in her last 2 decisions. Dark gray, medium gray, light gray and white cells indicate helping probabilities of 0, 1/3, 2/3 and 1, respectively.

Figure 9: Rewarder substrategies

In our experiment the terms “good” and “bad” reputation were not used. Instead, senders were given information about the last three help/pass decisions of their receiver, so that reputation was manifested through the number of visible helping actions. This yields several possible norms for what constitutes good reputation, from the strictest norm where all visible past actions must be help, to the weakest norm where at least one must be help. Consequently, a conditional strategy can be defined in different ways for different norms. As an example, consider a rewarder. Her behavior can be represented by three distinct deterministic substrategies: *REW1*, *REW2* and *REW3* (see Figure 9, panel A). *REWK* (where *K* is a variable) helps only if her receiver helped at least *K* times in the recent three opportunities. A receiver’s visible history can show 0-3 helps, and there can be 0-2 helps in the relevant history of the sender.[3]

[3]A sender that cares for own reputation should consider only her last two actions (not three), because her future reputation will show them next to the action she is going to take in her current interaction.

Any deterministic implementation of a substrategy in our experiment can thus be represented by a three-by-four table. In total we can define 19 substrategies: one defector, one altruist, three rewarder, two cautious, six cautious rewarder and six mild defector substrategies. We will also consider stochastic substrategies which relax the assumption that reputation is binary (either bad or good).

For each subject we can evaluate the fit between a specific substrategy and her actual behavior in different ways. A straightforward way is to count the number of game rounds in which her decision matches the substrategy prediction, given her information and reputation in that round. A subject is then associated with the substrategy that yields the best match. This is the approach taken by Seinen and Schram (2006). A subject's strategy is classified into one of the six strategy categories for which her decisions in the experiment are best captured by one of the corresponding substrategies. In case of a tie a subject is assigned the least complex strategy, with altruist and defector strategies being less complex than rewarder and cautious strategies, which are less complex than cautious rewarder and mild defector strategies. This biases the classification towards simple strategies. The approach does not account for noise, random choice, or poor fit, however. Given the evidence that subjects in experiments often apply strategies in a noisy or stochastic manner (Romero & Rosokha, 2019), this method may lead to misclassification as well as forced classification of subjects who should not be classified because they do not consistently apply any strategy. DFIT is a slight modification of Seinen and Schram's (2006) method that reduces this over-classification.

To account for noise or randomness, Ule et al. (2009) and Swakman et al. (2016) propose to instead measure correlations between information and decisions rather than the fit with theoretically predicted strategies. Both methods classify conditional strategies according to the logit regression analyses, whereby the issue of linear separation is handled with Bayesian regression. To facilitate separation between selfish and other-regarding behavioral strategies, Ule et al. (2009) include additional criteria related to the helping rates, for instance that a rewarder must help at least in 40% of the relevant rounds. We call this the SFIT method. Swakman et al. (2016), on the other hand, classifies subjects only according to the significance of the various effects in the regression, which we call the TREND method. This gives an advantage to conditional strategies. For instance, for a significant effect of reputation it may be sufficient to help only 20% of the receivers with perfect reputation and no one else, but such behavior is not consistent with any formal model of rewarding. Whether these two methods provide theoretically meaningful classifications is an open question which we would like to investigate.

SFIT is built around a sequence of tests, giving priority to the (conditional) rewarder and cautious strategies. In contrast, TREND bases the classification on a single logit regression (and a Bayesian regression in case of linear separation) with three explanatory variables

(sender's and receiver's reputation information, and the round), to avoid giving any strategy an exogenous advantage. A threshold is included only to separate defectors, altruists and subjects who do not use any of the six strategies. These two methods cannot isolate the sophisticated cautious rewarders and mild defectors but can identify subjects whose behavior is not consistent with any strategy which DFIT cannot.

None of these three classification methods simultaneously (i) provides a measure of fit with the theoretical set of strategies, (ii) accounts for noise, stochastic choice and non-strategic behavior, and (iii) is expandable with new strategies in a straightforward way. In this study we therefore adopt an alternative classification method that facilitates all that. Our MLFIT method is based on the Strategy frequency estimation method introduced by Dal Bó and Fréchette (2011) for detection of strategies in a population playing a repeated prisoners' dilemma game, and on its extension by Dvorak (2020b) that allows for estimation of individual strategies in simple repeated strategic games.

MLFIT determines best fit based on how well the sequence of actual observed choices matches the choices prescribed by a specific substrategy, accounting for stochastic or noisy behavior. It considers a predetermined set of substrategies K, described by deterministic or stochastic finite automata. Crucially, we include in this set all 19 standard deterministic substrategies, as well as the stochastic versions of all conditional and sophisticated strategies. To make our analysis tractable, we include only the most intuitive stochastic strategies, designed as a linear combination (average) over all corresponding deterministic substrategies. See Figure 9, panel B, for the example of stochastic rewarder that helps with probability *x/3* if receiver helped *x* times in the recent three rounds. To detect non-strategic behavior, we additionally include a *random* strategy that helps with probability 50% in every round, following Stahl (2013). This is an important step that reduces overestimation. Without this strategy a subject who behaved randomly or randomized between different strategies would be misclassified by MLFIT into one of the six behavioral types. Finally, we also include a *random8* strategy that helps with probability 1/8 in every round. It captures the behavior of an individual who in each round randomly chooses one of the eight options or who always chooses the same fixed option, e.g., option A (recall the design presented in Chapter 1; the sender is presented with one blue and seven green options). Such behavior is non-strategic in helping game but may be strategic in our deception game where it also corresponds to the equilibrium behavior. The inclusion of this strategy in our helping game is needed for Chapter 3 where we seek to prove that the equilibrium behavior detected in our deception game indeed stems from strategic motives (i.e., subjects randomize on purpose, e.g., to maximize their expected profit) and is not the result of randomization by subjects who are clueless about which behavior might be profitable.

MLFIT uses the maximum likelihood approach to estimate the frequency p_k with which each substrategy $k \in K$ is used in the sample, and a universal tremble probability τ, describing the

frequency with which subjects make mistakes or explore. To be more specific, each substrategy $k \in K$ can be written as a finite automaton with s_k states, characterized by the theoretical probabilities $\xi_{ks} \in [0,1]$ that substrategy k helps in state s. MLFIT then takes probabilities of given substrategies p_k and trembles τ as free parameters and returns values that best fit with (maximize the log-likelihood of) the sequences of actual binary choices of all subjects. Formally, the model returns values p_k^* and τ^* that maximize the following log-likelihood function subject to constraints $p_k \in [0,1]$, $\sum_{k\in K} p_k = 1$, $\tau \in [0,1]$:

$$\sum_{i=1}^{N} \ln\left(\sum_{k\in K} p_k \prod_{s=1}^{s_k} {\pi_{ks}}^{y^i_{ksh}} (1-\pi_{ks})^{y^i_{ksp}}\right),$$

where N is the number of experimental subjects, $\pi_{ks} \in [0,1]$ is the probability that substrategy k dictates help in state s:

$$\pi_{ks} = \begin{cases} \xi_{ks}(1-\tau) + (1-\xi_{ks})\tau, & \text{if } \xi_{ks} \in \{0,1\}, \\ \xi_{ks}, & \text{if } \xi_{ks} \in (0,1), \end{cases}$$

and y^i_{ksc} is the number of times that subject i using substrategy k chooses action $c \in \{h,p\}$ in state s. In the log-likelihood function, the expression in the parentheses is subject i's contribution to the log-likelihood function.

Since the expectation-maximization (EM) algorithm (Dempster et al., 1977) is used to find maximum likelihood estimates, the fitted model also stores for each subject i her posterior probability of using substrategy k (given the observed sequence of choices),

$$\theta_{ik} = \frac{p_k^* L(i,k)}{\sum_{k'\in K} p_{k'}^* L(i,k')},$$

where $L(i,k) = \prod_{s=1}^{s_k} {\pi_{ks}^*}^{y^i_{ksh}} (1-\pi_{ks}^*)^{y^i_{ksp}}$ is the likelihood that subject i uses substrategy k given the observed sequence of choices, and π_{ks}^* denotes π_{ks} evaluated at estimated τ^*. Based on these posterior probabilities we first calculate for each subject i her posterior probabilities of using strategy $j \in$ {DEF, ALT, REW, CAU, CR, MD, RAND, RND8}:

$$T_{ij} = \sum_{k\in K_j} \theta_{ik}$$

where $K_j \subseteq K$ is a subset of set K that contains all substrategies of strategy j. For example, K_{DEF}={DEF} and K_{REW}={REW1, REW2, REW3, SREW}.[4] A subject is assigned the strategy category $K_{j^*} \subseteq K$ when $T_{ij^*} = max_{h\in J}\, T_{ih}$ and $T_{ij^*} > 0.5$. If $T_{ij^*} \leq 0.5$ or if j^* is the random or random8 strategy, a subject goes to the category "unclassified". If assigned strategy category

[4]Note that K is a disjoint union of subsets K_j.

has substrategies, i.e., if $j^* \in$ {REW, CAU, CR, MD}, a subject is also assigned a substrategy. In particular, a subject is assigned the substrategy $k^* \in K_{j^*}$ when $\theta_{ik^*} = max_{h \in K_{j^*}} \theta_{ih}$.

In summary, the MLFIT is essentially a top-down classification method, involving one or two steps. On step 1 each subject is assigned one strategy (or is left unclassified) based on posterior probabilities of strategies, each of which is obtained by summing posterior probabilities of its corresponding substrategies. Subjects who are assigned a strategy that has at least two substrategies, enter step 2, where they are additionally assigned a substrategy with the highest posterior probability among all substrategies of the assigned strategy. For most subjects (around 80% in HREP) T_{ij^*} is above 0.8, indicating that they consistently apply one strategy throughout the experiment. It is worth mentioning that for a subject this does not mean that the assigned strategy is consistent with over 80% of her actual choices. It means that given her sequence of choices the probability that subject uses any other strategy is below 20%.

In contrast to MLFIT, DFIT evaluates only a fit with deterministic strategies and without trembles. Moreover, the original approach in Seinen and Schram (2006) assigns one of the six theoretical strategies to all subjects, including those that act randomly or use a strategy outside the predetermined set. All other methods allow for the possibility that a subject cannot be classified into any theoretical strategy category. For a meaningful comparison we therefore include in DFIT the requirement that at least 70% of actions of a subject must be predicted correctly by a single strategy, or else the subject is unclassified. While all four methods assume that people use fixed strategies, they provide distinct checks to detect consistency and otherwise leave a subject unclassified. Appendix A2 describes the detailed classification procedures for all four methods, including further remarks on the interpretations of parameters τ, π_{ks} and θ_{ik} in MLFIT.

MLFIT and DFIT consider all six theoretical strategy categories. SFIT considers only four categories; cautious rewarders are merged with rewarders and mild defectors are merged with cautious subjects.[5] TREND considers five categories: altruist, defector, rewarder, cautious and sophisticated, fusing cautious rewarders and mild defectors into a single class. All four methods add the unclassified category for subjects whose behavior is random or cannot be captured by a single behavioral strategy.

[5]Given that SFIT is based on logit regressions and overall helping rates and not on the best fit analyses like MLFIT (or DFIT) it is not straightforward to match the categories precisely. For our comparison we need to standardize the category set, however. We do that based on the interpretations offered in Ule et al. (2009), which we then related to the theoretical strategy categories. We found that most cautious rewarder substrategies are conceptually closest to rewarder substrategies, and most mild defector substrategies are conceptually closest to cautious substrategies. There are some exceptions (e.g., some substrategies of mild defectors are closest to rewarder substrategies), but they are not observed in our data, and hence do not affect the results.

2.3 Results

This section proceeds as follows. We first provide some general results on helping behavior because it is worth knowing, before exploring the behavior of individuals, what is happening on a treatment and group level. This gives us broader but less detailed picture. In this part, most of our results are just a confirmation of the results of the previous helping game studies, though. Then we apply the four methods to classify the subjects in HREP four times. From this we construct the consensus classification and observe that MLFIT provides the closest match. Next, we apply MLFIT to classify the subjects' behavior in HBASE and show that the classification improves substantially when we include the new experiential strategies in MLFIT*. We then return to HREP and reclassify all subjects with MLFIT*, finding a non-negligible share of experientials. Using this final classification, we find that defection is more profitable than rewarding. We also investigate the substrategies of subjects. The investigation reveals that experiential behavior is not driven by an immediate reaction to the most recent experience, but by an accumulated experience over many rounds. We also show that concern for own reputation diminishes in the final rounds of the experiment, which can explain the end-game decline in average helping rates. We conclude by showing that subjects' self-reports are not a reliable source for strategy classification.

2.3.1 General results on helping behavior

In our experiment subjects from the same matching group were paired with the same subject more than once (20 times on average), so it would be unreasonable to assume that the choices within a matching group are independent. On the other hand, the choices between matching groups are independent, since the subjects from different matching groups never interacted. Therefore, our independent observational unit is the matching group. For simple terminology, in this section we will call the average helping rate of the matching group the *group average* helping rate, and the average of group averages the *treatment average*. The treatment average is therefore a single number, calculated from 9 group averages each of which is calculated based on helping of six subjects from the same matching group.

Our first result concerns the reputation effect. We first test whether reputation information increases the treatment average helping across all rounds. We test this by comparing nine group average helping rates in HREP to that in HBASE. One-sided permutation test confirms that the treatment average helping rate in HREP was significantly higher than that in HBASE (43% in HREP vs. 27% in HBASE, $p<0.05$).[6] This result is consistent with previous results documented in the literature (e.g., Bolton et al., 2005; Seinen & Schram, 2006). Next, we turn to the group dynamics of helping. Figure 10 shows the 10-round group average helping rates in HREP (left

[6]Mann-Whitney tests yield the same statistical results.

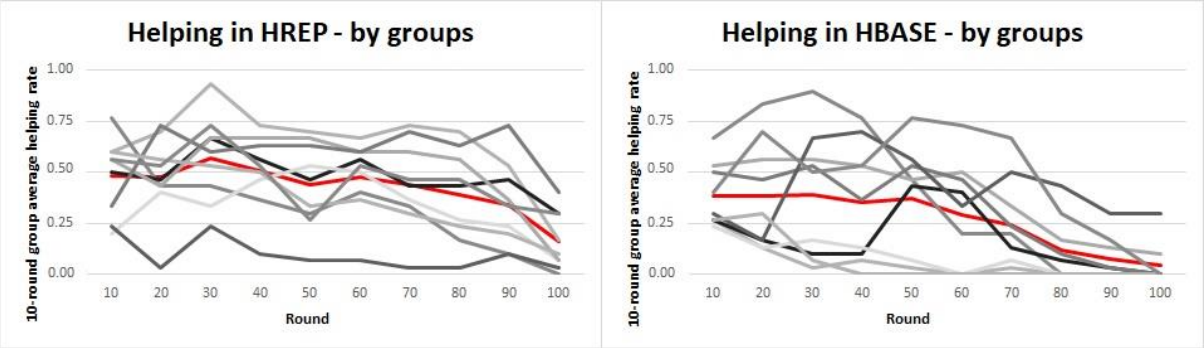


Figure 10: 10-round group average helping rates in HREP (left panel) and in HBASE (right panel). The red line corresponds to the 10-round treatment average.

panel) and in HBASE (right panel). The red line corresponds to the 10-round treatment average. For each treatment, the k-th point on the red line, $k \in \{1,\ldots,10\}$, is calculated as the treatment average over rounds 10k-9 to 10k. The figure shows a decreasing trend in both treatments. We formally tested for changes in helping rates over the rounds in HREP and HBASE by fitting logistic generalized linear mixed models (GLMM) to helping decisions. We included two fixed factors, "round" and a dummy variable for the last 10 rounds to control for the end-effect, and one random factor, "subject nested in matching group". The statistical analysis shows that the average helping rates were decreasing with rounds in both treatments, as the estimated regression coefficient of variable "round" has a negative sign (HREP: $p<0.001$; HBASE: $p<0.001$). In the analysis of a repeated helping game, this technique was used before by Swakman et al. (2016). In their version of the HREP treatment they also detected a decreasing pattern.

Figure 10 also reveals that different groups developed different dynamics and that differences in helping rates developed early. Similar observations were already made by Seinen and Schram (2006). We additionally tested whether the initial helping determines the long-run spread of helping by examining the correlation between the group average helping rates calculated over the first two rounds and the group average helping rates calculated over rounds 3-100. We compare the averages over the first two rounds (and not three or some other number of rounds) with the rest, because after two rounds, on average, everyone made their first decision as a sender.

Figure 11 displays the average helping rate for each group in the initial and the remaining rounds. The correlations were analyzed using a one-sided permutation Pearson's correlation test. In HREP, the correlation is positive, but very weak and insignificant ($r=0.09$, $p>0.1$), showing that the initial rounds do not have much impact on the overall group behavior. A very weak positive correlation also hints that experience-based behavior likely played a minor role

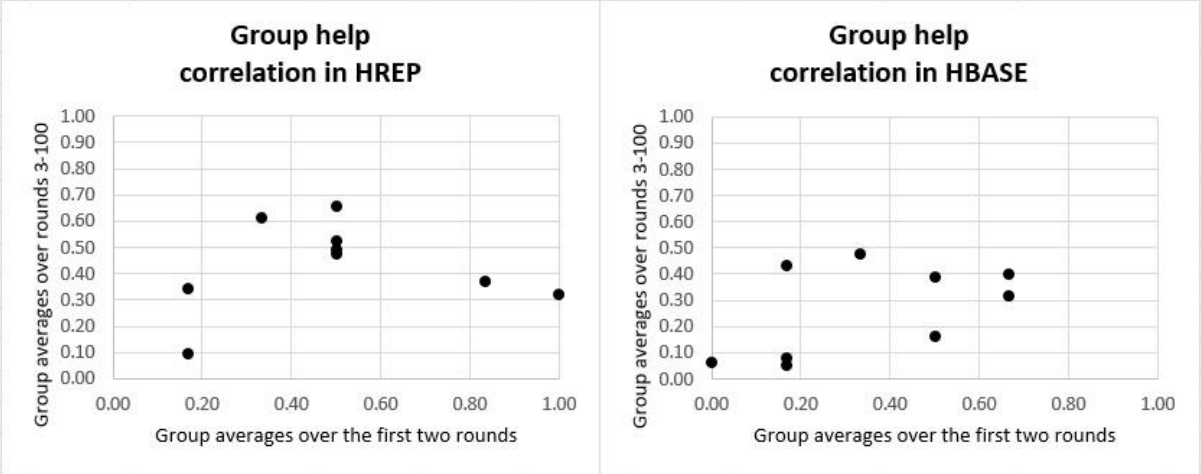


Figure 11: Group average helping rate in initial and remaining rounds. Left panel: HREP. Right panel: HBASE.

in HREP. We will investigate individuals' behavior more thoroughly in later subsections. In HBASE, the correlation is positive, moderate and marginally significant ($r=0.51$, $p<0.1$), meaning that the groups with higher overall help were more likely those that experienced more help in the initial rounds. The moderate positive correlation also indicates that many senders may be making decisions based on experience.

Therefore, at least in the completely anonymous environment, the results suggests that one may increase overall helping by promoting helping in the initial rounds. This may be relevant information for social planners who want to promote, for instance, gift-exchanges between buyers and sellers or compliance with authority recommendations, or for social planners who want to increase, for instance, charity donations or cleanliness in local environments and public places. The lack of help or cooperation within societies was particularly noticeable in the early phases of Covid-19 crisis when people could but did not completely adhere to preventive measures such as wearing face masks, washing hands, maintaining social distance and reducing gatherings, thereby curbing the spread of the coronavirus. This was psychologically costly for some people (e.g., some had difficulties with breathing while wearing a mask; some experienced anxiety and sadness due to social isolation) but beneficial for others because it reduced the chances of catching the coronavirus.

2.3.2 The four methods compared

We first classify the behavior of each subject in HREP four times, using the four classification methods that we described in the previous section. Each method either classifies a subject's behavior into one of the strategy categories or leaves it unclassified. Table 3 shows the distributions of behavioral strategies among the subjects in HREP according to the four methods. All methods used the same data. As a remark, since the random8 strategy has never been used in our helping game treatments, which is not surprising given that it is difficult to

rationalize, we omit it from the tables and further discussion in this chapter. While there are differences between the four distributions, all methods support the same general impression. The prevalent categories are rewarders and defectors, about 20% of subjects are sophisticated, about 10% are altruists, while purely cautious types are rare. Up to 20% of subjects do not classify into one of the six theoretical strategies, indicating that they either do not apply one strategy consistently enough, choose too erratically, or use a strategy outside our standard set of six strategies predicted by the theory.

<table>
<tr><th></th><th>DFIT</th><th>SFIT</th><th>TREND</th><th>MLFIT</th></tr>
<tr><td>altruist</td><td>13%</td><td>9%</td><td>9%</td><td>13%</td></tr>
<tr><td>defector</td><td>31%</td><td>28%</td><td>17%</td><td>22%</td></tr>
<tr><td>rewarder</td><td>22%</td><td rowspan="2">37%</td><td>30%</td><td>26%</td></tr>
<tr><td>cautious rewarder</td><td>9%</td><td rowspan="2">20%</td><td>13%</td></tr>
<tr><td>mild defector</td><td>9%</td><td rowspan="2">9%</td><td>6%</td></tr>
<tr><td>cautious</td><td>2%</td><td>6%</td><td>2%</td></tr>
<tr><td>unclassified</td><td>13%</td><td>17%</td><td>19%</td><td>19%</td></tr>
</table>

Table 3: Distributions of behavioral strategies among all subjects in HREP, according to the four classification methods.

To investigate the variations between the four classifications we first identify the subjects for which we can define a "consensus classification" and then check how much each method agrees with it. We classify a subject into a consensus category when at least three methods agree on this category. If three or four methods leave a subject unclassified, her consensus classification is also unclassified. For sophisticated strategies that SFIT and TREND do not distinguish, we permit any feasible precise classification in the consensus classification. A subject who is classified as rewarder by SFIT can be classified as rewarder or cautious rewarder in the consensus classification and a subject who is classified as cautious by SFIT can be classified as mild defector or cautious in the consensus classification.[7] A subject who is classified as sophisticated by TREND can be classified as cautious rewarder or mild defector in the consensus classification.[8] This approach slightly increases the chance that SFIT and TREND will agree with the consensus classification.

[7]Most subjects classified as cautious by SFIT turn out to be mild defectors in the consensus classification.

[8]So, if DFIT, SFIT, TREND and MLFIT classify a subject as mild defector, rewarder, sophisticated and cautious rewarder, respectively, the consensus classification classifies her as cautious rewarder. All methods except DFIT are considered consistent with the consensus classification for this subject. DFIT is not consistent because it does not classify the subject as cautious rewarder, which it can distinguish.

Table 4 shows for each method the proportion of subjects for which it agrees with the consensus classification. For 9 out of 54 subjects (16.67%) there is no consensus, and we omit them from this comparison.[9]

	DFIT	SFIT	TREND	MLFIT
Agreement with consensus	91%	93%	82%	98%

Table 4: Proportions of subjects for which the four methods agree with the consensus classification. Only the 45 subjects (83.33%) assigned the consensus classification are considered.

MLFIT agrees with the consensus classification for all but one subject. SFIT and TREND are less aligned with the consensus, despite having the advantage from a coarser set of strategies. On the other hand, DFIT and SFIT achieve relatively high agreement while using fewer parameters than MLFIT. TREND is the only method not directly fitting the behavior to the theoretical set of strategies, which may explain its relatively low agreement. In particular, for conditional strategies TREND requires a significant coefficient estimate in a subject's individual logit helping regression but ignores her overall helping rate, overestimating conditional behavior.

Having confirmed that MLFIT can be successfully used to classify individual behavior, we use it exclusively for our analysis below. MLFIT also deals relatively easily with stochastic and deterministic versions of theoretically postulated strategies and is more flexible than SFIT and TREND in that it can easily account for new strategies. This will help us identify an important new strategy in HBASE.

2.3.3 Strategies in HBASE

In HBASE the senders cannot access their receivers' reputations but might still remember their own past actions. The cautious strategy is therefore the only conditional strategy, considered in our strategy set, they can apply. From the strategic point of view, it is not meaningful, though, because it does not make any sense to invest in own reputation since it is not visible. The left panel of Table 5 shows the distribution of strategies in HBASE as estimated by MLFIT. Since no one played the cautious strategy, we omit it in Table 5. This is not surprising, though, and in line with our previous argument. The resulting classification is not very informative, leaving a large proportion of subjects (41%) unclassified. Perhaps the lack of reputational information hinders strategic and motivates random behavior and experimentation. But it is also possible that the theoretical set of strategies misses a popular behavioral type. For instance, it is

[9]This may happen for instance when each method assigns a subject to a different category. By including all subjects, the fit percentages in Table 4 are 16.67% lower, but the method ranking remains the same.

conceivable that in absence of reliable information about the receivers' past helping choices the senders react to their own experience as receivers of help. Such learning and experience-based behavior has been suggested as theoretically plausible by Boyd and Richerson (1989), Dilmé (2016) and Camera and Gioffré (2022), and observed in Bolton et al. (2005), Seinen and Schram (2006) and Swakman et al. (2016) but was never included in the experimental analyses of behavioral strategies in the repeated helping game. The most closely related experimental game in which it was included was the infinitely repeated prisoners' dilemma game with random matching (Camera et al., 2012). Based on this literature, our first estimations in HBASE and HREP, and tacit assumption of the previous literature that when the reputation is observable subjects use reputation rather than experience as a focal point, we hypothesize the following.

H1: *Subjects use experiential strategies in HBASE.*

H2: *Subjects do not use experiential strategies in HREP.*

To account for possible experience-based behavior, we introduce three new behavioral strategies under the umbrella term *experiential*. An experiential subject helps only if she had received sufficient help in her recent past. There can be several strategic formulations for such behavior, however. A key variation is in the scope of recall, which can be affected by the memory length and decay.

An extreme example is an individual who reacts only to her most recent experience as the receiver, driven for example by negative emotions such as anger or frustration, triggered after she did not receive help. She thus ignores or forgets all her previous experiences and can be described with a strategy of a minimum memory or maximal decay; the *EX1* strategy dictates help if the subject received help in her last interaction as receiver.

In contrast to emotions that trigger a strong immediate response, experiential behavior may be driven by learning or adaptation from experience over several rounds as a receiver. This can be modelled with limited memory or memory decay, and we consider one strategy of each kind. The subject using the *EX3* strategy considers the most recent three experiences as a receiver with equal weight and ignores all other experiences. We choose a strategy with memory length 3 for consistency with all our other reciprocal strategies (e.g., rewarders, cautious rewarders) which are also based on three recent actions. Similar longer memory strategies were studied in some earlier experimental studies involving infinitely repeated prisoners' dilemma games (Fudenberg et al., 2012; Camera et al., 2012). For classification we consider three deterministic substrategies *EX3K* which dictate help if the subject received help at least *K* times in the recent three opportunities, and one stochastic substrategy *EX3S* that dictates help with probability *x/3* if the subject received help *x* times in the recent three opportunities.

The subject using the *EXP* strategy, on the other hand, considers all received experiences, with weights depending on their recency via hyperbolic discounting of the past. Hyperbolic memory discounting is among the most considered memory discounting functions in psychology (Rubin & Wenzel, 1996; Yi et al., 2006; Findley, 2015). *EXP* constructs an index of received help and dictates help when this index is sufficiently high. Index h_P for *EXP* in round t is given by

$$h_P = \sum_{r=1}^{t-1} \frac{H(r)R(r)}{t-r} / \sum_{r=1}^{t-1} \frac{R(r)}{t-r},$$

where R and H are index functions: $R(r) = 1$ if an individual was receiver and $H(r) = 1$ if he received help in round $r < t$.[10] More distant experiences have a lower weight than more recent ones. We again consider three deterministic substrategies *EXPK* which dictate help if $h_P \geq K/4$, and one stochastic substrategy *EXPS* constructed as the linear combination (average) over all corresponding deterministic *EXPK* substrategies. In particular, *EXPS* dictates help with probability $x/3$ if $h_P \in [x/4, (x+1)/4)$, for $x \in \{0, 1, 2\}$, and with probability 1, for $x \geq 3/4$. As a remark, if such a (sub)strategy best fits a subject's behavior this just means that she uses one of the many possible long-memory experiential strategies, possibly with memory decay. It does not mean that subjects are actually doing such complex calculations consciously before choosing an action. Our proposed strategy is therefore just a representative of long-memory strategies that take into account that more distant past is more irrelevant due to forgetting.

The right panel of Table 5 shows MLFIT* - the MLFIT modified with the inclusion of experientials. The distributions presented in Table 5 are significantly different ($p<0.001$, Stuart-Maxwell homogeneity test), which was due to experientials, defectors and unclassified subjects as shown by the pairwise McNemar post hoc test with Bonferroni correction. The share of unclassified subjects drops to 11% and more than a half of subjects are classified as experientials, confirming our hypothesis H1 and that experiential strategy is an important behavioral strategy which was missing in our original set of strategies. It is worth mentioning that all 11% of subjects in category "unclassified" are those who consistently apply random strategy - in other words, in HBASE there are no inconsistent subjects who apply more than one strategy.

Result 1: *Subjects use experiential strategies in HBASE, where it is the modal behavior.*

[10] A similar index was introduced and used by Gong and Yang (2019) in a repeated prisoners' dilemma experiment. Note that our index h_P is defined for rounds $t \geq 2$ and that we have division by zero (i.e., 0/0) in h_P if a subject has not yet been a receiver in any round so far. For this case we can set the value of index equal to -1, for example. In fact, we can choose an arbitrary value (different from those we have already used), because the initial rounds with incomplete infomration are neglected in the main analysis anyway. We later show (subsection Robustness) that the exclusion of these initial rounds does not significantly alter the strategy distribution.

	MLFIT
altruist	7%
defector	52%
unclassified	41%

	MLFIT*
experiential	56%
altruist	4%
defector	30%
unclassified	11%

Table 5: Distribution of strategies in HBASE as estimated by MLFIT (left panel) and MLFIT* (right panel).

2.3.4 Learning in HREP

Given the prominence of experience in HBASE, we next test whether it drives behavior also when reputation information is available. This would invert the structure of indirect reciprocity. While rewarding is interpreted as "*I help you because you were helpful to others*" (downstream or reputation-based reciprocity), experiential behavior can be justified as "*I help you because others have helped me*" (upstream or experience-based reciprocity).

Table 6, column 1, shows that even with reputation information available, 7% of subjects were driven mainly by their own experience, contrasting our hypothesis H2. Accounting for them in MLFIT* reduces the unclassified category from 19% to 13% (i.e., by 30%) but does not substantially alter the distribution of other categories. In particular, MLFIT* and MLFIT distributions are not significantly different ($p>0.1$, Stuart-Maxwell homogeneity test). This validates the distributions of theoretically postulated behaviors found in previous studies on reputation-based helping which did not check for experience-based reciprocity and indicates that reputation mechanism is the key mechanism behind generosity. Nevertheless, experiential behavior explains a substantial part of previously unclassified subjects, indicating an improvement in classification.

Result 2: *Some subjects use experiential strategies even in HREP, where the reputation-based strategies (rewarder and cautious rewarder) and the defector strategy are the most common.*

Adapting to experience may be intuitive in absence of reputation but is curious when reputation facilitates conditional strategies such as rewarding. Experiential is also among the least profitable behaviors in HREP (see Figure 12 below). We can propose several possible explanations for experience-based reciprocity.

Ekeh (1974) suggests that experience-based reciprocity comes from an obligation to reciprocate, and Nowak and Sigmund (2005) interpret it as a misdirected act of gratitude. However, we observe a decline in generosity with rounds, and this suggests more negative explanations for experience-based reciprocity in helping games. By giving help and not receiving it, a subject may for instance feel she has been treated unfairly, triggering powerful negative emotions such as anger, frustration or distress, which she relieves by being selfish

towards strangers (Austin & Walster, 1975). Upstream reciprocity may also result from a sense of entitlement to behave selfishly after suffering injustice (Zitek et al. 2010). Finally, some subjects may simply conform to a perceived norm they learned about from their actual experience. Whether experientials are mainly driven by the most recent experience or learning and adaptation is further explored in Substrategies subsection below.

To conclude, column 2 in Table 6 shows average posterior probabilities *a.p.p.* (and standard deviations in parentheses) of all strategies, where a.p.p. of each strategy is calculated from posterior probabilities of assigned individual strategies. A.p.p. is a measure of certainty: the higher the value, the greater the certainty that subject plays a particular strategy (and not some other strategy from our strategy set). It can also be seen as a measure of consistency, as higher value indicates a consistent application of a particular strategy throughout the experiment. In particular, each of our seven strategies has high average posterior probability, indicating that most subjects consistently apply one strategy.

	MLFIT*	a.p.p.
altruist	13%	1.00 (0.00)
defector	22%	0.94 (0.11)
rewarder	24%	0.87 (0.16)
cautious rewarder	13%	0.87 (0.20)
mild defector	6%	0.98 (0.03)
cautious	2%	0.91 (/)
experiential	7%	0.90 (0.16)

Table 6: Strategy distribution in HREP, classified by the MLFIT* that includes experientials, and average posterior probabilities a.p.p. (and standard deviations in parentheses) of all strategies. Unclassified subjects are omitted.

2.3.5 Substrategies

In this section we explore subjects' behavior further by looking at the specific substrategies they used. This will tell us whether experientials are driven only by the most recent events or also by more distant ones. We will also learn which norms particular types of subjects adopt. For example, we will learn whether for rewarders, to help, is already enough that their receivers helped once in the last three occasions, or do they require that their receivers have perfect reputation. We will also see whether subjects' behavior is better captured by deterministic or stochastic substrategies.

Table 7 shows the substrategy distribution in HREP and HBASE, classified by the MLFIT* that includes experiential substrategies. One thing to note is that in both treatments no one is using EXP1 strategy which means that no one is reacting to the most recent experience only. This suggests that experiential subjects, rather than just imitating their most recent experience - which is cognitively simple, keep track of more distant encounters, learn from them and adapt to group norms. While in HBASE both long-memory strategies (EX3 and EXP) are equally common, in HREP the experiential strategy based on longer memory and hyperbolic memory discounting best captures the behavior of all experientials. This suggests that experientials indeed assign lower weight to more distant events and forget over time. In order to see whether EXP can capture all experiential behavior, which would lower the number of parameters that the model must estimate and facilitate the computations, we checked what happens with classification in HBASE if we exclude EX3 strategy from the model and only include experiential strategy with memory decay EXP. We found that almost all strategies that were classified as EX3 became EXP, resulting in the strategy distribution that was not significantly different from the strategy distribution displayed in Table 5, right panel ($p>0.1$, Stuart-Maxwell homogeneity test). This finding suggests that it is enough to include in the analysis of individual strategies only one type of experiential strategy (in terms of the scope of recall) - that is, the long-memory strategy with memory decay (EXP) - without fear of leaving any major experiential behavior undetected.

Result 3: *Experientials use long-memory strategies.*

In HBASE we found that many experientials were willing to provide help as long as they received at least one help recently. Anticipating (correctly) that it will be difficult to sustain help in a completely anonymous setup, the helping norm was set low. In our experiment most experientials used stochastic substrategies meaning that the probability of their help was increasing with the number of past helps received. The remaining experientials used deterministic experiential strategies and helped with probability 1 if their threshold (i.e., the number of helps they required to get in order to help) has been reached. In HREP, stochastic substrategies were also common, again showing that many subjects employed non-threshold strategies that react differently to different reputation information and experience. Among rewarder substrategies, the deterministic substrategy rewarder1 was often played, demonstrating that many rewarders were willing to help whenever their peers had provided help at least once in the last three occasions. This again shows that the norm was set low. From Table 7 one can also calculate and verify that in both treatments 50% of subjects played one of the deterministic substrategies and almost 40% of subjects played one of the stochastic substrategies.

	HREP		HBASE
altruist	13%	altruist	4%
defector	22%	defector	30%
stochastic rewarder	11%	stochastic experiential	13%
rewarder1	11%	experiential 1	11%
rewarder3	2%	experiential 3	2%
stochastic cautious rewarder	13%	stochastic minf	26%
stochastic mild defector	4%	minf 1	2%
mild defector10	2%	minf 3	2%
stochastic cautious	2%	unclass.	11%
stochastic minf	7%		
unclass.	13%		

Table 7: Substrategy distribution in HREP (left) and HBASE (right), classified by the MLFIT* that includes experiential substrategies.

2.3.6 Profitability in HREP

To investigate why some behavioral strategies are more popular than the others we look at their relative profitabilities. Figure 12 displays the profitability for each strategy in HREP, calculated from the average round payoffs for senders using this strategy, the average round payoffs for receivers using this strategy, and the overall average round payoff. The detailed description how we determine profitabilities is described in the next paragraph.

To calculate relative profitabilities we follow the procedure in Ule et al. (2009). For each subject, we first calculate her average payoff in rounds when she was sender and then again for the rounds when she was receiver. We then average these two values to calculate the expected earning for this subject, the average round payoff she would have received had she been in both roles equally often. These averages were calculated based on the first 90 rounds.

Our independent observational unit is a matching group. For each strategy used by at least one subject in a group we therefore determine its group payoff, calculated as the average expected earnings of all subjects in this group that used this strategy (Table 8). The strategy profitability is then the average of its group payoff over all groups where it was used, which yields also the standard errors. The relative profitability of a strategy is the difference between its profitability and the average profitability over all strategies divided by this average. Since not all strategies were used in all matching groups, we aggregate them for statistical analysis. We coalesce selfish strategies (defector, cautious and mild defector) and reciprocal strategies (rewarder and cautious rewarder).

	altruist	defector	rewarder	cautious rewarder	mild defector	cautious	experiential	unclass.	selfish	reciprocal
g1		5.86	4.05				-0.97	5.14	5.86	4.05
g2	-4.18	48.91	17.19					-15.09	48.91	17.19
g3	-13.75	52.33	11.96					21.21	52.33	11.96
g4		11.25	19.91	14.91	32.78		4.00		22.02	18.24
g5		44.65	29.64	25.38					44.65	27.08
g6	26.50		16.89	20.88	49.29		29.05	50.97	49.29	18.88
g7	20.00	38.89	32.19	43.71					38.89	37.95
g8	-8.03	34.55	18.82				9.39		34.55	18.82
g9	-8.15	47.24			55.13	23.75		31.89	42.04	
strategy profitability	2.07 (6.87)	35.46 (6.22)	18.83 (3.19)	26.22 (6.21)	45.73 (6.69)	23.75 (/)	10.37 (6.58)	18.82 (11.29)	37.61 (5.02)	19.27 (3.54)
average profitability over all strategies and groups	21.75									

Cell entries correspond to the expected round earnings (in francs) across all subjects assigned to a particular strategy in this group. Selfish types include defectors, cautious and mild defectors. Reciprocal types include rewarders and cautious rewarders. Strategy profitability is its average payoff over all groups, with standard errors in parentheses.

Table 8: Strategy payoffs per group in HREP.

The mild defector and defector strategies were the most profitable strategies while the altruist strategy was the least profitable, suggesting that other-regarding behavior is not sufficiently rewarded. Moreover, reciprocity itself was not a profitable behavior in HREP. Together, the reciprocal strategies were less profitable than the selfish strategies ($p<0.01$, paired permutation test). It is curious that despite its relatively poor payoff performance rewarding was nevertheless the most popular behavior. In contrast, mild defection is very profitable but rare, only partly confirming the theoretical prediction of its flourish (Leimar & Hammerstein, 2001). Finally, experiential behavior was not particularly profitable, which is unsurprising given that it reacts to the past and neglects own future reputation. Unsurprisingly, defectors earned the most also in HBASE, followed by experientials and altruists.

Result 4: *The selfish strategies are more profitable than the reciprocal strategies. The experiential and altruist strategies are the least profitable.*

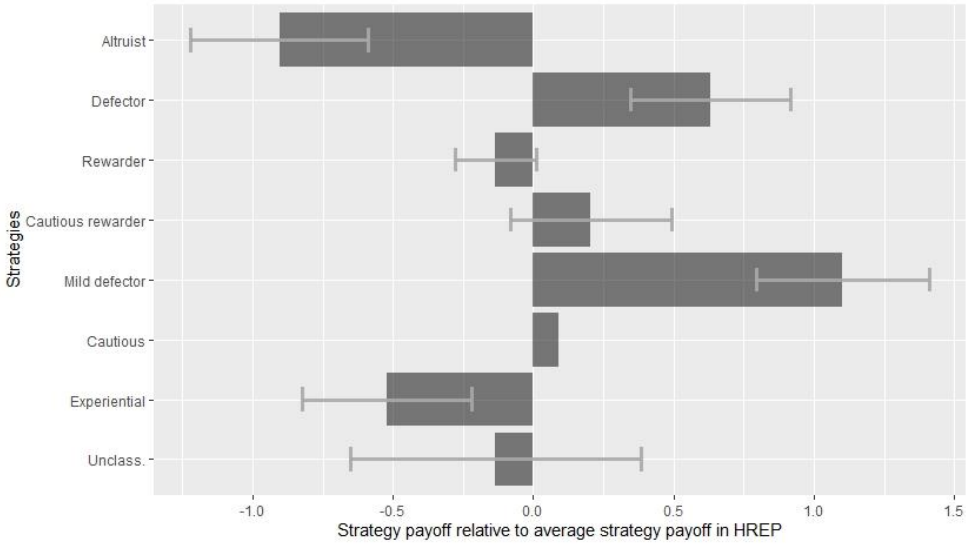


Figure 12: Bars show the average strategy payoffs relative to the average payoffs across all strategies in HREP, with ±1 SE shown by error bars.

