Three essays on indirect reciprocity

The importance of reputation

Andrea Marietta Leina

andrea.mariettaleina@gmail.com

Registration number: 100329119

A thesis presented for the degree of
Doctor of Philosophy

School of Economics and Centre for Behavioural and Experimental Social Science
University of East Anglia
United Kingdom

November 2024

Abstract

Cooperation among individuals is fundamental to societies and economies. This thesis comprises three essays exploring how indirect reciprocity can sustain cooperation in large populations with heterogeneous helping costs.

In the first essay, I generalise the Helping Game by introducing heterogeneity in individuals' helping costs. I investigate whether self-interested strategies can sustain indirect reciprocity in such a population. The findings suggest that, under certain parameters, cooperation can be an equilibrium even when helping costs vary.

The second essay experimentally compares two reputation-based mechanisms: "Image Scoring" (IS) and "Good Standing" (GS). IS was proposed by Nowak and Sigmund (1998) and tested by Engelmann and Fischbacher (2009); GS was introduced by Sugden (1986) and modified by Leimar and Hammerstein (2001). By adjusting participant information and introducing heterogeneous helping costs, I examine which mechanism more effectively fosters cooperation and reciprocity. The results indicate that while IS leads to higher overall cooperation, particularly in homogeneous cost settings, Hammerstein's version of GS better supports reciprocal helping by facilitating cooperative "clubs" among individuals with lower costs.

In the third essay, I experimentally test two new mechanisms: a binary version of IS to align structurally with GS and the Sugden's stricter version of GS. Additionally, I include a control condition without reputational information and elicit participants' beliefs using a novel, incentive-compatible method expressing beliefs as frequencies. The essay also contributes theoretically by developing an axiomatic framework for binary reputation mechanisms, identifying principles that effective mechanisms should satisfy. The findings suggest that Sugden's GS is most effective in sustaining reciprocal cooperation and that participants' beliefs significantly influence the functioning of reputation mechanisms.

Collectively, these essays advance our understanding of how indirect reciprocity and reputation mechanisms can sustain cooperation in heterogeneous populations. They highlight the importance of designing systems that account for variations in helping costs.

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisors: Professor Theodore T. Turocy, Dr. Amrish Patel, and Professor Robert Sugden. Ted, thank you for sharing your expertise in theoretical modelling, coding, and experimental design. Your guidance has been instrumental in shaping the practical aspects of my research. Amrish, I am extremely grateful for you teaching me how to organise my reasoning and approach problems systematically. Your insights have greatly enhanced the clarity and structure of my work. Bob, your generous allocation of time and exceptional mentoring have been truly inspiring. Your support and wisdom have profoundly influenced my development as a researcher. This dissertation marks only the beginning of my journey, and I know that your examples will continue to guide me in the years to come.

I am also indebted to everyone I met at various conferences. Your insightful feedback has significantly strengthened my critical thinking and inspired me to produce higher-quality papers.

To the examiners who will read this thesis, thank you for your time, patience, and thoughtful evaluation.

I extend my sincere appreciation to the School of Economics, the administrative staff, and the Centre for Behavioural and Experimental Social Science (CBESS). Your assistance and resources have been invaluable across my academic journey.

My heartfelt thanks go out to my friends in the Economics Department at Norwich: Aayushi, Amir, Alex, Deanna, Diego, Grace, Hector, Kensley, Israel, Nikita, Rui, and all those who have shared this route with me. You have been very important; completing this thesis would not have been possible without each and every one of you.

To my dear friends who have been in Norwich — Amy, Anna, Audun, Elien, and Marco — thank you for making my time here unforgettable.

To Maria and Riccardo — your unwavering support and encouragement from afar have been a constant source of strength. Despite the distance, you have remained close and integral to my life.

To my friends in Italy who have continued to lighten me up even from afar — Alessandro, Giulia, Pietro and Stefano — thank you for always being there.

I also wish to acknowledge my volunteering clubs, Norwich Soup Club and now Feed The People, which provided joy and meaningful experiences that kept me motivated and grounded.

To Victoria, my girlfriend: your help and boundless love have been my anchor and balance throughout this endeavour. Thank you for teaching me how to improve, for understanding my timing, and for standing by me in both big and small ways. Your constant presence has made all the difference. You are my everyday choice and a source of strength and inspiration every step of the way.

Finally — just in these lines and not in my heart — to my amazing parents and my sister, thank you for your endless love and to persistently pique my curiosity. Your belief in me has been my foundation. In loving memory of my grandmother, who passed away just a few months before this achievement became possible — you are always in my thoughts. This accomplishment is as much yours as it is mine.

Introduction

In a world where acts of generosity often defy the strict calculus of self-interest, one might wonder what compels individuals to help strangers without immediate benefit to themselves. This paradox lies at the heart of human society, where cooperation is both essential for collective well-being and a challenge to traditional notions of rational behaviour. From neighbours sharing resources to nations forging alliances, cooperative actions bind communities together. Yet, as societies become increasingly diverse, the mechanisms that sustain such cooperation amidst heterogeneity remain elusive.

The enigma of cooperation — why individuals choose to act benevolently towards others without direct personal gain — has long intrigued economists and social scientists. Traditional theories often emphasise direct reciprocity, where individuals are motivated to cooperate because they expect their actions to be reciprocated by the same individuals in future interactions. However, in large populations where interactions are infrequent or anonymous, direct reciprocity alone is insufficient to sustain cooperative behaviour. In such contexts, indirect reciprocity — where individuals base their cooperation on information about others — has been established as an effective mechanism for promoting and maintaining cooperation.

Real-world societies are far from uniform. Individuals differ not only in their preferences but also in their capacities and opportunities to contribute to the common good. While traditional theories provide explanations within homogeneous populations, they often fall short in accounting for the complexities of diverse societies. This thesis seeks to explore how cooperative behaviour can be sustained in such heterogeneous environments, examining the role of reputation-based mechanisms in fostering helping behaviours among individuals.

Central to this inquiry are questions about the effectiveness of reputation mechanisms in diverse societies. Are there mechanisms beyond those currently recognised in the literature that adequately encapsulate helpfulness? What principles define such mechanisms, and how do they function within the framework of repeated interactions?

Image Scoring and Good Standing are theoretically well-known mechanisms. Image Scoring (IS), first proposed by Nowak and Sigmund (1998) and experimentally tested by Engelmann and Fischbacher (2009), relies on first-order information about individuals' past actions. Good Standing (GS), introduced by Sugden (1986) and modified by Leimar and Hammerstein (2001), incorporates higher-order information by considering not only whether someone helped but also whom they helped.

This work comprises three interconnected chapters, each shedding light on different facets of cooperation in heterogeneous societies, from both theoretical and experimental perspectives. The first chapter extends the analysis of cooperative behaviour in helping games beyond the confines of homogeneous populations. We consider an infinitely repeated helping game within a society where individuals differ in their abilities and opportunities to help. Recognising that, in reality, individuals vary in their capacity to assist others — some may be unable to help due to constraints beyond their control — we investigate whether self-interested strategies can sustain indirect reciprocity when agents differ in their cost of helping. By characterising the conditions under which various cooperative strategies can be sustained in equilibrium, we illuminate the complexities that arise when heterogeneity in costs is introduced into this framework. The analysis reveals standing strategies have to account for these variations, emphasising the need for more inclusive models that reflect real-world complexities.

Building on these theoretical foundations, the second chapter employs experimental methods to investigate how different reputation-based mechanisms influence cooperative behaviour. Using a lab experiment, we compare the efficacy of IS and GS mechanisms. The experiment introduces a heterogeneous cost condition, reflecting scenarios where the cost of helping varies among individuals. Our findings suggest that while IS can lead to higher overall cooperation rates, particularly in homogeneous cost settings, it does so without participants necessarily conditioning their help on the reputations of others. Conversely, the GS mechanism appears to foster more reciprocal helping behaviour, with individuals more likely to help those who have also been cooperative. GS proves more effective in supporting reciprocal helping in heterogeneous populations, as it facilitates the formation of a cooperative 'club' among individuals with lower costs of helping. This indicates that the structure of the reputation mechanism significantly influences how individuals interpret and respond to reputational information.

The third chapter tests Sugden's version of the Good Standing mechanism and addresses the reputation asymmetries observed in the previous chapter by introducing what we call the "Binary Image Scoring" mechanism. This mechanism simplifies existing reputation systems and allows for a direct comparison of the two mechanisms' effects on cooperative behaviour. Additionally, we include a control condition without reputational information to establish baseline behaviour.

An essential innovation in this chapter is the incorporation of an incentive-compatible belief elicitation procedure. By measuring participants' expectations about the cooperative behaviour of others, we can disentangle the extent to which observed actions are driven by actual preferences versus beliefs about others' likely actions. This allows us to explore whether misunderstandings of the reputation mechanisms might be undermining their effectiveness.

Our experimental results indicate that Sugden's Good Standing mechanism leads to higher levels of both cooperation and reciprocity compared to Binary Image Scoring and the control condition. This suggests that stricter reputation criteria can more effectively promote cooperative behaviour by encouraging individuals to help those who are themselves cooperative.

Collectively, these chapters contribute to a deeper understanding of how indirect reciprocity can be harnessed to promote cooperation in diverse populations. They underscore the importance of designing reputation systems that account for heterogeneity in ability and cost, offering both theoretical and empirical insights into mechanisms that can effectively sustain cooperation in economic interactions.

In conclusion, this thesis advances the theoretical and experimental literature on indirect reciprocity by examining the ways in which reputation mechanisms influence cooperative behaviour. By integrating theoretical models with experimental evidence, it offers comprehensive insights into the design of systems that can sustain cooperation in the multifaceted landscape of human interactions.

Chapter 1

Standing Strategies in the Helping Game

We provide a theoretical analysis and full characterisation of standing strategies in the Helping Game. We further extend this framework by introducing heterogeneity in helping costs and examining its impact on equilibrium strategies and cooperative behaviour. We demonstrate that the distribution of helping costs significantly influences the number and nature of equilibria. Specifically, when the cost distribution is concave, there exists at most one equilibrium with a positive level of helping, whereas convex distributions can lead to multiple interior equilibria. Additionally, we generalise the helping game to include the possibility that participants are not always matched, reflecting more realistic social interactions.

1.1 Introduction

Cooperation among individuals is a fundamental aspect of human societies. The ability of people to work together has profound implications for the functioning of economies and the development of social norms. Economists recognise that cooperation problems are pervasive in many areas, including managerial economics (e.g., teamwork and hold-up problems; Odine, 2015), development economics (e.g., community governance and property rights; Redford, 2020), environmental economics (e.g., natural resource management and climate protection; Carattini et al., 2019), international economics (e.g., trade obstacles and treaty formation; Yarbrough and Yarbrough, 2014) and public economics (e.g., tax compliance and public goods provision; Kube et al., 2015).

Two principal mechanisms have been extensively studied as channels that facilitate cooperative behaviour: direct reciprocity and indirect reciprocity. Direct reciprocity involves mutual assistance between individuals, where one person's cooperative act is directly reciprocated by the other. For example, individual A helps individual B, and B helps A back in return (Trivers, 1971). This dyadic interaction fosters cooperation through repeated exchanges between the same individuals.

In contrast, indirect reciprocity refers to situations where a cooperative act is not reciprocated by the original recipient but by a third party (Alexander, 1987). In such settings, cooperation can be sustained through reputational mechanisms: individuals are motivated to help others to maintain a good reputation, which, in turn, increases the likelihood that they will receive help from others in the future. The role of reputation is particularly salient in environments where direct reciprocity cannot sustain cooperation, and indirect reciprocity has been proposed to explain the evolution of cooperation (Nowak, 2006).

A standard way to study indirect reciprocity is through the "Helping Game,"[1] in which a large population is randomly paired to address a helping decision. In each pair, one player (the helper[2]) can confer a benefit (b) upon the other player (the

[1] Also known as the "donor (or donation) game" (Nowak and Sigmund, 1998a; Leimar and Hammerstein, 2001) or the "indirect reciprocity game" (Ohtsuki, 2004). Other games that are technically identical include a repeated dictator game with random role allocation and matching, and Sugden's (1986) "mutual aid game," modified so that each recipient has a single donor.

[2] The roles in the game are: helper and helpee. In the literature you can find these roles named donor and recipient too.

helpee) at a personal cost (c), with $b > c > 0$. If there is only a single interaction (i.e., no repetition), the only equilibrium outcome is one in which no help is given. However, when the game is repeated, the Folk Theorem (Friedman, 1971) shows that helping can be sustained in equilibrium.

The link between cooperation and helping becomes clear when we recognise that the helper's action contributes to a pattern of reciprocal behaviour. The helper sacrifices immediate utility for future gains. This intertemporal trade-off embodies the cooperative essence: individuals act not solely out of immediate self-interest but with regard for the continuation of mutually beneficial interactions.

The existing literature has predominantly examined this game under the assumption of a homogeneous population — specifically, that all individuals possess the same ability to help. By contrast, in reality, abilities (and thus the costs of helping) often vary across individuals. In order to capture this heterogeneity, we introduce differences in the cost of helping into the model. Such an extension is critical, as diverse abilities characterize most real-world populations, and methods of sustaining cooperation must remain robust in the face of this diversity.

If all members of a society are perfectly homogeneous and derive identical benefits from helping one another, they may readily adopt a mutual strategy that resolves the helping problem. However, once individuals are heterogeneous — bearing different costs when helping — contentious issues can arise. In particular, when considerations of fairness or equity enter the picture, the precise structure of the cooperative strategy becomes critical for effective and sustainable implementation.

This chapter offers several contributions to the literature. First, it presents a complete characterization of standing strategies (Sugden, 1986; Leimar and Hammerstein, 2001) in the helping game. Second, it examines the evolution of these strategies under two modifications: (i) allowing for the possibility that some players are unmatched, and (ii) introducing heterogeneity in individuals' costs of helping.

The remainder of this chapter proceeds as follows: in Section 1.2, we review related literature. The general model is presented in Section 1.3. The heterogeneous cost analysis and some steady state considerations are provided in Section 1.4. We conclude in Section 1.5 with a discussion where we propose possible future improvements. The Appendix 1.A provides alternative proofs for the theorems and additional analyses.

1.2 Related Literature

Indirect reciprocity has been a subject of extensive study across various disciplines, including economics, biology, and sociology. It refers to a mechanism where individuals are willing to cooperate based on the previous actions of others (Trivers, 1971). The evolution of indirect reciprocity has been central to understanding cooperation in both human and animal societies (Sugden, 1986; Alexander, 1987; Nowak and Sigmund, 1998b), leading to the development of theoretical models and experimental investigations aimed at explaining why individuals choose to reciprocate.

In economics, two main strands of literature address reciprocity. The first focuses on modelling non-self-interested behaviour, often incorporating social preferences into individuals' utility functions. Early work by Sugden (1986) analysed reciprocity in the context of public goods games, proposing that individuals' decisions to contribute depend on their beliefs about others' contributions. Rabin (1993) formalised the concept of reciprocity using psychological game theory, suggesting that individuals are motivated to reward kind actions and punish unkind ones. This approach has been further developed by Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006), among others. These models highlight the role of perceived intentions in shaping utility.

The second explores cooperation as pure self-interested behaviour. This literature investigates how strategic considerations and reputational mechanisms maintain cooperation without appealing to social preferences. In this context, scholars have proposed various strategies to explain cooperation in the helping game or similar environments, including standing strategies (Sugden, 1986; Leimar and Hammerstein, 2001), costly punishment (Kandori, 1992), and image scoring (Nowak and Sigmund, 1998a).

Sugden (1986) introduced the concept of a standing strategy in the Mutual Aid Game, where individuals are labelled as being in "good standing" or "bad standing" based on their past actions. Cooperation is sustained by conditional strategies that prescribe helping those in good standing and withholding help from those in bad standing. This framework relies on shared reputational information and allows for the maintenance of cooperation through self-interested behaviour. Subsequent research has explored the evolution of such strategies in different settings. For example, Uchida and Sigmund (2010) studied systems incorporating various assessment rules similar to Sugden's, and Gaudeul et al. (2021) examined how these strategies reflect different moral judgments under indirect reciprocity.

Kandori (1992) developed the notion of community enforcement, showing that social norms and collective punishment strategies can support cooperation in random matching games. In his model, individuals are labelled as "innocent" or "guilty", and defection leads to a transition to the guilty state, triggering punishment by others. This mechanism is similar to standing strategies but emphasises the role of social norms and punishment in sustaining cooperation.

Nowak and Sigmund (1998a) proposed the concept of image scoring in the evolutionary biology literature, where individuals' reputations are updated based on their actions, and cooperation is directed towards those with higher scores. However, this approach has limitations, as it may not be robust to errors in perception or implementation. Moreover, Leimar and Hammerstein (2001) found that the evolution of cooperation using image scoring could only occur under restrictive conditions that consist of either the influence of genetic drift[3] or a very small cost of helping.

In related work on the evolutionary stability of strategies, Ohtsuki and Iwasa (2006) extended earlier analyses by classifying reputation-updating rules into first-, second-, and third-order assessments and identifying the "leading eight" strategies that seem to be able to reach high levels of cooperation in computer simulations. Their findings underscore the complexity of reputation mechanisms and the critical role of higher-order information in indirect reciprocity.

Despite the significant theoretical interest in indirect reciprocity, there is relatively little economic research on the topic compared to other forms of reciprocity. Recent studies, such as Berger (2011) and Berger and Grüne (2016), explore the dynamics of reputation and cooperation, but the field remains ripe for further exploration, particularly regarding the analysis of heterogenous populations.

In recent work, Camera and Gioffré (2014, 2017, 2022) have significantly advanced the theoretical understanding of heterogeneity and asymmetric interactions in repeated games. In Camera and Gioffré (2014), they develop a methodological framework to identify contagious equilibria in infinitely repeated games with random matching, leveraging key statistics of contagious punishment to derive closed-form

[3] The change in frequency of individuals' scores in the population due to random chance.

expressions for continuation payoffs. This approach not only enhances the tractability of equilibrium analysis but also extends beyond the standard helping game to broader settings, including the Prisoner's Dilemma.

Building on these insights, Camera and Gioffré (2017) investigate environments in which players experience stochastic productivity shocks, thereby generating heterogeneous payoffs across matches. They show that full cooperation can be sustained by publicly exposing defections when it is efficient, but occasional defections call for contagious punishments to prevent the collapse of cooperative norms. This highlights how heterogeneity can shape the conditions for cooperative equilibria.

More recently, Camera and Gioffré (2022) provide a comprehensive existence and characterization proof for cooperative equilibria under private monitoring in helping games. They generalise their earlier frameworks to accommodate random matching and asymmetric payoffs and emphasise the role of threshold discount factors in maintaining cooperation.

Whereas Camera and Gioffré primarily focus on heterogeneity arising from productivity shocks and employ contagious punishment to preserve cooperation, our work introduces heterogeneity through intrinsic differences in helping ability, captured by distinct cost functions. Rather than stochastic productivity variations that modify payoff matrices, we adopt a fundamental perspective whereby individuals differ in their inherent cost of providing help. This shift reveals how diverse abilities shape both the emergence and stability of cooperative equilibria.

Additionally, our generalized helping game includes periods in which certain participants remain unmatched, a feature absent in Camera and Gioffré, who assume all players are paired in an even-sized population. Allowing some participants to be unmatched mirrors real-world contexts in which not all agents engage in cooperative interactions continuously. We show that intermittent matching affects both the frequency of cooperation and the conditions necessary for equilibria to exist.

A further distinction is our focus on equilibrium structures under cost heterogeneity. While Camera and Gioffré examine heterogeneous populations via contagious equilibria, we demonstrate that multiple interior equilibria may arise when helping costs vary — particularly under specific cost distributions (e.g., quadratic). This result deepens our understanding of how heterogeneity influences cooperative outcomes in ways not previously emphasized.

In summary, although prior contributions have offered key insights into mechanisms that sustain cooperation, many have assumed homogeneity or have not fully accounted for heterogeneity in helping ability. By examining how cooperation endures among self-interested agents who differ in cost and may remain unmatched in certain periods, we fill an essential gap in the literature and further enrich the analysis of repeated games.

1.3 The Model

1.3.1 The Helping Game

The helping game is built out of a series of asymmetric stage games. In each stage, only one player (the helper) has a choice to make, whether or not to help the opponent (the helpee). When a player is chosen as the helper, he has to choose whether to pay a cost c (the cost of helping), in order to provide a benefit, b (the benefit of being helped), to the beneficiary, with $b > c > 0$. The intuition is that each player could eventually need support in a future scenario, and the benefit derived from receiving help exceeds the cost incurred by providing it.

Let N be a finite set of n agents, indexed by $i \in \{1, \dots, n\}$. Time is discrete and potentially infinite: $t = 1, 2, \dots$. In each period, the game continues with probability δ and otherwise terminates, so δ captures agents' valuations of future payoffs. For the moment, assume each agent observes all past actions taken by all n agents, though we suppress explicit dependence on the history in the notation.

In each period t, a random matching process forms a set $M(t)$ of paired agents $\{(j, k)\}$, where one is assigned as helper and the other as helpee. The number of such pairs, $|M(t)|$, ranges from 0, i.e., there are no matches at all ($|M(t)| = 0$, where $|M(t)|$ is the cardinality of the set), and the maximum is given by $|M(t)| = \frac{N}{2}$ if N is even, and $|M(t)| = \frac{N-1}{2}$ if N is odd.

If an agent i is among the matched pairs at time t, the probability of being chosen as helper in that pair is $\frac{1}{2}$. Let $\lambda = |M(t)|$ denote the total number of helpers each period. Define ω as the fraction of agents who are matched; hence $1 - \omega$ is the fraction of unmatched agents, who make no decisions in that period.