2.3.7 Robustness checks

Last rounds

MLFIT* strategy distributions presented above are based on individuals' data where the last rounds are excluded to avoid potential noise caused by the end-game effect. Similar has been done for example by Seinen and Schram (2006) and Ule et al. (2009). But what happens if we add last rounds? To our knowledge, this was not explored before, so this further analysis may yield some interesting new insights.

Our main findings are summarized in the following transition matrix (Table 9). The sum of the numbers on the diagonal (i.e., the trace of the matrix) represents the number of subjects who were estimated to apply the same strategy regardless of whether the data from the last rounds is included or not. In contrast, the sum of the numbers off the diagonal represents the number of subjects whose estimated behavior depends on whether the last rounds are included or not. The Stuart-Maxwell homogeneity test confirms that distribution does not significantly change if we include the last rounds in the analysis ($p>0.1$), which confirms the robustness of MLFIT*.[11] The sum 8 on off the diagonal however indicates that last rounds had a reasonable effect on eight subjects on which we focus next.

The transition matrix yields two interesting results. First, by adding the last rounds we get three experientials more. These were originally classified as either altruist, rewarder or cautious

[11]We obtain the same conclusion in HBASE.

rewarder, that is, as one of the most generous types. Second, cautious behavior diminishes substantially when we add last rounds, as more than half of cautious rewarders as well as the only purely cautious subject change their strategy. Cautious rewarders switch either to a similar strategy that is more selfish (one became mild defector) or to a strategy that completely neglects own reputation (two became rewarders and one became experiential), whereas the behavior of the cautious subject became unclassified.

These two results illustrate how end-game effect affects certain individuals. On the one hand it reduces cautious behavior, that is, strategic helping, and on the other hand, it increases experiential behavior. The experience however negatively affects the helping behavior as more subjects pass because others pass. Both results point in the same direction, namely, that end-game effect reduces helping which makes perfect sense.

	altruist	defector	rewarder	cautious rewarder	mild defector	cautious	experiential	unclass.
altruist	6						1	
defector		12						
rewarder			11				1	1
cautious rewarder			2	3	1		1	
mild defector					3			
cautious								1
experiential							4	
unclass.								7

Numbers on the diagonal represent subjects who are estimated to apply the same strategy regardless of the inclusion/exclusion of the last rounds. Numbers off the diagonal represent subjects who are estimated to apply different strategies when the last rounds are included. The numbers in specific cells also tell the exact number of subjects who switch to a particular strategy. For example, 1 in the first row and seventh column means that one subject who was classified as altruist when the last rounds were excluded became experiential when the last rounds were added.

Table 9: Transition matrix (HREP). Last rounds.

Initial rounds

Above we investigated what happens if we include the last rounds in the individuals' strategy analysis and found that the strategy distribution does not change significantly. In the original analysis we also exclude the initial rounds when subject's reputation and experience may have less information or be even empty. It would be curious to know whether our strategy classification is robust to the inclusion of the initial rounds or does the inclusion significantly alter the shape of strategy distribution. If yes, then one should not neglect the initial rounds in the analysis by default, because they may contain important information. Again, to our knowledge, this was not addressed in the previous literature.

Our main findings are summarized in the following transition matrix (Table 10). The sum of the numbers on the diagonal represents the number of subjects who were estimated to apply the same strategy regardless of whether the data from the first rounds is included or not. In

contrast, the sum of the numbers off the diagonal represents the number of subjects whose estimated behavior depends on whether the first rounds are included or not. The Stuart-Maxwell homogeneity test confirmed that the distribution did not significantly change when the initial rounds were included in the analysis - neither in HBASE nor HREP ($p>0.1$). This was expected, though, given that only few (less than 10%) of the first 90 rounds were skipped.

	altruist	defector	rewarder	cautious rewarder	mild defector	cautious	experiential	unclass.
altruist	7							
defector		9	1				2	
rewarder			10	2				1
cautious rewarder			1	6				
mild defector					3			
cautious						1		
experiential							4	
unclass.			1					6

Numbers on the diagonal represent subjects who are estimated to apply the same strategy regardless of the inclusion/exclusion of the first rounds. Numbers off the diagonal represent subjects who are estimated to apply different strategies when the first rounds are included. The numbers in specific cells also tell the exact number of subjects who switch to a particular strategy. For example, 2 in the second row and seventh column means that two subjects who were classified as defectors when the first rounds were excluded became experiential when the first rounds were added.

Table 10: Transition matrix (HREP). Initial rounds.

The transition matrix yields one main finding, namely that the inclusion of the first rounds results in three defectors less. Two were reclassified as experientials and one as a rewarder, implying that they were reclassified as a more generous type because they provided some help initially. The two subjects who became experientials helped only in the initial rounds until they were receiving help regularly. A subject who became a rewarder was in a very selfish group and reacted to reputation very weakly, and mostly in the first rounds. This finding suggests that in classifications that exclude the first rounds, the share of defectors may be slightly inflated due to the subjects who abandon their initial plan right after a few rounds. The exclusion of the first rounds, however, does not have a significant effect on the overall distribution.

In general, the issue with low number of certain observations[12] can be avoided by letting subjects play several supergames with different subjects which increases the chances that the subjects who are willing to help will meet more generous subjects.

2.3.8 Post-experimental self-reports

Our final analysis checks whether subjects' post-experimental self-reports are reliable descriptions of their behavior. At the end of the experiment subjects were asked to describe in

[12]For example, in a group of five defectors and one rewarder the rewarder will rarely help because the reputation of her opponents is always bad. Since the model takes the actual data as an input, it does not know what such subject would do after observing good reputation, making the estimations less precise. In our experiment this did not represent a major problem because we had 100 rounds and relatively heterogeneous groups.

words how they were making their decisions in the experiment. Their reports were then classified into nine (six in HBASE) categories of which: six (three) were for the standard strategies, one for the experiential strategy, one (i.e., category 'unclassified') for strategy descriptions that do not fit into previous categories, and one reserved for self-reports that did not describe a feasible strategy or described just a small part of it. Coding was performed by three incentivized and trained coders.[13] In particular, a coder earned 20 cents for each subject they classified into the same category as another coder. Coders therefore played a coordination game where the focal point was the content of the written descriptions (Houser & Xiao, 2011). Since earnings were based on coders' performance rather than on a predetermined flat fee (commonly employed in the content analysis), coders had an extra incentive to associate each written description with strategy that suits it the best. The content of the written description can therefore be thought of as a focal point that facilitates coders' classification and potentially improves their overall performance. From researchers' point of view, the main advantage of such approach is that it motivates coders to carefully examine and think about the content of written descriptions and search for hidden clues (focal points), which may ultimately lead to a more accurate classification. This classification method is also less subjective than the alternative method based on experimenters' self-classification, as coders did not interact during the task, did not know other classifications, and were not told the research hypotheses or the expected distribution of strategies. Coders completed the task in approximately four hours and earned on average 31.8 EUR.

To assess the agreement between the three coders, we calculated Fleiss' kappa statistic (Fleiss, 1971). In HBASE it was 0.35, and in HREP it was 0.44. This suggests a fair agreement in HBASE and moderate agreement in HREP (Landis & Koch, 1977). If two coders classified a particular subject's written description into the same category, then this category counted as final (81.5% such cases in both treatments); if they all disagreed then we selected one of the three suggested categories as final. We followed the procedure in Houser and Xiao (2011); if all coders disagreed, we selected one from the three suggested categories that we thought was the most appropriate. If we would consider such self-reports as unfeasible then our conclusion about unreliability would strengthen (around one third of self-reports would then be unfeasible in each treatment, though the match with MLFIT* would remain similar).

The classification of self-reports yielded a rather poor match with the MLFIT* classification of strategies based on actual decisions, especially in HREP which was more cognitively demanding treatment. First, a substantial number of subjects did not submit a meaningful strategy description or submitted an incomplete description (Table 11, left panel). Second,

[13]Coders were first trained on artificial self-reports and proceeded to the actual task only when they correctly classified all artificial self-reports and understood the concepts behind the strategy categories. Their instructions are shown in Appendix A3.

fewer than 70% of the remaining (feasible) self-reports matched the MLFIT* (Table 11, right panel).

Result 5: *Self-reports are not a very reliable source of information in the repeated helping game.*

Self-reports turned out a poor account for the actual helping behavior in our experiment, even if most subjects submitted descriptions of feasible behavioral strategies. No regularity was found in discrepancies between self-reports and MLFIT* classifications, suggesting that they are not a consequence of a bias in MLFIT*. We posit that many subjects are not aware of their own heuristics or any regularity in own behavior, even when their actions follow a standard behavioral rule. Subjects may also misattribute own choices due to fabrication (possibly related to the social desirability bias, Fisher, 1993) or forgetting (Russo et al., 1989), which is plausible in longer experiments. Our observation is therefore consistent with the economists' consensus on the non-veridicality of post experimental (retrospective) reports.

	BASE	REP		BASE	REP
Unfeasible self-reports	20.4%	16.7%	Match with MLFIT*	67.4%	53.3%

The left panel shows the proportions of reports that do not describe a feasible strategy. The right panel shows the proportions of feasible self-reports that match the MLFIT* classification.

Table 11: Evaluation of subjects' self-reports.

The result that classification based on self-reports provides worse fit to the actual sequences of choices than MLFIT* is an immediate by-product of the mixture model-based method. If the strategy distribution of the self-report classification, or at least some similar strategy distribution, fitted the data better than MLFIT*, then the mixture model-based method would return it as the best fit (according to AIC), because it would solve the maximization problem defined on page 34 – but it does not. Moreover, even the simplest counting method DFIT, where the actual choices are compared to the choices predicted by the strategy assigned to a subject, results in a distribution that is quite different from the self-report distribution. In fact, the self-report distribution is also quite different from the distributions of the other two methods that rely on regressions instead of on a simple count statistic, confirming that self-reports rather than statistical methods are problematic. Self-report classification is the only classification that is based on subjects' subjective perceptions, views and experiences and not on their actual observed choices, and as such may be exposed to greater bias than more formal objective statistical methods.

2.4 Conclusion

In this chapter we apply the mixture model-based estimation method to investigate the behavioral strategies that subjects apply in experimental repeated helping games between

strangers. We find that the method gives robust characterizations for almost 90% of our subjects and is closest to the consensus measure derived from four estimation methods from the literature. In contrast, our subjects' written self-reports were not a reliable source for strategy classification.

Next, we utilize the flexibility of mixture model estimation method to detect a new experiential strategy class, describing strategies of individuals who react to their own experience rather than to reputation. Previous statistical estimates in the literature did not consider these strategies and we find that this is an important omission. Experiential behavior is present in both experimental treatments, and it is even the modal behavior in the treatment without reputation. Moreover, our behavior analysis suggests that experientials do not react to the most recent experience only (short memory) but rather use strategies based on longer memory and memory decay. Experience-based behavior has been tangentially discussed in the theoretical literature on the evolution of indirect reciprocity. We show that it is sufficiently prevalent to merit more prominent discussion in both the theoretical and the empirical work. We also show that concern for own reputation diminishes in the final rounds of the experiment, which can explain the end-game decline in generosity.

Our final analysis shows that selfish strategies are more profitable than reciprocal strategies, and that reaction to experience and unconditional altruism do not pay off. This crafts a challenge to the evolutionary or economic explanations for the experience-based generosity, and we speculate about the alternative emotional or normative sources of experiential behavior that we uncover in our experimental study.

Chapter 3

Honesty and deception among strangers

3.1 Introduction

Trust in the advice of strangers is an increasingly important element of market and daily interactions. This matters when incentives of interactive parties are aligned but may be hard to achieve when they are in conflict. A common example of how misaligned incentives and information asymmetry may unravel a market was proposed by Akerlof (1970) in his analysis of the market for "lemons". He illustrated the problem using a used car market where there are used cars of different qualities (a "lemon" refers to a low-quality car), sellers who know the exact quality of their used car, and buyers who do not know the quality of used cars but have some information about the used car market so they can compute some market statistics. Since buyers know that cars can be of different qualities, their natural guess is that a particular car is of average quality (instead of average one could also assume median or modal quality). For such a car, a buyer will not be willing to pay a high price (i.e., the asking price for a high-quality car), so that sellers with high-quality cars will leave the market. This will reduce the quality of the cars in the market and consequently buyers will be willing to pay even less than before because the average quality of cars will degrade once the sellers of high-quality cars leave the market. After several repetitions of such "quality/price" reductions, only cars with the lowest quality will remain in the market that will either be sold at the lowest price or not sold at all. This simple yet important example is used as an illustration to show how, in general, low-quality goods can drive goods of higher quality out of the market if there is information asymmetry. Information asymmetry also results in inefficiency. For example, even if initially there is a seller that is selling a high-quality good and a buyer that is willing to pay a high price for a high-quality good, such trade will never be executed because a buyer will not recognize that a good is of high quality. In addition to that, the information asymmetry may also give rise to dishonest reporting of the quality of goods because the buyers cannot determine the quality of goods. Thus, dishonesty can drive honesty out of the market like low-quality goods drove high-quality goods out of the market. This again results in inefficiency as those buyers who are a priori willing to pay a higher price for a good of a higher quality would definitely not be willing to pay such price after finding out that sellers may exaggerate the quality of their goods.

The information transmission between an informed individual (e.g., a seller, an expert, a sender) and an uninformed individual (e.g., a buyer, a receiver) have attracted a lot of attention among scholars. A formal game-theoretic model was proposed by Crawford and Sobel (1982) who investigated how the information transmission between the rational individuals (in

equilibrium) varies with increasing misalignment (difference) in their preferences. Their main finding was that in equilibrium the more their incentives are misaligned the less information is transmitted. The perfect information transmission (i.e., always honestly revealing the private information) is possible only if individuals' preferences are perfectly aligned, whereas if preferences sufficiently differ no information transmission is expected. This and other standard theoretic models of strategic information transmission are built under standard game-theoretic assumptions including the assumption that individuals are selfish and do not have direct (psychological) costs associated with deception, meaning that they deceive whenever that is materially advantageous.

In principle deception can have different consequences on individuals' payoffs. It can either i) increase the payoff of a sender (informed individual) and decrease the payoff of a receiver (uninformed individual), ii) decrease the payoff of a sender and increase the payoff of a receiver, iii) increase the payoff of both a sender and a receiver, or iv) decrease the payoff of both a sender and a receiver. While all four types have been studied in the experimental literature (e.g., Erat & Gneezy, 2012; Sasaki et al. 2019), the first type gained the most attention because it captures the most natural situation where preferences of differently informed individuals are misaligned, and deception is advantageous for an informed sender. One such example is a used-car sale where an informed seller may try to deceive a buyer about the quality of her car to increase her profit. In this thesis we focus exclusively on the first type of deception.

Given its prominence in everyday life, economics and business, deception and other types of dishonest behavior have been extensively studied in laboratory experiments over the past twenty years. Experimental evidence is more optimistic about individuals' moral behavior than the theory, suggesting that some people have reservations against deceiving others. In an influential paper, Gneezy (2005) reported the results of a simple deception game where senders could increase their profit at the expense of their receiver by deceiving the receiver. He found that many senders refrained from deception. To confirm that this is due to deception aversion and not due to social preferences he ran an additional experiment with dictator games that had the same payoff allocations, i.e., monetary rewards, as his deception games. Then he compared the fraction of deceptive messages sent by senders in deception game with the fraction of "selfish" options chosen by senders in dictator game, where "selfish" refers to options that are materially advantageous for senders. He observed that the fraction of deceptive messages in deception games is significantly lower than the fraction of selfish options chosen in dictator games and attributed that to deception aversion. Using the same deception games, Sutter (2009) confirmed that many subjects send truthful messages, but warned, based on the results of his new experiment where he additionally elicited senders' beliefs about receivers' behavior, that the actual frequency of deception might be even higher, because some senders might engage in sophisticated deception where they tell the truth but believe that their receiver will not trust them. Hurkens and Kartik (2009) also studied deception games but unlike Gneezy investigated

the decisions of subjects playing both games (i.e., they employed a within-subject design) and classified them into four categories, based on whether they are selfish or generous and deceptive or honest. Their analysis revealed that many selfish subjects were honest despite the fact that honesty lowered their monetary payoffs. So, not everyone that acted selfishly also deceived, which again suggests that many subjects may be averse to deception. Deception aversion has been studied and seems to be present also in developing countries. In Leibbrandt et al.'s (2018) experiment in Bangladesh subjects could earn an amount equivalent to several months' income by deceiving, but many nevertheless refrained from deception. The result regarding the reservations against deception was recently supported by Vranceanu and Dubart (2019) who found i) that many subjects are honest when rewards for deceiving are reasonably small and ii) that some subjects remain honest, even when rewards for deceiving are high. Since they controlled for social preferences, their findings support the evidence that many people have some sort of psychological costs associated with deception, at least in one-shot interactions with strangers. Furthermore, as neuroimaging studies including Christ et al. (2009), Lisofsky et al. (2014) and Volz et al. (2015) have shown, deceiving seems to be more cognitive demanding than honesty, as during deception a greater activation in certain brain regions was observed than during honesty. Zuckerman et al. (1981) added that deception may require more effort because a deceiver must be consistent while deceiving and careful that she does not provide a statement that her reciever already knows is not true.

More recently Sasaki et al. (2019) ran a within-subject experiment where subjects made decisions in five deception games and five dictator games with different payoff schemes, of which three were such that deception or selfishness resulted in a benefit for a sender and a loss for a receiver (i.e., similar as in our experiment). Contrary to Gneezy (2005) and Hurkens and Kartik (2009) who both compared deception and dictator games, they found evidence for deception aversion in only one of the three payoff schemes, namely in one where deception was the least beneficial for a sender and the least costly for a receiver.

Despite mixed evidence provided by Sasaki et al. (2019), most existing studies agree that in one-shot interactions deception, or more generally dishonesty, is present but that some subjects have reservations against it. Do these reservations change, if subjects are explicitly told that deception is possible or if they are informed about their peers' behavior in similar past experiments? Such questions are addressed in the literature that studies the *contagion* of deception, or more broadly, dishonesty. These questions can again be explored using one-shot experiments in which for example subjects learn, before making their decision, that deception is an option or how many subjects deceived in similar past experiments. Fosgaard et al. (2013) found that simple awareness that certain environment gives rise to profitable dishonesty can increase dishonesty. They and several other studies (e.g., Innes & Mitra, 2013; Leib & Schweitzer, 2020) evidenced that subjects tend to engage in more dishonest behavior if they know that others are dishonest too (i.e., peers' effect), perhaps because it is easier to justify

their unethical behavior and thus maintain positive self-image. Contagion of deception and honesty along with contagion of selfishness and generosity has also been studied by Sasaki et al. (2019). They reported that deception and selfishness are both contagious, whereas honesty and generosity are not, thus revealing that information about socially undesirable behavior makes a higher impact on individuals than information about socially desirable behavior.

In general, many potential factors can influence dishonesty. Many of them are proposed in the meta-analysis by Gerlach et al. (2019), who separately discussed personal and situational factors. Personal factors include gender, age, student status (student vs. non-student) and study major (economists vs. non-economists). Situational factors include ethical reminders, peers' behavior, physical distance between the "perpetrator" and "victim" of dishonest act (e.g., whether they are both in laboratory or whether they are elsewhere and communicate online), reward size and externalities (e.g., how much damage/benefit dishonesty causes to others). The meta-analysis also indicates that dishonesty is higher in laboratory studies than in field studies.

So far, we have reviewed the results of experimental studies where experimental subjects engaged in one-shot interactions. While we have seen that these studies yielded many interesting insights, they did not provide answers to several important questions such as i) how honesty and deception evolve over time, and ii) what effect, if any, does the reputation mechanism have on the dynamics of honesty and deception. We also do not know whether deception aversion is a fleeting or a stable phenomenon as people learn and adapt to their social environments and social norms. These questions can be explored with a repeated experimental game. In addition, one of the advantages of running a repeated experimental game is that researchers are provided with multiple data for each subject, showing them how subjects reacted in different scenarios and giving them an opportunity to study what behavioral rules – strategies subjects use in the experiment. It would be interesting to know whether in deception games subjects condition their actions on the available information or is the provision of additional information worthless. Moreover, it would be interesting to know whether subjects' behavior can be captured by single strategies as it was the case in the helping game (see Chapter 2) and whether these strategies are analogous to those in the helping game. This will tell us whether honesty is promoted through indirect reciprocity. Another advantage of using a repeated game is also that it offers an opportunity to study social norms, for example reciprocity, or test the efficiency of different mechanisms such as reputation or punishment/reward mechanisms.

As explained in Chapter 1, a subject can play a repeated game under partners' or strangers' matching protocol. A direct comparison of these two matching protocols revealed that a partners' matching protocol is better at curbing dishonest behavior in completely anonymous environment, perhaps because partners can build a reputation through long-term relationships (Wilson & Vespa, 2020; Ben-Ner & Hu, 2021). A partners' matching protocol, however,

cannot be used in studies that explicitly want to avoid reputation building through long-term relationships (e.g., studies that want to estimate subjects' strategies in completely anonymous environments) or studies as ours that want to study indirect reciprocal honesty. Given that in our experiment we employ strangers' matching protocol, we focus next on experimental literature with strangers' matching.

Cai and Wang (2006) ran an experiment with sender-receiver games à la Crawford and Sobel (1982) and observed that senders transmit more information and receivers trust more than game-theoretic model predicts, which is evidence for excessive honesty and trust, respectively.[14] Since in their experiment subjects played multiple identical one-shot games (with different opponents), they had multiple data for each subject and could and did classify the behavior of subjects. However, since they considered a setup without reputation information, the subjects could not condition their behavior on opponent's or own reputation as in our game. To classify subjects, they followed Crawford's (2003) level-k reasoning model and assumed that subjects can be of different levels of reasoning:[15] as senders, they can be level-0, level-1, level-2, or sophisticated, whereas as receivers they can be level-0, level-1, level-2, sophisticated or equilibrium type. Level-0 senders are always honest, level-0 receivers best respond to level-0 senders and hence always trust, and for $k \in \{1, 2\}$, level-k senders best respond to level-(k-1) receivers, and level-k receivers best respond to level-k senders. Moreover, they also considered sophisticated sender's (receiver's) type who best responds to empirical distribution of receivers' (senders') behavior, and an additional receiver type that plays according to the game-theoretic prediction. They found that many subjects can be classified into one of the less sophisticated categories and propose that these are the reason for excessive honesty. Soon after, Sanchez-Pages and Vorsatz (2007) showed that excessive honesty is not necessarily due to the boundedly rational individuals. They demonstrated that this might be due to the so-called moral individuals that are driven by social norms such as honesty. At this point the following question arises: Is excessive honesty more likely due to individuals who gain utility by being honest (i.e., a preference for honesty) or due to individuals who lose some utility by deceiving (i.e., deception aversion)? To investigate this, Sanchez-Pages and Vorsatz (2009) conducted a new experiment where senders were given an alternative option to remain silent. Based on the new empirical data and further game-theoretic analysis they concluded that excessive honesty is more likely caused by deception aversion.

The above three studies examined the dynamics of honesty and deception but in a completely anonymous setup, where subjects did not know what their current counterparts had done in the past. Sanchez-Pages and Vorsatz (2007; 2009) additionally considered a treatment with costly punishment. Costly punishment proved to induce socially more desirable outcomes in the past

[14]Excessive honesty and trust simply mean that subjects are on average more truthful and trustful than the equilibrium predicts.

[15]Similar approach has later been taken by Kawagoe and Takizawa (2009).

experimental literature on cooperation and generosity (e.g., Fehr & Gächter, 2000; Ule et al. 2009), but did not induce more honesty in their experiments. It did induce more trust, though. As evidenced, for example, from the literature on indirect reciprocity, presented in Chapter 2, socially desirable behavior may be promoted through another channel, namely through reputation mechanisms that store and display the information about subjects' past behavior. A recent contribution to that strand of literature was made by Behnk et al. (2019) who investigated the role of reputation mechanisms in a repeated deception game similar to that used by Gneezy (2005). Behnk et al. (2019) showed that reputation mechanisms (which they called reporting systems) reduce deception and promote honesty. They compared two different reporting systems and found that the unbiased computerized (automatic) system is more reliable and efficient in deterring deception than the system based on individuals' subjective reports. When the computerized system is at work, the dynamics of deception is relatively stable over time, while trust slightly increases. In addition to that, they also found evidence that guilt aversion might be the reason behind observed honesty. Although Behnk et al. (2019), like us, studied the dynamics of honesty and deception, their study differs significantly from ours, including their main objective which was to compare two reporting systems in terms of reliability and efficiency in deterring deception. In addition, in Behnk et al. (2019) the roles were fixed during the experiment so each subject was either always a sender or always a receiver, making the indirect reciprocal honesty (e.g., "*be honest only to honest people*") and deception (e.g., "*deceive only deceivers*") impossible to study. They also did not investigate subjects' strategies or long-term persistence of deception aversion. Their design was also different, as only receivers were provided with reputation information about their current sender, whereas in our experiment receivers had no information whatsoever about their current sender. In our experiment, however, senders had information about their current receiver (i.e., what he did before in senders' role).

To date, little is known how honesty and any psychological costs associated with deception develop with time, especially in the presence of reputation mechanisms which are very important as they mimic real-life mechanisms such as gossiping and information sharing and spreading. On the one hand, an honesty norm may emerge with social sanctions imposed on deceivers. For example, subjects may sanction deceivers either by deceiving them (when deceivers are in receiver's role) or by not trusting them (when deceivers are in sender's role). On the other hand, reservations against deception may disappear after substantial experience of dishonesty, unravelling any previous trust in strangers in society. Related phenomena have been predicted and observed in the theoretical and experimental research on indirect reciprocity in helping games (Nowak & Sigmund, 1998a; 1998b; Seinen & Schram, 2006). In those, reciprocity may promote the development of altruism between strangers, but also lead to a vicious cycle of retaliation. Although honesty is not identical to generosity, with the right game structures and experimental design certain parallels between them may be drawn. However,

honesty may be driven by more than just social preferences, which Gneezy (2005) demonstrated comparing deception games and dictator games with identical payoffs. While generosity is motivated by social preferences, honesty is driven by both social preferences and an aversion to deception, which has been confirmed by a number of studies on deception (e.g., Sutter, 2009; Erat & Gneezy, 2012; Vranceanu & Dubart, 2019). The long run interaction of these two distinct behavioral motivations is still unclear, and one aim of this study is to investigate whether any aversion to deception in one-shot interactions is stable as groups learn and adapt.

By employing strangers' matching protocol and manipulating the access to reputation information between the treatments, we delve into indirect reciprocal honesty which has - at least to our knowledge - not been investigated before. There is recent evidence that deception is used as reciprocity device, at least in one-time deception opportunities. Namely, Alempaki et al. (2019) investigated the relationship between deception and direct reciprocity by employing a two-stage experiment in which subjects first played a dictator game and then a deception game in reverse roles. This gave deception game senders an opportunity to reciprocate selfish behavior by deceiving dictator game senders (as well as to reciprocate generosity with honesty). They found that deception in the second stage increases with selfishness experienced in the first stage. It is, however, unclear what would happen if such two-stage games would be played repeatedly or if deception game senders would play the game against a different subject in which case reciprocity will be indirect, as in our project.

As shown in Chapter 2, learning about the reputation of subjects' counterparts gives rise to a new class of strategies, namely reputation-based strategies such as the rewarder and cautious rewarder strategy. Therefore, one of the main goals of this chapter is to estimate strategies that individuals use in deception game. This will tell us, for example, if a simple conditional strategy "*be honest only to honest people*" is as popular as rewarding (in the helping game) and if it can sustain honesty (which was the main result from the literature on indirect reciprocal helping). We will also explore whether honesty is strategic in the sense that subjects apply a consistent type of behavior as group behavior evolves, and whether subjects adapt to their experience. To date, the literature is agnostic about behavioral strategies that subjects use in deception or sender-receiver games, especially when reputation information is observable, and we want to fill the gap in the literature with this project. Apart from learning how heterogeneous our experimental group is, and whether honesty/deception is more reputation or experience-driven, analysis of subjects' strategies will provide further insight into the dynamics of honesty and deception explored above.

To summarize, this study aims to investigate the dynamic of honesty and trust between strangers, long-term persistence of deception aversion, the role of reputation, and the strategic

basis of honesty. Behavioral strategies are presented in the next section. Section 3.3 presents our hypotheses and results. Section 3.4 is reserved for final remarks and our conclusion.

3.2 Strategies

In this section we introduce and describe behavioral strategies that could be representative of capturing the long-run behavior of most of our experimental subjects. The DREP treatment permits all strategies that we identified for the helping game with reputation (i.e., the HREP treatment). We also consider new strategies that take into account the strategic environment of the deception game that does not exist in the helping game. Finally, we introduce several strategies for decisions of receivers.

In our repeated deception game subjects in both roles actively participate and make decisions, one per round. Since a subject is sometimes a sender and sometimes a receiver she needs to think about her behavior in a role of a sender and in a role of a receiver. Therefore, if she uses a strategy, it will have two components – the first describing the behavior in sender's role, and the second describing the behavior in receiver's role. This is different from what we had in our repeated helping game, where strategies only described the sender's behavior. For simple terminology we will refer to the sender's and receiver's strategies as *sending* and *responding* strategies, respectively. The pair of a sending and a responding strategy of a subject will be called a *strategy pair*.

For a standard deception or sender-receiver game - apart from Cai and Wang (2006) and Kawagoe and Takizawa (2009) who under the assumption that subjects are boundedly rational classified subjects according to level-k reasoning model (Crawford, 2003) - we do not have, at least to our knowledge, the literature that would systematically analyze the behavior of experimental subjects. Although Cai and Wang's (2006) level-k classification is not the most suitable for our setup with reputation mechanism the study did report some findings relevant for our analysis of individuals' behavior. Namely, they found that i) there is heterogeneity in behavioral types, ii) the behavior of most senders (75%) can be described by one of the sender's behavioral types, iii) the behavior of most receivers (81.25%) can be described by one of the receiver's behavioral types, and iv) the subjects often do not use the same level of reasoning in the sender's and receiver's role. This last means that a subject can be for example L1-type as a sender and sophisticated type as a receiver. Point i) is important because it suggests that we should consider many different strategies. Points ii) and iii) are important because they suggest that subjects consistently use the same reasoning throughout the experiment. Point iv) is relevant because it suggests that subjects' behavior in sender's role might be independent of their behavior in receiver's role and can hence be analyzed separately.

Looking beyond the standard deception games, individuals' strategies in environments offering opportunities for profitable dishonesty were recently studied by Gneezy et al. (2013) in a repeated "die-rolling deception" game with random matching (without reputation information). This two-player game is a mixture of Gneezy's (2005) deception game and Fischbacher and Föllmi-Heusi's (2013) die-roll task. In this game a sender first privately observes a die roll and then sends to her receiver a message informing him about the outcome of the die roll. The receiver who does not see the outcome must then decide whether or not to follow the message. The sender's payoff depends entirely on her message (i.e., the message is binding): the number she writes in the message determines her payoff. Her payoff therefore depends neither on the outcome of the die roll neither on the receiver's decision. The receiver's payoff depends on his decision and potentially on the die roll. In particular, if the receiver follows the message, he gets a high payoff if the message is honest, and a low payoff if the message is deceptive. If, instead, the receiver does not follow the message, he receives a fixed medium payoff that does not depend on the actual outcome of the die roll. In the experiment Gneezy et al. (2013) employed the strategy method, that is, they asked each sender what she would do after observing each of the possible six die roll outcomes. This approach helped them identify several strategies. The three extreme strategies were to report 6, to be honest or dishonest independently of the outcome. The first two strategies were the most common, whereas the "always dishonest" strategy was rare. Their strategy set also included strategies that exaggerated only low numbers but were honest otherwise. These strategies were also common. They also included category "other" which was reserved for the strategies without specific pattern. Since they repeated this game multiple times, they were able to test whether subjects consistently apply the same strategy throughout the experiment. They found that across all rounds only 28% of senders are consistent, among which one half are consistently honest and one half consistently exaggerate the die rolls different than 6. The consistency increased to 53% in the final 5 rounds (out of 20), though. In addition to that, the authors found that preference for honesty persisted over time, as many subjects kept using the honest strategy even in the last rounds of the experiment. The popularity of honest strategy slightly decreased with experience, though. The strategies that Gneezy et al. (2013) considered are not suitable for our game, but the study is nevertheless informative for us, as it shows that in deception-like games there is heterogeneity in strategies and consistency may be relatively low because subjects try different strategies before their behavior stabilizes.

In absence of a formal theory of strategies of deception, and given similarities between our deception and helping game, we are inspired in our strategy formulation by strategies of helping. We will look for rewarding-type and defecting-type behavior, as well as experiential-type. To be more specific, given that i) we made the design of our helping game as similar as

possible to the design of our deception game,[16] and ii) that both honesty and generosity are acts of kindness aimed at maximizing both the receiver's payoff and the social welfare (i.e., the sum of the sender's and the receiver's payoff), the strategies proposed by the theory of indirect reciprocal helping, and found in the experiments with helping games, can be feasibly applied by senders in deception game. Having a similar set of strategies will also enable a direct comparison of the behavioral approaches between the two games, while the share of unclassified subjects will serve as a measure of success of this approach.

Experiments with helping games, including ours, have established that subjects use several different strategies (see Chapter 2). Some of these strategies are unconditional, while the rest are conditional on information that individuals receive, their experience, or their own past actions. The past actions of a subject reflect her own reputation, the actions that subject's current receiver made in the past against others reflect the reputation of subject's current receiver and the actions that subject's previous senders or receivers made against her reflect subject's personal experience. When subjects are provided with reputation information about their current receiver, the strategies that condition actions on subjects' own reputation and/or reputation of their current opponent are common, but not the only conditional strategies used, as has been assumed in previous experimental studies on helping game. As shown in Chapter 2, some subjects use strategies that condition actions on their personal experience. Even more, this is the modal strategy in the absence of reputation information.

To date, there is no theory about the behavioral strategies in deception game which allows indirect reciprocal honesty. Fortunately, the reasoning behind the models of indirect reciprocal helping apply promptly to the sender's behavior in our specific deception game, given the intentional similarity of its payoff and informational structure to our helping game. Unfortunately, the models of indirect reciprocal helping say nothing about the receiver's strategies because in helping games receivers make no decisions. There are some experimental studies evidencing that receivers are overly trusting and that some react to sender's behavior - especially if they were lied to,[17] but these studies did not systematically estimate and classify receivers' behavior (Forsythe et al., 1999; Sánchez-Pagés & Vorsatz, 2007). An exception is Cai and Wang (2006) whose setup allowed the classification of receivers according to level-k reasoning model, but because of their substantially different setup the most we can learn from them regarding the receivers is that more than 80% of subjects fits into one of the receiver's behavioral types, that there is heterogeneity in receiver's behavioral types and that receivers that always trust are rare. These findings are nevertheless useful because they justify our two

[16]Recall that in both games senders are provided with the same (reputation) information and payoff scheme, have almost identical screen, and make a similar decision.

[17]Either by reducing their trust in the future or, if possible, by punishing senders, e.g., by reducing the payoffs of both a sender and a receiver.

assumptions that subjects use fixed behavioral rules and that they use many different behavioral rules.

In our deception game, receivers were not given any reputation information about their senders in order to keep the same structure of information in both the deception and helping game. Although there are other possible (potentially more realistic) information setups that would be interesting to study - for example, a setup where receivers also know the reputation of their senders - they would not facilitate the comparison between long-run honesty and generosity, which is one of the aims of the present chapter. Our DBASE treatment, where neither senders nor receivers have any reputation information about others, can be seen as a stylized model of a used-car market or a consumer-to-consumer electronic commerce, where the reputation of traders is unknown, sellers of a good (e.g., a car, a painting) have private information (e.g., about the quality of a car, the effort put into painting) and buyers are uninformed. The DREP treatment, where only senders have reputation information about their current receiver, has stylized features of a consumer-to-business-like market with local intermediaries (Figure 13), where intermediaries who are known only locally buy products (e.g., apps, photographs, or paintings) from local unknown new market entrants (e.g., independent workers, freelancers, start-ups) and sell them to large companies whose reputation is globally known. In such markets intermediaries switch roles, i.e., they are buyers in transactions with local market entrants and sellers in transactions with global companies. On the one hand, since intermediaries are locally known, their selling reputation is known to new local market entrants when intermediaries are buying from them, while the reputation of entrants is unknown since they are new on the market. On the other hand, since intermediaries are globally unknown, their selling reputation is globally unknown when they are selling to large companies. The reputation of global companies is known because they are known worldwide.

Given that receivers have no reputation information about their senders, our analysis of receivers' behavior considers the most relevant unconditional and conditional (experience-based) strategies. To estimate the strategies of individuals, we employ the statistical method, introduced in Chapter 2, that relies on finite mixture models and maximum likelihood estimates. The MLE approach is now standard among scholars interested in the estimation of strategies used by experimental subjects in experimental repeated games (e.g., Dal Bó & Fréchette, 2011; and the references cited in Dvorak, 2020b). We have verified in Chapter 2 that

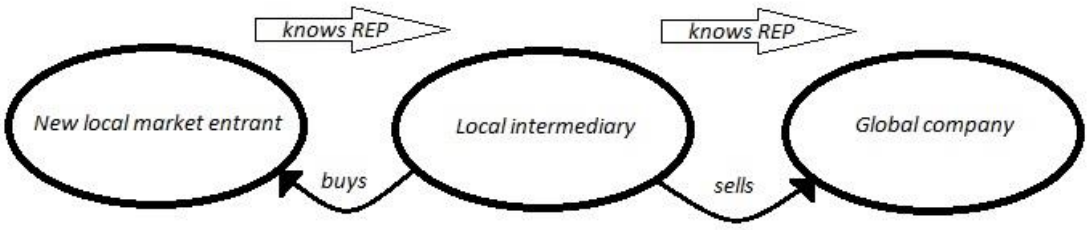


Figure 13: Consumer-to-business-like market with local intermediary

it provides a reliable classification of individuals' strategies. As in our analysis of strategies in the helping game in Chapter 2, we skip rounds 91-100 to avoid the end-game effect, and those earliest rounds when the subject's reputation information and experience are not yet complete. In particular, we consider only rounds where the subject has already made at least two decisions as a sender, has already been receiver at least three times and her receiver has already made at least three decisions in the past.

We next describe the strategies that we will consider in the analysis of our deception game. It is not obvious what pairs of sender's and receiver's strategies we should consider. For example, self-consistency may be a naïve criterion, where receivers use a best response strategy to their own behavior as senders. The empirical support, however, indicates that often this is not the case. For instance, Forsythe et al. (1999) and Sheremeta and Shields (2013) both found that many dishonest subjects tend to be gullible, even though gullibility is not the best response to dishonesty. In addition to that, Cai and Wang (2006) analyzed sender's and receiver's strategies separately, as if they were independent, and found the subjects often do not use the same level of reasoning in the sender's and receiver's role which suggests that subjects' behavior in sender's role is in general indeed independent of their behavior in receiver's role and can therefore be analyzed separately. Kawagoe and Takizawa (2009) also analyzed sender's and receiver's strategies separately and also used level-k reasoning model to explain subjects' behavior in sender-receiver games. Therefore, we will adopt the same approach as previous researchers and investigate the sender's and receiver's behavior separately but look for relations in separate analyses.

Strategies of senders

We first describe the sending strategies which are versions of the strategies we found in the helping game and introduced in Chapter 2. We rename some of them, for clarity, where the original names were specific to the altruistic framework of the helping game. We include two additional experience-based strategies that condition on actions of past receivers that subjects (when they were senders) had met. Such strategies are not feasible in the helping game where receivers have no agency. We will first describe all strategies for DREP and later indicate which

are relevant for DBASE. The summary of all sending strategies and their analogues in the helping game is provided in Table 12.

The simplest two behavioral strategies are unconditional strategies called an *honest* and a *deceptive* strategy. An honest sender always sends the honest message and a deceptive sender always sends a (random) deceptive message. These two strategies are analogous to the strategies used by altruists and defectors in helping game. Since in our deception game presented in Chapter 1, Figure 2, the sender in the Nash equilibrium sends an honest message with probability 12.5%, we add to our strategy set the unconditional *Nash* strategy that in every round sends an honest message with probability 12.5%.[18] The analogue of it was also considered in our helping game where it was called the random8 strategy (it was never used, though). The Nash strategy is theoretically meaningful because it is the Nash equilibrium strategy.

Next, we have two strategies that condition the sender's action on reputation. A *rewarder* strategy conditions honesty on receiver's reputation (the image score), while a *cautious* strategy conditions honesty on sender's own reputation. More precisely, a rewarder is honest only to receivers with good reputation, i.e., those who were themselves sufficiently honest in the recent past. In contrast, someone who is cautious is honest only when her own reputation is bad, i.e., when she was too deceptive in the recent past. Cautious subjects may be particularly detrimental for honest trustful societies, as their emergence may lead to the destruction of honesty, because they can, on the one hand, successfully invade trustful rewarders, and can, on the other hand, be easily invaded by deceivers. To illustrate this, consider a trustful society of rewarders who behave honestly towards individuals who were themselves sufficiently honest in the recent past. What "sufficiently honest" means depends on society and its norm: some societies may regard an individual as honest if she is honest all the time, some societies if she is honest most of the time, whereas some societies if she is honest occasionally. In such an honest and trustful society, social welfare, i.e., the sum of sender's and receiver's payoffs, is often maximized since senders are honest and receivers trust. If someone suddenly decides to experiment and adopts a cautious strategy, she starts behaving sufficiently honest only to fulfill a social norm, which guarantees her a good reputation and thus honesty of other rewarders towards her. Since a cautious individual has a good reputation, she can afford occasional deceit which increases her payoff, because deceit is a best response to trust, without substantially damaging her reputation. In other words, a good reputation offers a cautious individual an opportunity for the future exploitation. A cautious individual fares well in such society of rewarders because her level of honesty is above the honesty norm of the society and is thus not recognized by rewarders as a deviator who can destroy high level of honesty. Since a cautious

[18]This is the same as saying that sender in every round sends a random message or that in every round sends the same message, for example that "*option A will earn the receiver 250 francs*".

individual gains on average more than rewarders, because she is trusted to the same extent as rewarders but deceives more often than them, other society members adopt cautious strategy too. As a result, the number of cautious individuals increases, and more and more individuals start looking only at their own reputation. But as soon as sufficiently large number of individuals become cautious and start looking only at their own reputation, a deceptive strategy becomes advantageous, because deceivers, despite never telling a costly truth, occasionally benefit from honesty of those cautious subjects who try to improve their reputation because they want to be treated honestly by those rewarders who are still in the society. With fewer and fewer rewarders and more and more cautious individuals and deceivers the society becomes less and less honest which leads to full deception.

Next, we consider two sophisticated strategies, a *cautious rewarder* and a *mild deceptive* strategy, that condition actions on the reputations of both a sender and her receiver (Table 12). A cautious rewarder is honest either when the reputation of her receiver is good or when her own reputation is bad. A mild deceiver is honest only when both the reputation of her receiver is good and her own reputation is bad.

Finally, we include experience-based strategies that condition honesty on a subject's own experience. First, we include a strategy that conditions honesty on past senders' behavior. This strategy is analogous to the experiential strategy from our helping game, so we keep the name *experiential* strategy. An experiential sender is honest only if other senders were honest to her in the recent past.

In our deception game (but not in our helping game) receivers also make choices and so it is reasonable to consider that experience with previous receivers may also motivate the decisions of some senders. A selfish sender may for example deceive if her previous receivers trusted her but send an honest message if her past receivers doubted her message. In particular, if a sender wants her receiver to choose a green option but is sure that the receiver will not follow her advice then she should advise the receiver to choose the actual blue option. On the other hand, a prosocial subject would exhibit the opposite behavior. To capture this, we consider two additional strategies that have no analogue from the helping game, the *manipulative* and the *benevolent* strategy. A manipulative sender is honest only if previous receivers did not trust her in the recent past but deceives if previous receivers trusted her in the recent past. A benevolent sender is honest only if previous receivers trusted her in the recent past. So, while benevolent strategy rewards trust, manipulative strategy exploits it.

Finally, as in the helping game, we also include the *random* strategy which in every round sends an honest message with probability 50%. The random strategy is meaningful because it captures the behavior of non-strategic subjects who just randomly deceive or tell the truth. Such non-strategic behavior has been accounted for in the past literature. For example, it was considered in Axelrod first prisoner's dilemma tournament (Axelrod, 1980) and in Camerer et

al.'s (2004) cognitive hierarchy model, and also observed in the experiments (Nagel, 1995; Stahl, 2013). All strategies except those that condition actions on receiver's reputation (i.e., the rewarder, cautious rewarder and mild deceptive strategy) are also considered in DBASE. The strategies of DBASE (and HBASE) are denoted with "*" in Table 12.

Deception game sending strategy	Short description	Helping game analogue
honest*	always honest	altruist*
deceptive*	never honest	defector*
rewarder	honest if the receiver was sufficiently honest in the past	rewarder
cautious*	honest if the sender was sufficiently deceptive in the past	cautious*
cautious rewarder	honest if the receiver was sufficiently honest in the past *or* if the sender was sufficiently deceptive in the past	cautious rewarder
mild deceptive	honest if the receiver was sufficiently honest in the past *and* if the sender was sufficiently deceptive in the past	mild defector
experiential*	honest if the sender was treated sufficiently honestly in the past	experiential*
manipulative*	deceptive if the sender was trusted in the past	/
benevolent*	honest if the sender was trusted in the past	/
Nash*	honest with probability 1/8	random8*
random*	honest with probability 1/2	random*

The asterisk "*" indicates strategies that can be used in DBASE (HBASE) where reputation information is hidden.

Table 12: Deception game sending strategies and their helping game analogues.

Strategies of receivers

We now turn to the responding strategies. In our repeated deception game receivers have no information about their sender, so their actions are either unconditional or conditional on past experience. Table 13 lists all strategies along with short descriptions. As before, to account for any non-strategic or inconsistent behavior and to avoid overfitting, our strategy set includes the *random* strategy that in every round trusts the message with probability 50%. The simplest two behavioral strategies we consider are unconditional strategies called a *trustful* and a *sceptic* strategy. A trustful receiver always trusts and chooses the option advised by his sender. A sceptic never trusts and always chooses an option not advised by the sender. We consider another unconditional strategy for the reduced version of our deception game presented in Chapter 1, Figure 2, namely the *Nash* strategy. Since in the Nash equilibrium of our deception game a receiver trusts with probability 12.5%, the *Nash* strategy in every round trusts the message with probability 12.5%. This responding strategy complements the equally named sending strategy.

Receivers, just like senders, may be influenced by their experience with their previous senders, or may imitate receivers they had met in previous encounters as senders. Furthermore, a receiver may react to his own behavior as a sender, projecting it to all other senders in his

group. We capture these three types of experiential behaviors with three strategies: *reactive*, *conformist* and *projection* strategy. A *reactive* receiver trusts only if his previous (most recent) senders told the truth sufficiently often. Such a receiver simply best responds to the behavior of his past senders. There is experimental support for this type of behavior. Forsythe et al. (1999) found that receivers might respond to senders' behavior, in particular, they found that subjects who were frequently exposed to deception reacted with less gullibility. Sánchez-Pagés and Vorsatz (2007) found that receivers exposed to excessive honesty, correctly update their beliefs about the excessive honesty, and react to it with excessive trust (excessive with respect to game theoretic prediction). A *conformist* trusts only if his previous receivers trusted him sufficiently often recently when he was a sender. Such imitative learning through observation is one of the main forms of social learning and present in our lives since childhood. Imitation has also been observed in laboratory experiments (Huck et al., 1999; Apesteguia et al., 2007). Finally, a receiver using a *projection strategy* trusts only if he himself told the truth sufficiently often recently when he was a sender. In absence of any information about strangers people sometimes find it easiest to assume that strangers are similar to themselves, i.e., that their own behavior is relatively common, and then choose the best response. Such a false-consensus effect, introduced by Ross et al. (1977), has indeed been found in laboratory experiments (Irlenbusch & Ter Meer, 2013; Butler et al., 2015) and proposed as a plausible reason why in a one-shot trust game reciprocal individuals trusted more than selfish individuals (Altmann et al., 2008). Furthermore, Butler et al. (2015) showed that even if a setup offers an opportunity to learn from experience, the projection of subject's own behavior still persists (although weakens) over time.

Deception game responding strategy	Short description
trustful	always trust
sceptic	never trust
reactive	trust only if previous senders (of the receiver) told the truth sufficiently often
conformist	trust only if previous receivers (of the receiver) trusted her recently (when she was sender) sufficiently often
projection	trust only if the receiver herself told the truth sufficiently often recently when she was sender
Nash	trust with probability 1/8
random	trust with probability 1/2

Table 13: Deception game responding strategies with short descriptions.

Strategy pairs

In this last part of Section 3.2 we try to rationalize some of our strategy pairs. We first discuss strategy pairs (deceptive, trustful), (deceptive, sceptic) and (Nash, Nash) which can all be rationalized by the simple boundedly rational model derived from Crawford (2003) and Camerer et al. (2004). For example, suppose that our subject pool consists of subjects with four different levels of rationality – similar assumption has been made by Cai and Wang (2006). The least sophisticated (level-0) subjects are non-strategic subjects who randomly behave

honestly and randomly trust. More sophisticated (level-1) subjects believe that everyone in the subject pool is level-0 and best respond to such beliefs. A strategy pair of level-1 subject would in our game correspond to the (deceptive, trustful)-strategy pair, because the deceptive sending strategy is best response to random trust and the trustful responding strategy is best response to random honesty. Even more sophisticated (level-2) subjects believe that everyone in the subject pool is level-1 and best respond to such beliefs. A strategy pair of level-2 subject would in our game correspond to the (deceptive, sceptic)-strategy pair, because the deceptive sending strategy is best response to the trustful strategy and the sceptic responding strategy is best response to the deceptive strategy. Finally, there is a Nash type who plays Nash strategy in both roles. The unique feature of this strategy pair is that the sending strategy is the best response to the responding strategy and vice versa. Other strategy pairs do not possess this feature, because in our deception game the sender's and receiver's incentives are misaligned.

Next, the experience-based (experiential, reactive) strategy pair can also be easily rationalized, as this is the strategy pair where both the sending and the responding strategies react to the same information, namely a subject's experience as a receiver. The (benevolent, reactive) and (manipulative, reactive) strategy pairs can both be rationalized by the simple model of adaptive behavior. These two strategy pairs adapt to the environment and group norms, the first in the benevolent way, the second in the malevolent way. If a group is relatively honest and trustful, then the (benevolent, reactive)-subject will adapt to this and will trust and behave honestly which will result in maximum social welfare. If a group is deceptive and sceptic, the subject will react with scepticism and deception, as deception maximizes the social welfare. In contrast, the (manipulative, reactive)-subject will trust and deceive in a relatively honest and trustful group, because this maximizes her own expected payoff. In a deceptive and sceptic group, however, the subject will react with scepticism and honesty, as honesty is personally beneficial.

In addition to the above strategy pairs, the (honest, trustful)-strategy pair can also be rationalized. The (honest, trustful)-subject strives for social optimum as she always gives a counterparty a chance to coordinate on social optimum. This strategy pair, however, can also be rationalized by a simple model with naïve subjects (or subjects who lack any sophistication) who naively trust the senders who do not recognize that deception is possible or profitable in the short-term. Finally, as regards the reputation-based sending strategies we did not find compatible responding strategies, because a receiver in our deception game has no information about his sender and hence cannot condition his action on his sender's reputation. However, since reputation-based strategies are conditional, we expect that subjects using strategy pairs involving reputation-based sending strategies will likely use one of the experience-based responding strategies, as these are conditional.

We conclude this section with a remark on substrategies. In Results section we do not report the results regarding the substrategies (only regarding strategies) because they are not that important for the overall picture. For this reason, we do not formally describe them there. However, since the substrategies are the key for strategy estimation (recall from Chapter 2 how strategy classification works) we provide their description in Appendix A4 for completeness.

3.3 Results

3.3.1 Main hypotheses and results

Reviewing the experimental literature on deception and strategic information transmission we found that many experiments involve one-shot games. These experiments are the first step towards a better understanding of how people deal with new situations of which they have had no experience. They are also important for example for testing a theory, for measuring social preferences, or for eliciting individuals' beliefs. They, however, do not offer the opportunity to study social norms, for example reciprocity, or to study the effect of different mechanisms such as reputation or punishment/reward mechanisms. This can be studied by running an experiment with a repeated game. In this project, by running and comparing two treatments with the repeated deception game, DBASE and DREP, we will test whether reputation sharing increases the average honesty, and whether the average honesty and trust are stable over time. The average behavior of experimental groups might vary across groups, as evidenced for instance by Seinen and Schram (2006) in the repeated helping game and verified in our Chapter 2. If we find variation in honesty across groups, then by examining the correlation between the initial and average group behavior we will be able to tell whether or not this variation stems from the differences in the initial rounds. By comparing DBASE with HBASE and DREP with HREP we will test whether deception aversion, reported in most of the previous literature, exists when subjects are aware that they will play the game many times. We will also examine how deception aversion evolve over time. One additional strength of running an experiment with a repeated game is that researchers are provided with multiple data for each subject, showing them how subjects reacted in different scenarios and giving them the opportunity to study what behavioral strategies subjects use in the experiment. Our DREP treatment will reveal whether reputation-based and experienced-based types of reciprocity are present, whereas both, DBASE and DREP, will tell how important experiential behavior and indirect reciprocity in the context of honesty and deception are. In the following we motivate and present our hypotheses.

To date there is ample evidence that reputation information promotes socially desirable behavior. A recent study by Behnk et al. (2019) confirmed that reputation information reduces deception and promotes honesty in a setup where subjects were either always senders or always

receivers. Reputation information also increases generosity (e.g., Bolton et al., 2005; Seinen & Schram, 2006). Based on this we hypothesize that:

H1: *The average honesty across all rounds will be higher in DREP than in DBASE.*

This comparison is important because it will reflect differences in overall behavior as it considers all rounds. We will make another comparison, namely over the initial round, which will reflect differences in behavior when the subjects are not influenced by the learning effect yet. We expect the difference in average honesty also in the initial round because in DREP investing in reputation early may be profitable if a group has enough rewarders and cautious rewarders who reward honest subjects with honesty. Based on Chapter 2, and the fact that designs of our helping and deception game are analogous, we have a reason to believe that groups in DREP will have enough rewarders and cautious rewarders.

H2: *The average honesty in the initial round will be higher in DREP than in DBASE.*

The previous experiments on deception evidenced that senders are more honest than the game-theoretic predictions, and proposed several reasons for that: bounded rationality, social preferences, or deception aversion (Cai & Wang, 2006; Sánchez-Pagés & Vorsatz, 2007; Hurkens & Kartik, 2009). Based on this we hypothesize that:

H3: *The average honesty across all rounds is above the theoretically predicted (12.5%) in both treatments (DREP and DBASE).*

In our experiment subjects interact for 100 rounds and can thus learn about honesty of other group members either from their experience or from reputation information they were provided. Given that we have a relatively small groups of six anonymous individuals and that self-regarding subjects are common in experimental games, it is very likely that in our deception games subjects will frequently encounter deceivers. Exposure to frequent deception can make subject's own deception easier to justify, because others deceive as well. Such peers' effect has been observed in the experimental literature (Fosgaard et al., 2013; Innes & Mitra, 2013; Leib & Schweitzer, 2020). Also, people tend to conform to group norms, so if a group develops a norm for deception, then an individual may want to conform to that. Such a decreasing trend in a socially desirable behavior, helping in particular, has been reported in Swakman et al. (2016). Therefore, we hypothesize that:

H4: *The average honesty decreases over time.*

By analyzing the average honesty, we can learn about the dynamics of honesty but not about what is going on within each group, which is our independent observational unit. As evidenced for example by Seinen and Schram (2006) and Swakman et al. (2016) groups might develop different norms and hence the degree of socially desirable behavior may vary substantially across groups. But do these group differences develop over the course of the game or already

in the early rounds? Although these two studies did not formally test for positive correlation between the behavior in the initial rounds and overall behavior it seems, based on Figure 2 in Seinen and Schram (2006) and Figure 1 in supplementary material of Swakman et al. (2016), that initial rounds positively affect long-term group dynamics (for a related phenomenon of generosity). Based on this we expect that:

H5: *The greater the average group honesty in the initial two rounds, the greater the average group honesty in rounds 3-100.*

We compare the initial two rounds (and not three or some other number of rounds) with the rest, because after the first round only half of the subjects were senders, while after two rounds, on average, everyone made their first decision as a sender.

Now we turn to the behavior of receivers, trust in particular. Sánchez-Pagés and Vorsatz (2007) found that receivers exposed to excessive honesty correctly update their beliefs about the excessive honesty and react to it with excessive trust. Therefore, since we hypothesize that the reputation mechanism increases the average honesty (H1), we also expect that:

H6: *The average trust across all rounds will be higher in DREP than in DBASE.*

We will make another comparison, namely over the initial round, reflecting the one-shot behavior. In the initial round we expect the trust levels in DREP and DBASE to be similar, as receivers have learned nothing about the prevalence of honesty yet. In other words, they have not had a chance to update their beliefs yet.

H7: *The average trust in the initial round will be similar in DREP and in DBASE.*

The next two hypotheses, H8 and H9, are derived from the hypotheses H3 and H4, respectively, and the argument for hypothesis H6.

H8: *The average trust across all rounds is above the theoretically predicted (12.5%) in both treatments (DREP and DBASE).*

H9: *The average trust decreases over time.*

Our next hypothesis H10 is inspired by Gneezy (2005) and concerns the comparison of average honesty in our deception game and average helping in our helping game. He argued that honesty may be driven by more than just social preferences, which he demonstrated comparing deception games and dictator games with identical payoffs. He proposed that while generosity is motivated by social preferences, honesty is driven by both social preferences and aversion to deception. Since deception costs reduce selfish behavior, we expect that subjects will be more honest than generous, both in the initial rounds and on average.

H10: *Honesty will be higher than helping in treatments with reputation (DREP vs. HREP) and in treatments without reputation (DBASE vs. HBASE) – both, in the initial rounds and overall.*

The last part of Results section is devoted to the analysis of subjects' strategies. In DREP senders observed reputation information. Given that our results from HREP (Chapter 2) confirmed that reputation and not experience is the main driving force behind subjects' choices when reputation information is provided, and given that our treatments are designed to be as similar as possible, we expect that reputation information and not experience will serve as a focal point in DREP too. We also expect that in DREP the rewarding behavior will be as common as in HREP (which does not automatically imply that the average honesty in DREP should be similar to the average helping in HREP). Regarding the receivers' behavior we have no prior expectations, but given that they are provided the same information in both DREP and DBASE, we expect that the responding strategy distributions will not significantly differ. Thus, we have the following three hypotheses.

H11*: The share of reputation-based strategies in DREP is not significantly different from that in HREP.*

H12*: In DREP, the share of reputation-based strategies is significantly higher than that of experiential strategy.*

H13: *In DREP and DBASE the responding strategy distributions are not significantly different.*

Our main findings are as follows. We find that honesty is indeed higher when senders can access receiver's reputation. Interestingly and contrary to our expectations that were based on the general view in the literature on deception, we find in our experiments that honesty is not higher than generosity in the long run. This gives us no evidence in support of a stable cost of lying. We examine further the dynamics of honesty and generosity by investigating behavioral strategies that subjects apply in both games. We find slightly more deceivers (who are never honest) in the deception game than defectors (who never help) in the helping game, which might explain why the average honesty is lower than the average generosity.

The remainder of this chapter is devoted to the presentation and discussion of our results. Recall from our helping game (Chapter 2) that subjects from the same matching group were paired with the same subject more than once (20 times on average), thus the actions within a matching group are likely to be correlated across rounds. In contrast, subjects from different matching groups never met each other so actions between matching groups are independent. Therefore, our independent observational unit is a matching group, not a subject. For simple terminology, we will call the average honesty/trust rate of a matching group the *group average* honesty/trust rate, and the average of group averages the *treatment average*. The treatment average is therefore a single number calculated from 10 group averages, each of which is calculated based on honesty/trust of six subjects from the same matching group. As a final remark, whenever

we report that permutation test was used for hypothesis testing, we also perform Mann Whitney test which yields the same statistical results. For ease of notation, we denote the permutation test by PT.

3.3.2 Dynamics of honesty

Our first result concerns the reputation effect. We first test whether reputation information increases the treatment average honesty across all rounds (see Table 14). We test this by comparing ten group average honesty rates in DREP to that in DBASE. One-sided PT confirms that the treatment average honesty rate in DREP was significantly higher than the treatment average honesty rate in DBASE (32% vs. 17%, $p<0.05$). This supports our hypothesis H1. We also found that in DREP this treatment average honesty rate across all rounds is significantly higher than 12.5% ($p<0.05$, one-sided PT), which is the game-theoretic prediction for our deception game, whereas the treatment average honesty rate in DBASE is not ($p>0.1$, one-sided PT). This only partially supports our hypothesis H3.

	DREP	DBASE
all rounds	32% (0.25)	17% (0.12)
first round	47% (0.32)	23% (0.22)

Table 14: Treatment average honesty rates over all rounds and in the first rounds. Standard deviations of group averages are in parentheses.

Next, we examine the treatment average honesty in the first round by comparing the first-round group average honesty rates in DREP to that in DBASE. In DREP and DBASE, the first-round treatment average honesty rates were 47% and 23% (Table 14), respectively, and the difference although substantial was only marginally significant ($p<0.1$, one-sided PT). This only marginally supports our hypothesis H2. Table 14 shows that the variance in group average honesty rates is higher in the first round than overall. This is unsurprising, given that in the initial rounds subjects are usually experimenting and learning about the game and their group.

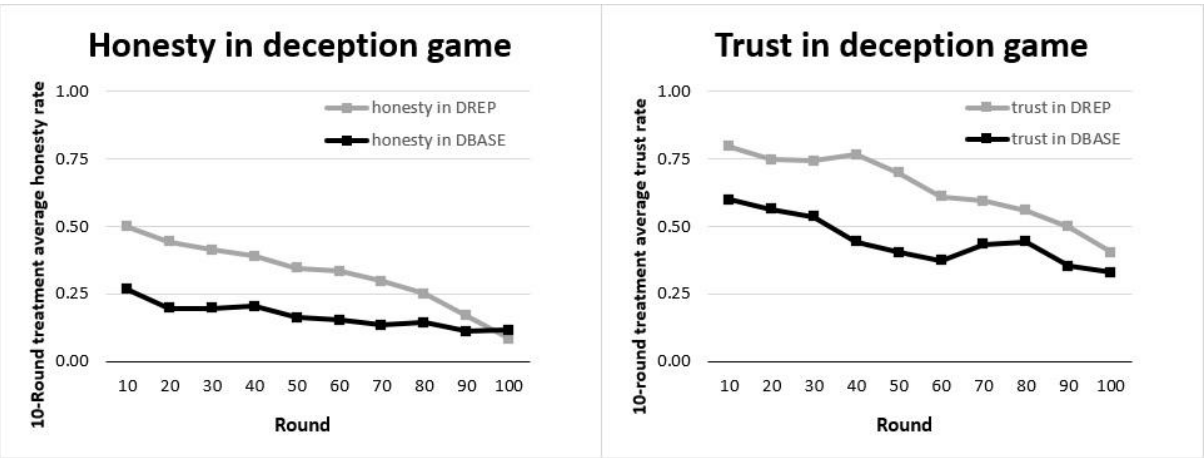


Figure 14: Left panel: Dynamics of honesty. Gray (black) line corresponds to the 10-round treatment average honesty rate in DREP (DBASE). Right panel: Dynamics of trust. Gray (black) line corresponds to the 10-round treatment average trust rate in DREP (DBASE).

The relatively substantial variance in the first-round honesty between groups can explain the low significance in the above difference. It also gives us an opportunity to later investigate hypothesis H5.

Result 1: *Reputation mechanism promotes honesty. In the first round, however, the average degree of honesty in DREP is only marginally higher than the average degree of honesty in DBASE. The average degree of honesty across all rounds is above the theoretically predicted (12.5%) only in DREP.*

In the first round, the reputation information in both treatments was the same - empty, because subjects have not made a decision yet. Still, the substantial difference was observed, which suggests that many subjects in DREP, knowing that their reputation will be visible to their future counterparts, invest in own reputation early.

Now we turn to the dynamics of honesty and trust. Figure 14, left panel, illustrates the 10-round treatment average honesty rates in our two treatments. For each treatment, the k-th point, $k \in \{1,\ldots,10\}$, is calculated as the treatment average over rounds 10k-9 to 10k. The figure shows a decreasing trend in both treatments. We formally tested for changes in honesty rates over the rounds in DREP and DBASE by fitting logistic generalized linear mixed models (GLMM) to honesty decisions.[19] We included two fixed factors, "round" and a dummy variable for the last 10 rounds to control for the end-effect, and one random factor, "subject nested in matching group".

[19]This technique was used before, for example, by Swakman et al. (2016) who tested for changes in helping rates in their helping game experiment.

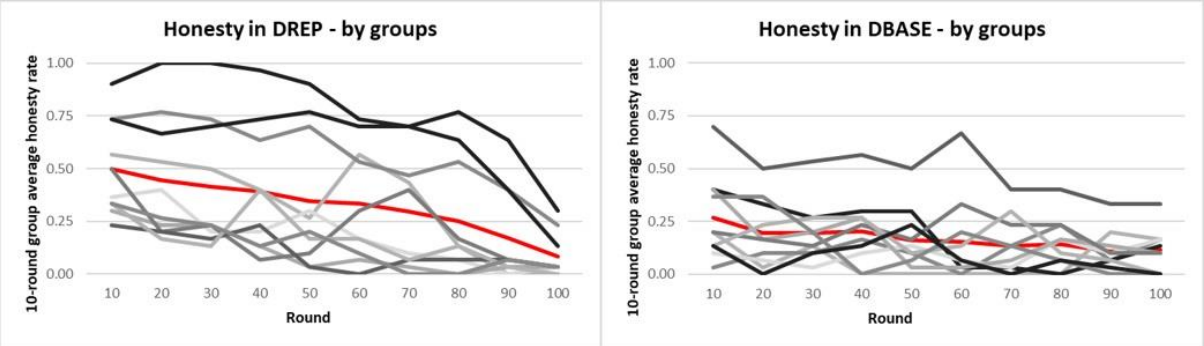


Figure 15: 10-round group average honesty rates in DREP (left panel) and in DBASE (right panel). The red line corresponds to the 10-round treatment average, which is also shown in Figure 14, left panel.

The statistical analysis confirmed our hypothesis H4 that the average honesty rates were decreasing with rounds in both treatments, as the estimated regression coefficient of variable "round" has a negative sign (DREP: $p<0.001$; DBASE: $p<0.001$).[20] Figure 14, left panel, also reveals that honesty rates ended up close to the equilibrium prediction, 12.5%. To be more precise, in both DREP and DBASE the last-round treatment average honesty rate was 7% and not significantly different from 12.5% ($p>0.1$, PT).

Result 2: *Honesty decreases over time in both treatments.*

The decline in honesty suggests that reservations against deception were disappearing over time. One causal mechanism that could be responsible for that is experience. It could also be that subjects simply learn by trial and error how to play the game with time, or that deception brings more money. In DREP, subjects could also infer from the opponents' (bad) reputation that over time more and more subjects deceive. Then, the decline in honesty could be caused by rewarders and mild deceivers who, after encountering dishonest subjects more often as the game progresses, become more deceptive themselves. This lowers rewarders' and mild deceivers' own reputation and leads to less honesty towards them in the future. We will further investigate whether honesty is more experience- or reputation-driven in our later sections where we will systematically examine individuals' strategies.

So far we have focused on treatment averages to learn about the dynamics of honesty. The treatment averages were calculated from the averages of our independent groups which were not given any special attention yet. We focus on groups in the following paragraphs. The standard deviations, displayed in Table 14, indicate heterogeneity in honesty across groups, especially in DREP. This variation is clearly visible in Figure 15 which also suggests that group

[20]The end-effect was significant only in DREP ($p<0.001$). In DBASE it was not ($p>0.1$), as the decline started way earlier. Similar observation was made below in the subsection *Dynamics of trust* where we examine the dynamics of trust.

differences in honesty levels develop in the early rounds rather than over the course of the game.

We formally test whether the initial honesty determines the long-run spread of honesty by examining the correlation between the group average honesty rates calculated over the first two rounds and the group average honesty rates calculated over rounds 3-100. We compare the averages over the first two rounds (and not three or some other number of rounds) with the rest, because after two rounds, on average, everyone made their first decision as a sender.

We know from the experiments on other games that differences across groups may exist (e.g., Seinen & Schram, 2006), but specifically regarding honesty we do not know whether higher initial honesty results in higher average honesty. That may be relevant information for anyone interested in promoting long-term honesty, because if it turns out that overall behavior strongly depends on the behavior in the initial rounds, then one may promote it by investing as much effort as possible in the initial rounds.

Figure 16 illustrates the average honesty for each group in the initial and the remaining rounds. The correlations were analyzed using a one-sided permutation Pearson's correlation test. In DREP, the correlation is positive, strong and significant ($r=0.77$, $p<0.01$), meaning that groups with higher overall honesty were indeed those that experienced more honesty in the initial rounds. Therefore, at least when the reputation information is observable, it seems that one may increase overall honesty by promoting honesty in the initial rounds. The high positive correlation also hints that many senders may be making decisions based on experience. This will be systematically investigated in later sections. In DBASE, the correlation is also positive, but not significant ($r=0.31$, $p>0.1$), showing that the initial rounds had less impact on the overall behavior than the initial rounds in DREP. These findings support our hypothesis H5 only for DREP. If many subjects relied on experience, then we would observe significant positive correlation in DBASE like we observed in HBASE. We did not observe that, though, which suggests that when the reputation is hidden either the experience is not as important for deception game as for helping game, or there is simply too little initial honesty in all groups to provide a long-term effect. It is difficult to expect that honesty norm will develop within groups if it is never noticed. There is however an outlier group in DBASE that exhibits a relatively high level of overall honesty. Its structure will be revealed later in strategy analysis.

Result 3: *The correlation between the initial and average honesty is positive but significant only when reputation is observable.*

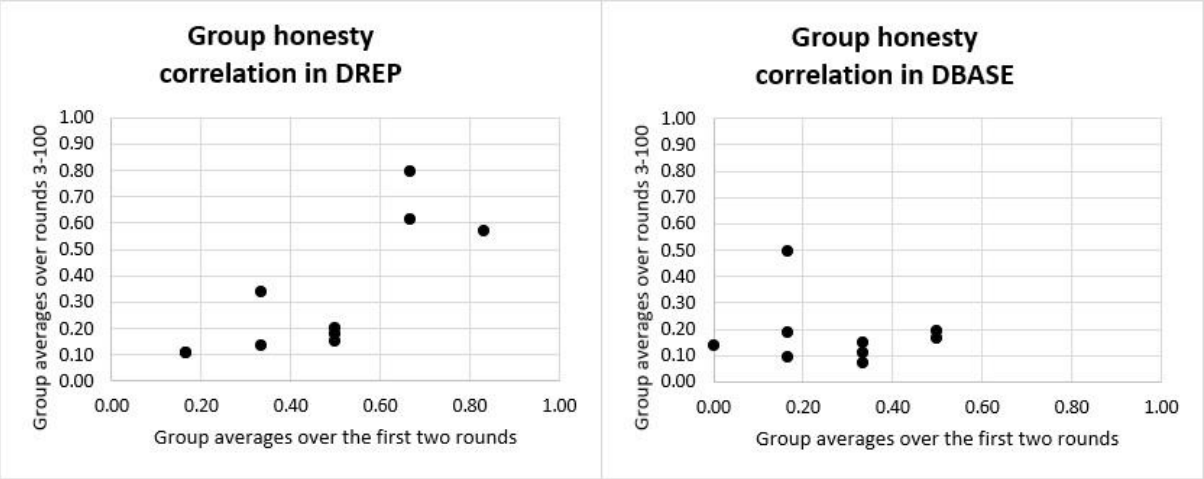


Figure 16: Group average honesty rate in initial and remaining rounds. Left panel: DREP. Right panel: DBASE.

3.3.3 Dynamics of trust

Our first result concerns the reputation effect. We first test whether reputation information increases the treatment average trust across all rounds (see Table 15). We test this by comparing ten group average trust rates in DREP to that in DBASE. One-sided PT confirms that the treatment average trust rate in DREP was significantly higher than the treatment average trust rate in DBASE (64% vs. 45%, $p<0.05$). This supports our hypothesis H6.

	REP	BASE
all rounds	64%	45%
first round	67%	57%

Table 15: Treatment average trust rates over all rounds and in the first rounds.

We additionally test whether these two treatment average trust rates are significantly higher than 12.5%, i.e., the Nash equilibrium prediction for a reduced deception game presented in Chapter 1, Figure 2, which would indicate that subjects are much more trusting than the theory predicts. We found that in both treatments the treatment average trust rates are significantly higher than 12.5% ($p<0.01$, one-sided PT). This supports our hypothesis H8.

Next, we examine the treatment average trust in the first round by comparing the first-round group average trust rates in DREP to that in DBASE. In DREP and DBASE, the treatment average trust rates were 67% and 57% (Table 15), respectively, and this difference was not significant ($p>0.1$, one-sided PT). This is not surprising, since in the beginning receivers have not updated their beliefs about the prevalence of honesty yet. This finding supports our

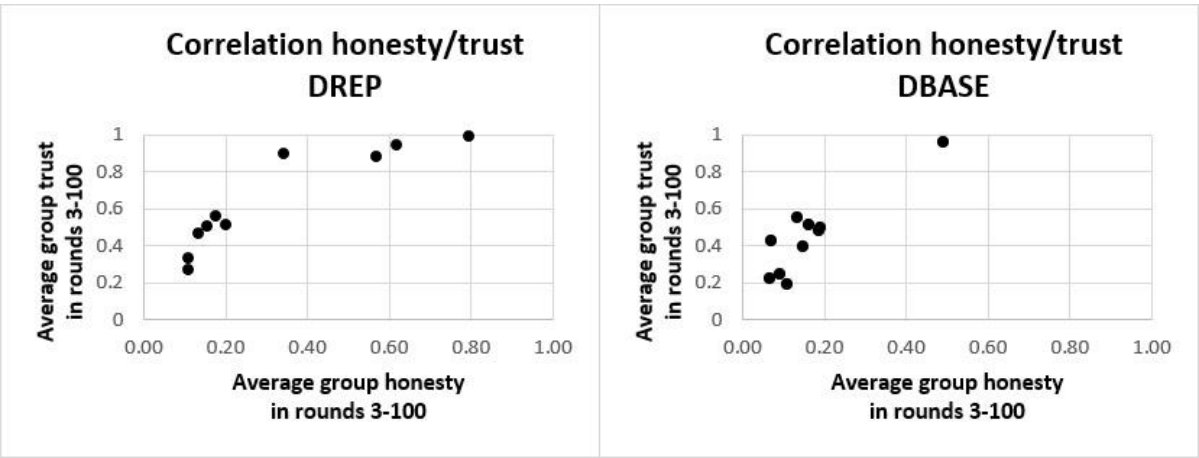


Figure 17: Correlation between the average group honesty and average group trust in rounds 3-100. Left panel: DREP. Right panel: DBASE.

hypothesis H7. Therefore, across rounds the receivers trusted more often in the DREP treatment even though in the initial rounds they exhibited similar trust in both DREP and DBASE.

Result 4: *Reputation mechanism promotes trust. The average degree of trust is above theoretically predicted in both treatments. In the first round, however, the average degree of trust does not significantly differ across treatments.*

Now we turn to the dynamics of trust which will give answer to our hypothesis H9. Figure 15, right panel, illustrates 10-round treatment average trust rates in our two treatments. For each treatment, *k*-th point, $k \in \{1,\ldots,10\}$, is calculated as the average over group averages over rounds 10k-9 to 10k. The figure shows a decreasing trend in both treatments. As above, we formally tested for changes in trust rates over the rounds by fitting logistic GLMM to trust decisions, where we included the same fixed factors ("round" and a dummy variable for the last 10 rounds to control for the end-effect) and random factor ("subject nested in matching group"). The statistical analysis confirmed our hypothesis H9 that trust rates were decreasing with rounds in both treatments, as the estimated regression coefficient of variable "round" has a negative sign (REP: $p<0.001$; BASE: $p<0.001$).

Result 5: *Trust decreases over time in both treatments.*

Coupled with Result 2, that honesty decreases over time, this result suggests that on average senders become less honest and receivers more skeptical with rounds, indicating a positive correlation between honesty and trust which we formally confirm below. We turn to group honesty and group trust to learn if group norms emerge where trust adapts to honesty. In particular, we explore whether there is a positive correlation between the average group honesty and average group trust in rounds 3-100, across all groups. Figure 17 plots the honesty/trust pairs for each group in DREP (left panel) and DBASE (right panel). We found that in both treatments the average honesty and trust rates of groups are strongly positively

correlated (REP: r=0.91, p<0.001; BASE: r=0.90, p<0.01, permutation Pearson's correlation test), which shows that more honest groups enjoy more trust. So, it seems that trust and honesty really go hand in hand.

Result 6: *Honesty and trust are strongly positively correlated.*

3.3.4 Honesty versus generosity

Gneezy (2005) studied honesty and deception in one-shot deception games and altruism and selfishness in one-shot dictator games with identical payoffs. His goal was to check if honesty in deception games is due to social preferences or if there are other motives. He found significantly less deception in deception games than selfishness in dictator games and concluded that honesty is driven by more than altruism. He proposed deception aversion that works on top of social preferences. Although Gneezy's (2005) influential study has inspired many new studies on the subject, nobody has yet investigated how stable are the differences between honesty and altruism. We will look at this by comparing altruism in HBASE and honesty in DBASE, as well as in HREP and DREP. Inspired by Gneezy, our experiment was designed to make the deception and helping game setups, interfaces and interactions as similar as possible, which facilitates direct comparison of one-shot helping and one-shot honesty in the long run, permitting learning and reciprocity in both games. We can therefore explore whether any specific aversion to deception persists or vanishes with time. Figure 18 illustrates 10-round average honesty and helping rates (gray and black, respectively) across all groups in REP (left panel) and BASE (right panel). For each treatment, the 10-round treatment averages are calculated as above. We formally test for the differences between honesty and helping (i.e., hypothesis H10) by comparing the average group honesty rates in DBASE to that in HBASE, as well as by comparing these rates in DREP to HREP. We make every comparison two times, once over the initial two rounds (in which, on average, everyone was sender once) and then across all rounds. Table 16 shows the observed rates.

	REP		BASE	
	honesty rate	helping rate	honesty rate	helping rate
initial two rounds	47%	50%	28%	35%
all rounds	32%	43%	17%	27%

Table 16: Treatment average honesty and helping rates in the first two rounds and over all rounds.

Surprisingly, and contrary to the findings of Gneezy's (2005) one-shot experiment, the average levels of honesty in the initial (two) rounds and over all rounds were slightly lower than the average levels of helping. For both averages, the one-sided PT failed to reject that the average honesty rate is less than or equal to the helping rate (p>0.1). The same conclusions hold when

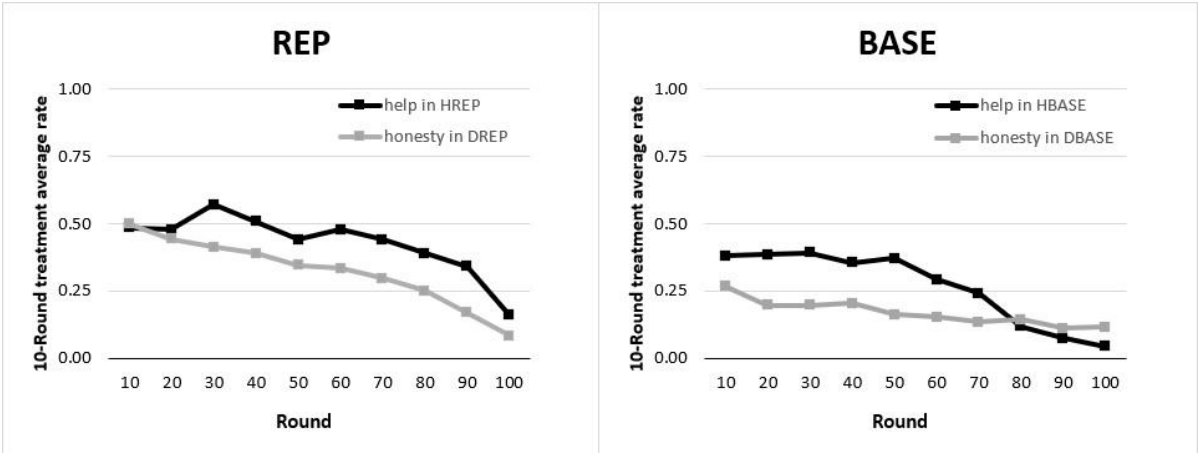


Figure 18: 10-round average honesty and helping rates in REP (left panel) and BASE (right panel) treatments.

comparing treatments with reputation and when comparing treatments without reputation (see Table 16). This does not support our hypothesis H10.

Result 7: *The average honesty is not higher than helping – not in the initial rounds nor overall.*

We therefore found no evidence that the average honesty is greater than the average helping (altruism), which would indicate that an average individual has deception costs. Our result is more consistent with Sasaki et al. (2019) than Gneezy (2005), as they found no evidence for deception aversion in two of the three deception/dictator games they considered. If anything, we found that average honesty levels are slightly below the average helping levels despite being closer together in the initial rounds. One possible explanation for our result may be that unconditional deceivers in DREP and DBASE are more common than defectors in HREP and HBASE. This may first explain slightly lower honesty than help in the first round and, if indirect reciprocal subjects (i.e., rewarding and experiential subjects) are found to be equally common across games, may then explain a higher difference between the average honesty and help over all rounds, because rewarding and experiential subjects will meet deceivers more often and consequently deceive more. The reason may also be the lower number of rewarding subjects in the deception game than in the helping game, which would make honesty difficult to sustain through indirect reciprocity. Furthermore, even if the games contain an equal number of deceivers/defectors, rewarding and experiential subjects, honesty may still be lower than helping, if more rewarders are assigned to a group with many deceivers. However, since DREP and HREP and DBASE and HBASE have similar initial-round average honesty and helping rates, the reason for the lack of deception aversion could simply be that the temptation for deception was too high to take deception costs into account. Namely, under our payoff scheme the monetary costs of honesty were 150 francs (i.e., 0.6 EUR), so if subjects valued the monetary costs of honesty higher than the psychological costs of deception, then they would deceive even if they were actually mildly deception averse.