1.3.2 Equilibrium Concept

We denote the possible actions for the helper j by H_j for helping the other player and D_j for not helping the other player. Therefore, the helper has two possible alternative actions: H_j, that means $-c$ for his payoff, or D_j, that means 0.

In general, let $a_{it} \in \{H_i, D_i, \emptyset\}$ be the realised actions player i faces at time t in the history, where $\emptyset$ means he was not a helper in that period (either a helpee or unmatched).

Helpee and the unmatched player do not have any action to make.

A strategy is a complete contingent plan. Here, the strategy of player j (the helper) tells him whether to choose H_j or D_j in his next encounter given the history of all previous play by all players.

The history $h(t)$ is the set of all the actions all players took in every previous period.

A strategy s_j indicates the action that player j makes in every circumstance in which he is a helper, conditional on every potential history.

The strategy profile s is the N-tuple $s = (s_1, ..., s_N)$ which specifies the strategy adopted by each player.

The strategy profile s_{-i} is the N-tuple $s_{-i} = (s_1, s_2, ..., s_{i-1}, s_{i+1}, ..., s_N)$ which specify the action sets of all players except player i.

We define $u_i(s^*)$ as the payoff function given the strategy profile s^*, and $u_i(s'_i, s^*_{-i})$ is the payoff function given by another possible strategy chosen by player i. In our game, a Nash equilibrium is a strategy profile s^* such that no player i has an incentive to deviate to a different strategy s'_i, i.e., each strategy is the best response to others' strategies. The strategy s^* is a Nash equilibrium if

$$u_i(s^*, s^*_{-i}) \geq u_i(s'_i, s^*_{-i}) \qquad \forall i \tag{1.1}$$

Moreover, we say that a strategy profile is a strict Nash equilibrium if

$$u_i(s^*, s^*_{-i}) > u_i(s'_i, s^*_{-i}) \qquad \forall i \tag{1.2}$$

1.3.3 Strategies

Stage Game

First, consider the one-shot (stage) game, corresponding to $\delta = 0$. In this scenario, each helper can either "help" or "not help" in a single interaction. Since helping yields a personal payoff of -c, whereas not helping yields 0 (and $0 >$ -c), the best response is always to refuse to help. Consequently, the unique and strict Nash equilibrium in this one-shot setting is that no one ever helps.

Repeated game

Next, we allow for the possibility of repetition by setting $\delta > 0$, meaning there is some probability that the game continues. One might suspect that if players value future interactions, everyone could agree to "always help" and thus sustain cooperation. However, if all other players follow "always help," any single individual can profitably deviate by refusing to help whenever they are the helper, thereby avoiding the cost -c. Because this unilateral defection increases the deviant's payoff with no immediate repercussions, "always help" cannot be a Nash equilibrium.

By contrast, "never help" is a Nash equilibrium under repetition, just as it was in the one-shot scenario: if you know that nobody else will ever help, refusing to help is your best response. However, the strategy "nobody helps" obviously does not capture the widespread cooperation observed in many real-world interactions.

To resolve this apparent contradiction between theoretical predictions and real-world outcomes, we introduce reputational mechanisms. Specifically, we adopt the notion of *status* from Sugden (1986), which provides individuals with information about others' past behaviour. If players can observe (or infer) how someone has behaved previously, it may become beneficial to help in order to preserve a favourable status, especially when the continuation probability δ is sufficiently high.

Below, we define status more formally as a label or piece of reputational information tied to each player's history of choices. We then propose and analyse two types of standing strategies based on the literature (Sugden, 1986; Leimar and Hammerstein, 2001)[4]. By examining these reputation-based strategies, we show how cooperation can be sustained in a repeated helping game, thereby offering a more plausible account of why cooperative behaviour arises in social contexts.

[4] In the evolutionary biology literature, these strategies take different names such as "stern-judging" or "simple-standing". See, for example, Santos et al. (2021).

1.3.3.1 Sugden (1986)

Player i assigns a *status* to each other player j at each time t, $S_i(j,t) \in \{GS, BS\}$, where GS ="good standing" and BS ="bad standing". All players start out at $t = 1$ with some arbitrary assignment of status. Status is "in the mind" of any given player i and it has the same definition for all the players. We assume that the initial assignment of status is a common and shared judgement, i.e., i's judgement of j's status is equal to k's judgement of j's status, $\forall i, j, k \in N$.

Sugden Updating Rule We update the status according to the following Sugden updating rule:

We iterate through each pair $(j,k) \in M(t)$, and update i's perception of j's status by

$$S_i(j,t+1) = \begin{cases} GS & \text{if } S_i(k,t) = GS \wedge a_{j,t} = H_j, \\ BS & \text{if } S_i(k,t) = GS \wedge a_{j,t} = D_j, \\ S_i(j,t) & \text{otherwise.} \end{cases}$$

Player k's status is unchanged, $S_i(k,t+1) = S_i(k,t)$.

The status of any player $z \neq j,k$ who is not matched at time t is unchanged, $S_i(z,t+1) = S_i(z,t)$.

This updating rule implies that if in successive periods player i is in BS, he can regain the GS just playing the action H_i with an opponent that is in GS.

Sugden Good Standing Rule The *good-standing rule strategy A.1* is a strategy for player i, $s(h)$ for each possible history h, which states that, if i is matched with j, then $s(h) = H_i$ (help) as long as $S_i(j,t) = GS$ and $s(h) = D_i$ if $S_i(j,t) = BS$ (i.e., opponent j is in BS, he plays D_i (defect)).

The Sugden strategy *A.1* says, irrespective of your own status, to help the other only if he is in good standing and not to help if the other is in bad standing.

Theorem 1.1. *The strategy profile in which all players play A.1 is a Nash equilibrium, and if all players play A.1, then players help if and only if* $\delta > \frac{2c}{[\omega(b-c)]+2c}$.

Proof. Given the infinite time horizon of the game and the constant discount rate, the model is stationary, i.e., players are playing now the same stage game it will be played in the future. Thanks to this property, it is sufficient to compare two strategies at the current time and generalise for the future periods.

We want to find if the strategy *A.1* is a Nash equilibrium and the critical value for which the helping behaviour occurs.

There are two probabilities we have to take into account now: the probability δ that the games continues to $t+1$ and the probability of the match ω. We have to remember that all the probabilities should be conditional on the current situation.

Let us denote U_H as the expected payoff from following strategy *A.1* for ever, given that you are matched as the helper in the current period.

As long as the game continues, this will give you the series of expected payoffs: $-c, \frac{\omega(b-c)}{2}, \frac{\omega(b-c)}{2}, ...$

At time $t+1$ you will get $\frac{\omega(b-c)}{2}$ with probability δ. The latter will be taken into account for all the other possible periods (we assume they could be infinite) discounted by the time period. The ω here comes from the probability of being matched.

Therefore, because of the probability that the game will end, becomes:

$$U_H = -c + \frac{\omega(b-c)}{2}\left[\delta + \delta^2 + ...\right] \tag{1.3}$$

knowing that[5]

$$\delta + (\delta)^2 + ... = \frac{\delta}{1-\delta} \tag{1.4}$$

then

$$U_H = -c + \frac{\omega(b-c)}{2}\left[\frac{\delta}{(1-\delta)}\right] \tag{1.5}$$

Let us denote U_D as the expected payoff from the strategy “defect at every stage” (myopic self-interest) for ever, given that you are the helper in the current period. Clearly, $U_D = 0$.

[5] This comes from a geometric series approximation. Indeed, when $|\delta| < 1$, as in our case, in the geometric series $\delta + \delta^2 + \delta^3 + ...$ the terms of the series approach zero in the limit (becoming smaller and smaller in magnitude), and the series converges to the sum $\frac{\delta}{1-\delta}$.

Therefore, $U_H > U_D$ if

$$\begin{aligned} -c + \frac{\omega(b-c)}{2}\left[\frac{\delta}{(1-\delta)}\right] &> 0 \qquad (1.6)\\ \omega(b-c)\delta + 2c\delta &> 2c \qquad (1.7)\\ \delta\left[\omega(b-c)+2c\right] &> 2c \qquad (1.8)\\ \delta &> \frac{2c}{[\omega(b-c)]+2c} \qquad (1.9) \end{aligned}$$

Therefore, if we assume 1.9 holds, *A.1* is an equilibrium strategy. If the helpee is following *A.1*, your best reply as helper is to help. Therefore, we can see that the Good Standing rule (*A.1*) is the best reply to itself. □

Theorem 1.1 is a generalisation of the special case studied in the literature (Sugden, 1986) in which everyone is matched ($\omega = 1$) randomly and players are even. In that case the condition for the helping behaviour would become

$$\delta > \frac{2c}{(b+c)} \qquad (1.10)$$

We can also show[6] that wherever the starting point is (in terms of standing in the population) your best reply as helper is to help if your opponent is in GS, but not if she is in BS. In fact, we made no assumption about the starting point. The only thing that matters is the *status* in which you are: everything about the $t+1$ period is fully determined by t status.

If I am the helper at time t and my opponent is in *BS*, my decision — whether to help or not — has no impact on anyone else's standing. In particular, it does not affect the opponent's status (since non-active player's status does not change) nor does it alter my own status.

However, starting from a period in which no one is in good standing, the Sugden Good Standing Rule strategy (*A.1*) tells everyone not to help: no one can regain the *GS status* given the *A.1* strategy that does not allow to help anyone that is not in *GS*.

Our generalisation, which includes the possibility of unmatched players, makes it somewhat more challenging for equation 1.9 to be satisfied compared to the special case (equation 1.10). As the variable ω moves to the denominator and

[6] See 1.A.3 in the Appendix 1.A.

as it decreases when the matches diminish, the requirements for the continuation probability δ become harder to meet. Consequently, the conditions for the helping behaviour are more difficult to satisfy.

The strategy described above is therefore a Nash equilibrium and the helping behaviour occurs when the value of δ is greater than the threshold we have derived.

1.3.3.2 Leimar and Hammerstein (2001)

Leimar and Hammerstein (2001) proposed a further strategy that allows players to regain their good standing in every period, simply by helping regardless of the status of their matches. In the following, we want to analyse whether this strategy is a Nash equilibrium or not and for which δ values.

Leimar and Hammerstein Updating Rule The main difference with the Sugden one arises when not everyone is in good standing. Indeed, under the Sugden rule, all helpers will rationally help when the helpee is in good standing, but not otherwise. The Hammerstein rule gives less incentive to helping people in GS because even if you are in BS you will benefit from people in BS trying to gain GS. But it allows a quicker return from BS.

We update the status according to the following Hammerstein updating rule:

We iterate through each pair $(j, k) \in M(t)$, and update helper j's status by

$$S_i(j, t+1) = \begin{cases} GS & \text{if } a_{j,t} = H_j, \\ BS & \text{if } S_i(k, t) = GS \wedge a_{j,t} = D_j, \\ S_i(j, t) & \text{otherwise.} \end{cases}$$

Player k's status is unchanged, $S_i(k, t+1) = S_i(k, t)$.

The status of any player $z \neq j, k$ who is not matched at time t is unchanged, $S_i(z, t+1) = S_i(z, t)$.

Leimar and Hammerstein Good Standing Rule The *good-standing rule strategy* A.2 is a strategy for player i, $s(h)$ for each possible history h , which states that, if i is matched with j , then $s(h) = H_i$ (help) as long as $S_i(i, t) = BS$ (i.e., player i is in BS) or $S_i(j, t) = GS$ (i.e., opponent j is in GS) and $s(h) = D_i$

if $S_i(j,t) = BS$ (i.e., opponent j is in BS, he plays D_i(defect).

Player i can recover GS by helping in any period, even if the helpee is not in GS (but if you are already in GS, you are permitted not to help someone who is not in GS).

Formally, strategy *A.2* is to help anyone when you are in BS, and to help (only) people in GS when you are in GS. As an intuition, it proposes to help anyone if you need to regain GS or if the opponent is in GS.

So, similarly to the Sugden one (*A.1*), you have the possibility to help every period someone in good standing, but, differently, in the case you are in bad standing, regain the good standing by helping the other every time you are an helper.

Theorem 1.2. *Strategy* $A.2$ *is a Nash equilibrium if and only if* $\delta > \frac{2c}{[\omega(b-c)]+2c}$.

Proof. As above, the model is stationary. Therefore, we can again compare two strategies at the current time and generalise for the future periods.
We denote U_H as the expected payoff from following strategy $A.2$ for ever, given that you are the helper in the current period.
As long as the game continues, this will give you the series of expected payoffs:

$$-c, \frac{\omega(b-c)}{2}, \frac{\omega(b-c)}{2}, ... \tag{1.11}$$

Because of the probability that the game will end, U_H becomes:

$$-c + \frac{\omega(b-c)}{2}\left[\frac{\delta}{(1-\delta)}\right] \tag{1.12}$$

On the other hand, we denote U_D as the expected payoff from the strategy "do not help at every stage" (myopic self interest) for ever, given that you are the helper in the current period. Clearly, $U_D = 0$. Therefore, $U_G > U_D$ if

$$-c + \frac{\omega(b-c)}{2}\left[\frac{\delta}{(1-\delta)}\right] > 0 \tag{1.13}$$

which rearranges to

$$\delta > \frac{2c}{[\omega(b-c)]+2c} \tag{1.14}$$

□

Similarities and differences between the two equilibria

First of all, the equilibrium condition for both strategies is the same, in fact equation 1.9 and equation 1.14 are identical.

But the two strategies have different implications for behaviour starting from a period in which not everyone is in good standing[7]. In the extreme case in which everyone at the start of the game (t) is provided with a bad standing, the Sugden rule strategy *A.1* induces a Nash equilibrium with no helping at all. Effectively, following *A.1*, all helpers would rationally help when the helpee is in good standing, but not otherwise. So at time $t+1$ no help would take place, and the same results in all the (possible) future periods.

Conversely, the Leimar and Hammerstein rule strategy *A.2* induces a Nash Equilibrium with 100% helping given that everyone would regain the GS status as soon as possible. The curious component is that in this period every helper helps and at $t+1$ half of the population times omega is in GS.

1.4 Heterogeneous Cost of Helping in Population

An individual's ability to help is inherently linked to the cost they incur when providing help. Specifically, higher ability often translates into greater efficiency or proficiency, thereby reducing the effort or resources required to help and consequently lowering the personal cost.

We extend our analysis by examining how variations in individuals' abilities, and therefore the costs of helping, influence the equilibrium strategies[8]. We find that the conditions required for helping to be sustained become even more stringent under such heterogeneity. However, our current focus is a theoretical examination of the steady states in equilibrium within this game. This steady-state analysis is particularly interesting as it reveals the long-term outcomes and stability of cooperation under different strategic settings.

In each period, a continuum of players is divided randomly into being helpers and helpees. If the helper helps, the helpee receives a benefit $b > 0$, and the helper incurs some cost c. Players differ in their costs of helping; we assume players' costs

[7] See full discussion in 1.A.3 in the Appendix 1.A.

[8] See 1.A.4 in the Appendix 1.A.

of helping are distributed according to a cdf $F(c)$, with support $[0, \infty)$, and a density $f(c)$.

Let us assume a fraction m of players are in good standing, and "the identities" of those players is public information. A player loses good standing if they are matched with another player in good standing, but they fail to help.

We ask what are the possible steady-state values of m, i.e., the proportion of players in good standing.

Proposition 1.1. *Given a steady-state proportion m of players in good standing, it is a best response for a player in good standing to help another player in good standing if*

$$c \leq \left[\frac{\delta m}{\delta m + 2 - 2\delta}\right] b \tag{1.15}$$

Proof. Because we are in a steady-state, the only two strategies to consider are "help if the helpee is in good standing" and "never help". The steady-state utility of a player who always helps when expected to at the start of a period (prior to knowing whether he is a helper or a helpee) is

$$U = m\left[\frac{1}{2}U_H + \frac{1}{2}U_R\right] + (1-m)\delta U \tag{1.16}$$

For the helper who chooses to help, $U_H = -c + \delta U$; and if the player is a helpee, $U_R = b + \delta U$. Therefore,

$$\begin{aligned} U &= m\left[(b-c)\delta U\right] + (1-m)\delta U && (1.17)\\ U &= \frac{1}{1-\delta} \cdot \frac{1}{2}m(b-c) && (1.18) \end{aligned}$$

Therefore, the utility of a player who is a helper and helps is

$$U_H = -c + \frac{\delta}{1-\delta} \cdot \frac{1}{2}m(b-c) \tag{1.19}$$

In the contingency where the player is a helper and does not help, his continuation payoff is zero. Therefore, helping is a best response if and only if

$$-c + \frac{\delta}{1-\delta} \cdot \frac{1}{2}m(b-c) \geq 0 \longrightarrow c \leq \left[\frac{\delta m}{\delta m + 2 - 2\delta}\right] b \tag{1.20}$$

□

Therefore, for a given m, the proportion of players for whom it is a best response to help other players in good standing is

$$F\left(\frac{b\delta m}{2-2\delta+\delta m}\right) \tag{1.21}$$

In order to have a proportion m of players in good standing in steady state, m must satisfy

$$F\left(\frac{b\delta m}{2-2\delta+\delta m}\right) = m \tag{1.22}$$

Proposition 1.2. *There is always an equilibrium with $m = 0$, i.e., nobody helps.*

Proof. For $m = 0$, equation (1.22) becomes $F(0) = 0$. Because 0 is the lower bound of the support of F and there are no mass points, this equation is satisfied. □

Special Case: Uniformly Distributed Costs

For now, we assume that c is distributed uniformly on $[0, 1]$, so $F(c) = c$ for $c \in [0, 1]$.

Proposition 1.3. *If c is uniformly distributed on $[0, 1]$, then there exists an interior equilibrium (that is, one in which $0 < m < 1$), if and only if $\frac{2}{\delta} - 2 < b < \frac{2}{\delta} - 1$, or, equivalently, $\frac{2}{b+2} < \delta < \frac{2}{b+1}$.*

Proof. From (1.22), an interior equilibrium must satisfy

$$\frac{b\delta m}{2-2\delta+\delta m} = m \tag{1.23}$$

This rearranges to

$$m = b - 2\left(\frac{1-\delta}{\delta}\right) \tag{1.24}$$

Therefore for an interior equilibrium we must have

$$0 < b - 2\left(\frac{1-\delta}{\delta}\right) < 1 \tag{1.25}$$

For each δ there exists a range of b compatible with an interior equilibrium, and vice versa. For a given δ, it follows that

$$\begin{aligned} -b &< -2\left(\tfrac{1-\delta}{\delta}\right) &< -b+1 \qquad (1.26)\\ \frac{b}{2} &> \left(\tfrac{1-\delta}{\delta}\right) &> \frac{b-1}{2} \qquad (1.27)\\ \frac{2}{b+2} &< \delta &< \frac{2}{b+1} \qquad (1.28) \end{aligned}$$

On the other hand, for a given b, we have

$$\begin{aligned} 2\left(\frac{1-\delta}{\delta}\right) &< b &< 2\left(\frac{1-\delta}{\delta}\right)+1 \qquad (1.29)\\ \frac{2}{\delta}-2 &< b &< \frac{2}{\delta}-1 \qquad (1.30) \end{aligned}$$

□

Proposition 1.4. *If c is uniformly distributed on $[0,1]$, then there exists an equilibrium with full helping, $m=1$, if and only if $b \geq \frac{2}{\delta}-1$, or equivalently $\delta \geq \frac{2}{b+1}$.*

Proof. For $m=1$, equation (1.22) becomes

$$\frac{b\delta}{2-\delta} \geq 1 \qquad (1.31)$$

Rearranging, we have

$$\delta(b+1) \geq 2 \qquad (1.32)$$

This means that fixing the two parameters, we can get the values for which

$$b \geq \frac{2}{\delta}-1 \quad \text{or, equivalently} \quad \delta \geq \frac{2}{b+1} \qquad (1.33)$$

□

Taken together, Propositions 1.2, 1.3, and 1.4 provide a full characterisation of the possible equilibria in the case of uniformly-distributed costs. Figure 1.1 plots the three relevant regions of parameters. Below and to the left of the solid line,