In this section we analyze the differences between honesty and generosity (altruism) by directly comparing the average honesty rates with the average helping rates. This is a conservative approach because it does not take into account that the equilibrium level of honesty in our deception game differs from the equilibrium level of help in our helping game. In our deception game the equilibrium level of honesty is 12.5%, while in our helping game the equilibrium level of helping is 0%. Since the game theoretic predictions of honesty and helping rates differ, the comparing rates should be adjusted for in our analysis. A better approach is therefore to first subtract the equilibrium levels of honesty and helping from the average group honesty and helping rates before making comparisons. Correcting for this equilibrium predictions, our Result 7 becomes even stronger, because the difference between the average honesty rates reduced by 12.5% and the average helping rates reduced by 0% becomes even higher than the simple difference between the average honesty and helping rates. On our particular data both approaches lead to the same conclusions, though, because our hypothesis H10 is directional (i.e., it says that honesty is *higher* than helping, not that honesty is *different* than helping) and we did not find support for it anyway, even without correcting for the equilibrium predictions.

3.3.5 Strategies

In this section we explore and estimate strategies that experimental subjects use in our repeated deception game. Apart from learning how heterogeneous our experimental group is and whether honesty/deception is more reputation or experience-driven, analysis of subjects' strategies will provide further insight into the dynamics of honesty and deception explored above. Before proceeding to the results of our analysis, it is worth mentioning again that in a few initial rounds of the repeated game a subject's reputation may have less information or be even empty, so to avoid complications with strategy definitions we exclude these rounds from the analysis of subjects' strategies. That has also been done in previous studies (Seinen & Schram, 2006; Ule et al., 2009).

Strategies in DBASE

Table 17 shows the estimated strategy distribution in DBASE. Strategy pairs that are not used by anyone are omitted. The rows correspond to the sending strategies, i.e., the strategies estimated from choices made when subjects were senders. The columns correspond to the responding strategies, i.e., the strategies estimated from choices made when subjects were receivers. The cells correspond to the (sending, responding) strategy pairs. For example, 15% in the second row and second column means that 15% of subjects consistently behave in line with the (deceptive, reactive)-strategy pair, sending deceptive messages while trusting only after meeting many honest senders recently. The (deceptive, reactive) strategy pair is also the modal strategy pair used by our subjects. Although it was not rationalized in Strategies section, it can be rationalized ex-post. For example, it may be used by subjects who i) expect that receivers are more trusting than the equilibrium prediction, and ii) try to maximize their

receiver's payoffs by adapting to the behavior of past senders they had met – if there is a momentum of honesty followed by a momentum of deception it is certainly better to adapt to this than to unconditionally trust or doubt. This strategy pair is followed by the (Nash, projection) and (Nash, trustful) strategy pairs. The (Nash, projection) strategy pair has a peculiar feature, namely that it is essentially equivalent to the (Nash, Nash)-strategy pair since projecting the Nash sending strategy means exactly playing the Nash responding strategy. Some subjects are therefore consistent across roles and choose sending and responding strategies that are best response to each other.

It is interesting that almost all subjects apply self-regarding sending strategies such as deceptive and Nash, but nearly one quarter of them apply responding strategies that permit frequent trust (i.e., conformist and trustful responding strategies). This is in fact consistent with Forsythe et al. (1999) and Sheremeta and Shields (2013) who both found that many dishonest subjects tend to be gullible. One possible reason for this finding is that deceptive senders believe that despite their deceptive behavior there are enough honest and trustful subjects in the experiment to make deception and trustful behavior profitable.

Of the remaining strategy pairs the (Nash, reactive) and (deceptive, projection) strategy pairs are relatively common, used by 8% and 7% of subjects, respectively. The (Nash, reactive)-subject is sporadically honest, but otherwise behaves the same as (deceptive, reactive)-subject. The (deceptive, projection)-subject is esentially equivalent to the (deceptive, sceptic)-subject, since the best response to subject's own unconditional deception is unconditional doubt (skepticism). According to our simple reasoning model presented in Strategies section, such subjects are level-2 thinkers. We also found that the responding strategy of 12% of deceivers could not be classified into one of our responding strategies, suggesting that as receivers they are not strategic (e.g., using the random strategy), mixing between the strategies from our set, or using a strategy that is not included in our set.

The existing literature suggests that rather than best responding to own strategy in another role, subjects often best respond to their beliefs about the behavior of others (Forsythe et al. 1999; Sheremeta & Shields, 2013). For example, in our experiment subjects may think that others are different (are more honest and trusting, not strategic). In our repeated game, these beliefs might get more and more accurate as the game progresses and subjects learn. In fact, in our game best responding to own strategy in another role could be far from rational if other subjects use different strategies. For example, skepticism is not optimal when there are even a few honest senders, but deception is optimal when there are a few skeptical receivers. Skepticism may be the optimal response to own deception but not to an environment with many deceptive and a few honest senders. The (deceptive, reactive) strategy pair is therefore a better response to an environment with an unknown distribution of types.

Since the relative frequencies of the remaining strategy pairs, including those rationalized in Strategies section, are low (mostly below 4%), we now turn to marginal distributions of sending and responding strategies to get some further insights into subjects' behavior. The last column and row of Table 17 show the relative frequencies of individual sending and responding strategies, respectively. Our estimations suggest that 85% of senders are mostly deceptive since they use either the purely deceptive (40%) or the Nash sending strategy (45%). This finding explains why the average honesty in DBASE is low (see Tables 14 and 16). In DBASE senders practically never used experience-based strategies: the experiential strategy was used by only one subject, whereas the benevolent and manipulative strategies that condition on received trust were never used. We can explain the lack of eexperientials by turning back to our helping game and comparing the distribution of sending strategies in DBASE to that in HBASE (see Table 18).

		Receiver					
		trustful	reactive	conformist	projection	unclass.	**%**
Sender	honest			2%	2%	2%	***5%***
	deceptive	3%	15%	3%	7%	12%	***40%***
	experiential				2%		***2%***
	Nash	13%	8%	2%	13%	8%	***45%***
	unclass.				8%		***8%***
	%	***17%***	***23%***	***7%***	***32%***	***22%***	

Rows correspond to strategies estimated from choices made when subjects were senders. Columns correspond to strategies estimated from choices made when subjects were receivers. Strategies that are not used by anyone are omitted.

Table 17: Strategy distribution in DBASE.

The Fisher exact test and pairwise tests corrected for multiple observations reveal that DBASE and HBASE distributions are significantly different and that this is due to the Nash and experiential strategies. One possible explanation why we did not find experientials in DBASE is because the average honesty in DBASE was so low (not significantly above 12.5%, which we showed above) that even if subjects were imitating the group behavior, they would be honest slightly more than 12.5% of the time which would make them closest to Nash and close to deceivers. Despite low levels of honesty, DBASE contains an exceptional honest group (see Figure 17, right panel) but experientials were not responsible for honesty. The key were two altruists and one unclassified subject with high average honesty rate who was mostly honest in the first 75th rounds but then completely gave up on honesty, possibly because the group also contained two Nash subjects and one deceiver.

DBASE	DBASE %	HBASE %	HBASE	sign.
honest	5%	4%	altruist	ns
deceptive	40%	30%	defector	ns
experiential	2%	56%	experiential	***
Nash	45%	0%	random8	***
unclass.	8%	11%	unclass.	ns

*** - significant at the 0.001 level
ns – not significant

Table 18: Comparison of sending strategy distributions, DBASE and HBASE.

Table 18 reveals that the shares of deceivers and defectors are not significantly different. The deceivers are however slightly more common than defectors which may explain why honesty is not higher than helping (Table 16). Moreover, Table 18 provides no evidence of deception averse subjects. There are completely honest subjects, but since their share is similar to the share of altruists in helping game, their honesty may simply be the result of their preference for the outcome stemming from honesty (e.g., because it maximizes the joint profit) rather than deception aversion. The lack of evidence for deception aversion is in line with recent results by Sasaki et al. (2019), but contrasts the results of others (e.g., Gneezy, 2005; Vranceanu & Dubart, 2019). There are several potential reasons for the stark difference between our and Gneezy's (2005) result. For instance, one reason might be the repetition (multiple interactions and decisions made) coupled with information about the environment, in particular, information about group behavior shared through reputation and experience. This might have caused that reservations against deception - that subjects who were initially deception averse had, disappeared after observing frequent deception, because it was easier for them to justify their own deception because others were deceiving too. It is also possible that subjects were more enraged and/or disappointed by deception than by selfishness which could have resulted in lower honesty compared to helping. However, both these reasons suggest the presence of rewarders and especially experientials but we found that only a few subjects responded primarily to experience or reputation in a standard way. Given that we had many unclassified behaviors it is possible, though, that subjects responded in some non-standard way which we did not account for. The reason might have also been the information about (misaligned) incentives which could give rise to sophisticated deception or honesty. However, we found no support for this either. Sophisticated deception would in our game result in higher honesty than helping, not vice versa, whereas expecting that honesty is low due to sophisticated honesty through deception (i.e., engaging in deception due to belief that receivers will not trust) seems unreasonable given that the existing literature documented that such behavior is extremely rare (Sutter, 2009). Finally, it is also possible that under our payoff scheme the reward from deception was simply too high compared to deception costs of otherwise mildly deception averse subjects.

Another insight from Table 18 comes from the comparison of Nash and random8 strategies. If in our deception game subjects in the sender's role on average send the honest message once per eight rounds it is not a priori clear whether they are doing this strategically (to maximize their expected payoff) or randomly (because they are clueless about which behavior could be profitable). If they are playing strategically, then given the similarity between DBASE and HBASE, no one should be playing random8 strategy in HBASE because it does not maximize the payoff. If they are instead playing randomly, then in HBASE the share of random8 strategy should be similar to the share of Nash strategy. Table 18 reveals that no one is playing random8 strategy in HBASE, suggesting that Nash subjects are indeed strategic.

Finally, we turn to the distribution of responding strategies. Table 17, last row, shows that more than 60% of subjects use one of the strategies that condition on experience or memory (i.e., reactive, conformist or projection strategy). This is unsurprising as receivers had no other information to rely on. Projecting own past sending choices to others and acting accordingly, and reacting to the past sending choices of other senders are particularly common behaviors. The unconditional trustful strategy is also relatively common and it is mostly used by senders playing the Nash strategy. Given that in more than half of the groups the average honesty is slightly above 12.5% (the game theoretic prediction), it is rational that Nash senders play trustful strategy, as it generates a higher expected payoff than the Nash or sceptic strategy would. In particular, if the average honesty is $p>12.5\%$, then in each round the expected payoff of trustful strategy is 250p (follows from Figure 2, Chapter 1), whereas the expected payoff of the Nash and sceptic strategy is 250/8 and 250(1-p)/7, respectively, which is less than 250p for $p>12.5\%$.

Strategies in DREP

Table 19 shows the estimated strategy distribution in DREP. The rows correspond to strategies estimated from choices made when subjects were senders while the columns correspond to strategies estimated from choices made when subjects were receivers. We begin by noticing that in DREP many sending and responding strategies are unclassified (27% and 40%, respectively) and elaborate on this first. This result suggests that many subjects, especially as receivers, did not behave strategically or used strategies we did not consider. It is also possible that receivers intentionally (and uniformly) randomized, trusting or doubting the advice with some fixed (possibly similar) probabilities. Such behavior could be justified by the lack of any information about their senders and would suggest that these subjects find the responding role harder to strategize. However, it is still puzzling why DREP has much higher share of unclassified behaviors than DBASE, because in both treatments receivers were equally (un)informed and could only condition trust on experience or they could trust unconditionally. The difference was, though, that in DREP the average levels of honesty were initially higher than in DBASE, perhaps because of the reputation building, and that honesty rates in DREP

declined faster over time than in DBASE. This indicates that the dynamics in our two treatments was different, which could at some point cause a sudden change in behavior (from more trustful to sceptic), experimentation or even switch between the strategies. For example, a subject who decided in advance not to trust in the first 40 rounds might realize after these 40 rounds that senders in her group are often honest, so she might change her strategy after 40^{th} round and trust if senders continue to treat her honestly. Such strategy would probably be left unclassified by our estimation method.

The relatively high percentage of unclassified sending strategies, which is however similar to that in Cai and Wang (2006), may also be a natural response to receivers' behavior which by large could not be explained by our responding strategies. It could also be explained by the fact that in our deception game honesty is not necessarily monetarily costly, whereas helping in our helping game is. In particular, in our deception game even a purely selfish sender may sometimes send the honest message if she believes that her receiver will not trust her. In contrast, in our helping game, a purely selfish sender will never choose the generous option, as that would definitely decrease her profit. In DBASE fewer strategies are unclassified than in DREP. One possible explanation is that in DBASE honesty levels were so low that most subjects did not have incentives to be honest themselves, to experiment or to try new strategies throughout the experiment because honesty would most likely not be returned. Another explanation may be the complexity of DREP. In fact, DREP was the most complex of our four treatments. First, it employed the deception game which is more complex than the helping game (which is actually a decision-making problem). The reason for increased complexity are receivers who are in the deception game also making decisions which can shape senders' beliefs and their subsequent actions. Second, DREP (contrary to DBASE and HBASE) provides reputation information which, on the one hand, can be used as a coordination device or a focal point, but on the other hand adds to the complexity since it gives rise to many new conditional behaviors. It is therefore not surprising that DREP has the most unclassified subjects, since many subjects might have trouble finding a profitable strategy and hence switched between strategies or experimented more.

The most common fully classified pair is the (deceptive, reactive) strategy pair which was also modal in DBASE. This finding suggests that many deceivers expect their current sender behaves as their past senders and best respond accordingly. The (deceptive, trustful)-strategy pair was also relatively common which according to our simple reasoning model from Strategies section corresponds to level-1 subjects. The relative frequencies of other strategy pairs, including those rationalized in Strategies section, are low, mostly below 4%. We can, however, gain insight from separate analyses of the sending and responding strategy distributions.

		Receiver						
		trustful	reactive	conformist	projection	Nash	unclass.	**%**
Sender	deceptive	7%	10%	3%		5%	13%	***38%***
	rewarder				2%		2%	***3%***
	cautious rewarder						3%	***3%***
	mild deceptive	2%						***2%***
	experiential			2%	3%		2%	***7%***
	manipulative				2%			***2%***
	benevolent						10%	***10%***
	Nash	3%		2%		2%	2%	***8%***
	unclass.	3%	3%	2%	8%	2%	8%	***27%***
	%	***15%***	***13%***	***8%***	***15%***	***8%***	***40%***	

Rows correspond to strategies estimated from choices made when subjects were senders. Columns correspond to strategies estimated from choices made when subjects were receivers. Strategies that are not used by anyone are omitted.

Table 19: Strategy distribution in DREP.

As regards the sending strategies, the marginal distribution (Table 19, last column) reveals that the most common strategy is the deceptive strategy. Table 20 shows that deceivers in DREP are more common than defectors in HREP, albeit not significantly, which might explain why honesty is not higher than helping (Table 16). Astonishingly, the reputation-based strategies are rare (8% compared to 44% in HREP) - even the rewarder strategy which was modal in HREP and together with the cautious rewarder strategy crucial for indirect reciprocal helping. Rewarders are also the main reason for the significant difference between DREP and HREP sending strategy distributions (see Table 20).[21] This finding contrasts our hypothesis H11. Moreover, the share of the reputation-based strategies is almost identical to the share of experiential strategy (7%), also contrasting our hypothesis H12. The presence of experientials confirms that experiential behavior is not unique to helping games and should not be neglected in the future experimental games.

Result 8*: In DREP, the reputation-based strategies are rare compared to HREP. Their share is similar to that of experiential strategy.*

The scarcity of the reputation-based strategies means that honesty is not promoted through indirect reciprocity (based on reputation) as often as helping is. This may also explain why honesty is not higher than helping. But why is reputation-based indirect reciprocity so rare in this game? One possible reason might be the strategic nature of the game, as the game gives receivers an opportunity to react to their sender's choice. The receivers' decisions even

[21]Even if in DREP we exclude from our strategy set the manipulative and benevolent sending strategies which were not in our strategy set in HREP, the same conclusions still hold: DREP and HREP distributions are still significantly different, mainly because subjects do not reward honesty with honesty.

determine the payoffs, so the power is distributed differently among senders and receivers in the deception game than in the helping game where receivers are powerless. This makes the game more involved and forces senders to form beliefs about the receivers' behavior. The sender's choice no longer directly results in her desirable outcome but at most influences the beliefs of receivers about the sender's private information which can consequently lead to the realization of the outcome that sender wants. In particular, in our helping game rewarders reward altruistic behavior with altruism (helping) and punish selfish behavior with selfishness (passing). In our deception game, neither honesty will necessarily reward honest individuals nor deception will necessarily punish deceivers, because sender's choice does not determine the outcome. A rewarder will reward honesty with honesty and punish deception with deception only if she beliefs that her receiver will trust her. If a rewarder believes that her receiver will not trust her, then she will reward honesty with deception and punish deception with honesty.[22] We included such inverted rewarder strategy in our strategy set and re-estimated strategies again, but the estimation procedure does not detect the presence of it. However, if rewarders were changing their beliefs about the receivers' behavior throughout the experiment, then it is possible that they were switching between the standard and inverted rewarder strategy in which case our model would most likely leave them unclassified because technically they were mixing between the two strategies. Another possible reason is that rewarding based on reputation was less prominent than rewarding based on experience in which case subjects would be classified as experientials. We checked this by excluding experientials from our strategy set and re-estimating the model but found no support for it. The third possible reason is that in DREP the rewarding was less intense than in HREP, possibly because honesty levels were slightly lower. In that case the rewarders' choices might be only weakly correlated with reputation and closer to 50% which would leave them unclassified. The fourth possible reason is that rewarders are present but are playing some other (stochastic) rewarder substrategy that is not included in our strategy set, in which case their strategy might be left unclassified or classified as some other non-rewarding feasible strategy.

[22]In both cases, however, whether or not the rewards and punishments are ultimately effective depends on receiver's actual choice and sender's belief about her receiver's choice. They are effective if a sender forms a correct belief about her receiver's choice.

DREP	DREP %	HREP %	HREP	sign.
honest	0%	13%	altruist	**
deceptive	38%	22%	defector	ns
rewarder	3%	24%	rewarder	**
cautious	0%	2%	cautious	ns
cautious rewarder	3%	13%	cautious rewarder	ns
mild deceptive	2%	6%	mild defector	ns
experiential	7%	7%	experiential	ns
manipulative	2%	0%	/	ns
benevolent	10%	0%	/	ns
Nash	8%	0%	random8	ns
unclass.	27%	13%	unclass.	ns

** - significant at the 0.05 level
ns – not significant

Table 20: Comparison of sending strategy distributions, DREP and HREP.

As regards the responding strategies (see Table 19, last row) we first notice that the behavior of 40% of the subjects remain unclassified, for which we have already provided some potential explanations above. Among the classified strategies, those that condition on experience or memory are the most common, as almost 40% of subjects use either the reactive, conformist or projection strategy. The most common among the three is the projection strategy. This suggests that many subjects in receiver's role act as if they expect others to act in sender's role as themselves. The least common among the three is conformist strategy. This suggests that most subjects do not blindly imitate actions of receivers they had met and rather condition their trust on some other information or trust unconditionally. Recall, the projection and conformist strategies were also, respectively, the most and least common conditional responding strategies in DBASE. This finding is therefore robust. Finally, we formally tested the similarity between DREP and DBASE strategy distribution (hypothesis H13) and found that they are significantly different ($p<0.05$, Fisher exact test). This suggests that the reputation mechanism triggers different behaviors in the same game, perhaps because the reputation information adds to the complexity of the social environment or because it increases the levels of honesty and trust, making experimentation and search for strategies oriented towards social goal (i.e., honesty and trust) less risky and more attractive.

Result 9: *DREP and DBASE responding strategy distributions are significantly different.*

3.4 Conclusion

In this chapter we investigate honesty/deception and trust among strangers using the repeated deception game. We first study the dynamics of honesty and trust under the simple reputation

mechanism which gives senders an access to information about the most recent decisions their current receiver has made. We find that reputation mechanism promotes honesty and trust, but both decline over time. Then we utilize the analogy between the decisions of senders in our deception and helping games - which had as similar interface as possible - to study deception aversion and its dynamics over time. We find no evidence of deception aversion, neither in the initial rounds nor overall, which contrasts the results of most one-shot experiments (e.g., Gneezy, 2005; Vranceanu & Dubart, 2019) but is in line with recent results by Sasaki et al. (2019). The potential reasons for our contrasting finding might be the dynamic version of our game coupled with reputation information (in one treatment) and experience, information about (misaligned) incentives or even the size of deception reward. To better understand our findings and to learn whether indirect reciprocal honesty based on reputation is as important as it was indirect reciprocal helping in our helping game, we look closely at our experimental subjects and classify their behavior (sending and responding separately) using the mixture model-based estimation method. We find that there are many deceivers, but more importantly, rewarders who were modal in our helping game are extremely rare in our deception game. The scarcity of reputation-based behavior indicates that honesty is not promoted through reputation-based indirect reciprocity as often as helping is. The lower level of this type of indirect reciprocity along with a slightly higher share of deceivers, as compared to the share of defectors in our helping game, may also explain why honesty is not higher than helping. In fact, in our deception game the reputation-based indirect reciprocity is as common as experience-based indirect reciprocity. We propose several potential reasons for the lack of indirect reciprocity, including the strategic nature of the game which gives power to receivers, and less intense rewarding that is only weakly correlated with reputation which makes the behavior closer to random. In deception game we detect some completely honest subjects, but since their share is not significantly higher than the share of altruists in helping game, their honesty may simply be due to their preference for the outcome corresponding to honesty rather than deception aversion. We also analyze the behavior of receivers and the analysis suggests that they most often project their own past sending choices to other senders and act accordingly. Reacting to past sending choices of other senders and unconditional trust are also relatively common behaviors. The behavior of many receivers remains unexplained, though, especially under reputation mechanism, suggesting that the reputation mechanism triggers different behaviors in the same game, perhaps because the reputation information adds to the complexity of the social environment or because it increases honesty and trust, making experimentation and search for social-oriented strategies less risky and more attractive. Finally, we also examine (sending, responding) strategy pairs and found that the most common fully classified pair is the (deceptive, reactive) strategy pair. It may be used by subjects who on the one hand adapt to an environment with an unknown sender's distribution and on the other hand expect that receivers are relatively trusting (which was indeed evidenced in our experiment).

Chapter 4

Cross-national study on sociality measures

This chapter is a minor modification of sections 2-6 in Velkavrh and Ule's (2022) paper entitled "*Indicators of human sociality in Slovenia and the Netherlands: Evidence from experiments with students*" that was published in Teorija in praksa journal (doi: 10.51936/tip.59.2.487-508). It explores socio-economic behavior along its many different dimensions. Its focus is on the question whether similar populations exhibit similar moral choices even when they have different geographical backgrounds. While Chapters 2 and 3 each focused in depth on a single dimension of prosociality or morality - generosity and honesty - this chapter investigates eight moral and economic problems. We replicated the same experiment in two cities in Slovenia and one in the Netherlands. Our experiment was built around a fixed sequence of eight economic tasks that induced different moral or economic phenomena: solidarity, cooperation, dominance, positive and negative reciprocity, trust, honesty, and risk aversion. In all three locations, we recruited local and international students in order to compare the behavioral characteristics of Slovenian students with those of international students and students from a distant European society: the Netherlands.

This part of the doctoral dissertation may be particularly insightful for Slovenian audience because it is, at least to our knowledge, one of the rare examples of a cross-national experimental study involving Slovenian subjects, where with "experimental" we mean laboratory or field experiments and not vignette studies or survey questionnaires. The only other study that we are aware of is an often-cited study by Roth et al. (1991) who compared the bargaining behavior in Israel, Japan, Slovenia and United States and found that Slovenes propose more generous offers than Japanese and Israelis. The main purpose of this project is to see whether significant differences between Slovenian and other groups exist, and if yes, on which dimensions.

Regarding the Slovenians and the Dutch, a comprehensive cultural study by Hofstede et al. (2010) documented that they differ in several dimensions. For instance, Slovenia is among the collectivistic and the Netherlands among the most individualistic countries. To date it is unclear whether these differences expand to student populations which are more homogeneous in values than the general populations.

Results presented in this chapter can be used in meta-analyses which often miss the data about Slovenia. They will also be useful for interpretation and general validity of results from future experiments in the new laboratory at the University of Primorska.

4.1 Theoretical concepts and contexts

The general social sciences aim to describe how the most common behavioral characteristics of human sociality, describing human moral or strategic interaction, vary within and across different contexts, which offers an insight into their drivers and evolution across history and geography. Sociality is "fundamentally dynamic and dialectical, subject to extension and contraction, and having both positive and negative valences, it is not only a resource but also a burden" (Sillander, 2021, pp. 1-2). Heterogeneity in sociality may, for instance, help explain differences in the dominant responses by people to social crises and conflicts. Culture is one source of contextual variance, and it is important to understand the extent to which it impacts the heterogeneity in sociality.

Societies are often compared on dimensions like trust, cooperation, honesty, fairness, reciprocity, and risk attitudes (e.g., Boer & Fischer, 2013; Thöni, 2019). These are among the key characteristics of human sociality and commonly viewed as positive. None is simple or one-dimensional, and we can find a wealth of related concepts in sociology and psychology. Trust, for instance, has emotive, behavioral and communication elements, even if it cannot be commanded, but only offered and accepted. Trust is not simply a rational act; it always contains an element of faith, but not blind faith. Trust therefore presupposes risk and may lead to disappointment and regret if expectations are unmet (Luhmann, 1988). Similarly, the display of solidarity or reciprocity in relationships spans positive and negative orientations. Solidarity may require social exclusion, while positive reciprocity often emerges in relationships that understand negative reciprocity. Demonstrations of solidarity, honesty and reciprocity in relationships also depend on the expressed strength of prosocial orientations and the wider social context (Smith & Sorrell, 2014).

Moreover, it is not merely the behavior that varies situationally for the same person; the core motivations to act also vary situationally within the same individual (Ross & Nisbett, 1991). For example, while the reciprocity of prosocial individuals does not strongly depend on the impressions of the other (honesty, intelligence and unintelligence, in particular), that of the proself individuals is chiefly promoted by impressions of honesty/sincerity and less by intelligence/unintelligence (Van Lange & Semin-Goossens, 1998).

In this study, laboratory experiments are used to explore whether the influence of dominant cultural patterns and national traditions can be detected over the variance in social behavior from personal aspects. We control for social and institutional factors by creating similar experimental incentives and environments in all geographical locations under study, and by controlling for our subjects' demographic characteristics. This would be difficult to control in a conventional public opinion survey. Survey responses are also often subject to prevailing stereotypes and prejudices in given national or social settings. One's personal sense about the

basic characteristics of sociality may be especially driven by prejudices and stereotypes that affect the social categorization of individuals or groups, such as those describing what is typical or atypical for the social functioning of people from one's own groups or from some foreign, especially marginal group (M. Ule, 2004). While every nation possesses stereotypes about how it compares to others, they can be misleading (Scheuch, 1993). There are hence few cross-national comparative studies of sociality, for example the regional analysis of 30 European countries by Koster (2013) and a comprehensive cross-cultural study by Hofstede et al. (2010) that compares over 75 countries and regions on several dimensions, including individualism/collectivism, power distance, and uncertainty avoidance.

Since sociality is most clearly expressed in practice in people's actual behavior in various social situations, we decided to conduct a comparative incentivized experimental study among Slovenian, Dutch and international students with respect to eight indicators of sociality: solidarity, trust, cooperation, positive and negative reciprocity, competition, honesty, and risk attitudes. These indicators were measured with eight standard tasks from experimental economics. In so doing, we are aware that "individual and cultural differences in game behaviors can reflect both the ways in which people perceive game situations and their general social preferences" (Yamagishi et al., 2013, p. 260).

Cross-cultural experimental comparative research is more commonly employed for individual tasks, although some studies have a larger scope. For example, Henrich et al. (2005) implemented three experimental economic tasks in 15 small-scale societies around the world, testing assumptions about economic rationality in the social behavior of people from different social and cultural backgrounds. The key results of this research were:

> *first, there is no society in which experimental behavior is fully consistent with the selfishness axiom; second, there is much more variation between groups than previously observed, although the range and patterns in the behavior indicate that there are certain constraints on the plasticity of human sociality; third, differences between societies in market integration and the local importance of cooperation explain a substantial portion of the behavioral variation between groups; fourth, individual-level economic and demographic variables do not consistently explain behavior within or across groups; and fifth, experimental play often reflects patterns of interaction found in everyday life.* (Henrich et al., 2005, pp. 797-798)

The scope of our study is broader as we cover eight classic economic tasks, yet it is narrower in geographic comparison given that our subjects come overwhelmingly from various European countries, primarily Slovenia and the Netherlands. Our working hypothesis is therefore that the sociality patterns in our samples are mostly similar, with the variance driven more by demographic characteristics than nationality.

4.2 Research methods

To gather the data, we organized a series of experiments with volunteers recruited from among students at various faculties in Koper and Ljubljana in Slovenia, and in Amsterdam in the Netherlands. In total, 128 subjects participated in the experiment, each once. Our sample contains 49 Slovenian students who study in Slovenia, 23 Dutch students who study in the Netherlands, and 56 international students who study in Slovenia or in the Netherlands but are neither Slovenian nor Dutch. All the Slovenian and Dutch students in our sample study in their home country. The experimental sessions were conducted between May 2017 and February 2018. The experiment was conducted through computers, using the Z-tree experimental software (Fischbacher, 2007). Statistical analysis was performed in the statistical software Program R (R Core Team, 2019) using stats and vgam packages (Yee, 2010).

Each subject participated in an identical sequence of eight experimental tasks at a laboratory dedicated to economic experiments at their university. After the experiment, the subjects completed a brief questionnaire that included demographic and background information. Communication between subjects was not allowed during the experiment. Anonymity was assured throughout the experiment by placing subjects randomly in private cubicles and making it obvious that the experimenters could not connect their decisions to their names.

In each task, the subjects could obtain points with their decisions. At the end of the experiment, we randomly selected one task and paid each subject 10 eurocents for every point they had obtained in the selected task. In this way, the decisions were not hypothetical but held real consequences for the subjects' earnings. Performance-based earnings are the key element of economic experiments, intended to increase the realism of every decision the subjects make (e.g., Hertwig & Ortmann, 2001). The subjects had the payment procedure explained to them before the experiment yet did not know which task would be paid, inducing them to consider each of the eight tasks as if it were one that would determine the earnings for all subjects. Participation fee and earnings from a disconnected post-experimental task were added on top of the money earned from the decisions and the total paid to the subject anonymously and in private before they left the laboratory. The average earnings were EUR 12 for an average duration of 50 minutes, a substantial premium over the average student wage. No other benefits were accrued from participation, except for the money earned from fees and decisions and this was advertised during the recruitment.

Each session began with instructions about laboratory conduct and then the subjects participated in the eight experimental tasks as described below. For each task, they first received the description written in a neutral language to avoid framing, and then everyone simultaneously submitted their decision. Subjects did not learn about the decisions of the other subjects until the end of the experiment to avoid any spill-overs between the tasks and to assure

we could analyze each task separately. All interactive tasks were therefore translated into simultaneous games.

The experiment comprised of six interactive tasks (two 3-player games, four 2-player games), and two individual tasks. Everyone completed the tasks in the sequence presented below, starting with the solidarity game and finishing with the risk task. In the interactive tasks, the subjects were randomly grouped in pairs or triplets. Identities of group members were not revealed to protect anonymity. We derive a simple prediction for each task using standard economic theory. We do not consider that subjects may randomize (use mixed strategies). Complete instructions are provided in Appendix B4.

Task 1: Solidarity game

The solidarity game investigates prosocial attitudes of fortunate individuals with regard to less fortunate others. It was developed by Selten and Ockenfels (1998) to measure the "willingness to help people in need who are similar to oneself but victims of outside influences such as unforeseen illness, natural catastrophes, etc." (Selten & Ockenfels, 1998, p. 518). In this game, donations are one-sided and there is no mechanism for explicit reciprocity.

The specific setup is as follows. Each subject in a group of 3 will play a lottery that gives either 60 points ("rich") with a 2/3 probability, or 4 points ("poor") with a 1/3 probability. Before a subject is told the outcome of anyone's lottery, they make two decisions that only become relevant if they later receive 60 points in their private lottery. First, they decide how many of their 60 points they would donate to a poor subject if there were just one in their group. Second, they decide how many points they would donate if both of the other subjects in their group were poor.

The final payoffs are as follows. If all three subjects are rich (poor), each gets 60 (4) points. If just two subjects are rich and donate $x_1 \in \{0, \ldots, 60\}$ and $y_1 \in \{0, \ldots, 60\}$ to the third poor subject, the former end with $60-x_1$ and $60-y_1$ points whereas the third ends with $4+x_1+y_1$ points. If only one subject is rich and donates $x_2 \in \{0, \ldots, 30\}$ to each other subject, she ends with $60-2x_2$ points, and the other two with $4+x_2$ points each.

A rich donor does not benefit financially from helping the poor. The standard prediction for the game is therefore that no donations will be made. However, a donor might donate some points if they dislike large inequalities (Fehr & Schmidt, 1999). Indeed, evidence from previous experiments suggests that many subjects commit to positive donations, leading to substantial average solidarity (Selten & Ockenfels, 1998). Solidarity can be affected by culture, however. For example, Ockenfels and Weimann (1999) found that subjects in the western part of Germany donate significantly more often (79% vs 52%) and higher average amounts (25%–31% vs 16%–20% of points) than those in the eastern part of Germany. As shown by Brosig-

Koch et al. (2011), these differences between West and East Germans were still visible in 2009 even after controlling for other variables such as education and gender.

Task 2: Public goods game

The public goods game models a problem of cooperation where the selfish interests of individuals conflict with the collective interest of the group as a whole (e.g., Andreoni, 1988). It exposes the free-riding problem that occurs when selfish individuals use and enjoy the benefits of publicly provided work, like clean environment and public facilities, but do not provide any work themselves. Widespread free-riding may destroy public good provision by the others (e.g., Marwell & Ames, 1979). Collective problems investigated with this game include teamwork, public space organization, donations to charities, and global pollution.

In our setup, the subjects are placed in groups of 3. Each must allocate 9 tokens between two projects: *private* and *common*. Any token allocated by any subject to the common project yields 2 points to each subject. Each token allocated by a subject to their private project yields 4 points to the subject and no points to the other two. A token in the common project is less profitable for the contributor, but more profitable for the group. Subjects can earn 54 points each if they invest all tokens in the common project. Yet, every subject can earn more by allocating their own tokens to their private project. They thereby earn points from both the common and private projects. Still, if everyone free-rides like this, there is no public good and the subjects earn just 36 points each. If three group members contribute (x_1, x_2, x_3) to the common project, subject *i* earns $\pi_i=2(x_1+x_2+x_3)+4(9-x_i)$ points.

The standard theory predicts no contributions to the common project, which is interpreted as an example of a free market failing to lead to efficient economic outcomes. In contrast, experimental evidence shows that many subjects contribute considerable amounts to the common project (40%–60% on average; see, e.g., Ledyard, 1995; Chaudhuri, 2011). Average contributions are similar in countries with highly integrated market economies (Brandts et al., 2004), yet vary from 22% to 65% in small-scale societies (Henrich et al., 2005).

Task 3: Trust game

The trust game is a simultaneous variant of the dynamic investment game that is used to measure both trust and trustworthiness among experimental subjects (Berg et al., 1995; Bohnet & Zeckhauser, 2004). The idea behind the model is that trust increases social welfare but may be prone to abuse and is therefore risky. The standard example is of two traders who can avoid lawyer fees if they trade without any contracts. One sends money to the other and the other should send goods back to the first after receiving the money. This exchange can be enforced with a contract. However, if one trusts the other to return the goods, the two can avoid the contract-associated costs. Related dilemmas emerge in many daily interactions and trust is an

essential element of functional societies. The trust game measures the fundamental level of trust in a society: towards anonymous strangers.

Our game involves a pair of subjects, a *trustor* and a *proxy* (trustee). Each has two available actions. The trustor (she) is given 40 points that she can either *hold* or *transfer*. The proxy (he) gets 0 points if the trustor holds. Yet, if the trustor transfers, the proxy takes the trustor's 40 points and turns them into 120 points that she can either *keep* or *share* equally with the trustor. In our task, the proxy decides whether to share without yet knowing the decision of the trustor. The final payoffs are shown in the table below. Each cell corresponds to a pair of actions and shows the resulting payoffs for the trustor (first number) and the proxy (second number).

		(proxy)	
		keep	*share*
(trustor)	*hold*	40 , 0	40 , 0
	transfer	0 , 120	60 , 60

Source: own analysis.

Table 21: Trust game.

The standard prediction is that the proxy will keep the points, to which the rational response of the trustor is to hold her points. This is obviously inefficient because both can earn 60 points if they transfer and share their points. The trustor would transfer her points only if she trusts that the proxy will share. A transfer therefore indicates trust and sharing indicates trustworthiness.

The common experimental finding from the sequential version of the trust game is that people generally show a substantial amount of trust, even to complete strangers, and that trust is often rewarded (Berg et al., 1995). This indicates that trust and reciprocity are both important economic primitives. There is some experimental evidence that trust varies across similar countries. For example, Willinger et al. (2003) found that in Germany subjects trusted more than in France, despite trustworthiness being similar in the two countries. Survey questionnaires, for comparison, may suggest greater variation in trust than what is observed in incentivized experiments. For instance, Holm and Danielson (2005) found similar levels of experimental trust between subjects in Tanzania and Sweden, despite significantly different responses to the survey's trust questions. Survey results concerning trust may measure social stereotypes or private trustworthiness rather than actual trust and depend on how respondents understand and interpret the questions as well as their subjective reference point (Glaeser et al., 2000; Sapienza et al., 2013; Banerjee, 2018). The fact that Eurostat (2013) and the World Values Survey (Inglehart et al., 2014) both found that respondents in Slovenia had less trust than those in the Netherlands makes it interesting to gather evidence about their actual trust in incentivized experimental exchanges.

Task 4: Ultimatum game

The ultimatum game is a simple model of bargaining (Güth et al., 1982). A *proposer* suggests a division of 100 points, while the *responder* then either accepts or rejects this division. In case of rejection, the two earn nothing. This game is used to investigate the prosocial attitudes of proposers and the negative reciprocity of responders. By rejecting a positive offer, the responder sacrifices a positive earning to indicate displeasure and punish the proposer. At the same time, a high offer indicates that the proposer understands the possibility of such negative reciprocity among the people in their community.

Our setup considers the simultaneous version of the originally sequential decision game (like, e.g., in Harrison & McCabe, 1996). In our pairs, the proposer (she) offers a number of points $P \in \{0, \ldots, 100\}$ to the responder (he) who at the same time indicates the minimum number of points $X \in \{0, \ldots, 101\}$ he is willing to accept. Here X=0 means "accept any proposal", while X=101 means "reject every proposal". Offer P is then compared to the minimum X. If $P \geq X$, the offer is accepted, the proposer earns 100-P points, and the responder earns P points. If $P<X$, the offer is rejected and both subjects earn 0 points.

The standard prediction for the dynamic game is that the responder will accept any positive offer and the proposer will offer either 0 or 1 point. Although the theory is less narrow for our simultaneous version of the game, the most plausible theoretic predictions are like those above. Choosing $X \leq 1$ means the responder will earn points whenever the proposer makes a positive offer (as they mostly do). Choosing $X>1$, on the other hand, risks losing positive earnings from low offers. A rational responder should therefore choose a higher minimum $X>1$ only when she is willing to incur a cost up to X to punish the proposer for an unfair offer.

In ultimatum game experiments across industrialized societies, the average offers are typically between 30% and 45% of the total, which are usually accepted. Offers below 20% are rare and often rejected (Camerer, 2003). Still, rejection patterns and the notion of a fair division might be country-specific (ultimatum bargaining in Israel, Japan, Slovenia, and the United States was compared by Roth et al., 1991; see also Oosterbeek et al., 2004). Henrich et al. (2005) found larger differences between small-scale non-industrialized societies, with average offers ranging between 26% and 58% and a related variance in rejection patterns.

Task 5: Chicken game

This simple game measures subjects' tendency to compromise and adopt a submissive role in society, which promotes hierarchical ranking. Subjects are paired and each chooses either option A (dominant) or option B (compromise). If one chooses A and the other chooses B, they earn 70 and 30 points, respectively. If both choose to dominate with A, they both earn 0 points. If both choose to compromise with B, they each earn 40 points. The table below shows how again the payoffs correspond to the chosen options.

	A	B
A	0 , 0	70 , 30
B	30 , 70	40 , 40

Source: own analysis.

Table 22: Chicken game.