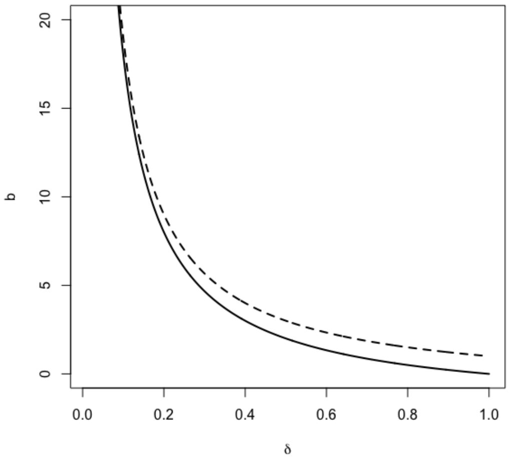


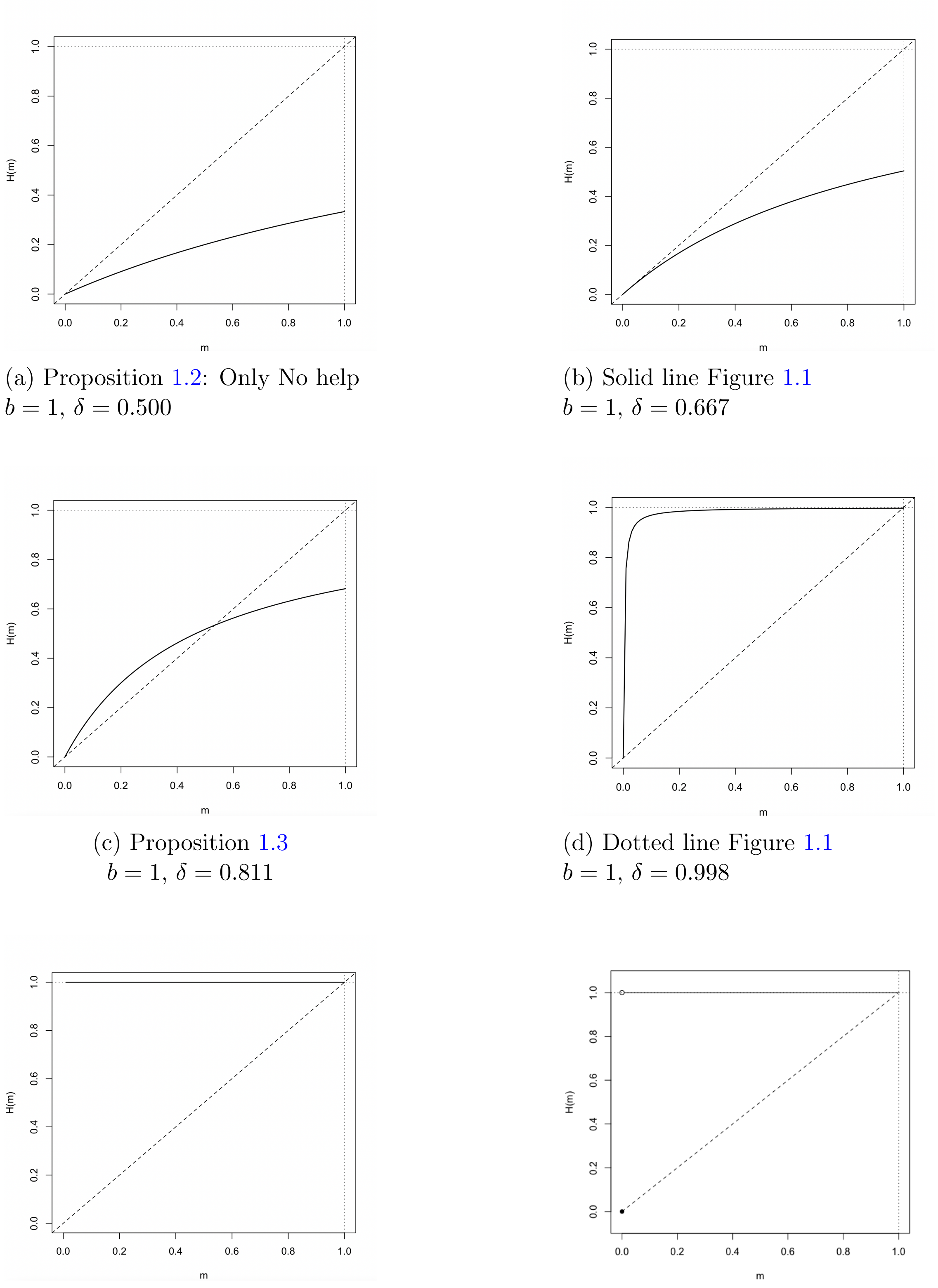
Figure 1.1: Regions of equilibria in the case of uniformly-distributed costs.

there is a unique equilibrium with no helping. Between the solid line and the dashed line, there is an equilibrium with no helping, and an interior equilibrium with partial helping, $0 < m < 1$. Above the dashed line, there is an equilibrium with no helping, and an equilibrium with full helping.

In the case with uniformly-distributed costs, there is at most one equilibrium which involves a positive amount of helping. To understand why, we plot the proportion of players for whom helping is a best response, $H(m) \equiv F\left(\frac{b\delta m}{2-2\delta+\delta m}\right)$, as a function of m, for selected values of b and δ in Figure 1.2.

Figure 1.2 shapes 6 possible interesting representations of equilibria in the game. All the cases in Figure 1.2, apart from the limiting case 1.2e, stand for a region of Figure 1.1. In the limiting case 1.2e, full help ($m = 1$) is the only equilibrium.

Figure 1.2a pictures the case of only "no help" as the only possible equilibrium in our game, i.e., proposition 1.2. Figure 1.2c shows us a situation where, as predicted by proposition 1.3, there is either an equilibrium at 0, or an internal equilibrium. On the other hand, Figure 1.2d perfectly depicts a case in which the parameters δ and b correspond to the dashed curve in Figure 1.1. Lastly, Figure 1.2f characterises a possible situation in which, as theorised in the proposition 1.4, there is an equilibrium at 0 or full help.



(a) Proposition 1.2: Only No help
$b = 1,\ \delta = 0.500$

(b) Solid line Figure 1.1
$b = 1,\ \delta = 0.667$

(c) Proposition 1.3
$b = 1,\ \delta = 0.811$

(d) Dotted line Figure 1.1
$b = 1,\ \delta = 0.998$

(e) Limiting case
$b = 1,\ \delta = 1$

(f) Proposition 1.4
$b = 2,\ \delta = 0.767$

Figure 1.2: Comparison of examples with uniformly-distributed cost.

More General Case

We note from Figure 1.2 that in all cases, $H(m)$ is increasing and concave as a function of m, which leads to there being at most one equilibrium with helping. Now we ask for what distributions of costs this result will continue to hold.

If we analyse in more detail the LHS of function 1.22, and look for the derivative, we can analyse the sign of the first derivative and the second derivative in order to better understand the behaviour of the function.

$$\begin{aligned}
\frac{\partial F\left(\frac{b\delta m}{2-2\delta+\delta m}\right)}{\partial m} &= F'\left(\frac{b\delta m}{2-2\delta+\delta m}\right)\cdot\frac{(2-2\delta+\delta m)b\delta - b\delta m\delta}{(2-2\delta+\delta m)^2} \quad (1.34)\\
&= F'\left(\frac{b\delta m}{2-2\delta+\delta m}\right)\cdot\frac{2b\delta-2b\delta^2+bm\delta^2-bm\delta^2}{(2-2\delta+\delta m)^2} \quad (1.35)\\
&= \underbrace{F'\left(\frac{b\delta m}{2-2\delta+\delta m}\right)\cdot\frac{2b\delta(1-\delta)}{(2-2\delta+\delta m)^2}}_{\geq 0} \quad (1.36)
\end{aligned}$$

All values in 1.36 are positive since $b > 0$, $\delta \in (0, 1]$ and $m \in [0, 1]$. So the first derivative is positive, which means that $H(m)$ is strictly monotone increasing.

Now we analyse the second derivative.

$$\begin{aligned}
\frac{\partial^2 F\left(\frac{b\delta m}{2-2\delta+\delta m}\right)}{\partial m^2} = F''\left(\frac{b\delta m}{2-2\delta+\delta m}\right)\cdot\left[\frac{2b\delta(1-\delta)}{(2-2\delta+\delta m)^2}\right]^2\\
\underbrace{+\,F'\left(\frac{b\delta m}{2-2\delta+\delta m}\right)\cdot\left[-\frac{2b\delta(1-\delta)\cdot 2(2-2\delta+\delta m)\delta}{(2-2\delta+\delta m)^4}\right]}_{\leq 0} \quad (1.37)
\end{aligned}$$

The second term in equation 1.37 is clearly negative, because it is multiplied by a negative sign while all other variables involved are positive. A natural question, then, is whether the first term can outweigh this negative component and ultimately render 1.37 positive.

From the cases analysed above, in which the second derivative is negative, it follows that there can be at most one interior equilibrium or a corner solution involving constant help; in such scenarios, two distinct equilibria cannot coexist.

Consequently, when the function in question is uniform, there are precisely three possible equilibria: zero (no help), an interior equilibrium, or full help. The calculations presented here lead us to the following proposition.

Proposition 1.5. *Let F be the cdf of the cost of helping. If $F'' < 0$, then there exists at most one equilibrium with a positive amount of helping.*

This proposition is illustrated by the examples in Figure 1.2, but it raises the question of whether F'' can ever be positive. To explore this possibility, we now generalize our analysis to determine whether the second derivative can indeed become positive. If so, two distinct equilibria may emerge.

From an economic standpoint, one way to motivate such a scenario is to consider a linearly increasing cost distribution, implying that certain individuals face higher costs. Mathematically, a straightforward way to investigate this is to assume an increasing cost function and examine how the resulting equilibrium outcomes behave under those conditions.

If we consider $F(c) = kc^{\alpha}$, therefore

$$F\left(\frac{b\delta m}{2-2\delta+\delta m}\right) = \frac{m^{1/\alpha}}{k} \tag{1.38}$$

Example Suppose $F(c) = c^2$, $b = 1$, and $k = 1$. Then in order for m to be an (interior) equilibrium,

$$\left(\frac{\delta m}{2-2\delta+\delta m}\right)^2 = m \tag{1.39}$$

Let us look for a value of δ such that $m = \frac{1}{2}$ is an equilibrium. Equation (1.39) becomes

$$\begin{aligned}
\left(\frac{\frac{1}{2}\delta}{2-2\delta+\frac{1}{2}\delta}\right)^2 &= \frac{1}{2} && (1.40)\\
\left(2-2\delta+\frac{1}{2}\delta\right) &= \frac{\sqrt{2}}{2}\delta && (1.41)\\
4\delta-\delta+\sqrt{2}\delta &= 4 && (1.42)\\
\delta &= \frac{4}{3+\sqrt{2}} \approx 0.906 && (1.43)
\end{aligned}$$

Now, we can show that given the δ just found, we can find another m which also satisfies the condition for an interior equilibrium.

$$\left(\frac{\frac{4}{3+\sqrt{2}}m}{2-2\frac{4}{3+\sqrt{2}}+\frac{4}{3+\sqrt{2}}m}\right)^2 = m \tag{1.44}$$

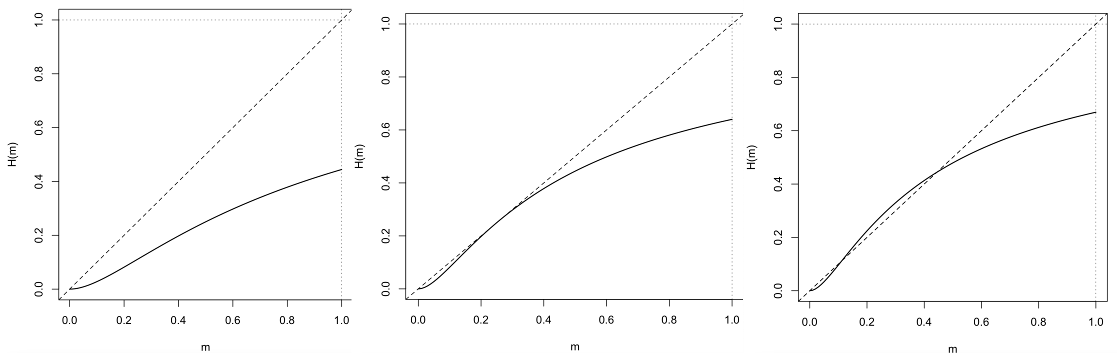
$$\left(\frac{4m^2}{(2m+\sqrt{2}-1)^2}\right) = m \tag{1.45}$$

$$m(2m+\sqrt{2}-1)^2 = 4m^2 \tag{1.46}$$

$$4m^2+4m(\sqrt{2}-2)+(-2\sqrt{2}+3) = 0 \tag{1.47}$$

$$\underbrace{(2m-1)(2m+2\sqrt{2}-3)}_{\text{this is a parabola}} = 0 \tag{1.48}$$

Therefore there are indeed two interior equilibria, $m=\frac{1}{2}$, and $m=\frac{3}{2}-\sqrt{2}\approx 0.0858$. This is illustrated in Figure 1.3c.



(a) No interior equilibria
$b=1$, $\delta=0.8$

(b) Double root
$b=1$, $\delta=0.888$

(c) 2 interior equilibria
$b=1$, $\delta=0.9$

Figure 1.3: Comparison of examples with quadratic cost function

Proposition 1.6. *Let $F(c)=c^2$, i.e., $\alpha=2$, and fix $b=1$. Then there exist exactly at maximum two interior equilibria with positive amounts of helping.*

Proof. The steps are as follows.

1. We can show that the first derivative of LHS of (1.39) is equal to 0 at $m=0$. In fact, it becomes

$$\frac{\partial F\left(\frac{\delta m}{2-2\delta+\delta m}\right)^2}{\partial m} = \frac{2\delta^2 m}{(\delta m-2\delta+2)^2}-\frac{2\delta^3 m^2}{(\delta m-2\delta+2)^3} \tag{1.49}$$

$$= -\frac{4(\delta^3 m-\delta^2 m)}{(\delta m-2\delta+2)^3} \tag{1.50}$$

Therefore, (1.50) is equal to 0 at $m = 0$.

2. Given the point above, the function is differentiable at 0 and therefore also continuous. Because this is a continuously differentiable function, this implies that there exists some m' such that, for $0 \leq m < m'$, $\frac{\delta m}{2-2\delta+\delta m} < m$.

3. Observe that $\left(\frac{\delta m}{2-2\delta+\delta m}\right)^2 - m = 0$ is quadratic in m. Therefore, there are either 0, 1, or 2 solutions (in the real numbers) for m.

$$m\left[\frac{\delta^2 m}{[\delta m + 2(1-\delta)]^2} - 1\right] = 0 \tag{1.51}$$

$$\delta^2 m - [\delta m + 2(1-\delta)]^2 = 0 \tag{1.52}$$

$$\delta^2 m - (\delta^2 m^2 - 4\delta^2 m + 4\delta^2 + 4\delta m - 8\delta + 4) = 0 \tag{1.53}$$

$$-\delta^2 m^2 + m(5\delta^2 - 4\delta) - 4(\delta^2 - 2\delta + 1) = 0 \tag{1.54}$$

Divide by $-\delta^2$ to get:

$$m^2 - \frac{m(5\delta^2 - 4\delta)}{\delta^2} - \frac{4(\delta^2 - 2\delta + 1)}{\delta^2} = 0 \tag{1.55}$$

Solve for m:

$$m = \frac{-(5\delta^2 - 4\delta) \pm \sqrt{16(\delta^2 - 2\delta + 1)\delta^2 + (5\delta^2 - 4\delta)^2}}{2\delta^2} \tag{1.56}$$

4. We know from algebra that the determinant of the quadratic function gives us the possible solution(s) of this equation.
In our formula the determinant is $16(\delta^2 - 2\delta + 1)\delta^2 + (5\delta^2 - 4\delta)$.
If the determinant is lower than 0, then we have no real equilibria. If the determinant is equal to 0, then we have exactly one equilibrium (double root). The latter corresponds, in our model, to the cutoff point for having at least one equilibrium. Finally, if the determinant is greater than 0, we have two solutions, i.e., two internal equilibria in our model.

We can find then this cutoff point:

$$16(\delta^2 - 2\delta + 1)\delta^2 + (5\delta^2 - 4\delta) = 0 \tag{1.57}$$

$$\delta(16\delta^3 - 32\delta^2 + 21\delta - 4) = 0 \tag{1.58}$$

$$\delta = \frac{1}{2}\left(8 - \frac{1}{\sqrt[3]{28 - 3\sqrt{87}}} - \sqrt[3]{28 - 3\sqrt{87}}\right) \approx 0.33^{9} \tag{1.59}$$

□

1.5 Discussion

In this chapter, we provided a theoretical characterisation of the standing strategies in the helping game and we extended the model by introducing heterogeneity in individuals' cost to help and incorporating the possibility that participants may not be matched in every period. By modelling the cost of helping as a continuous random variable with a specified distribution $F(c)$, we have analysed how differences in the ability to help influence the emergence and sustainability of cooperation in large societies.

Our analysis reveals that the shape of the cost distribution plays a pivotal role in determining both the number and the nature of equilibria. Specifically, when the cost distribution function is concave ($F''(c) < 0$), there exists at most one equilibrium with a positive level of helping. In contrast, if the cost distribution is convex ($F''(c) > 0$), multiple (up to two) interior equilibria may arise. This finding highlights the significance of heterogeneity in abilities: variations in individuals' costs of helping can lead to diverse equilibrium outcomes, affecting whether cooperation is sustained within the population.

By introducing a matching probability ω, we have accounted for the realistic scenario in which not all individuals are matched in every period. This modification acknowledges that opportunities for interaction and cooperation are not uniformly

[9] The cubic equation solution for a function $ax^3 + bx^2 + cx + d$:

$$p = \frac{-b}{3a}, \quad q = p^3 + \frac{bc - 3ad}{6a^2}, \quad r = \frac{c}{3a},$$

$$x = \left(q + \sqrt{q^2 + (r - p^2)^3}\right)^{\frac{1}{3}} + \left(q - \sqrt{q^2 + (r - p^2)^3}\right)^{\frac{1}{3}} + p \tag{1.60}$$

distributed, allowing us to examine how the frequency of interactions influences the conditions under which cooperative behaviour can be maintained.

Our work differs from previous studies, such as those by Camera and Gioffré (2014, 2017, 2022), which focus on heterogeneity arising from stochastic variations in payoffs due to productivity shocks and rely on private monitoring and contagious punishments to sustain cooperation. In contrast, we consider inherent differences in individuals' abilities to help, reflecting real-world variations such as disabilities or resource constraints. We employ reputational mechanisms based on standing strategies—specifically, those inspired by Sugden (1986) and further developed by Leimar and Hammerstein (2001) — to facilitate cooperation without relying on private monitoring or the threat of contagious punishment.

There remain several avenues for further research. One direction is to explore alternative forms of the cost distribution $F(c)$ including those with discontinuities or heavy tails, to understand how different types of heterogeneity affect equilibrium outcomes. Incorporating dynamic elements into the model — such as allowing individuals to invest in their ability to help or to adapt their strategies based on past experiences — could provide insights into how cooperative norms evolve over time. Additionally, examining the impact of asymmetric information, where individuals do not fully observe others' abilities or past actions, can be informative.

Empirical validation of the theoretical predictions through experimental studies could enhance our understanding of cooperative behaviour in heterogeneous populations. Investigating whether individuals behave in accordance with the proposed strategies and how variations in abilities influence their decisions to help would offer practical insights. Such studies could also assess the effectiveness of different reputational mechanisms in promoting cooperation among diverse agents.

In conclusion, our analysis demonstrates that cooperation can be sustained through self-interested strategies relying on reputational mechanisms, even in populations where individuals differ in their abilities to help. The presence of heterogeneity adds complexity to the strategic environment but also offers opportunities to design interventions that promote cooperative behaviour. By highlighting the conditions under which cooperative equilibria exist and how they are influenced by the distribution of helping costs, this chapter contributes to the theoretical foundations of cooperation in diverse societies.

References

Alexander, R. D.: 1987, *The biology of moral systems*, New York: Aldine de Gruyter.

Berger, U.: 2011, Learning to cooperate via indirect reciprocity, *Games and Economic Behavior* **72**(1), 30–37.

Berger, U. and Grüne, A.: 2016, On the stability of cooperation under indirect reciprocity with first-order information, *Games and Economic Behavior* **98**, 19–33.

Camera, G. and Gioffré, A.: 2014, A tractable analysis of contagious equilibria, *Journal of Mathematical Economics* **50**, 290–300.
URL: *https://www.sciencedirect.com/science/article/pii/S0304406813000670*

Camera, G. and Gioffré, A.: 2017, Asymmetric social norms, *Economics Letters* **152**, 27–30.
URL: *https://www.sciencedirect.com/science/article/pii/S0165176516305341*

Camera, G. and Gioffré, A.: 2022, Cooperation in indefinitely repeated helping games: Existence and characterization, *Journal of Economic Behavior & Organization* **200**, 1344–1356.
URL: *https://www.sciencedirect.com/science/article/pii/S0167268119303579*

Carattini, S., Levin, S. and Tavoni, A.: 2019, Cooperation in the climate commons, *Review of Environmental Economics and Policy* **13**(2), 227–247.

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

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

Friedman, J.: 1971, A non-cooperative equilibrium for supergames, *The Review of Economic Studies* **38**(1), 1–12.
URL: *https://EconPapers.repec.org/RePEc:oup:restud:v:38:y:1971:i:1:p:1-12.*

Gaudeul, A., Keser, C. and Müller, S.: 2021, The evolution of morals under indirect reciprocity, *Games and Economic Behavior* **126**, 251–277.
URL: *https://www.sciencedirect.com/science/article/pii/S0899825621000063*

Kandori, M.: 1992, Social norms and community enforcement, *The Review of Economic Studies* **59**(1), 63–80.
URL: *http://www.jstor.org/stable/2297925*

Kube, M., Hilgers, D., Koch, G. and Füller, J.: 2015, Explaining voluntary citizen online participation using the concept of citizenship: an explanatory study on an open government platform, *Journal of Business Economics* **85**, 873–895.

Leimar, O. and Hammerstein, P.: 2001, Evolution of cooperation through indirect reciprocity, *Proc. R. Soc. Lond. B.* **268**, 745–753.

Nowak, M.: 2006, Five rules for the evolution of cooperation, *Science* **314**(5805), 1560–1563.

Nowak, M. and Sigmund, K.: 1998a, The dynamics of indirect reciprocity, *Journal of Theoretical Biology* **194**(4), 561–574.

Nowak, M. and Sigmund, K.: 1998b, Evolution of indirect reciprocity by image scoring, *Nature* **393**, 573–577.

Odine, M.: 2015, Communication problems in management., *Journal of Emerging Issues in Economics, Finance & Banking* **4**(2).

Ohtsuki, H.: 2004, Reactive strategies in indirect reciprocity, *Journal of Theoretical Biology* **227**(3), 299–314.
URL: *https://www.sciencedirect.com/science/article/pii/S0022519303004260*

Ohtsuki, H. and Iwasa, Y.: 2006, The leading eight: social norms that can maintain cooperation by indirect reciprocity, *Journal of theoretical biology* **239**(4), 435–444.

Rabin, M.: 1993, Incorporating fairness into game theory and economics, *American Economic Review* **83**, 1281–1302.

Redford, A.: 2020, Property rights, entrepreneurship, and economic development, *The Review of Austrian Economics* **33**(1), 139–161.

Santos, F. P., Pacheco, J. M. and Santos, F. C.: 2021, The complexity of human cooperation under indirect reciprocity, *Philosophical Transactions of the Royal Society B: Biological Sciences* **376**(1838), 20200291.
URL: *https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2020.0291*

Sugden, R.: 1986, *The economics of rights, co-operation and welfare*, Oxford, UK: Basil Blackwell.

Trivers, R.: 1971, The evolution of reciprocal altruism, *The Quarterly review of biology* **46**, 35–57.

Uchida, S. and Sigmund, K.: 2010, The competition of assessment rules for indirect reciprocity., *Journal of theoretical biology* **263 1**, 13–9.
URL: *https://api.semanticscholar.org/CorpusID:29124254*

Yarbrough, B. V. and Yarbrough, R. M.: 2014, *Cooperation and governance in international trade: The strategic organizational approach*, Vol. 133, Princeton University Press.

1.A Appendix

Different approaches to deriving Theorem 1.1 and Theorem 1.2 are given below.

1.A.1 Alternative Proofs for Theorem 1.1

Proof 1

Start on period 0. Conditional on the game continuing to period 1, the probability for i to be a helper or a helpee is the same, $\frac{\omega}{2}$. This probability should also take into account the conditionality of *not* being a helper, i.e., the probability of being a helpee becomes $\frac{\frac{\omega}{2}}{1-\frac{\omega}{2}} = \frac{\omega}{2-\omega}$. Hence, the expected benefit you can get if the game continues to period 1 and you are not a helpee is $b\frac{\omega}{2-\omega}$.

An easy way to find the critical values, it is now to substitute in equation $\delta > \frac{2c}{(b+c)}$ (eq. 1.10) the new parameters.

Now, the equivalent for δ is $\delta(2-\delta)$, i.e., the probability that game continues to period 1. Whereas, the equivalent of b is $b\frac{\omega}{2-\omega}$, i.e., the benefit you get given that the game continues to period 1 and you are not a helper.

So, the formula becomes

$$\delta(2-\omega) > \frac{2c}{[(b\frac{\omega}{2-\omega}) + c]} \tag{1.61}$$

which rearranges to

$$\delta > \frac{2c}{[\omega(b-c)] + 2c} \tag{1.62}$$

Proof 2

At time t, all players apart from player i (the one under consideration now) are following strategy *A.1*. The helpee has not choices to make. So, we can see the alternatives for the helper. Your decision (as the helper) will determine if you'll be in GS in the next period(s) or not, until you'll be again a helper. What you care is therefore the benefit you can get in the (possible) next periods $t+1, t+2, .., t+\alpha$, where α is the number of periods in which you'll be a helpee before you'll be an helper again.

Two situations with two possible choices:

1. If the helpee is in *GS* and you do not help, the expected utility you get is 0: when you will be a helpee no one will help you because you didn't help when you were an helper.

   If the helpee is in *GS* and you decide to help, you incur in a loss c, the cost of helping at t. Moreover, there is a probability δ that the game continues to $t+1$. At the same time, conditional on the game continuing to period $t+1$, there is probability 1/2 that you are the helper again, so the analysis ends.

   Symmetrically, there is also probability 1/2 that you are the helpee: in this case there is the cumulative probability $\frac{\delta}{2}$ that you will be an helpee in the next period $t+1$ and you will get b. The same probability $(\frac{\delta}{2})$ will be taken into account for the other periods (we assume they could be infinite) discounted by the time period until you will be an helper again, to gain b. Therefore, your expected utility if you help is:

$$-c + \omega b \left[\frac{\delta}{2} + \left(\frac{\delta}{2}\right)^2 + ...\right] \tag{1.63}$$

   This expected utility is greater than zero if the following happen:

$$-c + \omega b \left[\frac{\delta}{2} + \left(\frac{\delta}{2}\right)^2 + ...\right] > 0 \tag{1.64}$$

   knowing that

$$\frac{\delta}{2} + \left(\frac{\delta}{2}\right)^2 + ... = \frac{\delta}{2-\delta} \tag{1.65}$$

   then

$$\begin{aligned} \omega b \left[\frac{\delta}{2-\delta}\right] &> c & (1.66) \\ \delta &> \frac{2c}{[\omega(b+c)] + 2c} & (1.67) \end{aligned}$$

   If the above result holds, then it is better to help than to defect if the helpee is in *GS*. Therefore, if δ is below the above threshold, the helping is not an equilibrium.

2. If the helpee is in BS, strategy *A.1* prescribes him not to help. His standing is uneffected by others strategy (they are playing A.1). Is it possible that he wants to deviate and help? That's impossible, in fact there is no reward for helping and he does not want to pay a cost for helping. He will not deviate and follow the strategy *A.1*.

1.A.2 Alternative Proof for Theorem 1.2

At time t, all players apart from player i (the one under consideration now) are following strategy *A.2*. The helpee has no choices to make. So, we can see the alternatives for the helper. Your decision will determine if you'll be in GS in the next period(s) or not, until you'll be again a helper. What you care is therefore the benefit you can get in the (possible) next periods $t+1, t+2, .., t+\alpha$, where α is the number of periods in which you'll be a helpee before you'll be an helper again.

Two situations with two possible choices:

1. If the helper is in GS but the opponent is not (he is in BS), the action used will not influence future payoffs. In fact, the best reply is to defect, without loosing the amount c (as in for strategy *A.1*).

2. In all the other combinations of matches (i.e., you are in BS or you are in GS and the helpee is in GS too), the helper should help the other to be in good standing immediately after the current period. If you decide to help, you incur in a loss c, the cost of helping at t. However, given that all the others are following the A.2 strategy, as soon as you are in GS as helpee you will always be helped. Therefore, every subsequent period in which you will be the helpee you will receive b. Therefore, the same mathematical proof used to study strategy *A.1*, i.e., studying $U_H > U_D$ applies here providing us the same results:
   if $\delta > \frac{2c}{[\omega(b+c)]+2c}$ than A.2 is a strict Nash equilibrium.

1.A.3 Population Dynamics for Strategy A.1 and Strategy A.2 under different Starting Points

Let us analyse different starting points.

If everyone at the start of the game is provided with a BS, the Sugden rule strategy $A.1$ induces a Nash equilibrium with no helping at all. So at time $t+1$, no help takes place, and the same results in all the possible rounds. In fact, $A.1$ prescribes helping only those who are in GS. Conversely, the Hammerstein rule strategy $A.2$ induces a Nash equilibrium with 100% helping, given that everyone would regain the GS status as soon as possible. The curious component is that in this period every helper helps, and at $t+1$, half of the population is in GS. Strategy $A.2$, in this particular case, leads people to help each other immediately and in all the following periods. This is evident from Figure 1.4.

Similarly, if everyone at the start of the game is provided with a GS, both strategies impose helping and are Nash equilibria. There is no incentive to deviate, given that all the other players are following the same strategy.

What happens if, in period t, some players (a fraction η) are in GS and some others are in BS (i.e., $1-\eta$)?

We would like to estimate not only the *status* of players period after period but also the helping behaviour of players in the game.

Let us start with strategy $A.1$.

Assuming that the population is large enough to allow us to use the Law of Large Numbers (LLN), the matches will be in the proportions:

GS helper, GS helpee: η^2. Helper helps and both players will be in GS in period $t+1$.

GS helper, BS helpee: $\eta(1-\eta)$. Helper does not help but he will be in GS in period $t+1$.

BS helper, GS helpee: $\eta(1-\eta)$. Helper helps and both players will be in GS in period $t+1$.

BS helper, BS helpee: $(1-\eta)^2$. Helper does not help and both players will be in BS in period $t+1$.

Therefore, in period $t+1$, the proportion of players in GS will be:

$$\begin{aligned}\eta_{t+1}^{A.1} = \eta^2 + \frac{\eta(1-\eta)}{2} + \eta(1-\eta) = \frac{2\eta^2 + 3\eta(1-\eta)}{2} = \\ = \frac{2\eta^2 + 3\eta - 3\eta^2}{2} = \frac{3\eta - \eta^2}{2} = \frac{\eta(3-\eta)}{2}\end{aligned} \tag{1.68}$$

Let us now consider the dynamics of strategy $A.2$.
Here, all the helpers will be in GS at $t+1$ given the prescription of the strategy. But let us analyse it more carefully starting at period t.
Assuming that the population is large enough to allow us to use the Law of Large Numbers (LLN), the matches will be in the proportions:
GS helper, GS helpee: η^2. Helper helps and both will be in GS in period $t+1$.
GS helper, BS helpee: $\eta(1-\eta)$. Helper does not help but he will be in GS in period 1. Helpee remains in BS in period $t+1$.
BS helper, GS helpee: $\eta(1-\eta)$. Helper helps and both will be in GS in period $t+1$.
BS helper, BS helpee: $(1-\eta)^2$. Helper helps and he will be in GS in period $t+1$.

Therefore, in period $t+1$, the proportion of players in GS will be:

$$\begin{aligned}\eta_{t+1}^{A.2} = \eta^2 + \frac{\eta(1-\eta)}{2} + \eta(1-\eta) + \frac{(1-\eta)^2}{2} = \frac{2\eta^2 + 3\eta(1-\eta) + (1-\eta)^2}{2} = \\ = \frac{2\eta^2 + 3\eta - 3\eta^2 + 1 - 2\eta + \eta^2}{2} = \frac{\eta+1}{2}\end{aligned} \tag{1.69}$$

Figure 1.4 compares the population dynamics of players in Good Standing (GS) under two strategies, $A.1$ (Sugden in red) and $A.2$ (Hammerstein in blue). The horizontal axis shows the proportion of players in GS at period t, and the vertical axis shows the proportion at period $t+1$. The 45° line indicates points where the proportion of GS remains unchanged from one period to the next. Under strategy $A.1$, any initial proportion in $(0,1)$ leads to a rise in GS each period, albeit at a rate converging to zero as t approaches infinity.

Under strategy $A.2$, however, the growth of GS occurs more rapidly. Indeed,

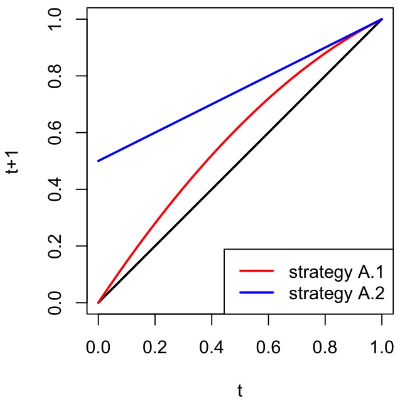


Figure 1.4: Dynamics of GS under strategies $A.1$ and $A.2$

when no one starts in GS ($\eta_t = 0$), the proportion immediately jumps to $\frac{1}{2}$. This difference arises because $A.2$ allows players in GS to help those in Bad Standing (BS), a possibility not afforded by $A.1$. Algebraically, one can compare

$$\eta_{t+1}^{A.1} = \frac{\eta_t(3 - \eta_t)}{2} \quad \text{and} \quad \eta_{t+1}^{A.2} = \frac{\eta_t + 1}{2},$$

and show that $\eta_{t+1}^{A.2} > \eta_{t+1}^{A.1}$ whenever $\eta_t \neq 1$. Both strategies nevertheless converge to $\eta_{t+1} = 1$ when $\eta_t = 1$, as substituting $\eta_t = 1$ into either equation yields an equilibrium at full GS. However, for any intermediate starting proportion η_t, the Hammerstein function (blue curve) exceeds the Sugden function (red curve), indicating that $A.2$ always generates a larger share of GS by the next period.

1.A.4 The Impact of Ability on Helping Behaviour - Heterogeneous Costs

We can further refine our model by mapping the ability to help directly to the cost of helping. Specifically, we want that an individual's ability inversely affects the cost they incur when providing help. That is, individuals with higher ability levels face lower costs when helping, while those with lower ability levels face higher costs. This relationship captures the notion that more capable individuals can help others more efficiently or with less effort, resulting in a lower personal cost.

We define ability as the inherent capacity to help when requested. This capacity is formally conveyed by the cost of helping. For simplicity, we consider a dichotomous population: individuals are either able (they know how to help) or disabled (they cannot help)[10].

An important factor in analysing games with asymmetric information is whether certain information is public (common knowledge) or private. Given the structure of players' statuses and the way they are updated from one period to the next, knowing who is able and who is not impacts the game's dynamics. Therefore, we find it essential to analyse both scenarios-where disability is public information and where it is private-and examine their implications on the game's equilibria.

We assume that the population size N is even and that all players are matched in each period. Suppose now that a fraction ϵ of the entire population is disabled and hence cannot help.

1.A.4.1 (Dis-)ability as Public Information: Honorary Good Standing

In this scenario, since individuals know who is incapable of helping, we assume that the disabled players maintain their GS (good standing) status throughout the game—they have what we call an "honorary GS". To find the critical values for which strategy $A.1$ (and similarly $A.2$) constitutes a Nash equilibrium, we employ the same approach as before.

We analyse the expected utility under strategy $A.1$. If an able individual is the helper at period t, they incur a cost c regardless of whether the helpee is able or disabled, since the strategy prescribes helping those in GS, and the disabled are always in GS. At $t+1$, with probability $\frac{1}{2}$, the individual will be a helper again and may incur an expected cost of $\frac{c}{2}$. With probability $\frac{1}{2}$, they will be a helpee and receive a benefit b only if matched with an able helper, which occurs with probability $1-\epsilon$. Therefore, as long as the game continues and the individual can help, the expected utility U_H becomes:

$$U_H = -c + \delta \left[\frac{1}{2} \left((1-\epsilon) b - c \right) \right] \frac{1}{1-\delta} \tag{1.70}$$

where $\delta \in (0,1)$ is the discount factor representing the probability that the game

[10] We use the terms "able" and "ability" although "expert" and "expertise" could express a similar meaning.

continues.

Let U_D denote the expected payoff from always defecting (never helping). Clearly, $U_D = 0$. Therefore, $U_H > U_D$ if:

$$-c + \delta \left[\frac{1}{2} \left((1-\epsilon)b - c \right) \right] \frac{1}{1-\delta} > 0 \tag{1.71}$$

Solving inequality (1.71) for δ, we obtain:

$$\delta > \frac{2c}{(1-\epsilon)b + c} \tag{1.72}$$

This condition is stricter than in the homogeneous ability case, indicating that heterogeneity in abilities raises the threshold discount factor δ required the strategy to be a Nash equilibrium stricter.

1.A.4.2 (Dis-)ability as Private Information

When disability is private information, the equilibrium conditions become even more stringent. We consider an able helper and assume that disabled individuals are in BS (Bad Standing) because they cannot help. Strategy $A.1$ requires players to help only those who are in GS.

As an able helper, the individual incurs a cost c when helping a helpee in GS, who is able with probability $1-\epsilon$. At $t+1$, with probability $\frac{1}{2}$, the individual will be a helper and may incur an expected cost of $\frac{c}{2}$. With probability $\frac{1}{2}$, they will be a helpee and receive a benefit b only if matched with an able helper, which occurs with probability $1-\epsilon$. Therefore, the expected utility U_H becomes:

$$U_H = -c + \delta \left[\frac{1}{2}(1-\epsilon)(b-c) \right] \frac{1}{1-\delta} \tag{1.73}$$

Again, $U_D = 0$, so $U_H > U_D$ if:

$$-c + \delta \left[\frac{1}{2}(1-\epsilon)(b-c) \right] \frac{1}{1-\delta} > 0 \tag{1.74}$$

Solving inequality (1.74) for δ, we obtain:

$$\delta \quad > \quad \frac{2c}{(1-\epsilon)(b-c)+2c} \tag{1.75}$$

This condition is even more restrictive than when disability is public information, reflecting the additional uncertainty introduced by private information.

We observe that this result resembles the case where some individuals are unmatched, with $(1-\epsilon)$ analogous to the matching probability ω. Games where one or both players are disabled have payoffs equivalent to those where players are unmatched.

In a game where all players are matched but some are disabled and remain in BS, strategy $A.1$ is a Nash equilibrium if and only if:

$$\delta > \frac{2c}{(1-\epsilon)(b-c)+2c} \tag{1.76}$$

A similar analysis applies to strategy $A.2$, and we derive analogous results by substituting $A.1$ with $A.2$ in the equations above.

Introducing heterogeneity in abilities affects the equilibrium strategies by making the requirements for the strategy to support helping behaviour stricter.

Chapter 2

Cooperation in the Helping Game: Image Scoring or Good Standing?

Using a lab experiment, we investigate the efficiency of two often studied reputation-based mechanisms — Image Scoring (IS) and Good Standing (GS) — in promoting cooperation in helping games. Both mechanisms assign a score to each individual, updating each time a choice is made. Under IS, a player's score (0 to 5) reflects their past help, updating only on the individual's choices. Under GS, a player's score (0 or 1) reflects both their help and who they helped, considering the individual's choices and the recipient score. First, we suggest that the centralised and recursive nature of GS encapsulates the entire history of interactions, departing from prior interpretations. Second, we conjecture that GS incentivises differentiation between "justified punishers" (those who do not help non-helpers) and "unjustified non-helpers" (those who do not help despite others' help). Our findings support our conjecture, highlighting reciprocal helping behaviour. However, IS leads to higher overall cooperation rates, regardless of reputation.

2.1 Introduction

Cooperation is an important part of everyday life and essential for achieving good societal goals. For instance, students helping each other with their coursework or a neighbour who watches your dog when you are on vacation exemplify cooperation on a small scale. At a larger level, cooperation between different countries can contribute to better economic outcomes, such as trade agreements, environmental policies, and global health initiatives. However, since groups sometimes cooperate and sometimes do not, understanding the conditions that support and foster cooperation is essential for economists, particularly in the context of institutional design.

Often, the short-term individual costs of cooperation exceed the benefits, limiting cooperation in one-shot interactions (Rand and Nowak, 2013). However, when individuals interact repeatedly with the same group, cooperation typically emerges (Trivers, 1971; Axelrod, 1984). This phenomenon can be explained by direct reciprocity, where the recipient of a helpful decision reciprocates by helping the individual who helped them in the past. For example, if Alice helps Bob, and then Bob reciprocates by helping Alice, the cumulative benefit of cooperation can outweigh the costs. The Folk Theorem (Friedman, 1971; Fudenberg and Maskin, 1986) provides a formal support for this: in repeated bilateral (or multilateral) interactions where decisions are observable, reciprocal behaviour can emerge as an equilibrium outcome, sustained by a variety of strategies, as long as interactions are sufficiently frequent and/or agents are sufficiently patient.

When individuals seldom meet the same partners more than once, as is the case in large populations, supporting cooperation becomes more challenging. While it is possible to construct schemes for enforcing cooperation as equilibria in these situations, these schemes often have implausible features. Moreover, such cooperation cannot be explained by direct reciprocity alone. Instead, it can be sustained through indirect reciprocity (Alexander, 1987), where the reciprocation of a helpful action comes, this time, from someone other than the original recipient of the help. For example, if Alice helps Bob, and then Caesar, who has information about this, helps Alice in return.

Several different types of indirect reciprocity mechanisms have been proposed.

They can be categorised into preference-based and reputation-based mechanisms. Researchers have primarily focused on preference-based mechanisms, proposing that reciprocity preferences can explain why prosocial behaviour is sometimes sustained in such settings. Some researchers define reciprocity as (reciprocal) fairness, where an individual is kind to those they perceive as having been kind to others (e.g., Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Hopp and Süß, 2024,[1]). Others describe it as (reciprocal) altruism, where an individual acts altruistically towards those they perceive as generally altruistic (Levine, 1998). We refer to these as "reciprocity preferences" in this chapter. Both fairness and altruism mechanisms typically rely on information about past behaviours to assess whether others have been kind or altruistic.

In contrast, we note that there is another class of mechanisms that can sustain prosocial behaviour without appealing to preferences. These are reputation-based mechanisms, which specifically focus on the structured use of shared information about individuals' past behaviours to sustain cooperation. Two main reputation-based mechanisms have emerged in this context: "Image Scoring" (IS) (Nowak and Sigmund, 1998a,b) and "Good Standing" (GS) (Sugden, 1986; Leimar and Hammerstein, 2001). IS and GS are both mechanisms (or information structures) for encoding and publicly displaying individuals' past behaviours in environments where mutually beneficial cooperation is supported by indirect reciprocity.

Reputation-based mechanisms like IS and GS are simplified or idealised rules that capture the essence of how reputation can influence cooperation in a way that is theoretically and experimentally tractable. These mechanisms are particularly useful in controlled environments where the goal is to understand which features of a scoring system might be most effective in encouraging efficient outcomes.

In real-world contexts, while no system may exactly mirror the IS or GS mechanisms, similar reputation-based systems are frequently employed. For instance, various institutions, such as voluntary organisations, trade unions, pressure groups, and social media platforms that use ratings and likes, can facilitate the recording and dissemination of reputation information. These systems do not require punitive or rewarding powers beyond updating reputations, thereby supporting cooperative

[1] An extension of the approach of Dufwenberg and Kirchsteiger (2004) to modelling direct reciprocity in the context of indirect reciprocity.

behaviour in practical settings.

These reputation-based mechanisms have potential practical applications in a wide range of economic contexts, where centralised reputation systems can significantly impact behaviour. For example, in crowd-sourcing platforms, contributions can be assessed based on subscribers' reputation scores, incentivising high-quality inputs. In e-commerce, these scores help evaluate the trustworthiness of sellers and buyers, improving transactional reliability[2]. Knowledge-sharing networks, such as online forums, also benefit from these mechanisms, because answering questions entails a cost (time) yet yields a benefit (satisfaction of helping others), thereby fostering higher participation and cooperation. Therefore, while IS and GS are theoretical constructs, they offer valuable insights for designing systems that enhance cooperative behaviours across various sectors.