It is best to choose A when the other chooses B, and to choose B when the other chooses A. The standard prediction is therefore that, despite facing a symmetric social situation, the subjects will make asymmetric choices, with the dominant subject earning much more than the compromising subject. While subjects may agree that specialization is efficient, they would disagree on who should profit from domination. In the absence of communication, like in our experiment, choosing A suggests a willingness to compete for a leading social position.

This game was recently experimentally studied in the Netherlands by de Heus et al. (2010) who found that compromise B is chosen by up to 87.5% of the subjects, but cross-country comparisons are scarce. Carment (1974), for example, found that in a repeated similar experiment Indian males initially compromise slightly more than Canadian males but the latter compromise more in the end.

Task 6: Reward game

In this task, we investigate positive reciprocity. Our reward game models an exchange of favors between two subjects in a pair, the *sender* (she) of a gift and its *recipient* (he). The sender's wealth is at risk of partial destruction. She can gift some of their wealth to the recipient who holds the power to prevent the destruction of the sender's remaining wealth. The recipient must pay to protect the sender but may do this as gratitude for the sender's gift. A sender may then send a positive gift if she expects such positive reciprocity from the recipient. This "gift-exchange" was proposed by Akerlof (1982) as a model to explain why wages are often above the bare minimum. Well-paid workers make a bigger effort which, through positive reciprocity, benefits workers and employees alike. Low wages may on the other hand be perceived as unfair and lead to low productivity and high unemployment (Akerlof & Yellen, 1990).

In our game, we pair the subjects and then each sender is given 90 points and their recipient is given 10 points. The sender chooses a number of points $G \in \{0, \ldots, 90\}$ to give to the recipient, while the recipient chooses the minimum gift $X \in \{0, \ldots, 91\}$ for which he will protect the sender's (remaining) points. Here, X=0 means "always protect the sender", and X=91 means "never protect the sender". Sender's gift G is then compared to the recipient's demand X. When $G \geq X$, the recipient pays 10 points and earns gift G, while the sender earns 90-G points. If $G<X$, the recipient earns G+10 points but does not protect the sender, who earns just one-third of their remaining points, (90-G)/3.

The standard prediction for the sequential version of our game is that the recipient will never protect the sender's points because this is costly. The sender will thus not send any gift to the recipient. The prediction for our simultaneous game is similar: the recipient's demand is so high ($X \geq 60$) that the sender prefers to give nothing and suffers the destruction, earning 30 points and losing 60, despite the protection costing just 10 points. The recipient in this case earns 10 points.

In contrast, most recipients in similar experiments appear to usually reciprocate gifts, which rationalizes gift sending. Senders in turn often send substantial gifts to the recipients, increasing the efficiency of their exchange (the average gift exceeded 40% of the total in Fehr et al., 1993). This efficiency does not substantially differ among industrialized countries, with Germany leading Israel, Japan and the United States, but Spain lagging behind with fewer gifts and lower reciprocity (Waichman et al., 2015).

Task 7: *Lying task*

In this individual task, a subject rolls a die in private and then reports a number from 1 to 6, which determines their payoff: 10 times the reported number. The subject is instructed to report the number of points they privately observe on their die. However, nobody can see their die, so they are free to report a high number even if they have thrown a lower number. The standard prediction is that everyone will report number 6, regardless of what they actually throw on their dice. There is no interaction between the subjects in this task and thus it can reveal the tendency to comply with instructions in the absence of any social context other than the relationship of authority between the experimenter and the subject. This task is hence used in the literature to investigate honesty by comparing the distribution of the numbers reported with the expected uniform distribution of the numbers observed.

Fischbacher and Föllmi-Heusi (2013) estimated that no more than 22% subjects lied by reporting the most profitable number, while almost 40% of subjects were potentially honest. Moreover, many subjects lied by reporting the second-most profitable number, perhaps trying to appear honest in order to maintain a favorable self-image. Experimental data from 47 countries show that honesty varies between countries, but on average only 23.4% of the potential profit from lying is actually taken (Abeler et al., 2019).

Task 8: Risk task

This individual task investigates risk attitudes in the absence of social interaction. A subject is presented with three choices, each between two options. Each choice concerns two options E and F. Option E is always the same lottery yielding either 80 or 20 points with equal probability. Option F is a sure payment, but the amount differs between the three choices, rising from 38 to 50. These three choices are:

a) E = [80p : ½ | 20p : ½] or F = 38

b) E = [80p : ½ | 20p : ½] or F = 44

c) E = [80p : ½ | 20p : ½] or F = 50

The expected payoff from choosing option E is 50 (calculated as 80/2 + 20/2). In pairs (a) and (b), this is better than the payoff from choosing option F. The standard economic theory assumes that people maximize their expected payoff and are therefore neutral with respect to risk (if they know the probabilities). This implies choosing E in both (a) and (b). On the other hand, we may have *risk-averse* subjects who would sacrifice some payoff to avoid risks. These might choose F over E even when F yields less than 50. Choosing F in (a) or (b) therefore suggests that the subject is risk-averse. We say below that those who always choose F exhibit high risk aversion. Those who choose F only in (b) and (c) exhibit moderate risk aversion. In contrast, a subject who seeks risks should always choose E, even in pair (c).

Our task is a simplified version of the classic risk aversion measure by Holt and Laury (2002) who estimate that the majority of people are risk-averse. However, Vieider et al. (2015) found significant cross-country differences in risk attitudes. While in developed countries the subjects are on average risk-averse, in others they can be risk-neutral (e.g., Brazil, Malaysia) and even risk-seeking (e.g., Ethiopia, Nicaragua, Saudi Arabia). Rieger et al. (2015) found that Slovenians are more risk-averse than the Dutch, although this was based on hypothetical choices.

To conclude the section, we provide motivation for our main hypothesis presented below. Since our study was carried out in two European countries, we expected that most of our experimental subjects will come from one of the European countries. Given this, and since the emergence of social networks has offered people the opportunity to chat with people from other countries, which may have caused student culture to cross borders, we hypothesize the following.

Hypothesis H1: *Nationality has no significant effect on sociality measures.*

4.3 Results

Our sample consists of 128 students, of whom 59% are female, 40% are economists (enrolled in finance, business, accounting, or economics tracks), and 62.5% had participated in economic experiments before. Subjects were divided into three cohorts based on their nationality: 38% were Slovenian nationals participating in Slovenia, 18% were Dutch nationals participating in the Netherlands, and 44% were international students from 30 countries, participating in either Slovenia or the Netherlands. Among the internationals, 62.5% came from Europe.

There are 12 decisions of interest in our experiment. In the solidarity game, subjects make two decisions: how much to donate if one group member is poor (Sol1), and how much to donate

if two group members are poor (Sol2). In the public goods game, we measure donation to the common project (PG). In the trust game, we check if the trustors transfer (Tr1) and if the proxies share (Tr2). In the ultimatum game, we measure the proposer's offer (Ult1=P) and the responder's minimum (Ult2=X). In the chicken game, we determine if the subject chooses the dominant action A (Chic). In the reward game, we measure the sender's gift (Rew1=G) and the recipient's minimum demand (Rew2=X). With the lying task, we measure the reported number following the roll of die (Die). In the risk task, the variable (Risk) is 0 if F is always chosen; 1 if (E,F,F) are chosen in (a,b,c); 2 if (E,E,F) are chosen in (a,b,c); and 3 if E is always chosen. Given that the choices for 6 subjects violate this framework, we exclude them from our analysis for this task. A higher value of (Risk) indicates more risky choices and therefore lower risk aversion. In the trust, ultimatum and reward games, only half the subjects choose for each role, resulting in 64 observations per variable.

	Sol1 (0–60)	Sol2 (0–30)	PG (0–9)	Tr1 (0/1)	Tr2 (0/1)	Ult1 (0–100)	Ult2 (0–101)	Chic (0/1)	Rew1 (0–90)	Rew2 (0–91)	Die (1–6)	Risk (0–3)
Standard prediction	0	0	0	0	0	≤1	≤1	0.5	0	≥60	6	2.5
Average observed	12.6	8.1	3.2	0.47	0.55	45.5	28.2	0.33	30.3	51.8	4.5	1.4
Normalized average (SD)	0.21 (0.18)	0.27 (0.24)	0.35 (0.29)	0.47 (0.50)	0.55 (0.50)	0.45 (0.16)	0.28 (0.25)	0.33 (0.47)	0.34 (0.17)	0.57 (0.29)	0.71 (0.29)	0.47 (0.26)

Source: own analysis.

The first row shows the variable and the range of values it can take. The second row shows its theoretically predicted value. The third row shows its observed average value. The fourth row normalizes this average to the interval [0,1]; standard deviation is shown in parentheses.

Table 23: Standard predictions and observed averages for all decisions in all tasks.

Table 23 shows the overall results for all 12 variables from our 8 experimental tasks. Additional distributional information of our economic tasks is presented in Appendix B1. The behavior in all our experimental tasks, unsurprisingly, differs substantially from predictions according to the standard theory, but is consistent with previous experimental evidence.

To investigate the similarities and differences among our three student cohorts, we ran a series of 12 regressions – one for each variable. In each regression, we investigate how the cohort affects a specific variable, controlling for familiarity with experiments, gender, study track (whether the subject is enrolled in one of the economics study programs) and whether the subject is a male economist. Male economists have, for instance, been observed in previous experiments to be significantly less solidary than other subjects (e.g., Selten & Ockenfels, 1998). For non-binary experimental tasks, we confirmed our results with additional Tobit regressions (see Appendix B2). To facilitate comparison across studies and models, we normalize each non-binary variable to the interval [0,1].

	Sol1	Sol2	PG	Chic	Die	Risk	Tr1	Tr2	Ult1	Ult2	Rew1	Rew2
INT	-0.05	0.41	-0.20	-0.26	0.46	-0.10	-1.22	0.24	0.07	-0.18	-0.37	0.75*
	(0.24)	(0.27)	(0.30)	(0.59)	(0.32)	(0.27)	(0.80)	(0.74)	(0.23)	(0.42)	(0.28)	(0.41)
NL	-0.80**	-0.71*	-0.53	1.07*	1.09**	-0.02	-1.10	0.08	-0.40	-0.91*	-0.76**	0.32
	(0.33)	(0.38)	(0.37)	(0.62)	(0.44)	(0.31)	(0.93)	(0.84)	(0.27)	(0.54)	(0.34)	(0.48)
Econ	-0.31	-0.64*	-0.40	0.66	-0.05	-0.11	-0.06	-0.61	-0.64**	0.07	0.05	-0.48
	(0.30)	(0.34)	(0.37)	(0.63)	(0.38)	(0.31)	(0.97)	(0.79)	(0.28)	(0.48)	(0.34)	(0.47)
Male	0.16	0.13	0.60**	-0.16	0.23	0.21	0.73	0.51	0.08	0.80*	0.30	-0.29
	(0.22)	(0.26)	(0.29)	(0.58)	(0.32)	(0.26)	(0.77)	(0.73)	(0.22)	(0.42)	(0.26)	(0.39)
Male	-0.39	-0.45	-0.98**	0.31	0.70	0.43	0.12	-0.75	0.39	-0.59	-0.15	1.12
Econ	(0.40)	(0.46)	(0.49)	(0.82)	(0.55)	(0.41)	(1.10)	(1.19)	(0.32)	(0.68)	(0.39)	(0.69)
Exper	-0.54**	-0.59**	-0.07	0.49	0.16	0.26	-0.10	-0.89	-0.18	0.81*	-0.45*	0.13
	(0.23)	(0.26)	(0.29)	(0.53)	(0.31)	(0.25)	(0.67)	(0.71)	(0.20)	(0.45)	(0.24)	(0.40)
N	128	128	128	128	128	122	64	64	64	64	64	64

Source: own analysis.

Explanatory variables are dummy variables: Econ=1 for economics student; Male=1 for male; MaleEcon=1 for male economics student; NL=1 for Dutch cohort; INT=1 for international cohort (Slovenian cohort is a reference group); Exper=1 if a subject attended at least one experiment in the past. Models for binary variables Chic, Tr1 and Tr2 are logit regressions, and other models are fractional logit regressions. Standard errors in parentheses. Coefficient of the constant omitted for brevity.

The coefficients' signs indicate whether a particular explanatory variable has a positive (+) or negative (-) effect on the dependent variable, and stars ** or * indicate whether this effect is significant at the 0.05 or 0.1 level.

Table 24: Regression results for variables of interest from our eight experimental tasks.

Table 24 summarizes the regression results for all 12 variables. The main observation is that the differences between our Slovenian and international cohorts are never significant at the 5% level. We find only one marginal difference between these two cohorts with respect to positive reciprocity in the reward game.[23] Still, this marginal significance disappears, when we compare the Slovenian students against the pooled Dutch and international cohorts, which suggests the lack of significance is not due to the small number of observations. We therefore conclude that there are no significant differences in any of our variables between the Slovenian and international students.

In contrast, we find significant differences for three measures of sociality between the Slovenian students and the Dutch ones. The Slovenian students give more in solidarity (Sol1 17.4 vs. 6.4), are more honest because they report lower die throws (Die 4.0 vs. 5.3), and send higher gifts in the reward game (Rew1 38.3 vs. 21.7). There are two further marginally significant differences (in Chic and Ult2) and, for a more statistically powerful comparison, we recheck them in new regressions comparing the Dutch against the pooled Slovenian and international cohorts (which do not differ significantly). This yields an additional significant difference (at the 5% level) between the Dutch and the pooled cohort in the chicken game,

[23]Tobit regression detects a significant difference, but we find this estimation more conservative, because Tobit treats our data as censored when in reality it is not and it is still based on normal distribution.

where the Dutch are more likely to choose the dominant action (Chic 0.57 vs 0.28). The Dutch students also show significantly less solidarity and honesty than the pooled Slovenian and international students. In our experimental tasks, the Dutch students therefore appear to be an outlier against the Slovenian and international student benchmark. Our hypothesis H1 is therefore only partially confirmed.

Given that we performed 12 regressions with the same explanatory variables it is possible that at least one of the significant effects is false positive. To further verify our results, we run Bayesian regressions and got similar results (presented in Appendix B2, Table B2). Namely, in Bayesian regressions explanatory variables that were most strongly associated with dependent variables were mostly those that were significant and marginally significant in Table 24.

In the regression models presented in Table 24 we only included independent variables that we measured in the experiment (with the questionnaire). To check whether our results correlate with some standard economic measures that are somehow related to our economic tasks and to further confirm the robustness of our results we re-ran all regressions and added external variables that measure the standard of living (economic development) and perceived corruption of a country. The standard of living can be measured with a well-known economic indicator GDP per capita, while the perceived corruption can be measured with the Corruption Perceptions Index (CPI). For both GDP per capita and CPI, we used 2018 data (World Bank, 2018; Transparency International, 2018) because the experimental sessions were conducted between May 2017 and February 2018. We added GDP per capita (expressed in constant 2015 US$), because our experiment involved economic tasks that mimicked real-life socio-economic situations, so it might have happened that subjects from more economically developed countries behaved differently than the others because they originate from different environments with different social and economic goals. For example, they might have had a different view on solidarity, cooperation and other sociality measures that we included. It might also have happened that subjects exposed to everyday corruption, perceived economic situations differently than the others and hence acted differently. For example, since the corruption is often driven by selfishness (private gain), subjects from more corrupt countries might have been more selfish and might have justified their selfishness (or lying) more easily. That is why we added CPI.

For each of 12 variables, we considered four different models: one with CPI, one with GDP per capita (GDPc), one with interaction term between INT and CPI (INT CPI), and one with interaction term between INT and GDP per capita (INT GDPc). In the following we summarize the main findings of these additional analyses. In the main text we provide only regression results with CPI, as an illustration (see Table 25). The rest can be found in Appendix B3, along

with brief explanation why we did not consider other possible combinations of those variables in the same model.

	Sol1	Sol2	PG	Chic	Die	Risk	Tr1	Tr2	Ult1	Ult2	Rew1	Rew2
INT	0.01	0.44	-0.14	-0.56	0.52	-0.10	-1.10	0.27	0.06	-0.12	-0.30	0.61
	(0.24)	(0.28)	(0.31)	(0.66)	(0.34)	(0.28)	(0.82)	(0.77)	(0.24)	(0.45)	(0.27)	(0.43)
NL	-0.99**	-0.83*	-0.71	1.74**	0.91*	-0.01	-1.90	-0.01	-0.36	-1.06	-1.18***	0.64
	(0.38)	(0.43)	(0.44)	(0.78)	(0.51)	(0.38)	(1.16)	(1.01)	(0.33)	(0.64)	(0.41)	(0.58)
Econ	-0.20	-0.56	-0.30	0.38	0.06	-0.12	0.41	-0.56	-0.67**	0.14	0.30	-0.62
	(0.31)	(0.37)	(0.39)	(0.68)	(0.41)	(0.33)	(1.05)	(0.83)	(0.30)	(0.51)	(0.36)	(0.49)
Male	0.17	0.14	0.61**	-0.21	0.24	0.21	0.69	0.53	0.08	0.83*	0.27	-0.35
	(0.22)	(0.26)	(0.29)	(0.58)	(0.32)	(0.27)	(0.78)	(0.74)	(0.22)	(0.44)	(0.26)	(0.39)
Male Econ	-0.39	-0.45	-0.99**	0.37	0.70	0.43	0.05	-0.75	0.39	-0.57	-0.19	1.13
	(0.41)	(0.47)	(0.50)	(0.83)	(0.56)	(0.41)	(1.12)	(1.19)	(0.32)	(0.69)	(0.38)	(0.70)
Exper	-0.58**	-0.62**	-0.10	0.56	0.12	0.26	-0.21	-0.91	-0.18	0.79*	-0.51**	0.18
	(0.23)	(0.27)	(0.29)	(0.53)	(0.32)	(0.26)	(0.68)	(0.71)	(0.20)	(0.46)	(0.23)	(0.40)
CPI	0.01	0.01	0.01	-0.03	0.01	-0.0003	0.03	0.004	-0.002	0.01	0.02*	-0.01
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.03)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
N	128	128	128	128	128	122	64	64	64	64	64	64

Source: own analysis.

Explanatory variables (except CPI) are dummy variables: Econ=1 for economics student; Male=1 for male; MaleEcon=1 for male economics student; NL=1 for Dutch cohort; INT=1 for international cohort (Slovenian cohort is a reference group); Exper=1 if a subject attended at least one experiment in the past. CPI is Corruption Perceptions Index. Models for binary variables Chic, Tr1 and Tr2 are logit regressions, and other models are fractional logit regressions. Standard errors in parentheses. Coefficient of the constant omitted for brevity.

The coefficients' signs indicate whether a particular explanatory variable has a positive (+) or negative (-) effect on the dependent variable, and stars ***, ** or * indicate whether this effect is significant at the 0.01, 0.05 or 0.1 level.

Table 25: Regression results for variables of interest from our eight experimental tasks. CPI included as an external variable.

Overall, the results and coefficients corresponding to the international and Dutch cohorts remain more or less similar. The models with CPI and GDPc do not detect any significant differences between our Slovenian and international cohort, not even at the 10% level (see Tables 25 and B3), making the results even stronger than before. These two models also strengthen the result of the chicken game (the Dutch cohort takes the dominant role significantly more often) and the reward game (Rew1 now significant at the 0.01 level) but weakens the result of the lying task (the Dutch cohort becomes only marginally less honest), compared to the results from Table 24. The results also confirm that the Dutch cohort does not have a significant effect on negative reciprocity (variable Ult2). In the model with GDPc (Table B3), the Dutch cohort has a negative marginal effect on trust, but we do not consider this as a robust finding, given that this effect is only marginal and insignificant in all other models. Finally, some changes in the significance levels are also detected in models with INT CPI and INT GDPc terms (Tables B4 and B5). These two models estimate that the international cohort sent significantly smaller gifts and trusts less than the Slovenian cohort (the latter result is only marginal in the model with INT CPI). However, these results must be taken with caution

because the significance and magnitude of those coefficients are not stable across the models. The differences summarized in this paragraph might have occurred because CPI and/or GDPc added explanatory power or because some important determinants are still not included in our regression (e.g., religion, the number of siblings).

Our next results concern the role of demographic characteristics in sociality. None of the characteristics we controlled for consistently affects behavior across different tasks. We do, however, find significant effects in specific tasks. Students familiar with previous economic experiments show less solidarity than other subjects. Their gifts are also smaller, but the result is significant only when we additionally controlled for CPI and/or GDP per capita. Men contribute more to the common good, but only if they are not economists. Economists offer less in bargaining, but the result becomes only marginally significant when GDP per capita is included in the regressions.

Our regression analyses indicate that GDPc has a significant effect only on gift-giving behavior in the reward game, whereas CPI is never significant at the 5% level. In particular, subjects from countries with higher GDP per capita send higher gifts. However, given that the Pearson's correlation coefficient between Rew1 and GDPc is very weak and that we performed 12 regressions it is possible that this result is a false positive.

4.4 Discussion

The overall behavioral patterns in our experiments are consistent with those seen in previous economic experiments, confirming the discrepancy between standard economic theory and actual behavior. The behavior of the Slovenian and international students is similar in our measures of sociality.[24] This is interesting given that the international students come from 30 different countries that vary in many dimensions like culture, individualism/collectivism, development, GDP (per capita), and corruption. Most of the international students are from European countries, however.

Having students for subjects is standard but not ideal. On one hand, students may behave slightly less socially desirable in the sense of generosity/altruism, cooperation, and trustworthiness/reciprocity than the general population, as evidenced for instance by Carpenter et al. (2008), Anderson et al. (2013), and Falk et al. (2013). On the other hand, the use of a student population in all of our experimental locations facilitates a level comparison and ensures a degree of homogeneity and internal validity as students are more likely to understand questions and experimental instructions.

[24]Some differences regarding trust and gift-giving behavior are detected in regressions with INT CPI and INT GDPc, but they must be taken with caution, since they are not confirmed by other regressions.

In contrast to the international students, the Dutch students show less solidarity, honesty,[25] generosity and compromise compared to the Slovenian benchmark. These differences indicate that the Dutch students are more self-oriented. The comprehensive cultural study by Hofstede et al. (2010) of 75 countries and regions may explain the nature of this contrast, showing that Slovenians and the Dutch differ in several dimensions. In particular, Slovenia is among the collectivistic and the Netherlands among the most individualistic countries. Collectivistic countries strive for loyalty and commitment to a group (e.g., extended family, organization) which consists of strong bonds and provides safety and protection. Individuals then feel responsible for other in-group members and act to promote the (relatively large) group goals. In comparison, individualistic countries emphasize independence, with a focus on oneself and one's closest family. In light of our experiment, Slovenian students may have considered the other subjects, most of whom were Slovenians, as members of their group and hence behaved socially desirable, whereas the Dutch students did not. This may be explained by the relative heterogeneity of our experimental sessions in the Netherlands, where 69% of the subjects were non-Dutch (international) compared to the relative homogeneity of the sessions in Slovenia where just 9% of the subjects were international. The Dutch students are therefore relatively unlikely to interact with another Dutch person, but the Slovenian ones are likely to interact with another Slovenian student. This does not explain the relatively high sociality among the international students, who are unlikely to share the experimental session with many subjects from their own country, except if they consider other international students as members of their group. Still, the nationality component was significant only in some experimental tasks.

We also observe a localized effect of our control variables on gender, study track, and general familiarity with experiments. This is similar to the result in Benndorf et al. (2017) that familiarity affects behavior in only one out of six tasks similar to ours'. In our experiment, familiarity significantly reduced only solidarity, next to two other marginal effects. When CPI and/or GDP per capita were added to the model, one of these marginal effects – the one corresponding to gift-giving behavior, became significantly negative. The "economist" (study track) variable significantly reduced only the offers in the ultimatum game, which is consistent with, e.g., Carter and Irons (1991). However, when we controlled for GDP per capita, the effect became marginal. Gender significantly affected only the public goods game, where men were more cooperative than women among non-economists, as also observed by Brown-Kruse and Hummels (1993). The marginal effect with men demanding more than women in the ultimatum game is also similar to Eckel and Grossman's (2001) finding that men reject offers more often than women do. When GDP per capita was added to the model, it had a significant positive

[25]The effect on honesty is marginal in regression with CPI or GDPc.

effect on gift-giving behavior, whereas the inclusion of CPI had only marginal positive effect on it.

Cross-country and cross-national experiments are attractive, yet challenging. Researchers must control for potential currency, experimenter and language effects (Roth et al., 1991; Thöni, 2019). We used the same payment schemes and experimenters in all locations. The instructions were given in the language of instruction at the university where we ran our sessions. In the Netherlands, they were given in English while in Slovenia they were translated into equivalent Slovenian. The experimenters were fluent in both languages.

4.5 Conclusion

Using incentivized experiments, we find that Slovenian students are similar to a sample of international students in economic tasks that measure various aspects of sociality. In contrast, we find that the Dutch students differ from the Slovenian students in several tasks. Our results are significant despite the relatively small differences established among the three cohorts of students. Indeed, the similarity between the Slovenian and internationally sampled students in our economic tasks is more telling than any odd difference might be. For our measures of sociality, the Slovenian students are similar to the mixed international student population, confirming the view that student culture crosses borders. On the other hand, the Dutch students appear as an outlier since they are less solidary, honest, generous, and less often adopt a submissive role than the Slovenian students. Therefore, the Dutch students appear more self-oriented and less prosocial. These results were observed despite the small Dutch sample (our statistical power was sufficient and confirmed by ex-post analysis) and are thus likely to extend to larger samples. Future studies should however explore whether this observation can be generalized to non-student populations. If our findings are not generalizable, they might be an effect of the variation in the local educational practices and systems. If they are, the differences in sociality may point to the existence of historically embedded cultural distinctions between a social democratic and a market-liberal society.

Chapter 5

Conclusion

The main results of this doctoral dissertation will broaden the knowledge on altruism (selfishness), honesty (deception), indirect reciprocity and deception aversion in dynamic setup using methods from the field of experimental and behavioral game theory and economics, particularly from the area of experimental repeated games with random matching. The doctoral dissertation will also highlight the usefulness and reliability of strategy classification method based on mixture model estimation that, at least to our knowledge, has not previously been used to estimate strategies in finitely repeated helping and deception games. The last part presents one of the few cross-national studies in the field of experimental economics (game theory) involving Slovenian subjects. It will contribute to the knowledge about sociality measures by comparing the behavior of Slovenian students with that of Dutch and other international students in standard economic tasks.

The dissertation starts with presentation of our (experimental) methodology and then continues with description of experimental games used in the main experiment on which Chapters 2 and 3 are based. Chapter 1 concludes with presentation of our experimental design and procedures of our main experiment.

In Chapter 2 we study helping behavior (altruism) among strangers using a repeated economic game of indirect reciprocity known as helping game. We first confirm the results of the previous literature that reputation increases helping (altruism). Then we investigate behavioral rules that subjects apply in our repeated helping game. Using the statistical mixture model-based method, we estimate that almost 90% of subjects use stable rules. In order to explain the nonstandard behaviors, we propose that previous estimations miss an important class of strategies that subjects use, motivated by subject's personal experience rather than reciprocity based on reputation. This describes the behavior of more than half of the subjects in one of our experimental treatments (HBASE). Moreover, our behavior analysis suggests that such experientials do not react to the most recent experience only (short memory) but rather use strategies based on longer memory and memory decay. This suggests that experientials are more likely driven by learning and adaptation to social environment and group norms rather than emotions that trigger a strong immediate response (e.g., gratitude, anger). We also show that concern for own reputation diminishes in the final rounds of the experiment, which can explain the end-game decline in generosity. Regarding the profitability of strategies we found that under such simple reputation mechanism, which stores first-order information about sender’s current receiver, selfish strategies are more profitable than reciprocal strategies, and

that reaction to experience and unconditional altruism do not pay off. We also find substantial deviations between subjects' self-reports and their actual behavior in the experiment.

In Chapter 3 we study honesty and deception among strangers using the repeated deception game. We first show that reputation increases honesty and trust, and that both decline over time. Then we investigate deception aversion and its dynamics over time by comparing the decisions of senders in our deception and helping games which had as similar interface as possible. We find no evidence of deception aversion, neither in the initial rounds nor overall, which contrasts the results of most one-shot experiments (e.g., Gneezy, 2005; Vranceanu & Dubart, 2019) but is in line with recent results by Sasaki et al. (2019). The reason for our contrasting finding might be the dynamic version of our game coupled with reputation information and experience, information about (misaligned) incentives or even the size of deception reward. A large part of Chapter 3 is devoted to the exploration of behavioral strategies that subjects apply in our deception game. The analysis of sending strategies reveals how important are rewarding and experiential behaviors (as well as deceptive behavior) in the context of honesty which is important for measuring indirect reciprocal honesty. The main finding is that rewarders who are modal in our helping game are rare in deception game. Deceivers are more common, as compared to defectors in our helping game, whereas experientials are present to similar degrees in both games. In our deception game, the reputation-based indirect reciprocity is as common as experience-based indirect reciprocity, which is a stark difference from helping game. We propose several potential reasons for the lack of indirect reciprocity, including the strategic nature of the game which gives power to receivers, less intense rewarding which makes the behavior closer to random, and exclusion of non-standard but important stochastic rewarder substrategy.

We also analyze responding strategies and found that subjects most often project their own past sending choices to other senders and act accordingly. Reacting to past sending choices of other senders and unconditional trust are also relatively common behaviors. The behavior of many receivers remains unexplained, though, especially when reputation is observable, suggesting that the reputation mechanism triggers different behaviors in the same game, perhaps because reputation information adds to the complexity of the environment. Finally, among strategy pairs the most common was the (deceptive, reactive) strategy pair that may be used by subjects who on the one hand adapt to an environment with an unknown sender's distribution and on the other hand expect that receivers are relatively trusting which was indeed evidenced in our experiment. Chapters 2 and 3 may be particularly interesting for psychologists, economists, mathematicians and other scientists interested in human behavior and evolutionary dynamics.

In Chapter 4 we present a cross-national study conducted in Slovenia and in the Netherlands. We report the results of an experiment designed to detect differences in behavioral characteristics among Slovenian, Dutch and international students. Using eight standard tasks

from experimental economics, we investigate the differences using experimental measures of solidarity, trust, cooperation, positive and negative reciprocity, competition, honesty, and risk attitudes. We found that our Slovenian and international cohorts are similar, but the Dutch students are found to exhibit lower levels of solidarity, generosity, honesty and compromise. This points to differences in sociality between institutionally similar yet ideologically distant countries like Slovenia and the Netherlands. The results of this chapter are informative not only for economists, psychologists and mathematicians, but also for other social scientists such as sociologists and anthropologists. Chapter 5 concludes the thesis.

Bibliography

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for Truth-Telling. *Econometrica, 87*(4), 1115-1153. doi:https://doi.org/10.3982/ECTA14673

Akaike, H. (1973). *Second International Symposium on Information Theory, chapter Information Theory and an Extension of the Maximum Likelihood Principle, 267-281.* Akademiai Kiado, Budapest, Hungary.

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics, 84*(3), 488-500. doi:https://doi.org/10.2307/1879431

Akerlof, G. A. (1982). Labor Contracts as Partial Gift Exchange. *The Quarterly Journal of Economics, 97*(4), 543-569. doi:https://doi.org/10.2307/1885099

Akerlof, G. A., & Yellen, J. L. (1990). The Fair Wage-Effort Hypothesis and Unemployment. *The Quarterly Journal of Economics, 105*(2), 255-283. doi:https://doi.org/10.2307/2937787

Alempaki, D., Doğan, G., & Saccardo, S. (2019). Deception and reciprocity. *Experimental Economics, 22*(4), 980-1001. doi:https://doi.org/10.1007/s10683-018-09599-3

Altmann, S., Dohmen, T., & Wibral, M. (2008). Do the reciprocal trust less? *Economics Letters, 99*(3), 454-457. doi:https://doi.org/10.1016/j.econlet.2007.09.012

Anderson, J., Burks, S. V., Carpenter, J., Götte, L., Maurer, K., Nosenzo, D., . . . Rustichini, A. (2013). Self-selection and variations in the laboratory measurement of other-regarding preferences across subject pools: evidence from one college student and two adult samples. *Experimental Economics, 16*(2), 170-189. doi:https://doi.org/10.1007/s10683-012-9327-7

Andreoni, J. (1988). Why free ride?: Strategies and learning in public goods experiments. *Journal of Public Economics, 37*(3), 291-304. doi:https://doi.org/10.1016/0047-2727(88)90043-6

Aoyagi, M., Fréchette, G. R., & Yuksel, S. (2021). Beliefs in repeated games. *ISER Discussion Paper, No. 1119, Osaka University, Institute of Social and Economic Research (ISER), Osaka*.

Apesteguia, J., Huck, S., & Oechssler, J. (2007). Imitation—theory and experimental evidence. *Journal of Economic Theory, 136*(1), 217-235. doi:https://doi.org/10.1016/j.jet.2006.07.006

Austin, W., & Walster, E. (1975). Equity with the World: The Trans-Relational Effects of Equity and Inequity. *Sociometry, 38*(4), 474-496. doi:https://doi.org/10.2307/2786362

Axelrod, R. (1980). Effective Choice in the Prisoner's Dilemma. *The Journal of Conflict Resolution, 24*(1), 3-25. doi:https://doi.org/10.1177/002200278002400101

Baker, W. E., & Bulkley, N. (2014). Paying It Forward vs. Rewarding Reputation: Mechanisms of Generalized Reciprocity. *Organization Science, 25*(5), 1493-1510. doi:https://doi.org/10.1287/orsc.2014.0920

Banerjee, R. (2018). On the interpretation of World Values Survey trust question - Global expectations vs. local beliefs. *European Journal of Political Economy, 55*, 491-510. doi:https://doi.org/10.1016/j.ejpoleco.2018.04.008

Behnk, S., Barreda-Tarrazona, I., & García-Gallego, A. (2019). Deception and reputation – An experimental test of reporting systems. *Journal of Economic Psychology, 71*, 37-58. doi:https://doi.org/10.1016/j.joep.2018.10.001

Benndorf, V., Moellers, C., & Normann, H.-T. (2017). Experienced vs. inexperienced participants in the lab: do they behave differently? *Journal of the Economic Science Association, 3*(1), 12-25. doi:https://doi.org/10.1007/s40881-017-0036-z

Ben-Ner, A., & Hu, F. (2021). Lying in a finitely repeated game. *Economics Letters, 201*, 109741. doi:https://doi.org/10.1016/j.econlet.2021.109741

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior, 10*(1), 122-142. doi:https://doi.org/10.1006/game.1995.1027

Biernacki, C., Celeux, G., & Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis, 41*(3-4), 561-575. doi:https://doi.org/10.1016/S0167-9473(02)00163-9

Bland, J. R. (2020). Heterogeneous trembles and model selection in the strategy frequency estimation method. *Journal of the Economic Science Association, 6*, 113-124. doi:https://doi.org/10.1007/s40881-020-00097-y

Boer, D., & Fischer, R. (2013). How and when do personal values guide our attitudes and sociality? Explaining cross-cultural variability in attitude–value linkages. *Psychological Bulletin, 139*(5), 1113–1147. doi:https://doi.org/10.1037/a0031347

Bohnet, I., & Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior & Organization, 55*(4), 467-484. doi:https://doi.org/10.1016/j.jebo.2003.11.004

Bolton, G. E., Katok, E., & Ockenfels, A. (2005). Cooperation among strangers with limited information about reputation. *Journal of Public Economics, 89*(8), 1457-1468. doi:https://doi.org/10.1016/j.jpubeco.2004.03.008

Boyd, R., & Richerson, P. J. (1989). The evolution of indirect reciprocity. *Social Networks, 11*(3), 213-236. doi:https://doi.org/10.1016/0378-8733(89)90003-8

Brandts, J., Saijo, T., & Schram, A. (2004). How Universal is Behavior? A Four Country Comparison of Spite and Cooperation in Voluntary Contribution Mechanisms. *Public Choice, 119*(3), 381-424. doi:https://doi.org/10.1023/B:PUCH.0000033329.53595.1b

Breitmoser, Y. (2015). Cooperation, but No Reciprocity: Individual Strategies in the Repeated Prisoner's Dilemma. *American Economic Review, 105*(9), 2882-2910. doi:10.1257/aer.20130675

Brief, A. P., & Motowidlo, S. J. (1986). Prosocial Organizational Behaviors. *Academy of Management Review, 11*(4), 710-725. doi:https://doi.org/10.5465/amr.1986.4283909

Brosig-Koch, J., Helbach, C., Ockenfels, A., & Weimann, J. (2011). Still different after all these years: Solidarity behavior in East and West Germany. *Journal of Public Economics, 95*(11-12), 1373-1376. doi:https://doi.org/10.1016/j.jpubeco.2011.06.002

Brown-Kruse, J., & Hummels, D. (1993). Gender effects in laboratory public goods contribution: Do individuals put their money where their mouth is? *Journal of Economic Behavior & Organization, 22*(3), 255-267. doi:https://doi.org/10.1016/0167-2681(93)90001-6

Bucciol, A., & Piovesan, M. (2011). Luck or cheating? A field experiment on honesty with children. *Journal of Economic Psychology, 32*(1), 73-78. doi:https://doi.org/10.1016/j.joep.2010.12.001

Butler, J. V., Giuliano, P., & Guiso, L. (2015). Trust, values, and false consensus. *International Economic Review, 56*(3), 889-915. doi:https://doi.org/10.1111/iere.12125

Cai, H., & Wang, J. T. (2006). Overcommunication in strategic information transmission games. *Games and Economic Behavior, 56*(1), 7-36. doi:https://doi.org/10.1016/j.geb.2005.04.001

Camera, G., & Gioffré, A. (2022). Cooperation in indefinitely repeated helping games: Existence and characterization. *Journal of Economic Behavior & Organization, 200*, 1344-1356. doi:https://doi.org/10.1016/j.jebo.2019.11.014

Camera, G., Casari, M., & Bigoni, M. (2012). Cooperative strategies in anonymous economies: An experiment. *Games and Economic Behavior, 75*(2), 570-586. doi:https://doi.org/10.1016/j.geb.2012.02.009

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction.* Princeton: Princeton University Press.

Camerer, C. F., Ho, T. H., & Chong, J. K. (2004). A Cognitive Hierarchy Model of Games. *The Quarterly Journal of Economics, 119*(3), 861-898. doi:https://doi.org/10.1162/0033553041502225

Capra, C. M. (2019). Understanding decision processes in guessing games: a protocol analysis approach. *Journal of the Economic Science Association, 5*, 123-135. doi:https://doi.org/10.1007/s40881-019-00074-0

Carment, D. W. (1974). Indian and Canadian choice behaviour in a maximizing difference game and in a game of chicken. *International Journal of Psychology, 9*(3), 213-221. doi:https://doi.org/10.1080/00207597408247105

Carpenter, J., Connolly, C., & Myers, C. K. (2008). Altruistic behavior in a representative dictator experiment. *Experimental Economics, 11*(3), 282–298. doi:https://doi.org/10.1007/s10683-007-9193-x

Carter, J. R., & Irons, M. D. (1991). Are Economists Different, and If So, Why? *Journal of Economic Perspectives, 5*(2), 171-177. doi:10.1257/jep.5.2.171

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization, 81*(1), 1-8. doi:https://doi.org/10.1016/j.jebo.2011.08.009

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics, 14*, 47-83. doi:https://doi.org/10.1007/s10683-010-9257-1

Christ, S. E., Van Essen, D. C., Watson, J. M., Brubaker, L. E., & McDermott, K. B. (2009). The Contributions of Prefrontal Cortex and Executive Control to Deception: Evidence from Activation Likelihood Estimate Meta-analyses. *Cerebral Cortex, 19*(7), 1557-1566. doi:https://doi.org/10.1093/cercor/bhn189

Crawford, V. P. (2003). Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions. *American Economic Review, 93*(1), 133-149. doi:10.1257/000282803321455197

Crawford, V. P., & Sobel, J. (1982). Strategic Information Transmission. *Econometrica, 50*(6), 1431-1451. doi:https://doi.org/10.2307/1913390

Dal Bó, P., & Fréchette, G. R. (2011). The Evolution of Cooperation in Infinitely Repeated Games: Experimental Evidence. *American Economic Review, 101*(1), 411–429. doi:10.1257/aer.101.1.411

Dal Bó, P., & Fréchette, G. R. (2019). Strategy Choice In The Infinitely Repeated Prisoners Dilemma. *American Economic Review, 109*(11), 3929-3952. doi:10.1257/aer.20181480

de Heus, P., Hoogervorst, N., & van Dijk, E. (2010). Framing prisoners and chickens: Valence effects in the prisoner's dilemma and the chicken game. *Journal of Experimental Social Psychology, 46*(5), 736-742. doi:https://doi.org/10.1016/j.jesp.2010.04.013

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1-22. doi:https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Dilmé, F. (2016). Helping Behavior in Large Societies. *International Economic Review, 57*(4), 1261-1278. doi:https://doi.org/10.1111/iere.12197

Dvorak, F. (2020a). stratEst: Strategy Estimation. R package version 1.0.1. Retrieved from https://CRAN.R-project.org/package=stratEst

Dvorak, F. (2020b). stratEst: Strategy Estimation in R. *TWI Working Paper Series, No. 119*.

Eckel, C. C., & Grossman, P. J. (2001). Chivalry and solidarity in ultimatum games. *Economic Inquiry, 39*(2), 171-188. doi:https://doi.org/10.1111/j.1465-7295.2001.tb00059.x

Ekeh, P. P. (1974). *Social Exchange Theory: The Two Traditions.* Cambridge, MA: Harvard University Press.