The IS mechanism assigns each individual an observable numerical score that updates with each cooperative decision: scores can increase when an individual cooperates and may decrease when they do not. Because these scores are visible to all, they form a reputation system in which people can base their cooperation decisions on others' scores.

However, IS captures only (a certain number of) first-order information — namely, whether someone cooperated or not — without distinguishing between "justified" and "unjustified" non-cooperation. "Justified" non-cooperation occurs when an individual refuses to cooperate with someone who previously did not cooperate with others, while "unjustified" non-cooperation occurs when an individual refuses to cooperate despite the other party's good cooperative history. This distinction matters because it affects perceptions of "deservingness,"[3] yet IS cannot account for it due to its limited information.

Under IS, the indirect reciprocal strategy for universal cooperation is to "*cooperate with those who have high scores.*" Hence, individuals are incentivized to maintain high scores by consistently cooperating, in the expectation that others will reciprocate in future interactions.

The GS mechanism assigns each individual a binary "status" — either "good

[2] For example, see Keser (2003), and Wibral (2015).

[3] "Deservingness" refers to whether individuals are deemed worthy of receiving help, based on their past behaviour.

standing" or "bad standing" — that reflects an individual's past cooperative behaviours. Similar to IS, each individual carries an observable numerical score that updates every time they make a choice. Differently from IS, the updating rule of the GS mechanism is not only based on the individual's behaviour but also on the players' status (second-order information). This recursive feature captures a richer history of interactions, enabling GS to more accurately track cooperative tendencies and distinguish between justified and unjustified non-cooperation.

Under GS, the indirect reciprocal strategy for universal cooperation is to "*cooperate with others who are in good standing and refuse cooperation to those who are in bad standing.*" This rule incentivizes individuals to maintain good standing by consistently cooperating, because only those with good standing can expect reciprocal cooperation in future interactions.

IS and GS have been theoretically explored in the so-called "helping game"(Nowak and Sigmund, 1998a,b). The game consists of a large population of players who interact repeatedly across many periods. Players are randomly matched in pairs during each period, with one randomly selected as the "active player" and the other as the potential receiver (named the "non-active player") of a helping decision[4]. The active player has a choice to make: whether to help the non-active player or not. If the active player chooses to help, they incur a cost c and confer a benefit b to the non-active player ($b > c > 0$). The benefit is only realised and the cost is only incurred if the active player chooses to help. Not helping results in the "status quo ex ante" being maintained. The receiver, or non-active player, does not have any choice to make.

The helping game has been experimentally studied to understand the dynamics of indirect reciprocity (Seinen and Schram, 2006; Engelmann and Fischbacher, 2008, 2009). Existing evidence suggests that IS can lead to high cooperation rates in the helping game, and it has therefore been considered a robust explanation for indirect reciprocity. However, IS has not been rigorously compared against alternative rating or scoring systems, which leaves us uncertain about the reasons that make IS effective. Moreover, as discussed earlier, IS does not provide information on

[4] In previous literature, these roles have been referred to by various names: "helper" and "recipient" or "helpee" (e.g., Erkut and Reuben, 2023); "donor" and "recipient" (e.g., Seinen and Schram, 2006); "mover" and "recipient" (e.g., Bolton et al., 2005); or "buyer" and "seller" (e.g., Camera and Casari, 2018). We refer to the same roles.

whether a non-active player's prior non-helping was justified or unjustified, a critical element in effectively implementing the logic of indirect reciprocity. This logic encompasses both a moral problem and an incentive problem. The moral problem arises from the principle of reciprocity itself, which suggests that people who do not help others do not deserve to be helped. The incentive problem, on the other hand, stems from the expectation that non-helpers will be punished; however, under IS, punishment is costly because it reduces the punisher's score. The advantage of GS is that it addresses these issues by ensuring that punishing non-helpers benefits the punisher, thereby better aligning with the principles of indirect reciprocity.

Our experiment introduces two important innovations: adjusting the information provided to individuals and testing a heterogeneous cost game. By varying the information about previous cooperativeness that individuals receive, we can differentiate between the IS and GS mechanisms and identify which mechanism is more successful in inducing indirect reciprocity. Since different mechanisms require distinct types of information, we modify the information each player receives to assess which mechanism is more effective in fostering cooperation. Specifically, our GS mechanism internally encodes higher order information into a binary score — 0 ("bad") or 1 ("good") — providing a simplified yet comprehensive representation of a player's cooperative behaviour. This approach departs from previous studies that claim to investigate GS by providing first- and second-order information separately (e.g., Wedekind and Milinski, 2000; Milinski et al., 2001; Bolton et al., 2005).

Additionally, a novel aspect of our study is the introduction of a heterogeneous cost game, which better reflects the complexities of real-world scenarios where individuals may face varying costs in helping others. In many social contexts, cooperation is not uniform across all members of a population; instead, it may be concentrated within certain "clubs" or subgroups where the costs and benefits of helping align more favourably. For example, in a population with heterogeneous costs, a small group of individuals with low costs may find it beneficial to help one another consistently, forming a cooperative club. However, sustaining such cooperation across the entire population, especially when some individuals face higher costs, requires a robust signalling mechanism.

In this context, Image Scoring (IS), which relies on helping others based on the number of times they have helped, may struggle to sustain cooperation among individuals with different costs. This is because IS does not adequately signal whether an individual with a low score is genuinely non cooperative or simply have the high cost of helping. On the other hand, Good Standing (GS), which updates a player's status based on both first-order and second-order information, can effectively signal membership in a cooperative club. In such a club, low-cost individuals help each other, while high-cost individuals neither help nor receive help. This arrangement requires a public marker indicating club membership. The GS mechanism provides such a marker endogenously, allowing individuals to opt in or out based on their cooperative behaviour.

In our heterogeneous cost game, each player is assigned either a high (c_H) or low (c_L) cost, where $c_H > c_L$. These costs remain constant throughout the game, with players being either high-cost or low-cost individuals. If an active player chooses to help, they incur their assigned cost (c_L or c_H) and confer a benefit b to the non-active player, where $b > c_H > c_L > 0$. The benefit is only realised and the assigned cost is only incurred if the active player chooses to help. Not helping results in the status quo ex ante being maintained. The receiver, or non-active player, does not have any choice to make. This setup allows us to explore how GS, compared to IS, might better support the emergence and stability of cooperation in populations with varying costs, reflecting the real-world challenges of sustaining prosocial behaviour in diverse groups.

To the best of our knowledge, this chapter is the first: (i) to investigate the differences between GS and IS in the helping game and (ii) to examine the differences in the responses of GS and IS in heterogeneous cost games.

There are two main research questions: (i) which mechanism is the most efficient in the helping game? (ii) do the mechanisms differ in their levels of reciprocity? To address the first question, we examine the frequency of helping, i.e., how much individuals help. For the second question, we operationalise "reciprocity" in terms of how one's propensity to help differs as a function of the non-active player's score. These questions will be analysed in both the homogeneous and heterogeneous cost games to provide a comprehensive understanding of the effectiveness and differences between the IS and GS mechanisms across different cost games.

We find that IS produces more cooperation, particularly in the homogeneous cost game. However, GS is more capable of supporting reciprocal helping, suggesting that GS may encourage more reciprocal behaviour than IS.

The rest of this chapter is structured as follows: we describe the literature in Section 2.2. Section 2.3 provides some theoretical considerations. Section 2.4 introduces our experimental design and procedure. Section 2.5 presents our conjectures and Section 2.6 shows our results. Section 2.7 concludes and offers some insights for future research.

2.2 Literature Review

In recent decades, significant effort has been made to understand the mechanisms underlying cooperation, particularly reciprocity. Social scientists have developed various game theoretic models to explore these mechanisms, providing valuable insights into how and why individuals cooperate. These models have been instrumental in explaining both direct and indirect reciprocity, highlighting the importance of structured interactions and information-sharing in fostering cooperative behaviour.

Reciprocity has been identified as a key determinant of individual behaviour in several experimental settings, including public goods games (Brandts et al., 2000; Greiff and Paetzel, 2020), prisoners' dilemma games (Andreoni and Miller, 1993; Cooper et al., 1996; Gong and Yang, 2019), centipede games (McKelvey and Palfrey, 1992), trust (investment) games (Berg et al., 1995; Charness et al., 2011), and gift exchange games (Fehr et al., 1997; Irlenbusch and Sliwka, 2005). The findings from these experiments consistently show that reciprocity is stable and robust, and that it can be effectively found in the laboratory.

Scholars have emphasised that strategic reputation plays a crucial role in the occurrence of reciprocity. Strategic reputation refers to the way individuals adjust their behaviour to build a reputation that encourages others to reciprocate. This concept is supported by theoretical models (Trivers, 1971; Axelrod, 1984) and empirical evidence, suggesting that even individuals who might not typically cooperate may do so to increase their chances of future reciprocation.

The motivation for reciprocal behaviour in economics has been linked to preferences that extend beyond self-interest. Most research has viewed reciprocity as an intrinsic motivation rather than a purely selfish behaviour. In particular, attempts have been made to define reciprocity as (reciprocal) fairness (e.g., Rabin, 1993; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), (reciprocal) altruism (Trivers, 1971; Levine, 1998; Dufwenberg and Kirchsteiger, 2004), or the pursuit of efficiency gains through cooperation (Brandts et al., 2000).

A game-setting that allows for the investigation of reciprocity and the importance of strategic reputation is the so-called "helping game". The helping game is a form of game used to study cooperative behaviour, particularly indirect reciprocity. In the game, individuals are randomly matched in pairs each period, with one serving as the active player and the other as the non-active player. The active player is given a set amount of money and has the possibility to incur a cost to provide a benefit to the non-active player. The non-active player does not make any decisions during the game.

Our main contribution is to enrich the experimental literature on helping games by expanding on three seminal studies: Bolton et al. (2005), Seinen and Schram (2006), and Engelmann and Fischbacher (2009). Each of these works experimentally examines reputational mechanisms akin to ours within a helping-game framework.

Bolton et al. (2005) (BKO) analyse the impact of first and second-order information on cooperation in a finite helping game by testing two mechanisms. The first mechanism incorporates only first-order information: the active player is informed about whether the non-active player helped in their last active period. This is similar to an image score with a one-period window. The second mechanism incorporates both first and second-order information: in addition to being informed about the non-active player's most recent decision, the active player also learns about the decision of the non-active player's previous partner.

BKO's findings suggest that second-order information can significantly impact cooperation. They argue that understanding whether a recipient's past decision is justified or not based on their partner's previous choice helps players make more informed decisions about whom to help and sustain cooperation. Their study highlights the importance of second-order information in fostering cooperation, which aligns with the theoretical basis of our GS mechanism.

The BKO second-order information differs from our GS in two main ways: (i) It uses two levels of information to create a player's score for a period, considering both the player's behaviour and the recipient's previous behaviour. (ii) It uses a binary score for each level. While this approach incorporates more levels than our GS, it provides less detailed information at the first level. Our GS mechanism improves upon this by compressing information about the entire history of interactions into a single binary score, thus providing a comprehensive and easily interpretable measure of cooperative behaviour.

Seinen and Schram (2006) (SS) investigate indirect reciprocity through an experimental helping game, where players are randomly assigned to the roles of active and non-active players. This experiment initiated the design that our study adopts and builds upon. Each player's previous six decisions as an active player are recorded and presented as an image score. SS compare the image scoring mechanism to a no-information condition, creating two treatments (IS vs. no information). They also vary the cost of helping, resulting in two cost treatments (high cost vs. low cost).

They find that helping behaviour is higher when the cost of helping is lower and when information about past cooperativeness is provided. Their findings also indicate that when such information is provided, active players tend to base their decisions to help on the non-active player's image score, confirming the presence of indirect reciprocity. However, their design cannot clearly distinguish to what extent cooperation is driven by reciprocity preferences (i.e., non-self-interested preferences) or by self-interest. This ambiguity arises because any decision that appears indirectly reciprocal also influences the active player's own reputation. Therefore, when an active player helps a non-active player with a high score, it remains unclear whether this choice is motivated by a desire to reward the non-active player or to boost their own reputation.

To disentangle these issues, Engelmann and Fischbacher (2009) (EF) extend SS's analysis by designing two treatments: one where individuals have a "private score" and another where they have a "public score". This setup allows for a focused examination of non-strategic reputation building. The score reflects the number of help decisions made by the non-active player in their last five periods as an active player. Public scores are visible to all active players, while private scores are not.

This design enables EF to compare the behaviour of active players with and without publicly visible scores when they encounter non-active opponents. When scores are private, active players have no self-interested incentive to help non-active players, since the helper's decision does not affect a publicly observed reputation (in fact, they cannot even see the opponent's score). As a result, the design isolates "pure reciprocity," in which helping decisions are motivated by a genuine desire to assist others, rather than by strategic concerns[5].

EF find a considerable amount of cooperation in the private score treatment and significantly more in the public score treatment. This seems to indicate that the visibility of active players' past decisions influences cooperative behaviour. Additionally, in both treatments, the probability of being helped increases with the non-active player's score. They also observe that players in the public score treatment maintain scores around 4, which is close to the self-interested optimum, while directing their "do not help" decisions at low-score non-active players. Approximately half of the helpful decisions in the public score treatment are driven by reputation-building motives, while the other half are attributed to reciprocity preferences.

There are two relevant differences between IS and GS. First, we argue that IS model may not fully capture indirect reciprocity in real-world scenarios because it does not account for the full history of interactions, which our GS mechanism encapsulates more comprehensively. Second, EF's explanation of public-score behaviour implies some degree of non-self-interested preference because an entirely self-interested player could maintain their score closer to the optimum by ignoring non-active players' scores. However, EF's results also suggest that strategic self-interest plays a role. This highlights the second difference between IS and GS: the IS mechanism requires some level of non-self-interested preference to induce indirect reciprocity, while the GS mechanism can function effectively without relying on non-self-interested preferences.

More in general, our study builds on these foundational works by further exploring the IS mechanism and introducing two novel aspects: the GS mechanism and an original heterogeneous cost condition. While previous studies primarily focus on homogeneous cost games, our inclusion of heterogeneous costs reflects more realistic scenarios where individuals face varying costs in helping others. This differentiation

[5] Engelmann and Fischbacher (2008) also present a model of entirely self-interested behaviour related to a simplified IS mechanism. In this model, differently from the experiment, when image scores are updated, the decision to be deleted is randomly picked from the five in the record, not necessarily the "oldest" one. They find equilibria in which active players are indifferent between adjacent scores, but these equilibria do not match the behaviour observed in their experiments.

is crucial as it affects the dynamics of cooperation and the applicability of GS and IS mechanisms.

Other significant contributions to the study of indirect reciprocity include the works of Wedekind and Milinski (2000) and Milinski et al. (2001). Wedekind and Milinski (2000) were the first to experimentally test the Nowak and Sigmund's model of indirect reciprocity, conducting their study over six periods. They study two treatments: in the information treatment, players have first-order information about the helping behaviour of their partner. Specifically, active players receive a detailed history of the non-active player's past decisions, displayed as a sequence of "H" for help and "N" for no help, representing whether the non-active player had helped or not in previous periods. The other treatment provides players with no information. Their findings indicate that cooperation increases when active players are aware of the non-active player's history of helping.

Milinski et al. (2001) extend their initial experiment by testing not only "image scoring" but also their intuition of "good standing" [6]. In their study, as in BKO, the treatment that studies the good standing provides not only first-order information but also second-order information, indicating whether the non-active player's behaviour was directed towards a recipient who had previously helped or not. They find limited evidence supporting good standing.

Our mechanisms differ from those tested by Milinski et al. (2001). Using similar reasoning to BKO, Milinski et al.'s interpretation of GS assumes that individuals need to remember a large amount of information about others' previous interactions. Their approach involves detailed tracking of first-order and second-order information separately. Therefore, the same differences highlighted previously between our GS and BKO's second-order information apply here as well, even more so since, unlike BKO, Milinski et al. do not limit the information to the last three interactions. We believe our understanding of GS and design summarises and simplifies this information provision to offer a better measure of cooperative behaviour.

Our study also takes inspiration from Swakman et al. (2016), who examine the value players place on second-order information in a repeated helping game. In their design, active players are endowed with non-active players' first-order information, which includes their previous three decisions as active players. In one treatment

[6] For further elaboration see Section 2.3.

of their experiment, each active player can request second-order information about non-active player for free. In another treatment, active players can purchase this information at a cost. Their findings indicate that second-order information is frequently sought after, particularly when first-order information reveals mixed helping behaviours. This preference for deeper insight into the motivations behind decisions supports the use of a comprehensive GS mechanism that encompasses higher order information.

Additionally, Ule et al. (2009) explore the role of punishment in helping games. By allowing active players to punish non-active players, they discover that active players are more inclined to punish when this action diminishes the earnings of the non-active players. Their helping game differs from ours because the action set is larger, allowing players not only to help or not help but also to punish. Although our study does not focus on punishment per se, punishment is embedded in the updating rule of GS mechanism, as it indirectly penalises non-helping decisions towards "good" players by affecting their reputation.

Finally, our research contributes to the broader experimental literature on indefinitely repeated games, particularly those exploring the conditions under which cooperation can be sustained. This body of work includes seminal contributions from Palfrey and Rosenthal (1994), Dal Bó (2005), Engle-Warnick and Slonim (2006), Aoyagi and Fréchette (2009), Camera and Casari (2009), Fudenberg et al. (2012), Arechar et al. (2017), Fréchette and Yuksel (2017), Camera and Casari (2018), Aoyagi et al. (2019), and Ghidoni and Suetens (2022). Notably, Bigoni et al. (2020) (BCC) investigate two record-keeping mechanisms in an indefinitely repeated helping game using two treatments: the "Memory treatment" and the "Money treatment". Both mechanisms are symmetric in their approach to tracking and incentivising cooperative behaviour, though they differ in how they are implemented.

In the BCC setup, the payoffs are structured such that if the active player chooses not to help, they receive a payoff of 6, while the non-active player receives 4. If the active player chooses to help, their payoff drops to 0, incurring a cost, while the non-active player receives a payoff of 20. This structure implies that the cost of helping for the active player is 6, and the benefit to the non-active player is 16, resulting in a cost/benefit ratio of 0.375.

In the Memory treatment, players alternate between the roles of active and non-active players. Each player starts with a "balance", initially set at 1 for non-active players and 0 for active players in the first period. The balance increases by 1 if a player helps as an active player and decreases by 1 if they receive help as a non-active player. Players are informed whether their recipient's balance is positive ($\geq$ 1) or not ($\leq$ 0), with a positive balance signalling that the player has helped more times than they have been helped[7]. Unlike IS, the total balance across all players remains constant, averaging 0.5 per player. This system ensures that balances shift dynamically in response to players' actions, with deviations from cooperative behaviour reflected in a negative or zero balance.

BCC describe a strategy supporting a helping equilibrium in the Memory treatment, where active players help only if the non-active player's balance is positive. This strategy, similar to the our GS strategy, highlights a conditional cooperation mechanism that effectively maintains cooperation in their setting.

In the Money treatment, a fixed supply of tokens can be transferred between players. Initially, each non-active player starts with one token, while active players have none. During each period, the non-active player can choose to keep their token, transfer one to the active player, transfer one conditionally on receiving help, or transfer one conditionally on not receiving help — provided they have at least one token. The active player, in turn, can choose to help or not and can also choose to help conditionally on receiving either one or no tokens. Importantly, the active player is informed whether the non-active player has any tokens available. This treatment effectively induces helping behaviour by allowing players to condition their actions on the actions of others, thereby mimicking a monetary exchange system. In this context, possessing at least one token during rounds as a non-active player serves as an analogue to having good standing in the traditional helping game. This mechanism illustrates how conditioning actions on others' behaviour can promote cooperation, even though it diverges from the traditional helping game structure.

BCC's findings underscore the importance of balance or record-keeping mechanisms in sustaining cooperation. The key distinction from our GS mechanism is that our mechanism does not require strict alternation of roles or a constant total

[7] From a GS perspective, indicating "good standing" or "deservingness".

of balances, which provides a fixed reference point for evaluating cooperative behaviour. Our GS mechanism offers a more flexible and comprehensive approach by encapsulating the entire history of interactions into a single binary score, potentially accommodating heterogeneous costs and dynamically adjusting to changing conditions.

A key contribution of our study is to provide further evidence on the effectiveness of GS and IS mechanisms in promoting indirect reciprocity while incorporating more realistic scenarios through heterogeneous costs. This approach builds on and extends the findings of previous experimental studies, offering new insights into the dynamics of cooperation and reputation in economic interactions.

2.3 Theoretical Considerations

In this section, we focus on providing some theoretical considerations derived from our understanding of Image Scoring (IS) and Good Standing (GS) and their comparison with previous research. As discussed in Section 2.2, past research has tested first-order and second-order information separately, without fully capturing the potential of the GS mechanism as a single score-based system. In particular, previous studies have effectively tested similar systems only in the context of IS.

In our study, we consider IS and GS as score-based mechanisms that publicly provide information about players' past behaviours. These scores are automatically updated by the mechanisms themselves following specific updating rules. IS keeps a record of first-order information, which tracks whether an individual has helped in their past interactions. The updating rule of the score depends not only on the current choice but also on the "oldest" choice in the record, which is being removed. Specifically, the numerical score increases by one if a player helps and their choice replaces a "do not help" choice, it decreases by one if they do not help and their choice replaces a "help" choice, or it does not change if the new and the old choices are the same.

GS, on the other hand, incorporates both first-order and second-order information. First-order information involves the immediate past behaviour of a player, while second-order information includes whether that behaviour was directed towards a player who had previously helped others. In the literature, Sugden (1986) and Leimar and Hammerstein (2001) propose two mechanisms that differ in their updating rules and the strategies they endorse.

Sugden's updating rule operates as follows: (1) if an active player is in good standing at the start of a round, they remain in good standing unless they fail to help a non-active player who is also in good standing — if they do not help, they fall into bad standing; (2) if an active player starts the round in bad standing, they stay in bad standing unless they help a non-active player who is in good standing — in which case, they regain good standing. Consequently, Sugden's strategy is to "cooperate with others who have good standing and do not cooperate with those who have bad standing", driven by self-interest to maintain or regain good standing.

Leimar and Hammerstein's updating rule operates as follows: (1) if an active player is in good standing at the start of a round, they remain in good standing unless they fail to help a non-active player who is also in good standing - if they do not help, they fall into bad standing; (2) if an active player starts the round in bad standing, they regain good standing if they help any other non-active player, regardless of that player's standing. Unlike Sugden's rule, which only allows moving from bad standing to good standing when helping another player in good standing, Leimar and Hammerstein's rule is more lenient. On the equilibrium path, Leimar and Hammerstein's strategy is the same as Sugden's one: "cooperate with others who have good standing and do not cooperate with those who have bad standing". The key difference lies off the equilibrium path; if a player is in bad standing, the Leimar and Hammerstein's strategy is to help *irrespective* of the standing of the recipient in order to regain good standing.

To visually understand the difference in the updating rules, please refer to Table 2.1, which provides a detailed representation of how active player score is updated.

Table 2.1: Updating rule differences

	Leimar and Hammerstein		Sugden	
Score at t	Help	Do not help	Help	Do not help
1, 1	1	0	1	0
1, 0	1	1	1	1
0, 1	1	0	1	0
0, 0	**1**	0	**0**	0

Notes: 1st column: active player, non-active player's score. 2nd & 4th columns: score of the active player at $t+1$ if they help. 3rd & 5th columns: score of the active player at $t+1$ if they do not help. Remember: the score of the non-active player does not change.

In this chapter, we focus on Leimar and Hammerstein's GS mechanism for two reasons. Unlike Sugden's GS mechanism, Hammerstein's allows easy recovery from a bad to a good standing; a player simply needs to help in one round to regain a good standing. Additionally, this mechanism avoids the issue of an absorbing state where all players in the population have a bad standing and cannot improve their standing, which could potentially stall cooperation entirely.

A critical aspect of our study is our interpretation of the GS mechanism as a centralised score-based system. This version contrasts with previous studies that considered the provision of first- and second-order information separately and criticised the GS mechanism. For instance, Milinski et al. (2001) test a "standing strategy" that should resemble GS but, to our eyes, misinterpret it. They assume this strategy required individuals to remember extensive information about all their past decisions. Hence, they criticised it for demanding excessive "memory capacity", interpreting this as the amount of information an individual must remember about others' past interactions. However, this critique, according to us, overlooks the actual GS model.

The GS mechanism only requires individuals to remember a single score for each other individual — their current standing[8]. We argue that the GS mechanism's updating rule ensures that the standing encapsulates a complete history of interactions in a single binary marker, thus efficiently summarising the necessary information without overburdening the individuals' memory capacity. This distinction is crucial for understanding the feasibility and effectiveness of the GS mechanism in real-world applications.

The recursive nature of the GS updating rule stems from the continuous reassessment of each player's standing based on their previous choice and towards who that choice was made (someone "good" or someone "bad"). This recursive process allows GS to encapsulate a comprehensive history of interactions within a single binary marker, making it more informative than IS. Moreover, the binary nature of the GS mechanism simplifies the decision-making process, particularly by avoiding the complexity of adjusting acceptable "high" scores based on the population's behaviour.

[8] In our experiment, a centralised system keeps track of them.

Based on these theoretical distinctions, we think it is very important to compare the effectiveness of IS and GS in a setting similar to those used by Seinen and Schram (2006) and Engelmann and Fischbacher (2009). We hypothesise that the recursive structure of GS offers two potential advantages over IS in supporting indirect reciprocity:

1. *Encoding Higher-Order Information*: IS captures only first-order information that allow partial justification for an individual's decision (e.g., Alice's reasoning might be: "I did not help Bob because Bob did not help Caesar"). This approach does not account for higher-order information that could allow a more comprehensive justification for an individual's decision (from the previous example, "I did not help Bob because Bob did not help Caesar, even though Caesar had helped Daniel"). In contrast, the recursive structure of GS allows a single binary marker to encapsulate the properties of an indefinitely long history. This means that GS can better reflect the complexity of social interactions and the rationale behind cooperative decisions, providing a more robust framework for indirect reciprocity.

2. *Facilitating Club Formation in Heterogeneous Cost Populations*: In populations where the costs of helping vary among individuals, Pareto-improving cooperation can be feasible within a "club" structure. In such a structure, low-cost individuals help each other, while high-cost individuals neither help nor receive help. This arrangement requires a public marker indicating club membership. The GS mechanism provides such a marker endogenously, allowing individuals to opt in or out based on their cooperative behaviour. Therefore, GS facilitates the formation of clubs, whereas IS does not. This feature is particularly important in heterogeneous cost condition, where individuals do not need to know the distribution of costs to make informed decisions about whom to help.

In summary, our theoretical considerations suggest that GS, with its recursive structure and ability to encapsulate higher-order information, provides a more effective mechanism for supporting indirect reciprocity compared to IS. This enhanced capability makes GS particularly valuable in settings with heterogeneous costs, offering a practical and robust solution for fostering cooperative behaviour.

2.4 Experimental Design

In this section, we introduce and justify the choices made in designing our experiment, and outline the structure of our helping games. We then describe the experimental procedure, providing detailed information on how the experiment was conducted.

2.4.1 Our Helping Games

The starting point for our experimental design is the paper by Engelmann and Fischbacher (2009) (EF), who study indirect reciprocity using a repeated helping game. Our design builds upon their framework with two major differences. Firstly, instead of assessing the relevance of a private or a public score as reputation mechanism, we test two mechanisms: Image Scoring (IS), and Good Standing (GS). This is implemented through a between-subject design to avoid contamination between the two scoring systems. Secondly, we investigate two conditions: a helping game with homogeneous costs and a variation with heterogeneous costs (specifically, two different costs, c_h and c_l). This is done using a within-subject design, which helps control for individual differences and allows us to detect any potential differences caused entirely by the two mechanisms.

Table 2.2 positions our experimental design in the context of existing literature: we generally replicate the "public score" feature of EF's experiment but introduce three new treatments to investigate reputation-based mechanisms within a laboratory setting.

Table 2.2: Experimental design

		WITHIN SUBJECT DESIGN	
		Homogeneous Cost ($b > c > 0$)	Heterogeneous Cost ($b > c_h > c_l > 0$)
BETWEEN SUBJECT DESIGN	Image Scoring (IS)	Engelmann and Fischbacher (2009)	
	Good Standing (GS)		

In each treatment (IS or GS), two cohorts of six subjects are matched randomly within each cohort. These cohorts remain the same across two sequences of helping game that subjects face in chronological order. These sequences are the homogeneous and the heterogeneous cost conditions, each consisting of an indefinite number of rounds. The order of the sequences is counterbalanced between and within treatments.

Interaction in a round: Within each sequence and each cohort, subjects are matched into pairs for each round and face a helping game. In each pair, one player is randomly assigned the role of the active player, while the other assumes the role of the non-active player. Both players begin the round with an account endowed with £7.

In each sequence, subjects are informed of their own cost of helping, which remains fixed for that sequence. They are also told that the benefit of being helped is £10 in each round. Additionally, subjects are informed that other players might have different costs, but that these costs would not exceed £6. However, the exact costs of other players are not disclosed, ensuring that decisions are made under some degree of uncertainty regarding others' costs.

In the homogeneous cost condition, if the active player chooses to help, the non-active player earns £10, which is added to the £7 in their account, while the active player loses £4, deducted from the £7 in their account. If the active player chooses not to help, both players retain the £7 in their accounts without any changes. The surplus from cooperation is £6[9]. The cost of cooperation to the active player is £4, and the benefit to the recipient is £10, resulting in a cost/benefit ratio of 0.4. We selected this ratio to align with Engelmann and Fischbacher's paper, as they use the same ratio.

In the heterogeneous cost condition, half of the active players are randomly assigned a high cost (£6) and the other half a low cost (£2), with this assignment remaining constant throughout the rounds. If the active player chooses to help, regardless of their cost, the non-active player earns £10, which is added to the £7 in their account. For the active player, helping results in either a £6 deduction (if they are a high-cost player) or a £2 deduction (if they are a low-cost player) from the £7 in their account. This creates cost/benefit ratios of 0.6 for high-cost players and 0.2 for low-cost players.

[9] The surplus is calculated as the difference between the sum of the accounts in the case of help versus no help:$[(7-4)+(7+10)]-[7+7]=$ £6 .

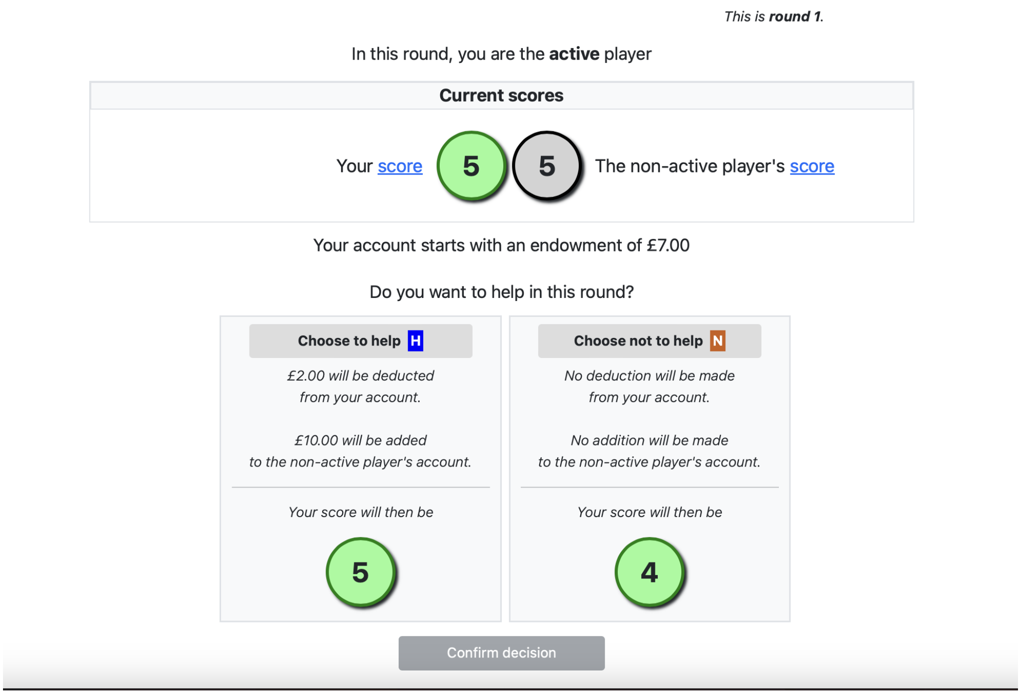
We chose these ratios to align with the two different cost treatments used in Seinen and Schram's study, where they were applied in separate games. However, in our experimental design, both costs are incorporated into the same game rather than being tested separately. This design allows us to investigate two key aspects: first, the responses of the GS and IS mechanisms to a game with heterogeneous costs, and second, whether a population with different costs of helping behaves differently compared to a population with uniform costs, depending on the mechanisms (i.e., scores) they are presented with. It is also important to note that the cost/benefit ratio in the homogeneous cost condition (0.4) is precisely halfway between the two cost/benefit ratios of the heterogeneous cost condition (0.6 and 0.2).

In each treatment, we provide subjects with the following information: (i) their own specific costs, one for the homogeneous cost condition and another for the heterogeneous cost one, (ii) their own value of benefit, (iii) that each session is divided into two cohorts of 6 subjects each, (iv) that the benefit is the same for all subjects within their cohort, and (v) that different members of their cohort may have different costs. The wording of (v) is crafted to be applicable to both cost conditions, ensuring that subjects are not explicitly informed about the difference between the homogeneous and heterogeneous cost conditions.

This setup was explained to the subjects as follows: the roles were labelled as "active player" and "non-active player". Following Engelmann and Fischbacher (2009), the helping choices were labelled as "help" and "do not help" to clearly define the choices available to the active player. Each active player was informed about their endowment, their score (see explanation below) and the potential outcomes based on their decisions. The active player's decision screen provided a clear breakdown of their possible choices, the associated costs, and how their score would update based on their decision (Figure 2.1[10]). Non-active players, on the other hand, were informed of their role and could see their own current score, but not the active player score. They had no decisions to make (see Figure 2.14 in the Appendix.).

We believe that labeling the players with this terminology, i.e., "active player" and "non-active player", rather than "helper" or "donor" and "helpee" or "recipient" has two advantages. First, this terminology does not alter the structure of the game

[10] The GS screen for the active player had score 0 and 1 and associated updating rule. You can see the decision screen in the Instructions for Good Standing in the Appendix.



*This is **round 1**.*

In this round, you are the **active** player

Current scores

Your score 5 5 The non-active player's score

Your account starts with an endowment of £7.00

Do you want to help in this round?

£2.00 will be deducted from your account.

£10.00 will be added to the non-active player's account.

Your score will then be

No deduction will be made from your account.

No addition will be made to the non-active player's account.

Your score will then be

Figure 2.1: Active player's decision screen in IS treatment

but clearly indicates which players have a choice to make and which do not. Second, terms like "helper" or "donor" suggest an expectation to help or donate, which we want to avoid to maintain neutrality.

At the end of each round, each player receives a summary screen providing essential information about that round. This screen displays the round number, the role they played (active or non-active), and a summary of their earnings if that round would have been chosen to be paid out for real. This summary includes the endowment, any deductions or additions resulting from the active player's choice, and the resulting balance. This comprehensive summary ensures that players are fully informed about the outcomes of that round and the potential implications in a clear and concise manner.

IS and GS treatments: Each subject is assigned a score, either IS or GS, and has the updating rule explained to them according to their treatment group. These scores are represented by circled numbers, providing a compact summary of their behaviour in previous rounds. Non-active players see only their own scores, while active players see both their own scores and those of the non-active players.

In the IS treatments, we follow Engelmann and Fischbacher's public IS. We decided to use their mechanism because it aligns well with our objective of comparing IS and GS. Their design effectively captures the dynamics of reputation and cooperation in a controlled setting, making it suitable for our experimental purposes.

A subject's score ranges from 0 to 5 and reflects the number of times they chose to help in their last five rounds as active players. Everyone starts with a score of 5. If a subject has fewer than five active rounds, the missing rounds are assumed to be instances where they helped, ensuring their initial score is 5. The score then updates according to the following rule: the most recent decision (help or no help) replaces the oldest of the last five decisions. This means the score increases by one if a helping decision replaces a non-helping decision, decreases by one if a non-helping decision replaces a helping decision, or remains the same if the new decision matches the oldest one.

In the GS treatments, we follow Leimar and Hammerstein's GS mechanism. A subject's score is binary (0 or 1) and represents the subject's standing (0 = bad, 1 = good). However, during the experiment, we did not use the labels "good" and "bad" to explain the scores to the subjects. Instead, we provided numerical scores (0 and 1) to maintain comparability with the IS mechanism.

Everyone starts with a score of 1. In every round, the score updates according to the following rule: if the active player helps, their score becomes 1. If they do not help, their score depends on the non-active player's score: it becomes 0 if the non-active player had a score of 1, or remains unchanged if the non-active player had a score of 0.

Subjects in both conditions start with the full score (5 for IS and 1 for GS) for two reasons. First, we aim to observe whether subjects can maintain helping behaviour from an initial state where everyone has the highest possible score. Second, based on the public goods game literature (e.g., Lugovskyy et al., 2017), where cooperation tends to decay over time, we wanted to start from a situation where cooperation is likely to be sustained for a while, allowing us to better investigate the dynamics of maintaining cooperation.

Both treatments allow active players to click on their scores to see detailed information on how the scores were computed, ensuring transparency and understanding of the scoring mechanisms.

We chose to test the Leimar and Hammerstein's GS mechanism for three main reasons. First, it is straightforward and easier to explain in an experimental setting. Second, unlike Sugden's GS mechanism, it allows easy recovery from a score of 0 to 1; a subject simply needs to help in one round to regain a score of 1. Third, this mechanism avoids the problem of an absorbing state where all players have a score of 0 and cannot improve their scores, which could potentially stall cooperation entirely.

Session: Each session involves 12 subjects in the lab at the same time, all exposed to the same information treatment. Six sessions use IS information, while another six use GS information.

In each session, participants play two distinct sequences: one under homogeneous costs and one under heterogeneous costs. The order of these sequences is counterbalanced across sessions: in six sessions, participants begin with the homogeneous-cost sequence, whereas in the other six they begin with the heterogeneous-cost sequence. Participants are informed that the cost condition might change between the first and second sequence, but they are not told the specific costs they would face. This design ensures that all participants are in a symmetrical position concerning their awareness of any role changes throughout the experiment.

We chose the cohort size of six players for several reasons. First, the use of six subjects per cohort balances the need for anonymity — important for simulating large population interactions — with the practicalities of laboratory-based experimentation. Second, the even-numbered cohort size aligns with the structure of the helping game, which involves pairs of players. The choice of six subjects enhances the robustness of our design, enabling random pairing within each cohort to allow for extensive interaction that the GS and IS mechanisms are designed to influence.

Within each cohort, subjects are randomly re-matched into pairs at the start of each round, resulting in a $\frac{1}{5}$ probability of meeting the same participant in two consecutive rounds. Subjects do not know with whom they are paired, nor do they know who is in their matching cohort in any sequence. Each round, the computer randomly assigns one subject to the non-active player's role and the other to the active player's role, with equal probability. Hence, in every round, half the subjects are non-active players and half are active players.

A random continuation rule determines the duration of each sequence (Roth and Murnighan, 1978). Each sequence has 40 fixed rounds, after which the sequence continues with a probability of 0.67. This design ensures a finite but indeterminate duration of interaction; beginning with round 40, the sequence is expected to continue for three further rounds. In the experiment, a computer simulates the roll of a six-sided die. If the roll iss 1 or 2, the sequence would end; otherwise, the sequence continues to round 41. At the end of each round, all subjects observe the number drawn, which informes them about the end or continuation of the sequence and also serves as a public coordination device. Sequences terminate simultaneously for both cohorts in every session.

2.4.2 Experimental Procedure

The experimental software was programmed in oTree (Chen et al., 2016) and the sessions were conducted in the Laboratory for Economic and Decision Research (LEDR) at the University of East Anglia. In total, 144 students participated in 12 experimental sessions (6 in IS treatment and 6 in GS). Each session had 12 participants. The subjects were recruited from the university's database.

The experiment lasted on average for roughly one hour and twenty minutes and participants earned on average £17.50[11]. This average amount is in accordance with common rules of the lab.

Each experimental session begins with participants being randomly assigned to their desks. Instructions[12] for the experiment are read aloud to ensure clarity and common knowledge, and participants complete a comprehension quiz to confirm their understanding of the games. Following this, participants engage in the helping games, which last approximately 50 minutes (see Section 2.4.1 for details).

After the game, participants complete a short survey that includes a set of socio-economic questions to capture individual characteristics, along with questions designed to elicit preferences using the Falk et al. (2018)'s validated survey. This survey measures time preferences, risk preferences, positive and negative reciprocity,

[11] Standard deviation: ± 8, median: £14. The minimum amount for a participant was £4, corresponding to £1 for the heterogeneous costs condition and £3 for the homogeneous cost condition. This case resembles the situation in which the player is paid for both rounds as active player and decided to help in both the conditions. The maximum amount was £34, corresponding to £17 for the heterogeneous costs condition and £17 for the homogeneous cost condition. This case resembles the situation in which the player is paid for both rounds as non-active player and was helped by the active players

[12] See *Instruction* in the Appendix 2.A.2. In particular, *Experimenter* (2.A.2) shows a detailed description used as a guide for the experimenters in every session.

altruism, and trust, linking participants' experimental choices to their underlying preferences. Upon completing the survey, the experiment concludes, participants are compensated, and they are free to leave.

2.5 Conjectures

Our primary objective is to examine differences in the extent of helping behaviour (i) between good standing (GS) and image score (IS) treatments and (ii) between homogeneous and heterogeneous cost conditions. We are interpreting these differences as comparisons between mechanisms rather than tests of theories.

Our analysis focuses on two primary concerns: the frequencies of helping behaviour and the ability to discern reciprocal cooperation from non-reciprocal cooperation under varying cost treatments.

Drawing on the theoretical considerations outlined in Section 2.3 and insights from previous literature (Section 2.2), we propose three conjectures:

Conjecture 0 - Replication of results:
In the homogeneous cost condition of the IS treatment, our average helping rates per round will broadly replicate the ones of Engelmann and Fischbacher (2009) in their corresponding treatment.

Our homogeneous cost condition in the IS treatment closely mirrors the public score treatment of their experiment. Given this replication, we anticipate that our results will be roughly similar to theirs. Validating this conjecture is important for establishing the robustness of our experimental design and ensuring that any observed differences in other treatments can be attributed to the mechanisms under investigation rather than methodological variations. Moreover, it allows us to assess whether differences in participant populations or experimental interfaces influence the outcomes.

Conjecture 1 - Frequency of Helping Behaviour:
There are differences in the frequencies of helping behaviour between IS and GS mechanisms across the two cost treatments (homogeneous and heterogeneous).