Engelmann, D., & Fischbacher, U. (2009). Indirect reciprocity and strategic reputation building in an experimental helping game. *Games and Economic Behavior, 67*(2), 399-407. doi:https://doi.org/10.1016/j.geb.2008.12.006

Erat, S., & Gneezy, U. (2012). White Lies. *Management Science, 58*(4), 723-733. doi:10.1287/mnsc.1110.1449

Eurostat. (2013). *Average rating of trust by domain, sex, age and educational attainment level (Year 2013)*, Last update: 08. 02. 2021. Retrieved 04. 04. 2022, from Eurostat: https://ec.europa.eu/eurostat/databrowser/view/ilc_pw03/default/table?lang=en

Falk, A., Meier, S., & Zehnder, C. (2013). Do Lab Experiments Misrepresent Social Preferences? The Case of Self-Selected Student Samples. *Journal of the European Economic Association, 11*(4), 839–852. doi:https://doi.org/10.1111/jeea.12019

Fehr, E., & Gächter, S. (2000). Cooperation and Punishment in Public Goods Experiments. *American Economic Review, 90*(4), 980-994. doi:10.1257/aer.90.4.980

Fehr, E., & Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics, 114*(3), 817-868. doi:https://doi.org/10.1162/003355399556151

Fehr, E., Kirchsteiger, G., & Riedl, A. (1993). Does Fairness Prevent Market Clearing? An Experimental Investigation. *The Quarterly Journal of Economics, 108*(2), 437-459. doi:https://doi.org/10.2307/2118338

Findley, T. S. (2015). Hyperbolic Memory Discounting and the Political Business Cycle. *European Journal of Political Economy, 40*(1B), 345-359. doi:https://doi.org/10.1016/j.ejpoleco.2015.08.002

Fischbacher, U. (2007). z-Tree: Zurich Toolbox for Ready-made Economic Experiments. *Experimental Economics, 10*(2), 171–178. doi:https://doi.org/10.1007/s10683-006-9159-4

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in Disguise—An Experimental Study on Cheating. *Journal of the European Economic Association, 11*(3), 525-547. doi:https://doi.org/10.1111/jeea.12014

Fisher, R. J. (1993). Social Desirability Bias and the Validity of Indirect Questioning. *Journal of Consumer Research, 20*(2), 303-315. doi:https://doi.org/10.1086/209351

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378-382. doi:https://doi.org/10.1037/h0031619

Forsythe, R., Lundholm, R., & Rietz, T. (1999). Cheap Talk, Fraud, and Adverse Selection in Financial Markets: Some Experimental Evidence. *The Review of Financial Studies, 12*(3), 481-518. doi:https://doi.org/10.1093/revfin/12.3.0481

Fosgaard, T. R., Hansen, L. G., & Piovesan, M. (2013). Separating Will from Grace: An experiment on conformity and awareness in cheating. *Journal of Economic Behavior & Organization, 93*, 279-284. doi:https://doi.org/10.1016/j.jebo.2013.03.027

Fudenberg, D., Rand, D. G., & Dreber, A. (2012). Slow to Anger and Fast to Forgive: Cooperation in an Uncertain World. *American Economic Review, 102*(2), 720-749. doi:10.1257/aer.102.2.720

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin, 145*(1), 1-44. doi:https://doi.org/10.1037/bul0000174

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring Trust. *The Quarterly Journal of Economics, 115*(3), 811-846. doi:https://doi.org/10.1162/003355300554926

Gneezy, U. (2005). Deception: The Role of Consequences. *American Economic Review, 95*(1), 384-394. doi:10.1257/0002828053828662

Gneezy, U., Rockenbach, B., & Serra-Garcia, M. (2013). Measuring lying aversion. *Journal of Economic Behavior & Organization, 93*, 293-300. doi:https://doi.org/10.1016/j.jebo.2013.03.025

Gong, B., & Yang, C.-L. (2019). Cooperation through indirect reciprocity: The impact of higher-order history. *Games and Economic Behavior, 118*, 316-341. doi:https://doi.org/10.1016/j.geb.2019.09.001

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin, 83*(2), 314-320. doi:https://doi.org/10.1037/0033-2909.83.2.314

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization, 3*(4), 367-388. doi:https://doi.org/10.1016/0167-2681(82)90011-7

Harrison, G. W., & McCabe, K. A. (1996). Expectations and fairness in a simple bargaining experiment. *International Journal of Game Theory, 25*(3), 303-327. doi:https://doi.org/10.1007/BF02425260

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., . . . Tracer, D. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences, 28*(6), 795–815. doi:10.1017/S0140525X05000142

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences, 24*(3), 383-403. doi:10.1017/S0140525X01004149

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and Organizations: Software of the Mind : Intercultural Cooperation and Its Importance for Survival.* (3rd ed.). New York [etc.]: McGraw-Hill Professional.

Holm, H. J., & Danielson, A. (2005). Tropic trust versus Nordic trust: experimental evidence from Tanzania and Sweden. *The Economic Journal, 115*(503), 505-532. doi:https://doi.org/10.1111/j.1468-0297.2005.00998.x

Holt, C. A., & Laury, S. K. (2002). Risk Aversion and Incentive Effects. *American Economic Review, 92*(5), 1644-1655. doi:10.1257/000282802762024700

Houser, D., & Xiao, E. (2011). Classification of natural language messages using a coordination game. *Experimental Economics, 14*, 1-14. doi:https://doi.org/10.1007/s10683-010-9254-4

Huck, S., Normann, H. T., & Oechssler, J. (1999). Learning in Cournot oligopoly – an Experiment. *The Economic Journal, 109*(454), 80-95. doi:https://doi.org/10.1111/1468-0297.00418

Hurkens, S., & Kartik, N. (2009). Would I lie to you? On social preferences and lying aversion. *Experimental Economics, 12*, 180-192. doi:https://doi.org/10.1007/s10683-008-9208-2

Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., . . . Puranen, B. et al. (Eds.). (2014). World Values Survey: Round Six - Country-Pooled Datafile 2010-2014. Madrid: JD Systems Institute. *Version: https://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp.*

Innes, R., & Mitra, A. (2013). Is dishonesty contagious? *Economic Inquiry, 51*(1), 722-734. doi:https://doi.org/10.1111/j.1465-7295.2012.00470.x

Irlenbusch, B., & Ter Meer, J. (2013). Fooling the Nice Guys: Explaining receiver credulity in a public good game with lying and punishment. *Journal of Economic Behavior & Organization, 93*, 321-327. doi:https://doi.org/10.1016/j.jebo.2013.03.023

Kawagoe, T., & Takizawa, H. (2009). Equilibrium refinement vs. level-k analysis: An experimental study of cheap-talk games with private information. *Games and Economic Behavior, 66*(1), 238-255. doi:https://doi.org/10.1016/j.geb.2008.04.008

Koster, F. (2013). Sociality in Diverse Societies: A Regional Analysis Across European Countries. *Social Indicators Research, 111*, 579-601. doi:https://doi.org/10.1007/s11205-012-0021-0

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159-174. doi:https://doi.org/10.2307/2529310

Ledyard, J. O. (1995). Public Goods: A Survey of Experimental Research. In J. Kagel, & A. Roth (Eds.), *Handbook of Experimental Economics* (pp. 111-194). Princeton: Princeton University Press.

Leib, M., & Schweitzer, M. (2020). Peer Behavior Profoundly Influences Dishonesty: Will Individuals Seek-out Information about Peers' Dishonesty? *Mimeo*.

Leibbrandt, A., Maitra, P., & Neelim, A. (2018). Large stakes and little honesty? Experimental evidence from a developing country. *Economics Letters, 169*, 76-79. doi:https://doi.org/10.1016/j.econlet.2018.05.007

Leimar, O., & Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 268*(1468), 745-753. doi:https://doi.org/10.1098/rspb.2000.1573

Lisofsky, N., Kazzer, P., Heekeren, H. R., & Prehn, K. (2014). Investigating socio-cognitive processes in deception: A quantitative meta-analysis of neuroimaging studies. *Neuropsychologia, 61*, 113-122. doi:https://doi.org/10.1016/j.neuropsychologia.2014.06.001

López-Pérez, R., & Spiegelman, E. (2019). Do economists lie more? In A. Bucciol, & N. Montinari (Eds.), *Dishonesty in behavioral economics* (pp. 143-162). Academic Press. doi:https://doi.org/10.1016/B978-0-12-815857-9.00003-0

Luhmann, N. (1988). Familiarity, Confidence, Trust: Problems and Alternatives. In D. Gambetta (Ed.), *Trust: Making and Breaking Cooperative Relations* (pp. 94-107). Oxford: Basil Blackwell.

Makalic, E., & Schmidt, D. (2016). High-Dimensional Bayesian Regularised Regression with the Bayesreg Package. *arXiv:1611.06649*. doi:https://doi.org/10.48550/arxiv.1611.06649

Marwell, G., & Ames, R. E. (1979). Experiments on the Provision of Public Goods. I. Resources, Interest, Group Size, and the Free-Rider Problem. *American Journal of Sociology, 84*(6), 1335-1360. doi:https://doi.org/10.1086/226937

Mazar, N., Amir, O., & Ariely, D. (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research, 45*(6), 633-644. doi:https://doi.org/10.1509/jmkr.45.6.633

McCloskey, D. N. (1983). The Rhetoric of Economics. *Journal of Economic Literature, 21*(2), 481-517.

Nagel, R. (1995). Unraveling in Guessing Games: An Experimental Study. *The American Economic Review, 85*(5), 1313-1326.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-259. doi:https://doi.org/10.1037/0033-295X.84.3.231

Nowak, M. A. (2006). Five Rules for the Evolution of Cooperation. *Science, 314*(5805), 1560-1563. doi:10.1126/science.1133755

Nowak, M. A., & Sigmund, K. (1998a). Evolution of Indirect Reciprocity by Image Scoring. *Nature, 393*, 573-577. doi:https://doi.org/10.1038/31225

Nowak, M. A., & Sigmund, K. (1998b). The Dynamics of Indirect Reciprocity. *Journal of Theoretical Biology, 194*(4), 561-574. doi:https://doi.org/10.1006/jtbi.1998.0775

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*, 1291-1298. doi:https://doi.org/10.1038/nature04131

Ockenfels, A., & Weimann, J. (1999). Types and patterns: an experimental East-West-German comparison of cooperation and solidarity. *Journal of Public Economics, 71*(2), 275-287. doi:https://doi.org/10.1016/S0047-2727(98)00072-3

Oosterbeek, H., Sloof, R., & Van De Kuilen, G. (2004). Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis. *Experimental Economics, 7*(2), 171-188. doi:https://doi.org/10.1023/B:EXEC.0000026978.14316.74

Primoratz, I. (1984). Lying and the "Methods of Ethics". *International Studies in Philosophy, 16*(3), 35-57. doi:https://doi.org/10.5840/intstudphil198416353

R Core Team. (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Reynolds, S. J., & Ceranic, T. L. (2007). The Effects of Moral Judgment and Moral Identity on Moral Behavior: An Empirical Examination of the Moral Individual. *Journal of Applied Psychology, 92*(6), 1610-1624. doi:https://doi.org/10.1037/0021-9010.92.6.1610

Rieger, M. O., Wang, M., & Hens, T. (2015). Risk Preferences Around the World. *Management Science, 61*(3), 637-648. doi:https://doi.org/10.1287/mnsc.2013.1869

Romero, J., & Rosokha, Y. (2019). Mixed Strategies in the Indefinitely Repeated Prisoner's Dilemma. *Working paper*. doi:https://dx.doi.org/10.2139/ssrn.3290732

Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. New York: Mcgraw-Hill Book Company.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13*(3), 279-301. doi:https://doi.org/10.1016/0022-1031(77)90049-X

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *The American Economic Review, 81*(5), 1068-1095.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred Years of Forgetting: A Quantitative Description of Retention. *Psychological Review, 103*(4), 734-760. doi:https://doi.org/10.1037/0033-295X.103.4.734

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition, 17*, 759-769. doi:https://doi.org/10.3758/BF03202637

Sánchez-Pagés, S., & Vorsatz, M. (2007). An experimental study of truth-telling in a sender–receiver game. *Games and Economic Behavior, 61*(1), 86-112. doi:https://doi.org/10.1016/j.geb.2006.10.014

Sánchez-Pagés, S., & Vorsatz, M. (2009). Enjoy the silence: an experiment on truth-telling. *Experimental Economics, 12*, 220-241. doi:https://doi.org/10.1007/s10683-008-9211-7

Sapienza, P., Toldra-Simats, A., & Zingales, L. (2013). Understanding trust. *The Economic Journal, 123*(573), 1313-1332. doi:https://doi.org/10.1111/ecoj.12036

Sasaki, S., Yamane, S., Mardyla, G., & Ohara, K. (2019). An experiment on conformity in deception. In A. Bucciol, & N. Montinari (Eds.), *Dishonesty in Behavioral Economics* (pp. 213-242). Academic Press. doi:https://doi.org/10.1016/B978-0-12-815857-9.00011-X

Scheuch, E. K. (1993). Theoretical Implications of Comparative Survey Research: Why the Wheel of Cross-Cultural Methodology Keeps on Being Reinvented. *Historical Social Research/Historische Sozialforschung, 18*(2), 172-195.

Schram, A. (2005). Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology, 12*(2), 225-237. doi:https://doi.org/10.1080/13501780500086081

Schram, A., & Ule, A. (Eds.). (2019). *Handbook of Research Methods and Applications in Experimental Economics*. Cheltenham, UK: Edward Elgar Publishing.

Seinen, I., & Schram, A. (2006). Social status and group norms: Indirect reciprocity in a repeated helping experiment. *European Economic Review, 50*(3), 581-602. doi:https://doi.org/10.1016/j.euroecorev.2004.10.005

Selten, R., & Ockenfels, A. (1998). An experimental solidarity game. *Journal of Economic Behavior & Organization, 34*(4), 517-539. doi:https://doi.org/10.1016/S0167-2681(97)00107-8

Sheremeta, R. M., & Shields, T. W. (2013). Do liars believe? Beliefs and other-regarding preferences in sender–receiver games. *Journal of Economic Behavior & Organization, 94*, 268-277. doi:https://doi.org/10.1016/j.jebo.2012.09.023

Sillander, K. (2021). Introduction: Qualifying Sociality through Values. *Anthropological Forum, 31*(1), 1-18. doi:https://doi.org/10.1080/00664677.2021.1893153

Smith, C., & Sorrell, K. (2014). On Social Solidarity. In V. Jeffries (Ed.), *The Palgrave Handbook of Altruism, Morality, and Social Solidarity* (pp. 219-247). New York: Palgrave Macmillan.

Smith, V. L. (1962). An experimental study of competitive market behavior. *Journal of Political Economy, 70*(2), 111-137. doi:https://doi.org/10.1086/258609

Smith, V. L. (1976). Experimental Economics: Induced Value Theory. *The American Economic Review, 66*(2), 274-279.

Sobel, J. (2020). Lying and Deception in Games. *Journal of Political Economy, 128*(3), 907-947. doi:https://doi.org/10.1086/704754

Stahl, D. O. (2013). An experimental test of the efficacy of a simple reputation mechanism to solve social dilemmas. *Journal of Economic Behavior & Organization, 94*, 116-124. doi:https://doi.org/10.1016/j.jebo.2013.08.014

Sutter, M. (2009). Deception Through Telling the Truth?! Experimental Evidence from Individuals and Teams. *The Economic Journal, 119*(534), 47-60. doi:10.1111/j.1468-0297.2008.02205.x

Swakman, V., Molleman, L., Ule, A., & Egas, M. (2016). Reputation-based cooperation: empirical evidence for behavioral strategies. *Evolution and Human Behavior, 37*(3), 230-235. doi:https://doi.org/10.1016/j.evolhumbehav.2015.12.001

Talwar, V. (2011). Moral Behavior. In S. Goldstein, & J. A. Naglieri (Eds.), *Encyclopedia of Child Behavior and Development* (pp. 965-967). Springer, Boston, MA. doi:https://doi.org/10.1007/978-0-387-79061-9_1829

Thöni, C. (2019). Cross-cultural behavioral experiments: potential and challenges. In A. Schram, & A. Ule (Eds.), *Handbook of Research Methods and Applications in Experimental Economics*. (pp. 349-367). Cheltenham, UK: Edward Elgar Publishing.

Transparency International. (2018). *Corruption Perceptions Index*. Retrieved 06. 08. 2022, from https://www.transparency.org/en/cpi/2018

Ule, A., & Živoder, A. (2018). Uporaba eksperimentalne ekonomije za oceno omejene finančne racionalnosti. *Teorija in Praksa, 55*(1), 27-40.

Ule, A., Schram, A., Riedl, A., & Cason, T. N. (2009). Indirect Punishment and Generosity Toward Strangers. *Science, 326*(5960), 1701-1704. doi:10.1126/science.117888

Ule, M. (2004). *Socialna psihologija.* Ljubljana: Založba FDV, zbirka Psihologija vsakdanjega življenja.

Van Lange, P. A. M., & Semin-Goossens, A. (1998). The boundaries of reciprocal cooperation. *European Journal of Social Psychology, 28*(5), 847-854. doi:https://doi.org/10.1002/(SICI)1099-0992(199809/10)28:5<847::AID-EJSP886>3.0.CO;2-L

Velkavrh, Ž., & Ule, A. (2022). Indicators of human sociality in Slovenia and the Netherlands: Evidence from experiments with students. *Teorija in praksa, 59*(2), 487-508. doi:10.51936/tip.59.2.487-508

Vespa, E. (2020). An Experimental Investigation of Cooperation in the Dynamic Common Pool Game. *International Economic Review, 61*(1), 417-440. doi:https://doi.org/10.1111/iere.12428

Vieider, F. M., Lefebvre, M., Bouchouicha, R., Chmura, T., Hakimov, R., Krawczyk, M., & Martinsson, P. (2015). Common Components of Risk and Uncertainty Attitudes Across Contexts and Domains: Evidence from 30 Countries. *Journal of the European Economic Association, 13*(3), 421-452. doi:https://doi.org/10.1111/jeea.12102

Volz, K. G., Vogeley, K., Tittgemeyer, M., von Cramon, D. Y., & Sutter, M. (2015). The neural basis of deception in strategic interactions. *Frontiers in behavioral neuroscience, 9*, 27. doi:https://doi.org/10.3389/fnbeh.2015.00027

Vranceanu, R., & Dubart, D. (2019). Deceitful communication in a sender-receiver experiment: Does everyone have a price? *Journal of Behavioral and Experimental Economics, 79*, 43-52. doi:https://doi.org/10.1016/j.socec.2019.01.005

Waichman, I., Siang, C. K., Requate, T., Shafran, A. P., Camacho-Cuena, E., Iida, Y., & Shahrabani, S. (2015). Reciprocity in Labor Market Relationships: Evidence from an Experiment across High-Income OECD Countries. *Games, 6*(4), 473-494. doi:https://doi.org/10.3390/g6040473

Webster Jr, M., & Sell, J. (Eds.). (2014). *Laboratory experiments in the social sciences.* Amsterdam [etc.]: Elsevier.

Wedekind, C., & Milinski, M. (2000). Cooperation Through Image Scoring in Humans. *Science, 288*(5467), 850-852. doi:10.1126/science.288.5467.850

Willinger, M., Keser, C., Lohmann, C., & Usunier, J.-C. (2003). A comparison of trust and reciprocity between France and Germany: Experimental investigation based on the investment game. *Journal of Economic Psychology, 24*(4), 447-466. doi:https://doi.org/10.1016/S0167-4870(02)00165-4

Wilson, A. J., & Vespa, E. (2020). Information Transmission under the Shadow of the Future: An Experiment. *American Economic Journal: Microeconomics, 12*(4), 75-98. doi:10.1257/mic.20170403

World Bank, World Development Indicators. (2018). *GDP per capita (constant 2015 US$).* Retrieved 06. 08. 2022, from https://data.worldbank.org/indicator/NY.GDP.PCAP.KD?end=2019&start=2010

Yamagishi, T., Mifune, N., Li, Y., Shinada, M., Hashimoto, H., Horita, Y., . . . Simunovic, D. (2013). Is behavioural pro-sociality game-specific? Pro-social preference and expectations of pro-sociality. *Organizational Behavior and Human Decision Processes, 120*(2), 260-271. doi:https://doi.org/10.1016/j.obhdp.2012.06.002

Yee, T. W. (2010). The VGAM Package for Categorical Data Analysis. *Journal of Statistical Software, 32*(10), 1-34. doi:https://doi.org/10.18637/jss.v032.i10

Yi, R., Gatchalian, K. M., & Bickel, W. K. (2006). Discounting of Past Outcomes. *Experimental and Clinical Psychopharmacology, 14*(3), 311-317. doi:https://doi.org/10.1037/1064-1297.14.3.311

Zitek, E. M., Jordan, A. H., Monin, B., & Leach, F. R. (2010). Victim Entitlement to Behave Selfishly. *Journal of Personality and Social Psychology, 98*(2), 245-255. doi:https://doi.org/10.1037/a0017168

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and Nonverbal Communication of Deception. *Advances in Experimental Social Psychology, 14*, 1-59. doi:https://doi.org/10.1016/S0065-2601(08)60369-X

Povzetek v slovenskem jeziku

V tržnih okoljih kot tudi vsakdanjem življenju igrata pomembno vlogo prosocialno in moralno vedenje. Prosocialno vedenje je vedenje katerega cilj je povečati korist oziroma zadovoljstvo družbe ali neke skupine posameznikov (na primer, Brief in Motowidlo, 1986), četudi je za to potreben lastni denarni vložek oziroma trud. Primera prosocianega vedenja sta sodelovanje in altruizem. Moralno vedenje je vedenje, ki je skladno s posameznikovimi moralnimi vrednotami, ki se običajno izoblikujejo na podlagi moralnih norm družbe iz katere posameznik izhaja (na primer, Reynolds in Ceranic, 2007; Talwar, 2011). Primer moralnega vedenja je poštenost. Čeprav se pojma velikokrat prepletata, nista ekvivalentna. Če na primer družbi prinese korist nepoštenost, je nepoštenost lahko prosocialno vedenje, ki pa ni moralno. Navkljub pomenu, ki ga prosocialnosti in moralnosti dajejo nekatere družbene vede, kot so na primer filozofija, psihologija in sociologija, prosocialnost in moralnost ne igrata bistvene vloge v klasični ekonomski teoriji in teoriji iger, ki se ukvarjata z odločanjem (racionalnih in sebičnih) posameznikov v interaktivnih okoljih.

Glavni namen doktorske disertacije je s pomočjo ekonomskih pristopov preučevati prosocialno in (ne)moralno vedenje, natančneje altruizem in (ne)poštenost, v interaktivnih ekonomskih situacijah, v katerih posamezniki sprejemajo odločitve, ki vplivajo na lasten dobiček in dobiček nekaterih drugih članov družbe. Temelj naših raziskav predstavljajo matematični modeli s področja teorije iger, za preučevanje dejanskega vedenja posameznikov pa uporabljamo eksperimentalni pristop. Naši rezultati temeljijo na empiričnih podatkih, zbranih s pomočjo nadzorovanih laboratorijskih ekonomskih poskusov, v katerih so (prostovoljni) preiskovanci postavljeni v realistične družbene ali ekonomske situacije, v katerih sprejemajo odločitve. V naših poskusih ima vsaka odločitev dejanske ekonomske posledice za vse vpletene, kar pomeni, da preiskovanci zaslužijo pravi denar in so plačani glede na svojo uspešnost (in uspešnost drugih). Tak način prinaša potencialno večjo notranjo in zunanjo veljavnost laboratorijskega poskusa za družboslovje (Hertwig in Ortmann, 2001; Schram, 2005).

Ekonomski poskusi predstavljajo pomemben, zanesljiv, učinkovit in pregleden način za pridobivanje spoznanj o družbi in, kar je še pomembneje, so ponovljivi in nadzorovani. V laboratorijskem poskusu imajo na primer raziskovalci nadzor nad okoljem (na primer, nad začetnim zneskom, ki ga imajo preiskovanci, nad možnimi odločitvami, ki jih preiskovanci lahko sprejmejo ter pripadajočimi zneski) in inštitucijami (na primer, nad pravili igre ter informacijami, ki jih imajo preiskovanci na voljo). Nadzor raziskovalcem omogoča, da neposredno preučujejo kako sprememba posameznega dejavnika (na primer, dostopnost do informacije o ugledu posameznika ali vrsta protokola, ki določa razvrščanje v pare ali skupine) vpliva na vedenje subjektov. Poskusi torej zagotavljajo učinkovite teste vzročnosti.

V doktorski disertaciji se v prvem poglavju po predstavitvi metodologije posvetimo podrobnemu opisu glavnega ekonomskega poskusa na katerem temeljita drugo in tretje poglavje. V tem delu med drugim natančno opišemo naši glavni eksperimentalni igri, zasnovo poskusa ter postopke izvedbe poskusa. Drugo poglavje je namenjeno preučevanju altruizma (oziroma pomoči) in sebičnosti med neznanci, tj. anonimnimi člani družbe. Prosocialno vedenje, kot sta altruizem in sodelovanje, ima v današnjem svetu pomembno vlogo, saj povečuje dobrobit družbe. Altruizem in sodelovanje sta si podobna v tem, da oba predstavljata strošek (bodisi denarni bodisi psihološki) za altruističnega posameznika oziroma posameznika, ki je pripravljen sodelovati, vendar korist za drugega posameznika ali skupino, pri čemer je korist večja od stroška. Razlika med njima je v tem, da sodelovanje zahteva dva ali več odločevalcev in je torej dvostransko (oziroma večstransko, če interakcija vključuje več kot dva odločevalca), kar pomeni, da je lahko posameznikov trud nagrajen s strani ostalih posameznikov v skupini, medtem ko altruizem vključuje enega odločevalca in enega ali več prejemnikov (pasivnih posameznikov, ki ne sprejemajo nobenih odločitev) in je torej enostranski, kar pomeni, da truda altruističnega posameznika ni mogoče takoj nagraditi.

V primeru enkratne priložnosti za altruistično dejanje, kot je na primer pomoč neznancu v težavah, altruizma med neznanci ni mogoče pojasniti z osnovnim ekonomskim ali biološkim modelom. Klasični ekonomski model namreč pravi, da posameznik ne bo altruističen, ker to zmanjšuje njegova denarna sredstva (dobiček), medtem ko klasični biološki model pravi, da posameznik ne bo altruističen, ker to negativno vpliva na uspešnost njegove reprodukcije. Nekoliko lažje je altruizem med neznanci pojasniti in razumeti v primeru, ko imajo posamezniki večkratne priložnosti za altruizem, pri čemer se občasno znajdejo tudi v vlogi prejemnika altruističnega dejanja. V tem primeru je altruizem lahko učinkovit, saj predstavlja relativno majhen strošek za altruističnega posameznika, a dragoceno korist za prejemnika (na primer, Nowak in Sigmund, 1998a). Raven altruizma oziroma pomoči v družbi je načeloma lahko odvisna od števila altruističnih posameznikov, ki brezpogojno nudijo pomoč. A ker je take posameznike enostavno izkoristiti, so v družbi neuspešni in posledično redki (Nowak in Sigmund, 1998a; Ule idr., 2009). Za raven altruizma v družbi je bolj pomemben pogojni altruizem, ki temelji na nagrajevanju neznancev, ki so sami storili altruistično dejanje. Takšno obliko vedenja, ki je recipročno na posreden način (»jaz *pomagam tebi, ker si ti pomagal nekomu drugemu*«), imenujemo posredno recipročno vedenje. Vloga posredne recipročnosti v družbi je postala še posebej pomembna v zadnjih letih, ko se je s silovitim vzponom spleta in socialnih omrežij število srečanj med anonimnimi neznanci preko spleta, v primerjavi s številom srečanj v živo, bistveno povečalo.

Za delovanje posredne recipročnosti mora biti v interaktivnem okolju omogočeno deljenje informacij o posameznikovih preteklih dejanjih, na primer prek izmenjave izkušenj, saj se na tak način lahko oblikuje posameznikov ugled, ki lahko v prihodnosti vpliva na odločitve članov družbe. Takšen mehanizem ugleda (angl. reputation mechanism) vsakemu posamezniku dodeli

oznako (na primer, pomoč prinese dober ugled, sebičnost pa slabega), s čimer posledično vsak posameznik pridobi potencialno koristno informacijo o svojih, sicer anonimnih, bodočih partnerjih. Opisana vrsta posredne recipročnosti, ki temelji na ugledu, je bila skupaj z vedenjskimi strategijami posameznikov, ki so zaslužna za posredno recipročnost, v preteklosti že predmet nekaterih ekonomskih laboratorijskih poskusov (Seinen in Schram, 2006; Ule idr., 2009; Swakman idr., 2016). Posredna recipročnost, ki temelji na ugledu, pa ni edini način, ki lahko vpliva na stopnjo altruizma v družbi. Altruizem je namreč možno spodbujati tudi z drugo vrsto posredne recipročnosti, ki temelji na osebnih izkušnjah in prilagajanju družbenemu okolju (Boyd in Richerson, 1989). Ta vrsta recipročnosti (*»jaz pomagam tebi, ker je nekdo drug pomagal meni«*) in s tem povezano izkustveno vedenje (angl. experiential behavior) nista bili sistematično preučevani v laboratorijskih poskusih zasnovanih za preučevanje posredne recipročnosti, kljub opažanjem, da so izkušnje pomembne (Bolton idr., 2005; Swakman idr., 2016).

V splošnem sta uspeh altruistično naravnanih posameznikov in razširjenost altruizma (pomoči) v družbi odvisna od vedenjskih pravil oziroma strategij, ki jih uporabljajo pripadniki družbe, pa tudi od vrste mehanizma ugleda in morebitnih izkušenj. Da bi lahko analitične rezultate uporabili za razumevanje odločanja v družbah, moramo razumeti, katere informacije bodo ljudje upoštevali pri svojem odločanju (na primer, izkušnje, ugled nasprotnika, lasten ugled), ali je vedenje posameznikov res mogoče zajeti z določenimi strategijami, kako priljubljene so določene stratgije in kdaj bo pomoč na dolgi rok donosna.

V doktorski disertaciji preučujemo altruizem (oziroma pomoč) ter sebičnost med neznanci s pomočjo ponavljane igre pomoči (na primer, Nowak in Sigmund, 1998a; Seinen in Schram, 2006). Najprej potrdimo ugotovitve preteklih študij, da mehanizem ugleda poveča pomoč oziroma altruizem. Nato se posvetimo preučevanju vedenjskih strategij posameznikov. Z uporabo statistične metode, ki temelji na modelu mešanic (angl. mixture model), ocenimo, da skoraj 90% preiskovancev dosledno uporablja eno od strategij iz naše množice strategij. Preiskovanci uporabljajo zelo različne strategije, ki se razlikujejo glede na raven altruizma kot tudi glede na informacijo na katero pogojujejo odločitve. Izmed vseh strategij je najbolj pogosto uporabljana ravno strategija nagrajevanja (angl. rewarder strategy), ki narekuje altruistično dejanje samo do posameznikov, ki so v preteklosti sami storili altruistično dejanje. Nagrajevalci torej z altruizmom nagradijo altruistično naravnane posameznike. Strategija nagrajevanja je skupaj s strategijo previdnega nagrajevanja (angl. cautious rewarder strategy) tudi ključni razlog za obstoj posredne recipročnosti ter vzdrževanja pomoči v našem poskusu. Velik delež preiskovancev uporablja sebično strategijo (angl. defector strategy), ki narekuje brezpogojno sebičnost. Nestandardno vedenje skušamo pojasniti z opažanjem, da prejšnje ocene zanemarjajo pomemben razred strategij, ki temelji na osebnih izkušnjah. Ta razred strategij pojasni vedenje več kot polovice preiskovancev v enem od naših eksperimentalnih okolij. Majhen delež vedenja pojasni tudi v okolju, kjer preiskovanci poznajo ugled

preiskovanca (potencialnega prejemnika pomoči) s katerim se srečajo. Z analizo podstrategij (angl. substrategies) posameznih strategij pokažemo, da ti »izkustveniki« (angl. experientials) uporabljajo strategije, ki temeljijo na daljšem spominu, kar nakazuje, da sta njihovo vodilo verjetno učenje ter prilagajanje družbenemu okolju, ne pa čustva, ki sprožijo močan takojšen odziv. Pokažemo tudi, da se skrb za lasten ugled v zadnjih ponovitvah igre zmanjša, kar lahko pojasni upad pomoči ob koncu igre. Rezultati analiz podstrategij tudi pokažejo, da številni nagrajevalci ter izkustveniki dopuščajo določeno stopnjo sebičnega vedenja, saj so pripravljeni ostalim preiskovancem pomagati že, če opazijo oziroma izkusijo zgolj eno pomoč v zadnjih ponovitvah igre. Glede dobičkonosnosti strategij ugotovimo, da so sebične strategije bolj dobičkonosne kot recipročne strategije in da se brezpogojni altruizem in odziv na izkušnje ne izplačata. Možen razlog za naš rezultat je lastnost našega mehanizma ugleda pod katerim vsaka pomoč (sebičnost) vodi v dober (slab) ugled, ne glede na to kakšen ugled ima nasprotnik, zaradi česar na primer vsak nagrajevalec, ki kaznuje sebično dejanje s sebičnim dejanjem, izgubi dober ugled, četudi je njegovo dejanje z vidika družbe morda pravično in celo zaželeno.

Ob koncu poskusa so preiskovanci izpolnili kratek vprašalnik, kjer so imeli tudi priložnost opisati svoje vedenje med poskusom. Ta samoporočila (angl. self-reports) preiskovancev analiziramo v zadnjem delu drugega poglavja. Ugotovimo, da samoporočila niso najbolj zanesljiv vir podatkov o vedenju posameznikov. Posamezniki se morda ne zavedajo regularnosti svojih dejanj ali pa jih samo ne znajo ubesediti. Do tega lahko prihaja tudi zaradi namernih izmišljotin - morda povezanih s pristranskostjo zaradi družbene všečnosti (angl. social desirability bias) ali pozabljanja (Russo idr., 1989; Fisher, 1993), saj je poskus vseboval 100 ponovitev igre.

Sodelovanje in altruizem sta pomembna, vendar ko gre za poslovanje, spletno trgovanje in osebne odnose, postaneta bistvena poštenost in zaupanje. Bolj kot na poštenost (oziroma splošneje moralno vedenje) se v laboratorijskih in terenskih poskusih raziskovalci osredotočajo na nepoštenost (oziroma nemoralno vedenje), saj ta v vsakdanjem življenju, ekonomiji in poslovanju predstavlja problem ter vodi v neučinkovitost. Nepoštenost je v eksperimentalni literaturi zelo širok izraz in nekako predstavlja nadpomenko za zavajanje, laganje in druge sorodne vrste vedenja. V kontekstu teorije iger, Sobel (2020) jasno poda razliko med laganjem in zavajanjem. Laž je preprosto izjava (o neki zasebni informaciji), za katero lažnivec verjame, da je napačna, medtem ko je zavajanje izjava, ki v drugih vzbudi napačna prepričanja o zasebni informaciji tistega, ki zavaja. V doktorski disertaciji se osredotočamo na zavajanje, ker je to prisotno v številnih vsakodnevnih situacijah, ki vključujejo dva neznanca, kjer je eden (na primer prodajalec) bolje obveščen kot drugi (na primer kupec), pri čemer so interesi neznancev (na primer trgovalcev) v konfliktu. Primeri vključujejo prodajo rabljenih avtomobilov ter domačih izdelkov in spletno trgovanje, kjer je interes prodajalcev, da oglašujejo nekoliko višjo kakovost prodajanih izdelkov, kar vodi v višji osebni zaslužek na račun kupcev.

V našem poskusu sta (ne)poštenost in (ne)zaupanje preučevana s pomočjo ponavljane igre vrste pošiljatelj-prejemnik (angl. sender-receiver game), imenovane igra zavajanja (angl. deception game). V tej igri je pošiljatelj o plačilni shemi bolje obveščen kot prejemnik, pri čemer oba vesta, da so njune preference neskladne, tj. v konfliktu. V poskusu obravnavamo dve okolji, ki se razlikujeta samo glede informacije, ki jo ima pošiljatelj na voljo o pretekli poštenosti trenutnega prejemnika. Do danes je le malo znanega o tem, kako se poštenost in morebitni psihološki stroški, povezani z zavajanjem, razvijajo s časom, zlasti ob prisotnosti mehanizmov ugleda, ki so zelo pomembni, saj posnemajo mehanizme iz vsakdanjega življenja. Po eni strani se lahko v družbi pojavijo poštene in zaupljive skupine ljudi, ki kaznujejo vsakršno nepoštenost. Po drugi strani je lahko zaradi redne izpostavljenosti zavajanju le-to enostavneje upravičiti, kar vodi do nastanka nepoštenih skupin, s čimer se posledično v družbi izgubi vsakršno zaupanje v neznance. Sorodni pojavi so bili napovedani in opaženi v teoretskih in eksperimentalnih raziskavah posredne recipročnosti v igrah pomoči (Nowak in Sigmund, 1998a; 1998b; Seinen in Schram, 2006). V teh igrah lahko posredna recipročnost spodbuja razvoj pomoči oziroma dobrodelnosti med neznanci, a tudi vodi v začaran krog maščevanja, kjer se sebičnost kaznuje s sebičnostjo.

Velik del tretjega poglavja je posvečen prepoznavanju in raziskovanju vedenjskih strategij, ki jih preiskovanci uporabljajo v naši igri zavajanja, zlasti v okolju kjer je zagotovljen dostop do informacije o ugledu. Za razliko od igre pomoči, v igri zavajanja preiskovanci v obeh vlogah (tj. v vlogi prejemnika in pošiljatelja) aktivno sprejemajo odločitve, in sicer v vsaki ponovitvi igre sprejmejo eno odločitev. Ker je preiskovanec včasih pošiljatelj včasih pa prejemnik, se mora odločiti kako se bo vedel kot prejemnik in kako kot pošiljatelj. Če torej preiskovanec uporablja strategijo, ima ta dve komponenti – prva opisuje vedenje v vlogi pošiljatelja, medtem ko druga opisuje vedenje v vlogi prejemnika. Ker se ta definicija strategije nekoliko razlikuje od definicije strategije, ki smo jo imeli v ponavljani igri pomoči, kjer so strategije opisovale samo vedenje pošiljatelja, smo v izogib zmedi v disertaciji strategijo ponavljane igre zavajanja poimenovali (pošiljatelj, prejemnik) par strategij. Do danes v literaturi ni bilo zaslediti raziskav, ki bi preučevale strategije v igri zavajanja, ki omogoča posredno recipročno (ne)poštenost. Zanimivo bi bilo vedeti ali se posredno recipročno vedenje, zaznano v okoljih z altruističnimi priložnostmi, prenese v okolja, ki ponujajo priložnosti za (ne)poštenost. V igri zavajanja raziskujemo obe vrsti posredne recipročnosti, tj. tiste, ki temelji na ugledu, ter tiste vezane na pretekle izkušnje posameznikov. Poleg strategij pošiljateljev podrobneje raziskujemo tudi strategije prejemnikov in (pošiljatelj, prejemnik) pare strategij.

Čeprav se poštenost in altruizem oziroma pomoč v svoji osnovi razlikujeta, lahko med njima povežemo določene vzporednice in ju primerjamo, če so poskus in vključene osnovne igre ustrezno zasnovane. V ozadju poštenosti je sicer lahko več kot le družbene preference, kar je Gneezy (2005) pokazal s primerjavo iger zavajanja in iger diktatorja z enakimi izplačili. Ugotovil je, da na poštenost poleg družbenih preferenc vpliva tudi odpor do zavajanja, kar so

nato potrdile številne študije o zavajanju (na primer, Sutter, 2009; Erat in Gneezy, 2012; Vranceanu in Dubart, 2019). Dolgoročna interakcija teh dveh različnih vedenjskih motivacij je še vedno nejasna in eden od ciljev te študije je raziskati, ali se morebiten odpor do zavajanja v enkratnih interakcijah ohrani skozi čas, ko se skupine učijo in prilagajajo.

Na podlagi preteklih ugotovitev, da informacija o ugledu povečuje pomoč in da stroški zavajanja zmanjšujejo sebičnost, pričakujemo, da bo zavajanja manj, ko bodo pošiljatelji poznali ugled prejemnika, in da bo poštenost višja od pomoči. Naši rezultati podpirajo naše prvo pričakovanje, ne pa tudi drugega. Poštenost ni bila višja od pomoči niti v prvi ponovitvi igre, kar ni v skladu z ugotovitvami večine prejšnjih študij zavajanja, v katerih so bile igre odigrane enkrat (na primer, Gneezy, 2005; Vranceanu in Dubart, 2019), a bolj konsistentno z ugotovitvami Sasaki idr. (2019). Naš rezultat ne nakazuje, da bi imel povprečni preiskovanec odpor do zavajanja. Morebitni razlogi za naš presenetljiv rezultat so lahko dinamična različica naše igre, informacija o (neskladnih) preferencah, velikost nagrade povezane z zavajanjem, informacija o ugledu in izkušnje ali celo večja čustvena reakcija na zavajanje kot pa na sebičnost, pri čemer se zdita slednja dva razloga po analizi posameznikovih strategij manj verjetna, saj je tako nagrajevanje na podlagi ugleda kot tudi izkušenj prisotno zgolj v manjši meri. To pa niso edini rezultati naše študije. Potrdimo tudi, da informacija o ugledu vodi do večjega zaupanja in da sta poštenost in zaupanje znotraj posameznih skupin preiskovancev močno pozitivno korelirana. Ugotovimo tudi, da je v okolju, ki zagotavlja informacijo o ugledu, poštenost višja od teoretske napovedi, medtem ko je zaupanje v obeh naših obravnavanih okoljih višja od teoretske napovedi. Slednje na nek način nakazuje naivnost prejemnikov - vsaj v okolju, kjer preiskovanci nimajo na voljo nobene informacije o svojem trenutnem prejemniku (čeprav se lahko še vedno učijo iz svojih preteklih izkušenj). Tako zaupanje kot poštenost s časom padata.