While we conduct a two-tailed test to examine these differences, we believe that the helping behaviour will be greater under the GS mechanism compared to the IS mechanism. Our belief is grounded in theoretical predictions about the effectiveness of the GS mechanism in fostering cooperation. The GS mechanism incorporates a norm of justified defection, where individuals who refuse to help not-helpers do not suffer reputational harm. This feature is expected to sustain higher levels of cooperation compared to the IS mechanism, which updates reputations based only on the helping actions.

Conjecture 2 - Reciprocity in Helping behaviour:
There are differences in the correlation coefficient between IS and GS mechanisms across the two cost treatments (homogeneous and heterogeneous).

We anticipate a stronger positive correlation between the score of non-active players and the relative frequency of being helped in GS, indicating indirect reciprocity. In this case, reciprocity refers to the desire to help people with a higher score, which reflects past cooperative behaviour. If our prediction is accurate, we ought to observe a stronger correlation between this two variables (helping and non-active player's scores) in GS. This could imply that GS is more capable of facilitating reciprocal helping than IS.

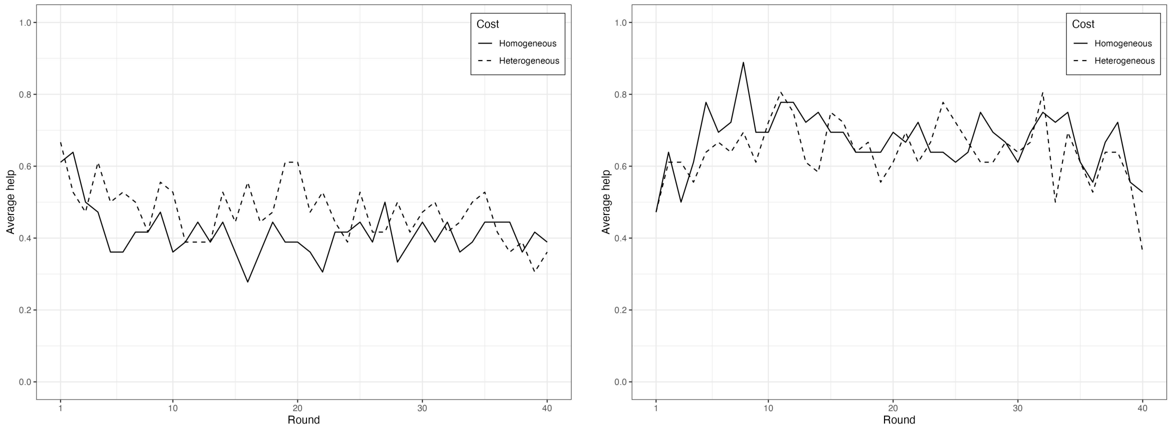
By clearly distinguishing between the concepts of frequency of helping and reciprocity, we aim to increase our understanding of the effectiveness of GS and IS mechanisms in fostering cooperative behaviour in helping games.

2.6 Results

In total, our study included 144 subjects[13], equally divided into two groups of 72 individuals each, corresponding to the GS and IS scoring systems. For every scoring system, we had a consistent team of 36 active players available during each round of both sequences.

Figure 2.2 and Figure 2.3 provide an overview of the experimental outcomes across different rounds for both the GS and IS treatments, illustrating trends in helping behaviour and scoring over the course of 40 rounds for both homogeneous and heterogeneous cost conditions.

[13] See summary statistics in Table 2.18 in the *Appendix*.



(a) Good Standing treatment

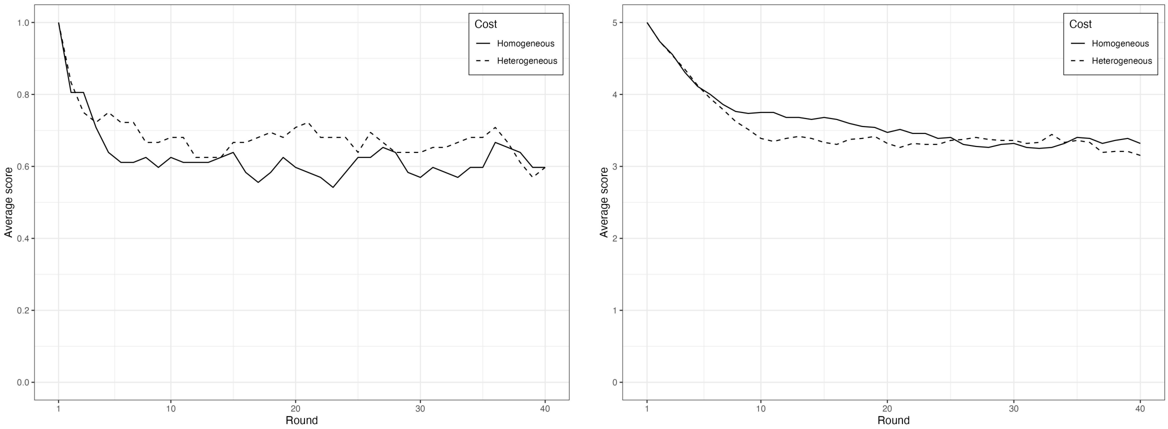
(b) Image Scoring treatment

Figure 2.2: Trends average helping per round
Thick lines display homogeneous cost. Dash lines display heterogeneous cost.

Figure 2.2 depicts the average helping rates per round, highlighting the differences between the GS and IS mechanisms. In both panels of Figure 2.2, the solid line represents the homogeneous cost condition, while the dashed line denotes the heterogeneous cost condition. In the GS treatment, shown in Figure 2.2a, helping rates decrease initially for around 5 rounds but tend to stabilise around 40% average help after the initial rounds, suggesting that players develop consistent behaviour patterns. A similar stabilisation trend is observed in the IS treatment depicted in Figure 2.2b where helping rates also stabilise after some rounds, albeit at generally higher initial levels than in the GS treatment. This stabilisation indicates that, despite the different initial helping rates, players in both treatments reach a relatively steady state in their cooperation levels as the game progresses.

Result 0: *The average helping rate per round in the homogeneous cost condition of our IS treatment replicates the ones of Engelmann and Fischbacher (2009) in their treatment in which both players have public scores.*

The thick line in Figure 2.2b, representing the homogeneous cost condition in our IS treatment, exhibits average helping levels primarily ranging between 0.5 and 0.8. This pattern of helping behaviour closely replicates the findings of Engelmann and Fischbacher (2009) in their treatment where both participants maintained public scores. A direct comparison between the trends observed in Figure 2.2b and their data reveals a striking similarity[14]. Additionally, both our sessions and theirs demonstrate comparable end-game effects, reinforcing the consistency between the two studies. This parallel suggests that our findings robustly reproduce their results within the context of their public score treatment.



(a) Good Standing treatment

(b) Image Scoring treatment

Figure 2.3: Trends average score per round
Thick lines display homogeneous cost. Dash lines display heterogeneous cost.

Figure 2.3 presents the trends in average scores over the rounds, reflecting how players' scores evolve under the GS and IS mechanisms. The GS treatment, illustrated in Figure 2.3a, uses a strictly binary scoring system with only two possible values: 0 or 1. Both cost conditions display stabilisation in scores after the initial 5 rounds, indicating that players' standings become stable as they adapt to the game. In contrast, the IS treatment, shown in Figure 2.3b, uses the IS scoring system ranging from 0 to 5, reflecting the cumulative number of helping actions. Scores under the IS treatment also stabilise, in this case between round 5 and round 10. This occurs slightly later than GS, aligning with the scoring mechanism's reliance on a moving window of the last five decisions. This stabilisation is crucial as the score in IS becomes a meaningful representation of a player's behaviour only after each active player has made at least five decisions, which realistically happens around round 10.

The observed stabilisation of helping rates and scores in both treatments justifies our focus on rounds 10 to 35 for the main analysis. By round 10, players would have made enough decisions to establish their behaviour patterns, making their scores un-

[14] See Figure 3 in page 25 of Engelmann and Fischbacher (2008).

der both mechanisms meaningful and reflective of their strategy. Analysing rounds beyond round 35 might introduce biases related to end-game effects, as participants might alter their behaviour in anticipation of the game ending. Focusing on rounds 10 to 35 ensures that we capture the stable, steady-state behaviour of participants, reducing the impact of early learning effects and late-game strategic adjustments.

We begin our analysis at the cohort level, as this approach provides independent data points that allow for a clearer and more reliable examination of the overall helping behaviour. In Subsection 2.6.1, we present results related to our first conjecture by analysing the average helping rates across different cohorts. Following that, we move into individual-level data to investigate more specific behavioural patterns. In Subsection 2.6.2, we present the findings from our second conjecture, as well as a deeper investigation of individual decision-making processes in the helping game.

2.6.1 Cohort-level Analysis

In this subsection, we evaluate rounds 10 to 35 of each game and use the average choices of each cohort in a sequence as the unit of observation[15]. We restrict the sample to rule out rounds where there is learning and endgame effects. This approach provides us with 24 independent data points, corresponding to the 24 independent cohorts in our experiment.

2.6.1.1 Cooperation rates between GS vs IS in helping games

A preliminary examination of the data suggests that helping rates are very different in each cohort. Figure 2.4 presents both the box plot and the dot plot of the average helping rates for each cohort in the homogeneous cost helping games. Each dot represents a cohort and the black line inside each box denoting the median helping rate. The box itself represents the interquartile range (IQR), encompassing the middle 50% of the data. Figure 2.5 provides the corresponding information for the heterogeneous cost helping games.

[15] Results are robust if we use the full dataset.

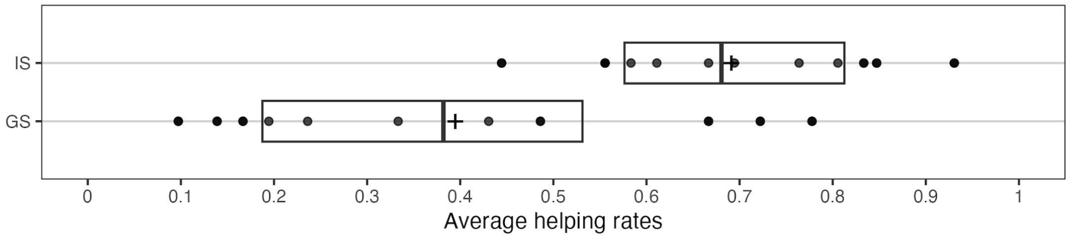


Figure 2.4: Average helping rate in homogeneous cost helping games
Each dot represents one cohort. Each cross represents the average for each treatment.

In the homogeneous cost condition (Figure 2.4), the Good Standing (GS) mechanism exhibits a median helping rate of 0.382, whereas the Image Scoring (IS) mechanism demonstrates a higher median helping rate of 0.681. In the heterogeneous cost condition (Figure 2.5), the Good Standing (GS) mechanism exhibits a median helping rate of 0.396, whereas the Image Scoring (IS) mechanism demonstrates a higher median helping rate of 0.743.

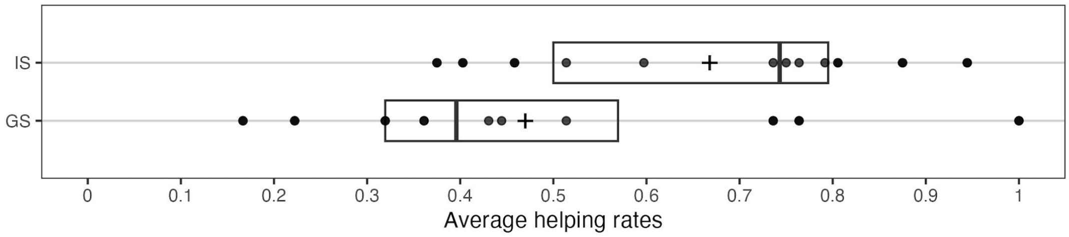


Figure 2.5: Average helping rate in heterogeneous cost helping games
Each dot represents one cohort. Each cross represents the average for each treatment.

The variance in helping rates is notably higher for the GS group compared to the IS group in both conditions. Specifically, in the homogeneous cost condition the variance for GS is 0.056, while for IS, it is 0.022 (Fligner-Killeen test, $\tilde{\chi}^2 = 3.477$, $p = 0.062$). In the heterogeneous cost condition, the variance for GS is 0.060, compared to 0.037 for IS (Fligner-Killeen test, $\tilde{\chi}^2 = 0.086$, $p = 0.769$).

Result 1: *The Image Scoring mechanism induces higher levels of cooperation compared to the Good Standing mechanism.*

The crosses (+) in Figures 2.4 and 2.5 indicate the mean helping rates for each mechanism under both cost conditions. The average helping rates differ between GS and IS in the homogeneous cost condition, with mean helping rates of 0.395 for

GS and 0.691 for IS (Fisher-Pitman, $Z = -2.958$, $p = 0.002$). In the heterogeneous cost condition, the mean helping rates are 0.470 for GS and 0.668 for IS, although the difference here is smaller (Fisher-Pitman, $Z = -2.037$, $p = 0.041$).

We find that the Image Scoring mechanism generally promotes greater cooperation than the Good Standing mechanism. These findings appear to disconfirm Conjecture 1, which posited that the Good Standing mechanism would induce higher levels of cooperation.

2.6.1.2 High cost and low cost players in heterogeneous game

We further analyse the dynamics of helping behaviour under heterogeneous cost conditions to explore whether the cost differences affect cooperation levels within each mechanism. Figure 2.6 and figure 2.7 present the same information as Figure 2.5, but focus on the low-cost and high-cost players within each cohort in the heterogeneous cost games. In particular, the crosses illustrate the average helping rates for subjects facing low and high costs under the GS and IS mechanisms.

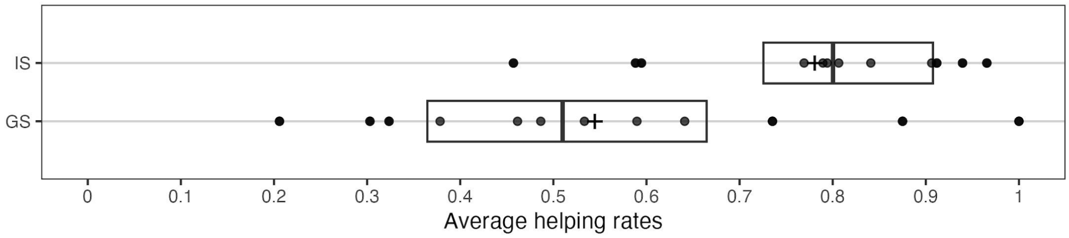


Figure 2.6: Average helping rate in heterogeneous cost helping games for players with a low cost
Each dot represents one cohort. Each cross represents the average for each treatment.

For the low-cost players (Figure 2.6), the average helping rate was higher in the IS treatment (0.780) compared to the GS treatment (0.544) (Fisher-Pitman, $Z = -2.497$, $p = 0.010$). The variance in helping rates was lower in IS (0.025) than in GS (0.057) (Fligner-Killeen test, $\tilde{\chi}^2 = 1.969$, $p = 0.161$).

For the high-cost players (Figure 2.7), the average helping rate was also higher in the IS treatment (0.562) compared to the GS treatment (0.396) (Fisher-Pitman, $Z = -1.349$, $p = 0.181$). The variance in helping rates was 0.097 for GS and 0.078 for IS (Fligner-Killeen test, $\tilde{\chi}^2 = 0.650$, $p = 0.420$).

Within the GS treatment, the variance increased from 0.051 for low-cost players

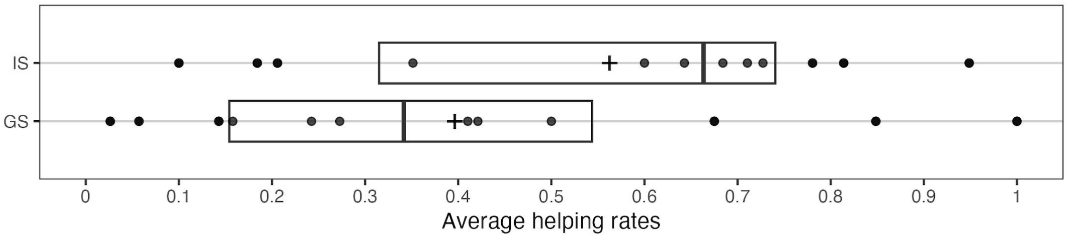


Figure 2.7: Average helping rate in heterogeneous cost helping games for player with a high cost

Each dot represents one cohort. Each cross represents the average for each treatment.

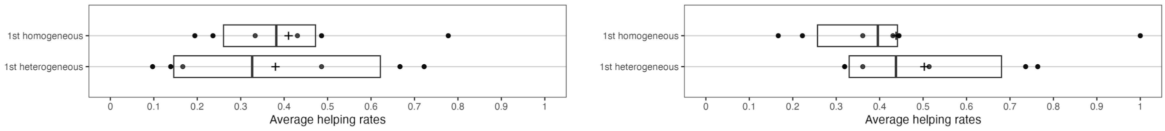
to 0.084 for high-cost players (Fligner-Killeen test, $\tilde{\chi}^2 = 0.211$, $p = 0.646$), while the mean helping rates did not show a large difference (Fisher-Pitman, $T = 0.189$, $p = 0.0884$). In the IS treatment, the variance increased from 0.028 to 0.069 when comparing low to high-cost players (Fligner-Killeen test, $\tilde{\chi}^2 = 2.007$, $p = 0.157$), and there was a noticeable difference in mean helping rates (Fisher-Pitman, $T = 0.194$, $p = 0.0432$).

These findings reinforce the general trend observed in the homogeneous cost condition, where the IS mechanism generally induced higher cooperation rates than the GS mechanism. This pattern persists in the heterogeneous cost condition, particularly among low-cost players, suggesting that IS might be more effective in sustaining cooperation across different cost structures. However, the observed differences in variability between high-cost and low-cost players within both treatments indicate that cost asymmetry plays a role in influencing cooperative behaviour. This finding is consistent with the broader trend that IS tends to generate more consistent cooperation, while GS shows more variability.

2.6.1.3 Order effect

We examine the potential impact of order effects on cooperation. To ensure robust data collection, we implemented a counterbalancing strategy to evenly distribute the first condition played across the sessions. Specifically, we have 24 cohorts divided between the IS and GS mechanisms. For each mechanism, 6 cohorts played the homogeneous cost condition sequence first, followed by the heterogeneous cost condition sequence, while the other 6 cohorts played the sequences in the opposite order. This setup allows us to assess whether individuals exhibit higher or lower rates of helping when they play the homogeneous cost condition first, as opposed to starting with the heterogeneous cost condition. This comparison is illustrated in Figure 2.8, and Figure 2.9.

Figure 2.8 illustrates the average helping rates under the GS mechanisms across the homogeneous (2.8a) and heterogeneous cost conditions(2.8b). Fisher-Pitman permutation tests and Fligner-Killeen tests were performed to assess differences in mean and variance between cohorts that played the homogeneous cost condition first and those that played the heterogeneous cost condition first within each mechanism.

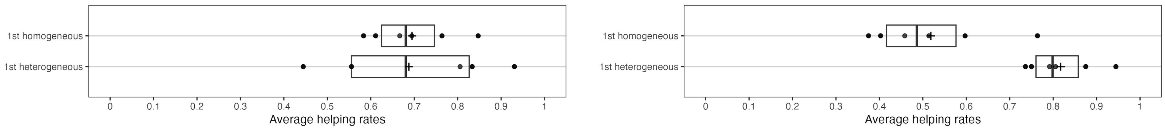


(a) Homogeneous cost

(b) Heterogeneous cost

Figure 2.8: Average helping rates in GS
Each dot represents one cohort. Each cross represents the average for each starting condition.

In the GS mechanism under the homogeneous cost condition, the mean helping rate was 0.410 for cohorts that played the homogeneous cost condition first and 0.380 for those that played the heterogeneous cost condition first (Fisher-Pitman, $Z = -0.219$, $p = 0.838$). The variances were 0.045 and 0.079, respectively (Fligner-Killeen test, $\tilde{\chi}^2 = 3.477$, $p = 0.062$). Similarly, under the heterogeneous cost condition, the mean helping rate was 0.438 for the cohorts that started with the homogeneous cost condition and 0.502 for those that started with the heterogeneous cost condition (Fisher-Pitman, $Z = 0.457$, $p = 0.697$), with variances of 0.088 and 0.042, respectively (Fligner-Killeen test, $\tilde{\chi}^2 = 0.124$, $p = 0.725$).



(a) Homogeneous cost

(b) Heterogeneous cost

Figure 2.9: Average helping rates in IS
Each dot represents one cohort. Each cross represents the average for each starting condition.

For the IS mechanism, Figure 2.9 presents the average helping rates across both the homogeneous (2.9a) and heterogeneous cost conditions (2.9b). In the homogeneous cost condition, the mean helping rate was 0.694 for cohorts that played the homogeneous cost condition first and 0.688 for those that played the heterogeneous cost condition first (Fisher-Pitman, $Z = -0.082$, $p = 0.957$). The variances were 0.010 and 0.038, respectively (Fligner-Killeen test, $\tilde{\chi}^2 = 4.429$, $p = 0.035$). Under the heterogeneous cost condition, the mean helping rate was 0.519 for cohorts that started with the homogeneous cost condition and 0.817 for those that started with the heterogeneous cost condition (Fisher-Pitman, $Z = 2.702$, $p = 0.007$), while the variances were 0.021 and 0.006, respectively (Fligner-Killeen test, $\tilde{\chi}^2 = 1.220$, $p = 0.270$).