V nadaljevanju poglavja podrobneje preučujemo strategije posameznikov. Naša glavna ugotovitev je, da so v naši igri zavajanja nagrajevalci, ki so v naši igri pomoči najbolj pogosti, redki. Njihov delež je podoben deležu izkustvenikov. V okolju brez mehanizma ugleda izkustveno vedenje praktično ni zaznano. Ena od možnih razlag za slednjo ugotovitev je dejstvo, da je bila raven poštenosti v tem okolju tako nizka, da tudi če bi preiskovanci posnemali preteklo vedenje v skupini, bi bili pošteni nekoliko več kot 12.5% časa, kar bi njihovo vedenje uvrstilo najbližje Nashevi strategiji, ki v vsaki ponovitvi igre narekuje poštenost z verjetnostjo 12.5%, ali strategiji, ki narekuje brezpogojno zavajanje (angl. deceptive strategy). Tudi v tem okolju smo sicer zabeležili pošteno skupino, a za poštenost niso bili zaslužni izkustveniki temveč dva altruista in en relativno pošten preiskovanec, katerega vedenje ni bilo mogoče zajeti z nobeno od naših standardnih strategij. Omenjeni preiskovanec je bil v prvih 75 ponovitvah večinoma pošten, nato pa je poštenost popolnoma opustil, verjetno zato, ker sta bila v skupini tudi dva preiskovanca, ki sta uporabljala Nashevo strategijo, in en preiskovanec, ki je brezpogojno zavajal.

V obeh preučevanih okoljih, v katerih so preiskovanci igrali igro zavajanja, opažamo, da je bilo brezpogojno zavajanje (kot strategija) nekoliko bolj pogosto kot je bila brezpogojna sebičnost v okoljih z igro pomoči. Čeprav razlike niso bile signifikantne, lahko to, skupaj z ugotovitvijo, da je pogojna recipročnost, ki temelji na ugledu, v igri zavajanja redka, pojasni naš rezultat, da poštenost ni višja od pomoči. Analiza strategij posameznikov razkrije tudi nekaj popolnoma poštenih preiskovancev, vendar ker njihov delež ni večji od deleža altruistov v igri pomoči, ne moremo zaključiti, da je glavni razlog za njihovo poštenost odpor do zavajanja. Lahko so naši pošteni preiskovalci zgolj preiskovanci, katerim je alternativa, ki pripada poštenosti, bolj priljubljena kot alternative, ki maksimizirajo lasten dobiček na račun prejemnika, zato ker na primer maksimizira skupni dobiček. Poleg tega še pokažemo, da se preiskovanci, ki uporabljajo Nashevo strategijo, vedejo strateško (na primer, da povečajo svoj pričakovani dobiček) in ne naključno (na primer, ker ne bi vedeli, katero vedenje bi lahko bilo donosno). Analiza strategij prejemnikov nakazuje, da ti najpogosteje projicirajo lastno vedenje v vlogi pošiljatelja na druge pošiljatelje in ravnajo v skladu s tem. Prilagajanje oziroma reagiranje na vedenje preteklih pošiljateljev in brezpogojno zaupanje sta prav tako relativno pogosti vedenji prejemnikov.

V okolju z mehanizmom ugleda vedenje številnih preiskovancev ostaja nepojasnjeno, še posebej vedenje prejemnikov, kar kaže na to, da mehanizem ugleda sproži drugačno vedenje v isti osnovni igri, morda zato, ker informacije o ugledu prispevajo h kompleksnosti družbenega okolja ali pa ker zvišajo raven poštenosti in zaupanja, zaradi česar postane eksperimentiranje in iskanje strategij, ki temeljijo na poštenosti in zaupanju, manj tvegano in privlačnejše. Kar se tiče (pošiljatelj, prejemnik) parov strategij je najbolj pogost tisti par strategij, ki narekuje redno zavajanje v vlogi pošiljatelja in prilagajanje okolju v vlogi prejemnika (tj. zaupanje po tem, ko preiskovanec izkusi zadostno raven poštenosti in nezaupanje po tem, ko izkusi zadostno raven zavajanja). Tak par strategij na primer ustreza posamezniku, ki po eni strani verjame, da so ostali naivni oziroma bolj zaupljivi kot predpostavi teorija, po drugi strani pa poskuša čim bolj povečati svoj dobiček v vlogi prejemnika tako, da se prilagodi vedenju preteklih pošiljateljev, ki jih je srečal – če obstajajo obdobja poštenosti, ki jim sledijo obdobja zavajanja, je prilagoditev okolju zagotovo boljši odgovor kot brezpogojno zaupanje oziroma nezaupanje. S tretjim poglavjem se zaključuje osrednji del doktorske disertacije, ki je posvečen raziskovanju altruizma, poštenosti in zaupanja med neznanci, katerim so ponujene večkratne priložnosti za pomoč oziroma poštenost in zaupanje.

Zadnji del disertacije opisuje meddržavno študijo, ki s pomočjo osmih standardnih nalog (iger) s področja eksperimentalne ekonomije oziroma teorije iger preučuje in primerja kakšne so vedenjske značilnosti slovenskih študentov v primerjavi z vedenjskimi značilnostmi nizozemskih ter mednarodnih študentov, ki obiskujejo študij izven države rojstva. Osredotočamo se na eksperimentalna merila solidarnosti, zaupanja, sodelovanja, pozitivne in negativne recipročnosti, tekmovanja, poštenosti in odnosa do tveganja. Ugotavljamo, da so slovenski in mednarodni študenti zelo podobni, medtem ko so nizozemski študenti v primerjavi

s slovenskimi manj solidarni, radodarni in pošteni. Prav tako so nizozemski študenti pripravljeni pogosteje prevzeti dominantno vlogo. To nakazuje na razlike v socialnosti med institucionalno podobnimi a ideološko različnimi državami, kot sta Slovenija in Nizozemska. Podobnost rezultatov slovenskih in mednarodnih študentov dodatno potrjuje splošno veljavnost rezultatov pridobljenih v novem laboratoriju za teorijo iger Univerze na Primorskem.

Appendices

APPENDIX A *Helping and honesty (deception) among strangers*

A1 *Complete list of substrategies in the helping game experiment*

Strategy Category	**Substrategy**	**Description**
altruist	ALT	always helps
defector	DEF	never helps
rewarder	REW1	helps only if receiver helped at least once in the recent three rounds
	REW2	helps only if receiver helped at least twice in the recent three rounds
	REW3	helps only if receiver always helped in the recent three rounds
	SREW	helps with probability *H/3* if receiver helped *H* times in the recent three rounds, where H ∈ {0, 1, 2, 3}
cautious	CAU0	helps only if sender never helped in the recent two rounds
	CAU1	helps only if sender helped at most once in the recent two rounds
	SCAU	helps with probability *1-H/2* if sender helped *H* times in the recent two rounds, where H ∈ {0, 1, 2}
cautious rewarder	CR10	helps whenever REW1 or CAU0 helps
	CR11	helps whenever REW1 or CAU1 helps
	CR20	helps whenever REW2 or CAU0 helps
	CR21	helps whenever REW2 or CAU1 helps
	CR30	helps whenever REW3 or CAU0 helps
	CR31	helps whenever REW3 or CAU1 helps
	SCR	helps with probability *X/6* where X ∈ {0, 2, 3, 4, 5, 6} is the number of deterministic CR strategies that help at a particular input (sender (own) # help, receiver # help)
mild defector	MD10	helps whenever REW1 and CAU0 helps
	MD11	helps whenever REW1 and CAU1 helps
	MD20	helps whenever REW2 and CAU0 helps
	MD21	helps whenever REW2 and CAU1 helps
	MD30	helps whenever REW3 and CAU0 helps
	MD31	helps whenever REW3 and CAU1 helps
	SMD	helps with probability *X/6* where X ∈ {0, 1, 2, 3, 4, 6} is the number of deterministic MD strategies that help at a particular input (sender (own) # help, receiver # help)
random	RAND	in each round helps with probability 1/2
random8	RND8	in each round helps with probability 1/8
experiential	EX1	helps only if sender received help in the last round
	EX31	helps only if sender received help at least once in the recent three rounds
	EX32	helps only if sender received help at least twice in the recent three rounds
	EX33	helps only if sender always received help in the recent three rounds
	EX3S	helps with probability *H/3* if sender received help *H* times in the recent three rounds, where H ∈ {0, 1, 2, 3}
	EXP1	helps only if $h_P \geq 1/4$
	EXP2	helps only if $h_P \geq 1/2$
	EXP3	helps only if $h_P \geq 3/4$
	EXPS	helps with probability x/3 if $h_P \in [x/4, (x+1)/4)$, for x ∈ {0, 1, 2}, and with probability 1, for $x \geq 3/4$.

Table A1: The complete list of substrategies in the helping game.

A2 *Classification methods*

MLFIT Classification method – further details on implementation

MLFIT is based on Dvorak's (2020b) extension of Dal Bó and Fréchette's (2011) and Breitmoser's (2015) mixture model estimation of strategies in repeated prisoners' dilemma games. Our entire statistical analysis was performed in statistical software *Program R* (R Core Team, 2019) using stratEst package developed by Dvorak (2020a). The parameter estimation procedure uses the expectation-maximization (EM) algorithm (Dempster et al., 1977) which is an iterative method that starts with randomly chosen initial parameters p_k and τ and then, at each iteration step, updates them accordingly until convergence is achieved. To avoid local optima to which EM algorithm may converge, Biernacki et al.'s (2003) procedure is used (for further technical details see Dvorak, 2020b). Furthermore, our final model avoids over-fitting of the data by using the Akaike Information Criterion, AIC (Akaike, 1973), to eliminate the worst fitting subset of substrategies. In our particular case, this procedure essentially eliminates all those strategies that are not used by anyone in the complete model.

MLFIT Classification method – additional remarks

Parameter τ: We report tremble τ, as do Dvorak (2020b) and Bland (2020). In contrast, Dal Bó and Fréchette (2019), Vespa (2020) and Aoyagi et al. (2021) report $\beta=1-\tau$, i.e., the probability that a subject does not make an error. Some earlier literature, such as Dal Bó and Fréchette (2011), Fudenberg et al. (2012) and Gong and Yang (2019), estimate and report equivalent parameter γ that measures noise. There is a one-to-one correspondence between τ and γ: $\tau = 1/(1 + e^{1/\gamma})$. Regardless of the parameter chosen to be estimated and reported, it is common when using mixture model-based method to consider it as independent from the subject, strategy and state.

Probability π_{ks}: The intuition behind π_{ks} in connection with the log-likelihood function and maximum likelihood estimates is as follows: for states where a substrategy dictates one of the choices with certainty, i.e., with probability 1, the model assumes that subjects in fact make mistakes and additionally estimates tremble τ, whereas for states where both choices are assigned a positive probability, tremble is assumed to be already incorporated in those non-zero-one probabilities and is not estimated again. See Dvorak (2020b) for further explanation.

Probability θ_{ik}: The formula for θ_{ik} is the standard way of computing posterior probabilities in Bayesian statistics (Dvorak, 2020b). It is used in the repeated prisoners' dilemma literature to estimate subjects' posterior probabilities (Stahl, 2013; Breitmoser, 2015).

DFIT Classification method

This method is based on Seinen and Schram (2006). Each substrategy is first modeled as a deterministic automaton and then we let the computer play the game on behalf of a subject against receivers that the subject had met in the actual experiment. For each subject we then determine, by comparing computer choices with subject's actual choices, which substrategy (and corresponding strategy category) best explains or fits her behavior. If more than one strategy category best explains the behavior of a particular subject, the subject's strategy is classified into a less cognitively complex category. Altruists and defectors are considered as the simplest categories since they are unconditional. They are followed by rewarders and cautious, which condition on one parameter, i.e., either on the receiver's or own reputation. The most complex categories are cautious rewarders and mild defectors which depend on two parameters, i.e., on both the receiver's and own reputation. Furthermore, if the strategy category that best explains subject's behavior correctly predicts less than 70% of her actual decisions, subject's strategy is left unclassified. This last step is not considered in the original paper by Seinen and Schram (2006). We add it to make this classification method more in line with other three methods that leave random behavior and strategies which appear to be mixed unclassified.

SFIT Classification method

This method is based on Ule et al. (2009). It differs substantially from MLFIT and DFIT, as it classifies conditional strategies according to logit regression analyses in combination with some additional criteria related to helping rates. The issue of linear separation occurs rarely and is handled with Bayesian regression.

Under this classification method strategies are classified into one of the four categories: altruist, defector, rewarder and cautious. Unlike other methods, this method does not distinguish sophisticated strategies that condition on both the sender's and receiver's reputation. In particular, it does not distinguish cautious rewarders and mild defectors. Based on the interpretations offered in Ule et al. (2009), however, we believe that cautious rewarder and mild defector strategies are the closest to and hence can be joined with rewarder and cautious strategies, respectively.

SFIT is built under the tacit assumption that rewarder strategies are prevalent in population and therefore classifies these strategies first. To classify rewarder strategies, we run a logit regression (and a Bayesian regression in case of linear separation) for each subject, in which the binary dependent variable is sender's decision to help, and the explanatory variable is the number of helping choices that sender's current receiver has made in the previous three rounds when he was a sender. A subject is classified as rewarder if: i) her individual logit regression yields a significantly positive (5% 1-tail) coefficient estimate on the number of helps chosen

by her receivers, and ii) she helps at least 40% of the time overall. In the second step cautious strategies are classified. In these regressions, the binary dependent variable is again sender's decision to help. However, we now include two explanatory variables, namely 1) the number of helping choices that sender's current receiver has made in the previous three rounds when he was a sender, and 2) the number of own (i.e., sender's) helping choices made in the previous two rounds when she was a sender. A subject is classified as cautious if: i) she has not been already classified as rewarder, ii) her individual logit regression yields a significantly negative (5% 1-tail) coefficient estimate on the number of own helping choices, and iii) her overall helping rate is between 15% and 85%. In the third step subjects who do not fit into one of the above categories are classified. In particular, a subject is classified as defector when her overall helping rate is lower than 20% and as altruist when her overall helping rate is higher than 80%. Otherwise, she is left unclassified.

TREND Classification method

This method is based on Swakman et al. (2016). It is closest to SFIT as it also classifies subjects based on logit regression analyses. However, unlike SFIT, it does not impose any additional criteria related to helping rates. Under TREND subjects are classified into one of the five categories: altruist, defector, rewarder, cautious and sophisticated. In the first step, rewarder, cautious and sophisticated strategies are classified. We run one logit regression for each subject (and a Bayesian regression in case of linear separation), in which the binary dependent variable is sender's decision to help, and three explanatory variables are the number of helping choices that sender's current receiver has made in the previous three rounds, the number of own (i.e., sender's) helping choices made in the previous two rounds, and the round (to control for changes in helping rates over time). A subject is classified as rewarder if her individual logit regression: i) yields a significantly positive (5% 1-tail) coefficient estimate on the number of helps chosen by her receivers, and ii) does not yield a significantly negative (5% 1-tail) coefficient estimate on the number of own helping choices. A subject is classified as cautious if her individual logit regression: i) yields a significantly negative (5% 1-tail) coefficient estimate on the number of own helping choices, and ii) does not yield a significantly positive (5% 1-tail) coefficient estimate on the number of helps chosen by her receivers. If a subject's individual logit regression yields a significantly positive (5% 1-tail) coefficient estimate on the number of helps chosen by her receivers and a significantly negative (5% 1-tail) coefficient estimate on the number of own helping choices, the subject is classified as sophisticated. In the second step the subjects who do not fit into one of the above categories are classified. In particular, a subject is defector when her overall helping rate is below 10% and an altruist when her overall helping rate is above 90%. Otherwise, she is left unclassified.

A3 *Instructions*

A3.1 Instructions for the original helping game experiment

You are about to participate in a decision making experiment. The instructions are simple. If you follow them carefully, you may make a considerable amount of money. Your earnings will be paid to you privately in cash at the end of today's session. In the experiment, earnings are denoted in 'francs'. At the end of the experiment, francs will be exchanged into Euro. The exchange rate will be **1 Euro for 250 francs**. In other words, for every 1000 francs, you will receive 4 Euro. Your decisions are anonymous. They will not be attached to your name in any way. You are not allowed to speak with other participants or communicate in any other way. If you want to ask a question, please raise your hand.

Rounds and Pairs
This experiment consists of 100 rounds. At the beginning of every round the participants will be randomly divided into pairs. In each round new pairs will be made. You are equally likely to form a pair with any other participant. However, the probability that you will form a pair with the same participant in two consecutive rounds is small. The two matched participants in a pair will have different roles. One will be a 'sender', and the other will be a 'receiver'. Which role you have in your pair will be determined randomly at the start of every round.

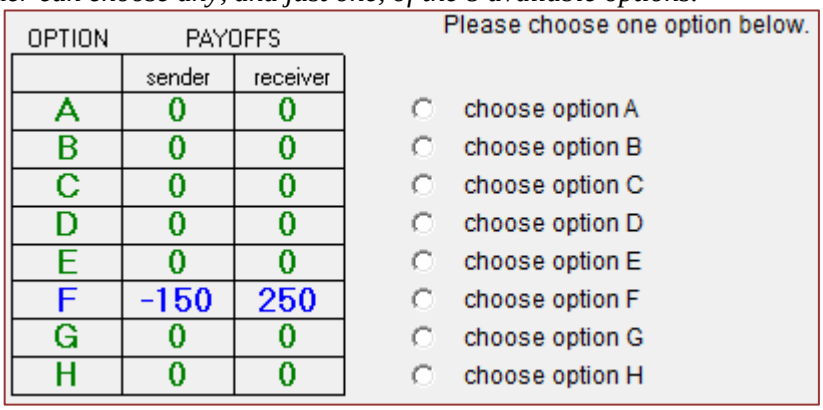
Options
In every round there will be eight available options to choose from: **A , B , C , D , E , F , G ,** and **H**. One of them will be the '**blue**' option. The other seven options will be '**green**' options.

- If the **blue** option is chosen, then the receiver will **earn 250 francs** and the sender will **lose 150 francs**.
- If a **green** option is chosen, then **neither participant in the pair will gain or lose** money.

At the start of every round the computer will randomly select which of the eight options is **blue**. The sender will see what the computer has selected, but the receiver will not see this. The receiver will not know which option is blue and which options are green. The sender will then choose an option. The receiver will make no decision. The sender's choice will determine the earnings for both paired participants in that round.

Sender's choice
If you are the sender, you will see on your screen a table showing the color and earnings for every option. You will then choose one of the options A-H. There will be eight available choices, one for the **blue** option and seven for different **green** options. You will be able to choose any, and only one, of these eight options.

*Example: Below you can see an arbitrary example table with colors and earnings for the 8 options, shown on the sender's computer screen. In this example the computer selected option F as the **blue** option. All the other options are **green**. Only the sender can see this table. The sender can choose any, and just one, of the 8 available options.*

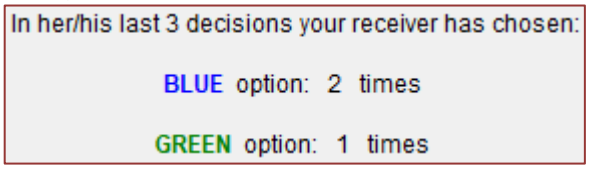


--- [*The following section shown only in HREP treatment*] ---

Sender's information about the receiver

At the start of every round and before making a choice, the sender will receive information about her/his current receiver. If you are the sender, you will see a summary of the **3 most recent** choices that your receiver has made in the past. The information will show how often your receiver chose the **blue** or the **green** option in the 3 most recent rounds when she/he was in the role of the sender. In early rounds of the experiment, the receiver may not have made 3 choices yet. In that case the information will describe all (that is, 0, 1, or 2) of her/his previous choices. Only the sender will get information about the receiver. The receiver will not get any information about the sender.

*Example: Below you can see an arbitrary example information that the sender may have about the receiver. In this example you can see that the receiver has chosen the **blue** option twice and a **green** option once, in the 3 most recent rounds when she/he was a sender.*



---*[ ]*---

Receiver's choice

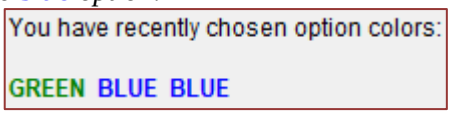
If you are the receiver, you will make no choice in this round. You will be informed about the sender's choice.

End of the round

A round will end after the sender has made a choice. You will then be informed about the **blue** option, and the chosen option. You will also see the resulting earnings.

At the bottom of the screen you will see a review of your own last 3 choices. It will show the colors of the three most recent options you have yourself chosen - in the last 3 rounds you were a sender.

*Example: Below you can see an arbitrary example review of your own past choices. Your most recently chosen option color is shown on the left. In this example you would have last chosen a **green** option, before that you would have chosen the **blue** option, and before that you would have chosen the **blue** option.*



After you press button "continue", the next round will begin: you will be rematched into a new pair with a new participant, you will be randomly assigned a new role, and the computer will randomly select a new **blue** option (any option A-H) for your new pair.

Final instructions

At the start of the experiment, we will provide you with a **starting capital of 3000 francs**. Throughout the experiment, you can see the current round number and your total earnings at the top of your screen. This brings you to the end of these instructions. When you think that you understood everything, please click the 'Ready' button on your computer screen. This will let us know that you are ready. When you are finished, you might have to wait a while until all others are ready. Please wait silently and patiently until we continue with the experiment.

A3.2 Instructions to the coders of subjects' self-reports in HBASE

Welcome! You are about to participate in a content analysis exercise. In this exercise you can earn substantial amount of money, thus read carefully these guidelines for your task. During the exercise you are not allowed to speak with other coders or communicate in any other way. If you want to ask a question, please raise your hand. There will be four similar tasks in today's exercise. You can earn money in each of these tasks. You will be paid the money you earn in all tasks in private at the end of all parts in today's exercise. You will receive the relevant guidelines separately at the start of each task. Below are guidelines for Task 1.

Guidelines: Task 1
The guidelines for Task 1 are simple. You will be asked to classify into categories some written text we received from 54 subjects – these are 54 people who took part in a laboratory experiment earlier this year. The amount of money you will earn in this task depends on the quality of your classification. The experiment was about repeated decision making in groups and took place in a computer laboratory at the University of Amsterdam earlier this year. The subjects first took part in the experiment and then we asked each one of them to write down the strategy she/he used in the experiment. We would like you to classify each of these strategy descriptions into one category. You will find the list and descriptions of 6 categories below. For each subject you have to classify their written description into exactly one category. If you do not classify a subject's description or if you classify it into more than one category, we will count your classification for this subject as invalid. To understand the strategy descriptions and the nature of the experiment we will first ask you to read the experimental instructions that our subjects received at the start of their experiment. Then you will read our descriptions of the 6 categories. Afterwards you will have to classify 10 imaginary descriptions of strategies, which we prepared ourselves, to check that you have understood your task. If you classify all our 10 imaginary descriptions correctly, you will start with the actual task of classification of our subjects' real written descriptions. On the next pages you can read the instructions for the original experiment with our 54 subjects. Please read the instructions carefully. It is very important that you understand the logic and dynamics of the original experiment, as only then will you be able to appropriately classify the subjects' strategy descriptions.

---[Here instructions for HBASE]---

Strategy categories
Our theoretical analysis suggests four viable strategies in this experiment: Selfish, Generous, Experience, Appearance. You will find the description and examples for each strategy below. We would like you to estimate whether our subjects describe any of these 4 strategies in their written descriptions. We also add two additional categories, Other and None, for cases when descriptions cannot be classified into one of the above 4 strategies. We will therefore ask that you classify each of the 54 individual descriptions into 4+2 = 6 categories: Selfish, Generous, Experience, Appearance, Other, None. Here are the descriptions and an illustration examples for each category:
(S) Selfish strategy predominantly and indiscriminately chooses **green** option and does not send points to receivers.

- For example, one written description corresponding to Selfish would be the following: "*I have chosen green most of the time.*"

(G) Generous strategy predominantly and indiscriminately chooses **blue** option and does send points to receivers.

- For example, one written description corresponding to Generous would be the following: "*I have selected blue most of the time.*"
- or another would be "*I tried to be generous.*"

(E) Experience strategy discriminately chooses **green** or **blue**, depending on its own experience with senders. This strategy tries to behave similar to other senders it had met in the past. In particular, the

Experience strategy is more likely to choose blue option if its past senders (when this player was a receiver) chose the blue option.

- For example, one written description corresponding to Experience would be the following: "*I played green except when I saw many nice senders.*"
- Another one would be: "*Blue increases group points so I played blue in early rounds to motivate the others but I switched to green because they didn't follow my choices.*"

(A) Appearance strategy prefers green choices but cares that its review list (three recent own choices) contains some blue choices. It does not react to its past experience with other senders, however. Instead, an Appearance strategy conditions its choices mostly on its own recent decisions and is more likely to choose blue when its review list contains many green choices. In summary, this strategy switches between green and blue, in order to keep some blue in its own review list; but does not discriminate based on other subjects it meets.

- For example, one written description corresponding to Appearance strategy would be the following: "*I have tried to have at least one blue option in my last 3 choices.*"
- Another one would be: "*I switched between green and blue.*"

(O) Other category contains feasible descriptions of strategies that cannot be classified into one and only one of the above four categories, but still contains enough instruction to replicate most of its play.

- For example, one written description corresponding to Other would be the following: "*My decisions were random.*"
- Another one would be: "*I sometimes opt for blue, and sometimes for green.*",
- or: "*I have chosen blue option in the first 30 rounds, and green option in the remaining 70 rounds.*"

(N) None category contains statements that do not describe a feasible strategy or describes just a small part of it.

- For example, the description "*I started with blue but everyone was choosing green.*" or "*I am too tired.*" or "*I do not want to respond.*" all fall into this category.

Note that these strategy categories need not be uniformly distributed in subjects' descriptions. Some may appear more often than the others, and some may never be described by the subjects.

Your task

At the end of the experiment, each of the 54 subjects who participated in the original experiment, had to briefly describe the strategy they used in the experiment. Your task is to classify each individual description in *exactly one* category, using the 6 categories above. If you classify a subject into no category or more than one category, you will not earn any money from this subject classification. You will receive a list of all 54 written descriptions in a table, one description per subject. For each description please enter the letter corresponding to the most suitable strategy category in the field next to the description in the table – one letter per subject.

Your earnings

Your earnings in this task depend on how well your classification will match the classifications of the other two coders today. In particular, you will receive 20 cents for your classification of one subjects' written description if it matches the classification by at least one other coder here today. That is, for each subject you will earn 20 cents if you classify their description into the same category as another coder. You will earn nothing for invalid individual classifications. This means that if you match other coders in your classifications for all 54 subjects, you will earn 10.8 € in this task.

End of the guidelines for Task 1

Please raise your hand when you have read and understood the experiment and the guidelines for coders. We will then ask you to classify 10 imaginary strategy descriptions that we have written ourselves to test whether you can correctly classify different strategy descriptions. If you classify all our 10 imaginary descriptions correctly, you will start with the actual task of classification of our subjects' real written descriptions.

A3.3 Instructions to the coders of subjects' self-reports in HREP

The guidelines for Task 3 are similar to that of Task 1, but this time you will be asked to classify into categories the strategy description we received from a different group of 54 subjects. These 54 people took part in a different, second laboratory experiment, also earlier this year. The amount of money you will earn in this task again depends on the quality of your classification. The second experiment also took place in a computer laboratory at the University of Amsterdam earlier this year and was related to the experiment you saw in Task 1. The main difference was that in the second experiment every sender would know the last three colors their receiver had chosen when in the role of a sender. **Each sender was therefore told, before making any choice, about the three recent color choices of their receiver.** A sender could thus learn something about the past behavior of their current receiver and use this information in their strategy. Our theoretical work suggests there are more plausible behavioral strategies for this experiment. You will see the new list of strategies and complete experimental instructions for the second experiment below. The subjects first took part in the experiment and then we asked each one of them to write down the strategy she/he used in the experiment. We would like you to classify each of these strategy descriptions into one of 9 categories described below. To understand the strategy descriptions and the nature of the experiment we will next ask you to read the experimental instructions for the second experiment. Most of the text is identical to the instructions for the first experiment and we show the identical parts in grey color for your convenience. The part of the instructions which differs between the first and second experiments are shown in normal colors. Our subjects had read these instructions, all in the normal colors, at the start of their experiment. Afterwards you will read our descriptions of the new categories and classify 15 imaginary descriptions of strategies, which we prepared ourselves, to check that you have understood your task.

---*[Here instructions for HREP]*---

Strategy categories

Our theoretical analysis suggests seven viable strategies in the second experiment: Selfish, Generous, Experience, Pretender, Rewarder, Benevolent, Cautious. You will find the description and examples for each strategy below. The first three strategies are similar to those from the first experiment because they do not consider the receivers' past choices (which are shown to the senders in the second experiment). The Pretender strategy is related to the Appearance strategy, but its appeal is different in the second experiment. Namely, senders can see the last three choices of the receivers, so a Pretender will want to show some blue choices in order to appear nice when it is a receiver. We would like you to estimate whether our subjects describe any of these 7 strategies in their written descriptions. We also add two additional categories, Other and None, for cases when descriptions cannot be classified into one of the above 7 strategies. We will therefore ask that you classify each of the 54 individual descriptions into 7+2 = 9 categories: Selfish, Generous, Experience, Pretender, Rewarder, Benevolent, Cautious, Other, None. Here are the descriptions and an illustration examples for each category:

(S) Selfish strategy predominantly and indiscriminately chooses **green** option and does not send points to receivers. It does not consider the information about its receivers.

- For example, one written description corresponding to Selfish would be the following: "*I have chosen green most of the time.*"

(G) Generous strategy predominantly and indiscriminately chooses **blue** option and does send points to receivers. It does not consider the information about its receivers.

- For example, one written description corresponding to Generous would be the following: "*I have selected blue most of the time.*"
- or another would be "*I tried to be generous.*"

(E) Experience strategy discriminately chooses **green** or **blue**, depending on its own experience with senders. This strategy tries to behave similar to other senders it had met in the past. In particular, the Experience strategy is more likely to choose blue option if its past senders (when this player was a receiver) chose the blue option. It does not consider the information about its receivers.

- For example, one written description corresponding to Experience would be the following: "*I played green except when I saw many nice senders.*"

(P) Pretender strategy prefers green choices but cares that its review list (three own recent choices) contains some blue choices. Similarly to Appearance strategy, Pretender is selfish but wants to appear generous, hoping that other senders will then choose blue more often in return. Pretender does not react to its past experience and, crucially, does not consider the information about its receivers. Instead, Pretender conditions its choices only on its own recent decisions and is more likely to choose blue when its own review list shows too many previous green choices. In summary, this strategy switches between green and blue, in order to keep some blue in its own review list, irrespective of what it gets from other senders or learns about its receivers.

- For example, one written description corresponding to Pretender strategy would be the following: "*I have tried to have at least one blue in my last 3 choices so that senders would send me blue.*"
- Another one would be: "*I switched between colours.*"

(R) Rewarder strategy considers the information about past behaviour of its receivers and rewards those who chose blue options in the past. It therefore conditions its choices only on its receiver's past behaviour. It prefers the blue option but will choose green to punish receivers with poor reputation. In particular, Rewarder is more likely to choose blue if its receiver has chosen some blue in the past.

- For example, one written description corresponding to Rewarder would be the following: "*I have chosen blue option if receiver has at least one blue option in his history.*"
- Another one would be: "*I punished only my partners who have chosen a lot of green.*"

(B) Benevolent strategy considers the information about the past behaviour of receivers it meets, but also its own past choices. It will punish selfish behaviour, except when this would hurt its own reputation. It will therefore reward nice receivers, but will sometimes also choose blue to avoid having many greens in its own review list. In particular, a Benevolent strategy will choose blue when its receiver has chosen some blue in the past, or when its own review list shows too many previous green choices.

- For example, one written description corresponding to Benevolent would be the following: "*I have chosen blue for receivers that had at least one blue option in their history. I also tried to keep at least one blue in my history.*"
- Another would be: "*I decided to have one blue option in my history, but I always gave money to generous receivers.*"
- or: "*I sometimes punished those that are selfish, but made sure I always appeared nice.*"

(C) Cautious strategy also considers both the information about the past behaviour of its receivers and its own past choices. It will mostly be selfish, but will reward a nice receiver when its own reputation is poor. It will therefore occasionally reward those with a good reputation. In particular, a Cautious strategy will only choose blue when its receiver has chosen some blue in the past and its own review list shows too many previous green choices.

- For example, one written description corresponding to Cautious would be the following: "*I have always selected green because receivers had many green options in their history. Only sometimes I would choose blue and only for those that had shown some blue.*"
- Another one would be: "*I have never chosen blue option if my history already contained two blue options or if partner's history contained at least two green options.*"

(O) Other category contains feasible descriptions of strategies that cannot be classified into one and only one of the above seven categories, but still contains enough instruction to replicate most of its play.

- For example, one written description corresponding to Other would be the following: "*My decisions were random.*"
- Another one would be: "*I gave green if others had three blue or three green choices in their summary, but otherwise blue.*",
- or: "*I have chosen blue option in the first 30 rounds, and green option in the remaining 70 rounds.*"

(N) None category contains statements that do not describe a feasible strategy or describes just a small part of it.

- For example, the description "*I started with blue but everyone was choosing green.*" or "*I dislike people who are selfish.*" or "*I would punish sometimes.*" all fall into this category.

Note that these strategy categories need not be uniformly distributed in subjects' descriptions. Some may appear more often than the others, and some may never be described by the subjects.

Your task

At the end of the experiment, each of the 54 subjects who participated in the original experiment, had to briefly describe the strategy they used in the experiment. Your task is to classify each individual description in *exactly one* category, using the 9 categories above. If you classify a subject into no category or more than one category, you will not earn any money from this subject classification. You will receive a new list of all 54 written descriptions in a table. For each description please enter the letter corresponding to the most suitable strategy category in the field next to the description in the table.

Your earnings

Again, you will receive 20 cents for each individual strategy classification in which you match at least one other coder. These earnings will be added to those you have earned in Tasks 1-2.

End of the guidelines for Task 3

Please raise your hand when you have read and understood the experiment and the guidelines for coders. We will then ask you to classify 15 imaginary strategy descriptions that we have written ourselves to test whether you can correctly classify the new strategy descriptions. If you classify all our 15 imaginary descriptions correctly, you will start with the actual task of classification of our subjects' real written descriptions.

A3.4 Instructions for the original deception game experiment

You are about to participate in a decision making experiment. The instructions are simple. If you follow them carefully, you may make a considerable amount of money.
Your earnings will be paid to you privately in cash at the end of today's session. In the experiment, earnings are denoted in 'francs'. At the end of the experiment, francs will be exchanged into Euro. The exchange rate will be **1 Euro for 250 francs**. In other words, for every 1000 francs, you will receive 4 Euro.
Your decisions are anonymous. They will not be attached to your name in any way. You are not allowed to speak with other participants or communicate in any other way. If you want to ask a question, please raise your hand.

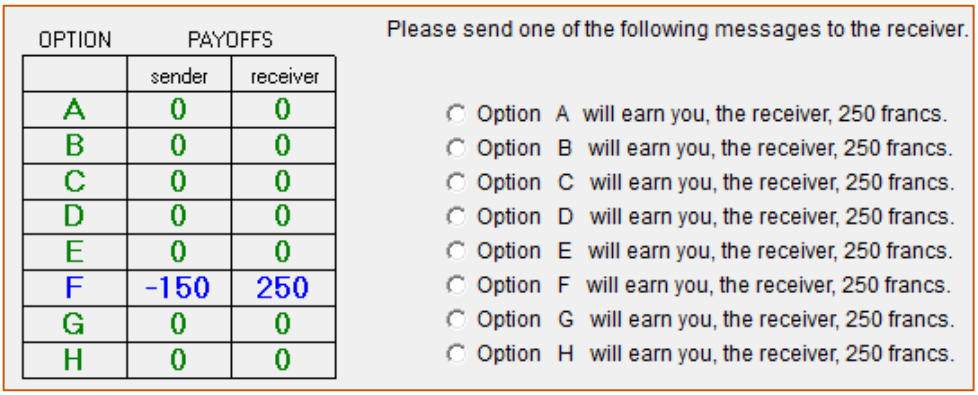
Rounds and Pairs
This experiment consists of 100 rounds.
At the beginning of every round the participants will be randomly divided into pairs. In each round new pairs will be made. You are equally likely to form a pair with any other participant. However, the probability that you will form a pair with the same participant in two consecutive rounds is small.
The two matched participants in a pair will have different roles. One will be a 'sender', and the other will be a 'receiver'. Which role you have in your pair will be determined randomly at the start of every round.

Options
In every round there will be eight available options to choose from: **A , B , C , D , E , F , G ,** and **H**. One of them will be the '**blue**' option. The other seven options will be '**green**' options.

- If the **blue** option is chosen, then the receiver will **earn 250 francs** and the sender will **lose 150 francs**.
- If a **green** option is chosen, then **neither participant in the pair will gain or lose** money.

At the start of every round the computer will randomly select which of the eight options is **blue**. The sender will see what the computer has selected, but the receiver will not see this. The receiver will not know which option is blue and which options are green.
The sender will send a message to the receiver and the receiver will then choose an option. The receiver's choice will determine the earnings for both paired participants in that round.
Sender's message
If you are the sender, you will see on your screen a table showing the color and earnings for every option. You will then recommend one of the options A-H to the receiver. There will be eight available messages. One message will recommend the **blue** option and seven will recommend different **green** options. Each message will state that the recommended option will earn the decider 250 points. You will be able to send any, and only one, of these eight messages.

> *Example: Below you can see an arbitrary example table with colors and earnings for the 8 options, shown on the sender's computer screen. In this example the computer selected option F as the **blue** option. All the other options are **green**. Only the sender can see this table. The sender can choose any, and just one, of the 8 available messages, and send it to the receiver.*



--- [*The following section shown only in DREP treatment*] ---

Sender's information about the receiver

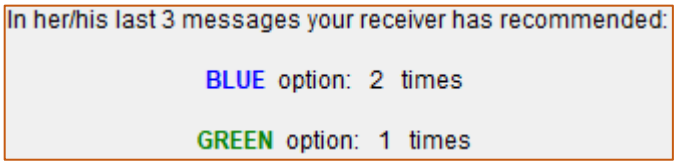
At the start of every round and before making a choice, the sender will receive information about her/his current receiver.

If you are the sender, you will see a summary of the **3 most recent** messages that your receiver has sent in the past. The information will show how often your receiver recommended the **blue** or the **green** option in the 3 most recent rounds when she/he was in the role of the sender.

In early rounds of the experiment, the receiver may not have sent 3 messages yet. In that case the information will describe all (that is, 0, 1, or 2) of her/his previous messages.

Only the sender will get information about the receiver. The receiver will not get any information about the sender.

> *Example: Below you can see an arbitrary example information that the sender may have about the receiver. In this example you can see that the receiver has recommended the **blue** option twice and a **green** option once, in the 3 most recent rounds when she/he was a sender.*



--[]--

Receiver's choice

If you are the receiver, you will see on your screen a table showing options A-H, but the table will not show the earnings or colors of the options. You will first receive one message from the sender, and then you will choose one of the options A-H. You will be able to choose any, and only one, of the eight options. Your choice will determine your own earnings and the earnings of your receiver in that round.

> *Example: Below you can see an arbitrary example table with the 8 options, shown on the receiver's computer screen. The receiver does not know which is the **blue** option. The receiver can choose any, and just one, of the 8 available options.*

OPTION	PAYOFFS		Please choose one option below
	sender	receiver	
A	?	?	choose option A
B	?	?	choose option B
C	?	?	choose option C
D	?	?	choose option D
E	?	?	choose option E
F	?	?	choose option F
G	?	?	choose option G
H	?	?	choose option H

End of the round

A round will end after the receiver has made a choice. You will then be informed about the **blue** option, the recommended option, and the chosen option. You will also see the resulting earnings.

At the bottom of the screen you will see a review of your own last 3 messages. It will show the colors of the three most recent options you have yourself recommended - in the last 3 messages you have sent as a sender.

> *Example: Below you can see an arbitrary example review of your own past messages. Your most recently recommended option color is shown on the left. In this example you would have last recommended a **green** option, before that you would have recommended the **blue** option, and before that you would have recommended the **blue** option.*

You have recently recommended option colors:

GREEN BLUE BLUE

After you press button “continue”, the next round will begin: you will be rematched into a new pair with a new participant, you will be randomly assigned a new role, and the computer will randomly select a new **blue** option (any option A-H) for your new pair.

Final instructions

At the start of the experiment, we will provide you with a **starting capital of 3000 francs**. Throughout the experiment, you can see the current round number and your total earnings at the top of your screen. This brings you to the end of these instructions. When you think that you understood everything, please click the 'Ready' button on your computer screen. This will let us know that you are ready. When you are finished, you might have to wait a while until all others are ready. Please wait silently and patiently until we continue with the experiment.

A4 *Substrategies in the deception game experiment*

A4.1 Substrategies of sending and responding strategies

In the main text we use the terms good and bad reputation to motivate our strategies and facilitate their descriptions. In our experiment, however, these words were not used to avoid the framing effect. As in our helping game, senders were given information about the last three honest/deceptive messages that their receivers sent, so that reputation was manifested through the number of visible honest messages. This yields several possible norms for what constitutes good reputation, from the strictest norm where all visible past messages must be honest, to the weakest norm where at least one must be honest. Consequently, a reputation-based conditional (sending) strategy can be defined in different ways for different norms, analogously to how we defined them in our helping game. For example, a rewarder strategy covers three distinct deterministic substrategies: REW1, REW2 and REW3. A rewarder using substrategy REWK (where *K* is a variable) is honest only if her receiver sent at least *K* honest messages in the recent three opportunities. The complete list of substrategies can be found in Tables A2 and A3.

Now we turn to (sub)strategies that condition on experience or memory. In principle, there are many different potential strategies, not only in terms of information on which they condition but also in terms of the scope of recall (i.e., memory length and how the past is discounted). Regarding the scope of recall, we consider only the type that has proven to be representative in our helping game. In particular, in the analysis of our helping game we have seen that it is sufficient to include just the substrategies of EXP strategy into the mixture model-based estimation to capture the experiential subjects. This facilitates and simplifies the subsequent analysis of strategies in our repeated deception game. We justified our decision by adding other experiential strategies to our strategy set in DBASE and found no effect on the final DBASE strategy classification, confirming that the *EXP*-like strategies are flexible enough to detect any experiential behavior.

In the following paragraphs we formally present experiential (sub)strategies considered in our deception game treatments. All these (sub)strategies consider all received experiences, with weights depending on their recency via hyperbolic discounting of the past. We first present the (sub)strategies of senders. The *EXP* strategy is analogous to that considered in our helping game. *EXP* constructs an index based on the received honesty and dictates honesty when this index is sufficiently high. As in Chapter 2, index h_P for *EXP* incorporates memory decay and its value in round $t \geq 2$ is calculated as follows:

$$h_P = \sum_{r=1}^{t-1} \frac{H(r)R(r)}{t-r} \Big/ \sum_{r=1}^{t-1} \frac{R(r)}{t-r},$$

where *R* and *H* are index functions: $R(r) = 1$ if an individual was receiver and $H(r) = 1$ if he received the honest message in round $r < t$. We again consider three deterministic substrategies

EXPK which dictate honesty if $h_P \geq K/4$, and one stochastic substrategy *EXPS* constructed as a simple linear combination (average) over all corresponding deterministic *EXPK* substrategies. Namely, *EXPS* dictates honesty with probability *x/3* if $h_P \in [x/4, (x+1)/4)$, for $x \in \{0, 1, 2\}$, and with probability 1, for $x \geq 3/4$.

The benevolent strategy *BEN* constructs an index based on the received trust and dictates honesty when this index is sufficiently high. Index h_B for *BEN* also incorporates memory decay and its value in round $t \geq 2$ is calculated similarly as the value of h_P:

$$h_B = \sum_{r=1}^{t-1} \frac{T(r)S(r)}{t-r} / \sum_{r=1}^{t-1} \frac{S(r)}{t-r},$$

where *S* and *T* are index functions: $S(r) = 1$ if an individual was sender and $T(r) = 1$ if her receiver trusted her in round $r < t$. We consider three deterministic substrategies *BENK* which dictate honesty if $h_B \geq K/4$, and one stochastic substrategy *BENS* constructed as a linear combination over all corresponding deterministic *BENK* substrategies. Namely, *BENS* dictates honesty with probability *x/3* if $h_B \in [x/4, (x+1)/4)$, for $x \in \{0, 1, 2\}$, and with probability 1, for $x \geq 3/4$.

The manipulative strategy *MAN* is exactly the opposite of *BEN* as it dictates honesty when the index h_B is sufficiently low and dictates deception when it is sufficiently high. Deterministic substrategies *MANK* dictate deception if $h_B \geq K/4$, while a stochastic substrategy *MANS* dictates deception with probability *x/3* if $h_B \in [x/4, (x+1)/4)$, for $x \in \{0, 1, 2\}$, and with probability 1, for $x \geq 3/4$.

Now we turn to (sub)strategies of receivers. The reactive strategy *REA* uses the same information as *EXP* and dictates trust when index h_P is sufficiently high. Deterministic substrategies *REAK* dictate trust if $h_P \geq K/4$, while a stochastic substrategy *REAS* dictates trust with probability *x/3* if $h_P \in [x/4, (x+1)/4)$, for $x \in \{0, 1, 2\}$, and with probability 1, for $x \geq 3/4$. The conformist strategy *CON* uses the same information as *BEN* and dictates trust when index h_B is sufficiently high. Deterministic substrategies *CONK* dictate trust if $h_B \geq K/4$, while a stochastic substrategy *CONS* dictates trust with probability *x/3* if $h_B \in [x/4, (x+1)/4)$, for $x \in \{0, 1, 2\}$, and with probability 1, for $x \geq 3/4$.

Finally, the projection strategy *PRO* constructs an index based on sender's own previous honesty and dictates trust the sender when this index is sufficiently high. Index h_O for *PRO* in round $t \geq 2$ is given by

$$h_O = \sum_{r=1}^{t-1} \frac{H(r)S(r)}{t-r} / \sum_{r=1}^{t-1} \frac{S(r)}{t-r},$$

where *S* and *H* are index functions: $S(r) = 1$ if an individual was sender and $H(r) = 1$ if she was honest in round $r < t$. Deterministic substrategies *PROK* dictate trust if $h_O \geq K/4$, while

stochastic substrategy *PROS* dictates trusts with probability *x/3* if $h_O \in [x/4, (x+1)/4)$, for $x \in \{0, 1, 2\}$, and with probability 1, for $x \geq 3/4$.

As a final remark, all our three indices (h_P, h_B *and* h_O) are defined for rounds $t \geq 2$ and we have division by zero (i.e., 0/0) in h_P (h_B and h_O) if a subject has not yet been a receiver (sender) in any round so far. For these cases we follow the same indexation procedure as in Chapter 2, i.e., we assign the value -1.

A4.2 Complete list of substrategies in the deception game experiment

Strategy Category	Substrategy	Description
honest	HON	always honest
deceptive	DEC	never honest
rewarder	REW1	honest only if receiver was honest at least once in the recent three rounds
	REW2	honest only if receiver was honest at least twice in the recent three rounds
	REW3	honest only if receiver was always honest in the recent three rounds
	SREW	honest with probability *H/3* if receiver was honest *H* times in the recent three rounds, where $H \in \{0, 1, 2, 3\}$
cautious	CAU0	honest only if sender was never honest in the recent two rounds
	CAU1	honest only if sender was honest at most once in the recent two rounds
	SCAU	honest with probability *1-H/2* if sender was honest *H* times in the recent two rounds, where $H \in \{0, 1, 2\}$
cautious rewarder	CR10	honest whenever REW1 or CAU0 was honest
	CR11	honest whenever REW1 or CAU1 was honest
	CR20	honest whenever REW2 or CAU0 was honest
	CR21	honest whenever REW2 or CAU1 was honest
	CR30	honest whenever REW3 or CAU0 was honest
	CR31	honest whenever REW3 or CAU1 was honest
	SCR	honest with probability *X/6* where $X \in \{0, 2, 3, 4, 5, 6\}$ is the number of deterministic CR strategies that were honest at a particular input (sender (own) # honesty, receiver # honesty)
mild deceptive	MD10	honest whenever REW1 and CAU0 were honest
	MD11	honest whenever REW1 and CAU1 were honest
	MD20	honest whenever REW2 and CAU0 were honest
	MD21	honest whenever REW2 and CAU1 were honest
	MD30	honest whenever REW3 and CAU0 were honest
	MD31	honest whenever REW3 and CAU1 were honest
	SMD	honest with probability *X/6* where $X \in \{0, 1, 2, 3, 4, 6\}$ is the number of deterministic MD strategies that were honest at a particular input (sender (own) # honesty, receiver # honesty)
random	RAND	in each round honest with probability 1/2
Nash	NASH	in each round honest with probability 1/8
experiential	EXP1	honest only if $h_P \geq 1/4$
	EXP2	honest only if $h_P \geq 1/2$
	EXP3	honest only if $h_P \geq 3/4$
	EXPS	honest with probability x/3 if $h_P \in [x/4, (x+1)/4)$, for $x \in \{0, 1, 2\}$, and with probability 1, for $x \geq 3/4$.
manipulative	MAN1	deceive only if $h_B \geq 1/4$
	MAN2	deceive only if $h_B \geq 1/2$
	MAN3	deceive only if $h_B \geq 3/4$
	MANS	deceive with probability x/3 if $h_B \in [x/4, (x+1)/4)$, for $x \in \{0, 1, 2\}$, and with probability 1, for $x \geq 3/4$.
benevolent	BEN1	honest only if $h_B \geq 1/4$
	BEN2	honest only if $h_B \geq 1/2$
	BEN3	honest only if $h_B \geq 3/4$
	BENS	honest with probability x/3 if $h_B \in [x/4, (x+1)/4)$, for $x \in \{0, 1, 2\}$, and with probability 1, for $x \geq 3/4$.

Table A2: The complete list of sending substrategies in the deception game.

Strategy Category	**Substrategy**	**Description**
trustful	TRU	always trust
sceptic	SCE	never trust
reactive	REA1	trust only if $h_P \geq 1/4$
	REA2	trust only if $h_P \geq 1/2$
	REA3	trust only if $h_P \geq 3/4$
	REAS	trust with probability x/3 if $h_P \in [x/4, (x+1)/4)$, for $x \in \{0, 1, 2\}$, and with probability 1, for $x \geq 3/4$.
conformist	CON1	trust only if $h_B \geq 1/4$
	CON2	trust only if $h_B \geq 1/2$
	CON3	trust only if $h_B \geq 3/4$
	CONS	trust with probability x/3 if $h_B \in [x/4, (x+1)/4)$, for $x \in \{0, 1, 2\}$, and with probability 1, for $x \geq 3/4$.
projection	PRO1	trust only if $h_O \geq 1/4$
	PRO2	trust only if $h_O \geq 1/2$
	PRO3	trust only if $h_O \geq 3/4$
	PROS	trust with probability x/3 if $h_O \in [x/4, (x+1)/4)$, for $x \in \{0, 1, 2\}$, and with probability 1, for $x \geq 3/4$.
Nash	NASH	in each round honest with probability 1/8
random	RAND	in each round honest with probability 1/2

Table A3: The complete list of responding substrategies in the deception game.

APPENDIX B *Sociality measures: cross-national study*

B1 Additional distributional information

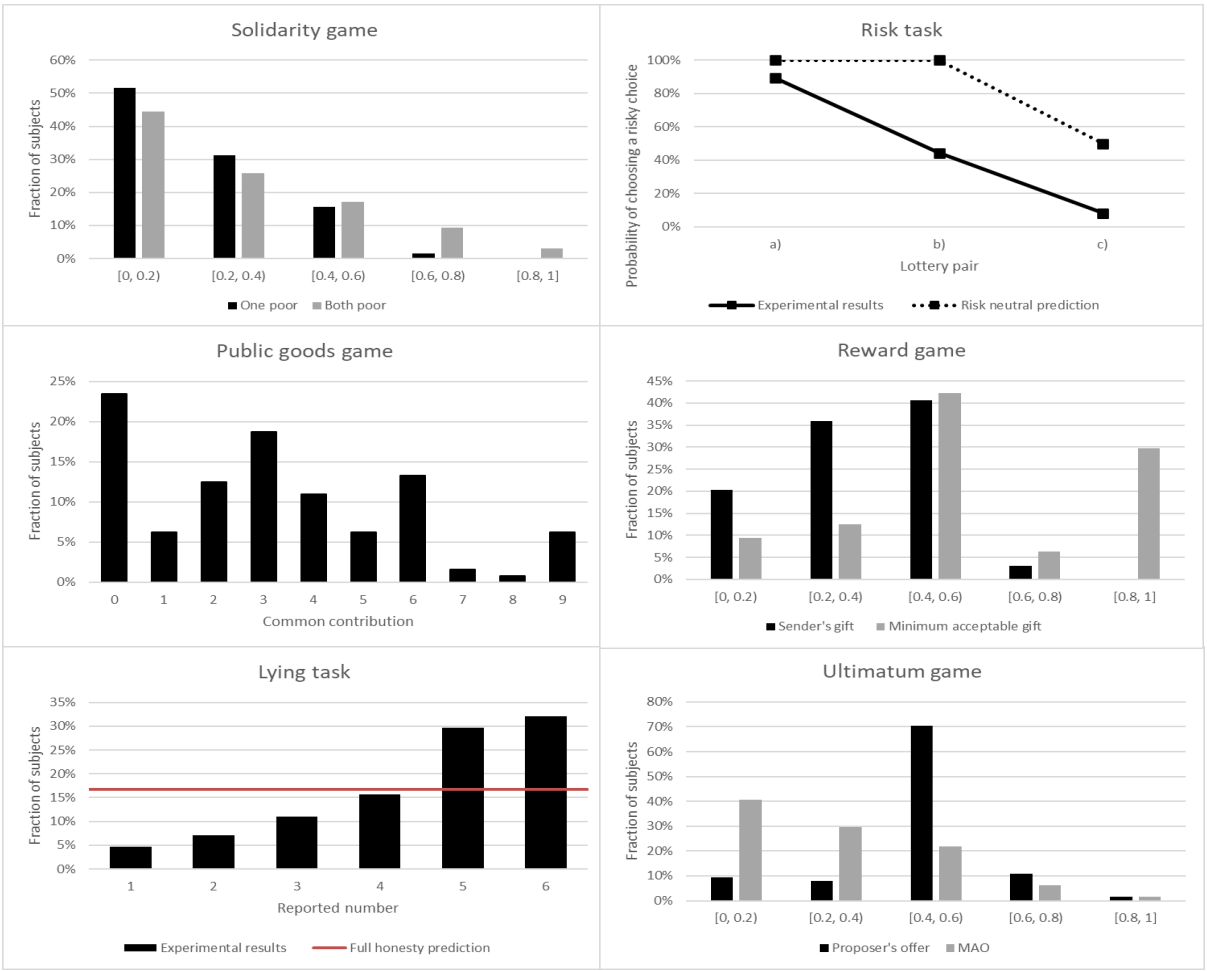


Figure B1: Additional distributional information.

B2 Robustness checks

B2.1 Tobit regressions

	Sol1	Sol2	PG	Die	Risk	Ult1	Ult2	Rew1	Rew2
INT	-0.01	0.09	-0.06	0.15	-0.03	0.02	-0.05	-0.09	0.23**
	(0.05)	(0.07)	(0.09)	(0.10)	(0.08)	(0.05)	(0.11)	(0.06)	(0.11)
NL	-0.17***	-0.20**	-0.18*	0.28**	-0.00	-0.10	-0.24*	-0.19**	0.13
	(0.06)	(0.08)	(0.11)	(0.12)	(0.09)	(0.06)	(0.13)	(0.07)	(0.13)
Econ	-0.03	-0.11	-0.12	-0.01	-0.03	-0.16**	0.04	0.00	-0.15
	(0.06)	(0.08)	(0.10)	(0.12)	(0.09)	(0.06)	(0.12)	(0.07)	(0.13)
Male	0.04	0.04	0.18**	0.07	0.05	0.02	0.25**	0.07	-0.10
	(0.05)	(0.07)	(0.09)	(0.10)	(0.08)	(0.05)	(0.10)	(0.06)	(0.11)
Male Econ	-0.12	-0.16	-0.33**	0.15	0.14	0.10	-0.22	-0.03	0.32*
	(0.07)	(0.10)	(0.14)	(0.15)	(0.12)	(0.07)	(0.17)	(0.08)	(0.19)
Exper	-0.12***	-0.16**	-0.01	0.05	0.06	-0.05	0.23**	-0.11**	0.00
	(0.05)	(0.06)	(0.08)	(0.09)	(0.07)	(0.04)	(0.11)	(0.05)	(0.11)
N	128	128	128	128	122	64	64	64	64

Source: own analysis.

Explanatory variables are dummy variables. Econ=1 for economics student; Male=1 for male; MaleEcon=1 for male economics student; NL=1 for Dutch cohort; INT=1 for international cohort (Slovenian cohort is a reference group); Exper=1 if a subject attended at least one experiment in the past.

All models are Tobit regressions. Standard errors are displayed in parentheses. Coefficient of the constant omitted for the brevity.

We also pooled the data from Sol1 and Sol2 (resulting in 2 observations per each individual) and ran random-effects Tobit regression. This yielded stronger results, as NL and Exper were statistically significant at the 0.01 level, and MaleEcon became marginally significant ($p<0.1$).

***, **, * indicate significance at the 0.01, 0.05 and 0.10 levels, respectively.

Table B1: Results of Tobit regressions (non-binary experimental tasks).

B2.2 Bayesian regressions

In Table B2 we do not display all the outputs from Bayesian regressions but only asterisks that indicate the strength of the association between each explanatory variable and each dependent variable, obtained from Bayesian regression analyses. "**" indicates strong association and "*" indicates a weaker association. ** (*) means that the 95% credible interval (75% credible interval) for the corresponding explanatory variable does not include value 0.

Bayesian regressions were performed in Program R using bayesreg package (R Core Team, 2019; Makalic & Schmidt, 2016). Table B2 shows that explanatory variables which have the strongest association with dependent variables are mostly those that were significant and marginally significant in Table 24, confirming our results in the main text.

	Sol1	Sol2	PG	Chic	Die	Risk	Tr1	Tr2	Ult1	Ult2	Rew1	Rew2
INT												
NL	**	*		*	*						*	
Econ	*	*	*						*			
Male			*									
Male Econ			*									
Exper	**	**									*	
N	128	128	128	128	128	122	64	64	64	64	64	64

Source: own analysis.

Explanatory variables are dummy variables. Econ=1 for economics student; Male=1 for male; MaleEcon=1 for male economics student; NL=1 for Dutch cohort; INT=1 for international cohort (Slovenian cohort is a reference group); Exper=1 if a subject attended at least one experiment in the past.

Models for binary variables Chic, Tr1 and Tr2 are Bayesian logit regressions, and other models are Bayesian linear regressions. ** (*) indicates that the 95% credible interval (75% credible interval) for the corresponding explanatory variable does not include 0.

Table B2: Association (based on ranks) between explanatory variables and dependent variables

B3 Additional regression models with CPI and GDP per capita

As mentioned in the main text, we considered four different models for each of 12 variables: one with CPI, one with GDPc, one with INT CPI, and one with INT GDPc. We did not include both CPI and GDPc or both interaction terms in the same model due to high correlation between them and high VIF (variance inflation factor). Due to high VIF we also did not include in the same model CPI and INT GDPc, and GDPc and INT CPI. Interaction terms between NL and CPI and NL and GDPc were not included due to perfect multicollinearity (they were directly derived from NL dummy variable by multiplying it by a constant). Since we kept INT in all models, CPI and INT CPI and GDPc and INT GDPc were also not included in the same model due to perfect multicollinearity. Below we provide the summary of our regressions that included GDPc, INT CPI and INT GDPc.

	Sol1	Sol2	PG	Chic	Die	Risk	Tr1	Tr2	Ult1	Ult2	Rew1	Rew2
INT	-0.02	0.41	-0.17	-0.50	0.47	-0.08	-1.21	0.19	0.07	-0.14	-0.35	0.67
	(0.24)	(0.28)	(0.31)	(0.65)	(0.33)	(0.27)	(0.84)	(0.75)	(0.23)	(0.43)	(0.27)	(0.42)
NL	-1.06**	-0.79	-0.78	2.04**	1.01*	-0.14	-2.55*	0.35	-0.45	-1.13	-1.39***	0.70
	(0.42)	(0.48)	(0.49)	(0.94)	(0.57)	(0.43)	(1.36)	(1.14)	(0.37)	(0.72)	(0.46)	(0.66)
Econ	-0.19	-0.60	-0.28	0.34	-0.01	-0.06	0.70	-0.70	-0.61*	0.16	0.37	-0.62
	(0.32)	(0.38)	(0.40)	(0.69)	(0.42)	(0.34)	(1.12)	(0.85)	(0.31)	(0.52)	(0.37)	(0.50)
Male	0.17	0.14	0.61**	-0.21	0.24	0.22	0.72	0.47	0.08	0.83*	0.28	-0.33
	(0.22)	(0.26)	(0.29)	(0.58)	(0.32)	(0.27)	(0.79)	(0.74)	(0.22)	(0.44)	(0.26)	(0.39)
Male	-0.40	-0.45	-1.00**	0.38	0.70	0.42	-0.04	-0.75	0.38	-0.58	-0.21	1.14
Econ	(0.41)	(0.47)	(0.50)	(0.83)	(0.55)	(0.41)	(1.13)	(1.18)	(0.32)	(0.69)	(0.38)	(0.70)
Exper	-0.58**	-0.60**	-0.11	0.55	0.15	0.24	-0.24	-0.87	-0.19	0.78*	-0.52**	0.18
	(0.23)	(0.27)	(0.29)	(0.53)	(0.32)	(0.26)	(0.68)	(0.71)	(0.20)	(0.46)	(0.23)	(0.40)
GDPc	0.0000	0.0000	0.0000	-0.0000	0.0000	0.0000	0.0001	-0.0000	0.0000	0.0000	0.00002**	-0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
N	128	128	128	128	128	122	64	64	64	64	64	64

Source: own analysis.

Explanatory variables (except GDPc) are dummy variables: Econ=1 for economics student; Male=1 for male; MaleEcon=1 for male economics student; NL=1 for Dutch cohort; INT=1 for international cohort (Slovenian cohort is a reference group); Exper=1 if a subject attended at least one experiment in the past. GDPc is GDP per capita. Models for binary variables Chic, Tr1 and Tr2 are logit regressions, and other models are fractional logit regressions. Standard errors in parentheses. Coefficient of the constant omitted for brevity.

The coefficients' signs indicate whether a particular explanatory variable has a positive (+) or negative (-) effect on the dependent variable, and stars ***, ** or * indicate whether this effect is significant at the 0.01, 0.05 or 0.1 level.

Table B3: Regression results for variables of interest from our eight experimental tasks. GDPc included as an external variable.

	Sol1	Sol2	PG	Chic	Die	Risk	Tr1	Tr2	Ult1	Ult2	Rew1	Rew2
INT	-0.50	0.11	-0.63	1.21	0.03	-0.08	-3.02*	0.02	0.16	-0.50	-1.32**	1.47*
	(0.51)	(0.59)	(0.65)	(1.15)	(0.69)	(0.54)	(1.73)	(1.44)	(0.48)	(0.85)	(0.59)	(0.83)
NL	-0.80**	-0.71*	-0.53	1.10*	1.09**	-0.02	-1.20	0.08	-0.39	-0.92*	-0.80**	0.32
	(0.33)	(0.38)	(0.37)	(0.62)	(0.45)	(0.32)	(0.95)	(0.84)	(0.27)	(0.55)	(0.34)	(0.48)
Econ	-0.20	-0.56	-0.30	0.38	0.06	-0.12	0.41	-0.56	-0.67**	0.14	0.30	-0.62
	(0.31)	(0.37)	(0.39)	(0.68)	(0.41)	(0.33)	(1.05)	(0.83)	(0.30)	(0.51)	(0.36)	(0.49)
Male	0.17	0.14	0.61**	-0.21	0.24	0.21	0.69	0.53	0.08	0.83*	0.27	-0.35
	(0.22)	(0.26)	(0.29)	(0.58)	(0.32)	(0.27)	(0.78)	(0.74)	(0.22)	(0.44)	(0.26)	(0.39)
Male Econ	-0.39	-0.45	-0.99**	0.37	0.70	0.43	0.05	-0.75	0.39	-0.57	-0.19	1.13
	(0.41)	(0.47)	(0.50)	(0.83)	(0.56)	(0.41)	(1.12)	(1.19)	(0.32)	(0.69)	(0.38)	(0.70)
Exper	-0.58**	-0.62**	-0.10	0.56	0.12	0.26	-0.21	-0.91	-0.18	0.79*	-0.51**	0.18
	(0.23)	(0.27)	(0.29)	(0.53)	(0.32)	(0.26)	(0.68)	(0.71)	(0.20)	(0.46)	(0.23)	(0.40)
INT CPI	0.01	0.01	0.01	-0.03	0.01	-0.0003	0.03	0.004	-0.002	0.01	0.02*	-0.01
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.03)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
N	128	128	128	128	128	122	64	64	64	64	64	64

Source: own analysis.

Explanatory variables (except INT CPI) are dummy variables: Econ=1 for economics student; Male=1 for male; MaleEcon=1 for male economics student; NL=1 for Dutch cohort; INT=1 for international cohort (Slovenian cohort is a reference group); Exper=1 if a subject attended at least one experiment in the past. INT CPI is the interaction term between INT and Corruption Perceptions Index. Models for binary variables Chic, Tr1 and Tr2 are logit regressions, and other models are fractional logit regressions. Standard errors in parentheses. Coefficient of the constant omitted for brevity.

The coefficients' signs indicate whether a particular explanatory variable has a positive (+) or negative (-) effect on the dependent variable, and stars ** or * indicate whether this effect is significant at the 0.05 or 0.1 level.

Table B4: Regression results for variables of interest from our eight experimental tasks. INT CPI included as an external variable.

	Sol1	Sol2	PG	Chic	Die	Risk	Tr1	Tr2	Ult1	Ult2	Rew1	Rew2
INT	-0.27	0.34	-0.41	0.41	0.39	-0.20	-2.43**	0.44	0.03	-0.34	-0.89**	1.05*
	(0.32)	(0.38)	(0.41)	(0.74)	(0.44)	(0.36)	(1.19)	(0.96)	(0.31)	(0.56)	(0.38)	(0.54)
NL	-0.81**	-0.71*	-0.53	1.10*	1.09**	-0.02	-1.29	0.09	-0.40	-0.92*	-0.83**	0.32
	(0.33)	(0.38)	(0.37)	(0.62)	(0.45)	(0.32)	(0.97)	(0.84)	(0.27)	(0.55)	(0.34)	(0.48)
Econ	-0.19	-0.60	-0.28	0.34	-0.01	-0.06	0.70	-0.70	-0.61*	0.16	0.37	-0.62
	(0.32)	(0.38)	(0.40)	(0.69)	(0.42)	(0.34)	(1.12)	(0.85)	(0.31)	(0.52)	(0.37)	(0.50)
Male	0.17	0.14	0.61**	-0.21	0.24	0.22	0.72	0.47	0.08	0.83*	0.28	-0.33
	(0.22)	(0.26)	(0.29)	(0.58)	(0.32)	(0.27)	(0.79)	(0.74)	(0.22)	(0.44)	(0.26)	(0.39)
Male	-0.40	-0.45	-1.00**	0.38	0.70	0.42	-0.04	-0.75	0.38	-0.58	-0.21	1.14
Econ	(0.41)	(0.47)	(0.50)	(0.83)	(0.55)	(0.41)	(1.13)	(1.18)	(0.32)	(0.69)	(0.38)	(0.70)
Exper	-0.58**	-0.60**	-0.11	0.55	0.15	0.24	-0.24	-0.87	-0.19	0.78*	-0.52**	0.18
	(0.23)	(0.27)	(0.29)	(0.53)	(0.32)	(0.26)	(0.68)	(0.71)	(0.20)	(0.46)	(0.23)	(0.40)
INT	0.0000	0.0000	0.0000	-0.0000	0.0000	0.0000	0.0001	-0.0000	0.0000	0.0000	0.00002**	-0.0000
GDPc	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
N	128	128	128	128	128	122	64	64	64	64	64	64

Source: own analysis.

Explanatory variables (except INT GDPc) are dummy variables: Econ=1 for economics student; Male=1 for male; MaleEcon=1 for male economics student; NL=1 for Dutch cohort; INT=1 for international cohort (Slovenian cohort is a reference group); Exper=1 if a subject attended at least one experiment in the past. INT GDPc is the interaction term between INT and GDP per capita. Models for binary variables Chic, Tr1 and Tr2 are logit regressions, and other models are fractional logit regressions. Standard errors in parentheses. Coefficient of the constant omitted for brevity.

The coefficients' signs indicate whether a particular explanatory variable has a positive (+) or negative (-) effect on the dependent variable, and stars ** or * indicate whether this effect is significant at the 0.01, 0.05 or 0.1 level.

Table B5: Regression results for variables of interest from our eight experimental tasks. INT GDPc included as an external variable.

B4 Instructions

Welcome

In a few minutes we will start the experiment. In the experiment you can earn money. The amount of money you will earn depends on your decisions and the decisions of the other participants. At the end of the experiment we will pay the earnings to each participant individually and in private. Your decisions are associated only with the number of your computer. Because they are not associated with your name, your decisions will remain anonymous.

The experiment consists of 8 parts. In every part of the experiment you will make one or more decisions. Different parts are not connected and your decisions in one part will not affect the rest of the experiment. In every part you can earn points. At the end of the experiment we will throw a 8-sided dice to randomly select one of the parts. Each participant will then be paid in money for the points she/he earned in this part. Each participant will earn 1 euro for every 10 points that she/he earned in the selected part of the experiment.

All participants will earn money from just one, the same part of the experiment. Any part of the experiment can be selected, and you should consider each part as if it is the only part that brings you money.

Your decisions will not be disclosed during the experiment to the other participants. Also, the decisions of the other participants will not be disclosed to you. At the conclusion of all 8 parts we will reveal only those decisions that will determine your earnings, but even then we will not reveal which of the participants has made those decisions.

At the beginning of each part of the experiment you will receive detailed instructions for that part. For simplicity we will use in the instructions the masculine form (“he”) when we refer to another person, male or female. In some parts of the experiment you will be paired with one other participant, in some parts you will be a member of a group of three participants, and in the remaining parts you take decisions on our own. After you complete the 8 parts you will get another separate task in which you will not be connected with other participants.

During the experiment you are not allowed to talk or communicate in any way with other participants. If you have any questions, please raise your hand and one of us will come to your desk. Do not ask your questions aloud.

When you finish reading this introduction, press on the "Next" button at the bottom of the screen. Then please wait until all other participants finish reading the introduction. Once everyone is ready, you will receive detailed instructions for the first part of the experiment.

1. SOLIDARITY

\Everyone:\
In this part of the experiment, you are in a group with two other randomly selected participants. Your group therefore has three members: you and two other participants. For each member of your group separately the computer will simulate a throw of a 6-sided dice. If the number thrown is 1 or 2, this member will receive 4 points. If the number thrown is 3, 4, 5 or 6, this member will receive 60 points.
In your group, each member with 60 points can donate some of his points to another member with only 4 points (if there is such a member in your group). Before the computer throws the dices, you must decide how many points you would donate if you had 60 points and one, or two, other members of your group had 4 points.
Please enter two numbers in the fields below.
In the upper field, enter the number of your points (0-60) that you would donate to a third member if you and the second member of your group both receive 60 points, but the third member receives 4 points. Also the second member can donate some of his points. If you donate X points and the second member donates Y points, then you would earn 60-X points, the second member would earn 60-Y points, and the third member would earn 4+X+Y points.
In the bottom field, enter the number of your points (0-30) that you would donate to each of the other two members of your group if you receive 60 points, but both other members receive 4 points. If you donate X points to each of them, then you would earn 60-2X points, and they would each earn 4+X points.
The other two members of your group will also now decide how to donate their points. To confirm your decision press the button "Confirm".

Decision:
Enter the number of points that you would donate to one other member with 4 points (0-60):
Enter the number of points that you would donate to each of the two other members with 4 points (0-30):

2. PG

\Everyone:\
In this part of the experiment, you are in a group with two other randomly selected participants. Your group therefore has three members: you and two other participants. Each member of your group has 9 tokens that he can allocate between two projects, named O and S. Project O is an own personal project of each member. Project S is a joint project of all three group members.
You must divide all 9 tokens between projects O and S. You can allocate into each project any integer number of tokens between 0 and 9, but the sum of both allocations must equal 9. If you allocate X tokens into your personal project O, then you must allocate 9-X tokens into your joint group project S. Each token allocated by any member of your group to his project O will bring 4 points only to this member himself. Each token allocated by any member of your group to the joint project S will bring 2 points to every member of your group.
A token that you allocate to O will therefore bring 4 points only to you. A token that you allocate to S will bring to you and to both other members of your group each 2 points.

Please decide how you want to allocate your tokens between projects O and S. The other two members of your group will also now allocate their tokens. Enter below the numbers of tokens that you will allocate into projects O and S. The sum of your two allocations must be equal to 9. To confirm your decision press the button "Confirm".

Decision:
Enter the number of tokens that you will allocate into project O (0-9):
Enter the number of tokens that you will allocate into project S (0-9):

3. TRUST

\Player A:\
In this part of the experiment, you are paired with one other randomly chosen participant, whom we will call your coparticipant. Your coparticipant has 0 points, and you have 40 points available. You can hold all 40 points, or you can transfer them to your coparticipant. If you transfer the points, they will be tripled, and your coparticipant will receive 120 points. Your coparticipant can then keep them all, or share them equally between both of you.
If you hold your points, then your earning is 40 points and your coparticipant earns 0 points.
If you transfer your points and your coparticipant keeps them, then your earning is 0 points and your coparticipant earns 120 points.
If you transfer your points and your coparticipant shares them, then your earning is 60 points and your coparticipant earns 60 points.
Please decide now whether you will hold or transfer your points. You make your decision by pressing either the “Transfer” or the “Hold” button below. To confirm your decision press the button "Confirm".
Decision:
(Transfer) (Hold)

\Player B:\
In this part of the experiment, you are paired with one other randomly chosen participant, whom we will call your coparticipant. Your coparticipant has 40 points available, and you have 0 points. Your coparticipant can hold all 40 points, or can transfer them to you. If he transfers the points, they will be tripled, and you will receive 120 points. You can then keep them all, or share them equally between both of you.
If your coparticipant holds the points, then you earn 0 points and the earning of your coparticipant is 40 points.
If your coparticipant transfers the points and you keep them, then you earn 120 points and the earning of your coparticipant is 0 points.
If your coparticipant transfers the points and you share them, then you earn 60 points and the earning of your coparticipant is 60 points.
Please decide now whether you will keep or transfer your points. You make your decision by pressing either the “Share” or the “Keep” button below. To confirm your decision press the button "Confirm".
Decision:
(Share) (Keep)

4. ULTIMATUM

\Player A:\
In this part of the experiment, you are paired with one other randomly chosen participant, whom we will call your coparticipant. Together you have 100 points available that can be divided between you two. You will propose a division and your coparticipant will accept or reject your proposal. If it is rejected then you will both receive no points.
Choose the number of points that you will offer to your coparticipant. Your coparticipant will in the meantime choose the minimum number of points he would accept. Then we will compare your offer and his number. Your offer will be rejected if it is below the number chosen by your coparticipant.
If your offer of P points is accepted, then your coparticipant will receive P points and you will receive the remaining 100-P points. If your offer is rejected, then your coparticipant will receive 0 points and you will receive 0 points.
Please enter below the number of points that you are offering to your coparticipant. To confirm your decision press the button "Confirm".
Decision:
Enter the number of points that you are offering to your coparticipant (0-100):

\Player B:\
In this part of the experiment, you are paired with one other randomly chosen participant, whom we will call your coparticipant. Together you have 100 points available that can be divided between you two. Your coparticipant will propose a division and you will accept or reject this proposal. If it is rejected then you will both receive no points.
Your coparticipant will choose the number of points that he will offer to you. You will in the meantime choose the minimum number of points you would accept. Then we will compare his offer and your number. The offer will be rejected if it is below your chosen number.
If the offer of P points is accepted, then you will receive P points and your coparticipant will receive the remaining 100-P points. If the offer is rejected, then you will receive 0 points and you coparticipant will receive 0 points.
Please enter below the lowest offer (number of points) that you would accept. Your coparticipant will not see your number. If you enter X then you will reject all offers below X. If you enter 0 then you will accept every offer. If you enter 101 then you will reject every offer. To confirm your decision press the button "Confirm".
Decision:
Enter the minimum number of points that you would accept (0-101):

5. CHICKEN

\Everyone:\
In this part of the experiment, you are paired with one other randomly chosen participant, whom we will call your coparticipant. Each have two possible moves: A and B. Each must choose one of these two options. You both choose your options simultaneously. The numbers of points earned depend on the options select by you and your coparticipant, as explained below:

If you select A and your coparticipant chooses B, you earn 70 points and your coparticipant earns 30 points.
If you select A and your coparticipant chooses A, you earn 0 points and your coparticipant earns 0 points.
If you select B and your coparticipant chooses B, you earn 40 points and your coparticipant earns 40 points.
If you select B and your coparticipant chooses B, you earn 30 points and your coparticipant earns 70 points.
Please choose option A or option B. You make your decision by pressing either the “A” or the “B” button below. Your coparticipant will also now choose one of these two options. To confirm your decision press the button "Confirm".
Decision:
(A) (B)

6. GIFT EXCHANGE

\Player A:\
In this part of the experiment, you are paired with one other randomly chosen participant, whom we will call your coparticipant. You have 90 points, and your coparticipant has 10 points. You can gift some of your points to your coparticipant, and your coparticipant can protect you or not protect you. If he protects you, then you keep all the points that you have not gifted. If he does not protect you, then you keep only one-third (1/3) of non-gifted points (so you keep only one out of every three points that you did not gift). Your participants pays 10 points for your protection.
Choose the number of points G that you will gift to your coparticipant. If your coparticipant protects you, he would earn G points and you would earn 90-G points. If your coparticipant does not protect you, he would earn G+10 points and you would earn (90-G)/3 points.
In the meantime, your coparticipant will enter the minimum number of points that he wishes to receive from you in order to protect you. We will then compare your gift and his number. If your gift is at least as large as the number entered by your coparticipant, you will be protected.
Please enter below the number of points that you will donate to your coparticipant. To confirm your decision press the button "Confirm".
Decision:
Enter the number of points you will donate (0-90):

\Player B:\
In this part of the experiment, you are paired with one other randomly chosen participant, whom we will call your coparticipant. You have 10 points, and your coparticipant has 90 points. Your coparticipant can gift some of his points to you, and you can protect or not protect him. If you protect him, then he keeps all the points that he did not gift. If you do not protect him, then he keeps only one-third (1/3) of non-gifted points (only one out of every three points that he did not gift). You pay 10 points to protect your coparticipant.
Choose the minimum number of points that you would like to receive from your coparticipant in order to protect him. In the meantime, your coparticipant will enter the number of points G that she will gift

to you. We will then compare his gift and your number. If his gift is at least as large as the number you have entered, then you will protect him.
If you protect your coparticipant, he would earn 90-G points and you would earn G points. If you do not protect your coparticipant, he would earn (90-G)/3 points and you would earn G+10 points.
Please enter below the smallest gift (number of points) that you wish to receive in order to pay 10 points for protection of your coparticipant. Your coparticipant will not see this number. If you enter X then you will protect him as long as he gifts you at least X points. If you enter 0 then you will always protect him. If you enter 91 then you will never protect him. To confirm your decision press the button "Confirm".
Decision:
Enter the minimum number of received points where you would protect your coparticipant (0-91):

7. LYING

\Everyone:\
In this part of the experiment, your earnings depend only on your throw of a dice. You will throw a dice and receive ten times as many points as the number shown on the dice.
You will receive a dice in a paper cup. Please cover the cup and the dice with your hand, shake it, and then put it covered on your desk so that the dice rotates a little under the cup. There is a small hole in the cup, through which you can look to see the number on the top of the dice. Enter this number below. You will earn 10 times as many points as the number that you have entered.
Throw the dice only once, and do not show your throw to anyone else. After you have entered the number, please shake the cup again, turn it around, and put the dice in the cup. To confirm your decision press the button "Confirm".
Decision:
Enter the number that you saw on top of the dice (0-6):

8. RISK

/Everyone:/
In this part of the experiment, your earnings will depend on your decisions and on chance. Below you can see 6 different options written in three rows. In each row you can choose between option E and option F.
In all three rows option E is the same: with 50% probability it brings you 80 points, and with 50% probability it brings you 20 points. With this option you therefore have the same chance to earn 80 or 20 points.
Option F brings you a fixed earning. The number of points you earn with this option is different in each row, however.
In each row you must choose your preferred option. In each row you must therefore choose either option E or option F. You therefore need to make three decisions, by pressing in each row either button E or button F.
Once you have confirmed all three decisions, the computer will randomly choose one of the three rows. If in this row you have chosen option E, then the computer will again randomly determine (with equal probability) whether you will receive 80 points or 20 points. If in this row you have chosen option F, then you will receive the number of points written next to the button F on this row.

Please choose now one of the two options in each row, by pressing one of two buttons in this row. When you choose one option in each row, confirm your decisions by pressing the "Confirm" button.
Decision:
50% for 80 points and 50% for 20 points (E), (F) 38 points
50% for 80 points and 50% for 20 points (E), (F) 44 points
50% for 80 points and 50% for 20 points (E), (F) 50 points