This difference in helping rates, particularly in the heterogeneous cost condition, suggests a potential order effect within the IS mechanism. Participants appear to adjust their helping behaviour depending on whether they first encountered the homogeneous or heterogeneous cost structure. As discussed in Section 2.6.1.4, under IS, players need to adjust their behaviour continuously to maintain a good score. Transitioning between cost conditions requires recalibrating their strategies to match the new cost-benefit structure, which likely explains the observed shifts in cooperation levels.

These findings indicate that the order of conditions has a minimal effect on average helping rates within the GS mechanism. In contrast, the IS mechanism exhibits more complex dynamics, particularly under the heterogeneous cost condition. The difference in mean helping rates and the observed variation in behaviour across conditions highlights the IS mechanism's sensitivity to both the sequence of conditions and the cost structure. This contrasts with earlier results, where GS showed greater variability in helping behaviour, particularly when costs were heterogeneous. Together, these observations emphasise the context-dependence of IS, where players must adjust strategies based on the cost structure, whereas GS induces more stable, albeit variable, patterns of behaviour.

2.6.1.4 Correlation between average helping rates within IS and GS

Figure 2.10 presents two scatter plots illustrating the relationship between average helping rates in the homogeneous and heterogeneous cost conditions under the Good Standing (GS) and Image Scoring (IS) mechanisms. To quantify these relationships, we calculate Pearson correlation coefficients for each mechanism.

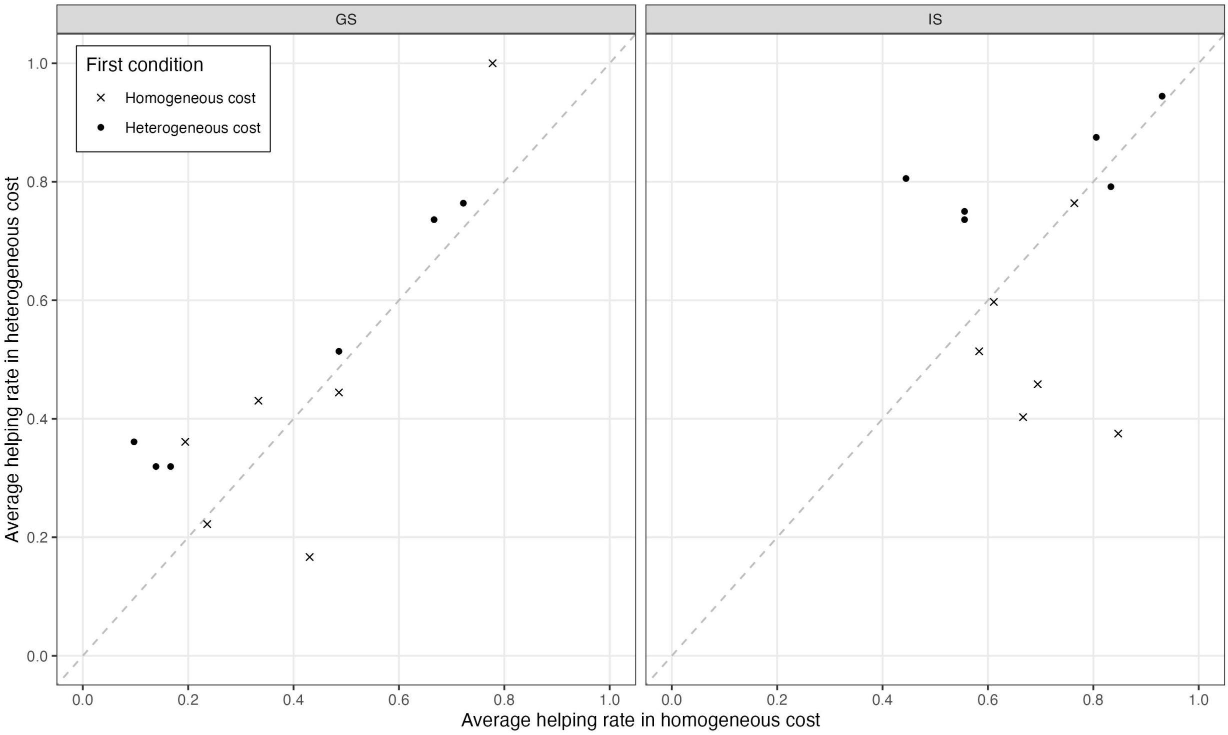


Figure 2.10: Correlation between average helping rates in homogeneous and heterogeneous cost condition

Each symbol corresponds to a cohort (× cohorts started with the homogeneous cost condition; • cohorts started with the heterogeneous cost condition). Left panel: Good Standing. Right panel: Image Scoring.

For the GS mechanism, there is a strong positive correlation between helping rates in the two cost conditions (Pearson correlation, $r(12) = 0.827$, $p = 0.001$). This statistically significant result suggests that cohorts who help more in the homogeneous cost condition also tend to help more in the heterogeneous cost condition. The high correlation indicates a consistent pattern of behaviour across different cost structures under the GS mechanism. This consistency aligns with our earlier findings that GS encourages more stable cooperation across various scenarios, as evidenced by the similar variances and helping rates observed between the homogeneous and heterogeneous cost conditions.

One possible explanation for this stability is that under the GS mechanism, players internalise the rule of cooperation, leading to behaviour that is less sensitive to changes in external conditions. This might be due to the binary nature of the GS score, where being in "good standing" consistently promotes reciprocal behaviour across different contexts. The wider range of helping rates observed under the GS mechanism can also be explained by reciprocity theory, which suggests the existence of multiple equilibria depending on cohort-specific dynamics, such as the proportion of cooperative individuals and the random initial conditions during the first few rounds. These dynamics can set the tone for the entire sequence, leading to the observed positive correlation in helping rates between the two conditions.

In contrast, the IS mechanism shows virtually no correlation between helping rates in the homogeneous and heterogeneous cost conditions (Pearson correlation, $r(12) = 0.138$, $p = 0.670$). The lack of statistical significance implies that helping rates in one condition do not predict helping rates in the other condition within the IS mechanism. This suggests that the IS mechanism may lead to more context-dependent behaviour, where participants' helping rates are influenced more by the specific cost structure than by their general propensity to help.

This context-dependence in IS is consistent with our earlier findings of a noticeable order effect under the IS mechanism, particularly in the heterogeneous cost condition. In the IS treatment, players must continuously adjust their helping behaviour to maintain a score that is perceived "as good". When participants transition from one cost condition to another, they need to recalibrate their strategies to align with the new cost-benefit structure, which may lead to significant shifts in cooperation levels. This adaptation process likely contributes to the lack of predictive power of helping rates between conditions in IS, as each game presents a distinct challenge that requires a fresh strategic approach. For further insights into this phenomenon, see Figure 2.15 in the Appendix, where differences in low and high scores are highlighted on the y-axis, illustrating how initial conditions might influence the trajectory of cooperation within cohorts.

By contrast, the strong positive correlation observed under the GS mechanism indicates that players tend to remain reciprocal irrespective of the cost condition. One explanation might be that the GS rule encourages a focus on sustaining reciprocal relationships rather than merely optimizing scores. Once such relationships are established, they are likely to persist across different cost environments.

2.6.2 Individual-level Analysis

In this subsection, we shift our focus to the individual-level data to further explore the dynamics of helping behaviour. We will specifically examine the results related to Conjecture 2, which involves analysing the frequency of being helped as a function of an individual's score under both the IS and GS mechanisms. Additionally, we will investigate other aspects of scoring, such as how scores correlate with helping behaviour across different cost conditions. Finally, we will review the data from control questions to ensure that participants understood the instructions correctly and to confirm that our experimental design was implemented without any significant errors.

Our selected rounds ensured that we maintained a minimum number of observations for each combination of scoring system and cost condition, which amounted to 36 active players multiplied by 24 rounds, yielding 864 individual helping data points for each treatment condition[16].

2.6.2.1 Analysis of frequency of being helped

Figures 2.11 and 2.12 analyse how a non-active player's standing (in the GS mechanism) and image score (in the IS mechanism) influence their likelihood of being helped under two different cost structures: homogeneous costs (Figure 2.11) and heterogeneous costs (Figure 2.12). The key distinction between the Good Standing (GS) and Image Scoring (IS) mechanisms is evident in the frequency with which non-active players with low scores receive help. These figures aim to highlight this critical difference, illustrating how each mechanism treats low-scoring individuals under different cost conditions.

In both figures, the left panels show how good standing affects helping rates. In the homogeneous-cost condition, non-active players with a score of 0 are helped only 11.5% of the time, whereas those with a score of 1 enjoy a notably higher 59.2% helping rate. Similarly, in the heterogeneous-cost condition, the corresponding figures (12.2% vs. 64.3%) further confirm the strong and stable influence of good standing on the likelihood of receiving help.

[16] The analysis does not change substantially in case we use full data. For the homogeneous cost condition we have 1480 observations, and for the heterogeneous cost condition we have 1512 observations.

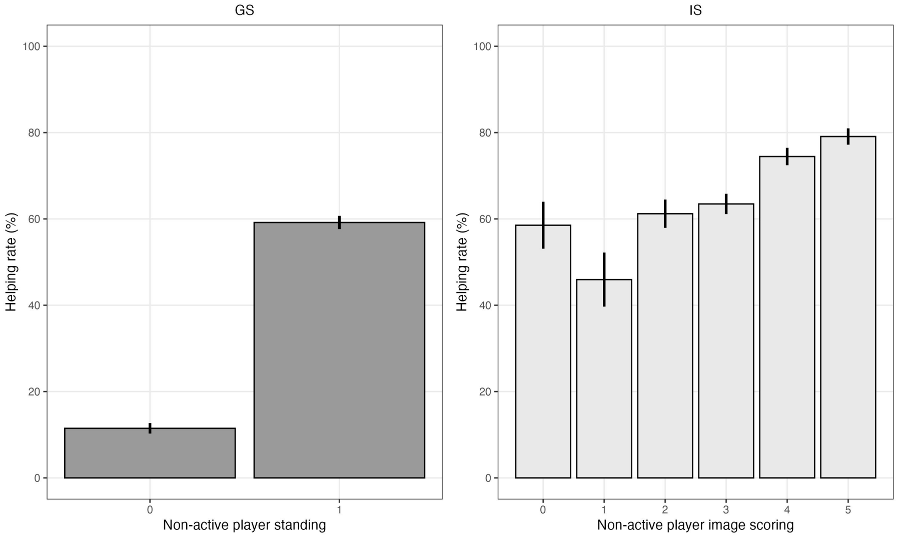


Figure 2.11: Homogeneous cost, frequency of being helped by score
Left panel: Good Standing. Right panel: Image Scoring.

The right panels of Figures 2.11 and 2.12 explore the relationship between the non-active player's image scoring and the frequency of being helped. Under heterogeneous costs (Figure 2.12), there is a clear monotonic increase in helping rates as the image score rises. Non-active players with a score of 0 are helped 36.5% of the times, which steadily increases to 75.8% and 78.1% for non-active players with scores of 4 and 5, respectively. This pattern suggests that when costs are heterogeneous, a higher image score is strongly associated with a greater likelihood of being helped.

However, when the cost of helping is homogeneous (Figure 2.11), the relationship between the non-active player's score and helping rates becomes less linear. While there is still an overall increase in helping rates with higher image scores, the progression is not as smooth. For instance, the frequency of being helped for a non-active player with a score of 1 drops to 46.0%, lower than the 58.5% observed for a score of 0, before increasing again with higher scores. The frequency of being helped eventually reaches 74.5% and 79.1% for scores of 4 and 5, respectively. This variability at lower scores may indicate a more complex decision-making process when costs are homogeneous, possibly reflecting a combination of strategic considerations and the perceived value of the non-active player's image. Additionally, the lower helping rates at certain scores, such as 1, could be driven by a smaller number of observations at these specific scores, which might influence the stability of the

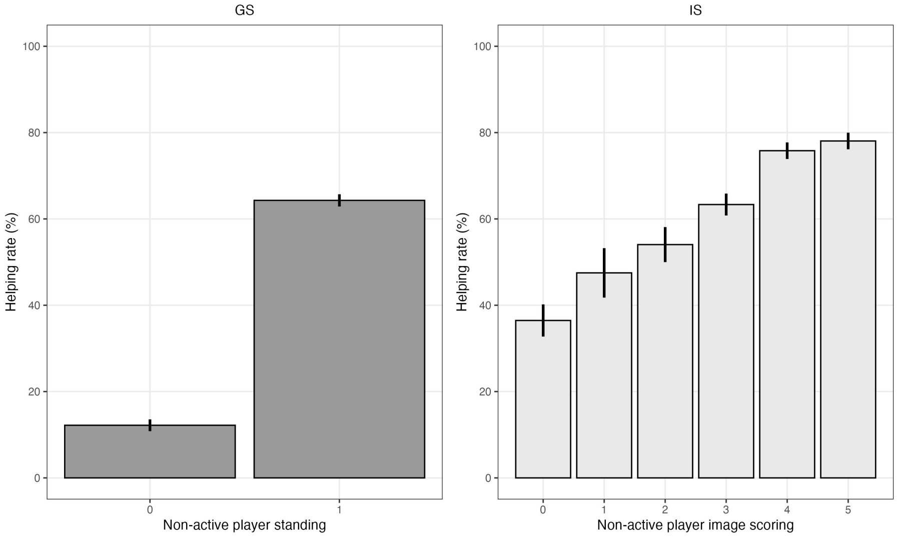


Figure 2.12: Heterogeneous cost, frequency of being helped by score
Left panel: Good Standing. Right panel: Image Scoring.

observed trends.

To quantify these relationships, we conduct Pearson and Kendall correlation tests for both the GS and IS treatments in both the cost conditions. These tests provide a statistical measure of the strength and direction of the association between non-active player standing and non-active player image scoring with the frequency of being helped.

In the GS condition, we observe strong positive correlations between the non-active player's standing and the frequency of being helped in both the homogeneous and heterogeneous cost games. Specifically, in the homogeneous cost game, the Pearson correlation is $r(864) = 0.480$, and the Kendall correlation is $\tau(864) = 0.480$ (both with $p < 0.001$). In the heterogeneous cost game, the correlations remain similarly strong, with a Pearson correlation of $r(864) = 0.492$, and a Kendall correlation of $\tau(864) = 0.492$ (both with $p < 0.001$). These results indicate that players in good standing are more likely to be helped, and the consistency between the Pearson and Kendall correlations suggests that this relationship is both linear and monotonic across both cost structures.

In contrast, the IS condition shows notably weaker correlations between the non-active player's image score and the frequency of being helped. In the homogeneous cost game, the Pearson correlation is $r(864) = 0.176$, and the Kendall correlation is $\tau(864) = 0.163$ (both with $p < 0.001$). In the heterogeneous cost game, the Pearson correlation increases slightly to $r(864) = 0.282$, and the Kendall correlation to $\tau(864) = 0.236$ (both with $p < 0.001$). These results suggest that while a higher image score does increase the likelihood of being helped, the effect is notably less pronounced compared to the influence of good standing.

These results confirm our Conjecture 2 and suggest that IS, as operationalized in this experiment, exerts only a modest influence on being helped, casting doubt on its robustness as a reputation-based mechanism for explaining indirect reciprocity in this context.

Result 2: *The Good Standing mechanisms induces higher levels of reciprocity compared to the Image Scoring mechanism.*

The difference in correlation strength between the GS and IS conditions suggests that good standing may be more strongly tied to helping behaviour than image score. In particular, the higher correlation observed in the GS condition indicates that good standing could play a larger role in shaping cooperative actions. These results imply that although reputational factors matter in fostering cooperation, their influence may depend on the specific facet of reputation under consideration. Within this experimental context, GS appears to be a stronger predictor of helping behaviour than IS.

There are important limitations to note regarding these correlation measures. In the GS condition, we are comparing the correlation between standing (which is binary, either 0 or 1) and helping behaviour. In the IS condition, however, we are measuring the correlation between helping behaviour and a more granular score, ranging from 0 to 5. This discrepancy in score range limits the direct comparability of the correlations between the two mechanisms. The broader range of scores in IS introduces more variability, which likely weakens the strength of the correlation. Additionally, the binary nature of the GS mechanism may make it easier for participants to interpret and act upon, leading to clearer and stronger relationships between standing and helping rates.

2.6.2.2 Analysis of helping choice

In this section, we examine the factors influencing participants' decisions to help in both the Good Standing (GS) and Image Scoring (IS) treatments. Using probit regression models, we analyse how the characteristics of both active and non-active players affect the likelihood of helping. The models include individual fixed effects to control for unobserved heterogeneity among participants. Standard errors are clustered at the session, cohort, and round levels to account for potential dependencies within the data. Our analysis focuses on rounds 11 to 34, allowing us to observe behaviour after participants have become familiar with the experimental setting.

Table 2.3 provides a comprehensive summary of the key variables used in the regression analysis. Each variable is listed alongside a brief description of its meaning or role within the context of the study. This summary serves to clarify the interpretation of the variables and their relevance to the model, offering readers a clear understanding of the constructs and measurements underlying the empirical analysis. The table includes a combination of game-specific factors, individual characteristics, and behavioral measures derived from prior literature, ensuring a holistic representation of the variables utilized.

Table 2.3: Summary of regressions' variables and their meanings

Variable	**Meaning**
`player_helped`	Our dependent variable. Dummy: 1 if help
`other_status`	Status of the non-active player matched with the active player.
`help_change_status`	Dummy: 1 if the active player's help action changes their status.
`round`	Round of the game.
`male`	Dummy: 1 if the participant is male.
`Falk_favour`	Positive reciprocity (Falk et al., 2018).
`Falk_future`	Patience (Falk et al., 2018).
`Falk_giving`	Altruism (Falk et al., 2018).
`Falk_intentions`	Trust (Falk et al., 2018).
`Falk_revenge`	Negative reciprocity (Falk et al., 2018).
`Falk_risk`	Risk-taking (Falk et al., 2018).
`other_image_X`	Dummy: 1 if non-active player's image score is $X \in \{0, 1, 2, 3, 4\}$.
`lagY_player_helped`	Lagged choice of active player's help action.

Good Standing (GS)

Table 2.4 presents the results of probit regressions for the helping choices of active players in the GS treatment under both homogeneous and heterogeneous cost conditions. The dependent variable is a binary indicator of whether the active player chose to help the non-active player.

In the homogeneous cost condition (columns 1–3), the coefficient on *other_status* is positive and highly significant across all specifications. This indicates that active players are more likely to help non-active players who have a higher status. This behaviour aligns with the principles of indirect reciprocity inherent in the GS mechanism, where helping those in good standing is beneficial for maintaining one's own reputation.

The variable *help_change_status* also has a strong positive and significant effect on the likelihood of helping. This suggests that active players are particularly motivated to help when doing so will improve their own status from 0 to 1. The substantial magnitude of this coefficient reflects the strong incentive embedded in the GS mechanism for participants to attain or maintain good standing.

In model (2), we include the variable *round* to test for any temporal trends. The coefficient is positive but not statistically significant, implying that there is no significant change in helping behaviour over time within this treatment.

The inclusion of the variable *male* in model (2) reveals that male participants are significantly more likely to help than female participants in the homogeneous cost condition. However, when we include the *Falk_* variables in model (3), the coefficient on *male* decreases and remains significant only at the 10% level.

The *Falk_* variables in model (3) capture individual differences in social preferences. The coefficients on *Falk_favour* and *Falk_future* are positive and highly significant, indicating that participants who place a higher value on favouring others and who consider future consequences are more inclined to help. The coefficient on *Falk_risk* is also positive and significant, suggesting that more risk-tolerant individuals are more likely to help.

Table 2.4: Probit models for the help choice of active players in GS

Individual fixed effects are used in all regressions. Clustered standard errors take the dependence of the data within sessions, cohorts, and rounds into account. Only data from round 11 to round 34 are included.

	Homogeneous Cost			*Heterogeneous Cost*		
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	-1.452^{***} (0.101)	-1.808^{***} (0.218)	-4.787^{***} (0.543)	-1.319^{***} (0.105)	-1.424^{***} (0.205)	-3.758^{***} (0.449)
other_status	1.597^{***} (0.117)	1.629^{***} (0.121)	1.725^{***} (0.134)	1.610^{***} (0.118)	1.614^{***} (0.118)	1.765^{***} (0.142)
help_change_status	6.724^{***} (0.111)	6.829^{***} (0.133)	7.329^{***} (0.190)	6.495^{***} (0.122)	6.515^{***} (0.126)	7.270^{***} (0.236)
round		0.007 (0.008)	0.005 (0.008)		0.005 (0.007)	0.006 (0.007)
male		0.348^{***} (0.103)	0.190^{*} (0.115)		−0.039 (0.097)	-0.425^{***} (0.112)
Falk_favour			0.184^{***} (0.041)			0.203^{***} (0.039)
Falk_future			0.105^{***} (0.031)			0.166^{***} (0.031)
Falk_giving			0.005 (0.031)			-0.092^{***} (0.032)
Falk_intentions			0.016 (0.023)			-0.054^{**} (0.023)
Falk_revenge			0.041 (0.029)			0.029 (0.024)
Falk_risk			0.066^{**} (0.030)			0.034 (0.030)
Observations	864	864	864	864	864	864
Log Likelihood	−410.656	−404.457	−370.632	−442.432	−442.065	−405.343
Akaike Inf. Crit.	827.312	818.913	763.263	890.865	894.130	832.686

Note: *p<0.1; **p<0.05; ***p<0.01

In the heterogeneous cost condition (columns 4–6), the patterns are similar. The coefficients on *other_status* and *help_change_status* remain positive and significant, though the coefficient on *help_change_status* is slightly lower than in the homogeneous cost condition. This may reflect the increased cost of helping for some participants in this treatment.

Interestingly, the coefficient on *male* becomes negative and significant in model (6), indicating that male participants are less likely to help under heterogeneous costs when controlling for individual preferences. The *Falk_* variables continue to have significant effects, with *Falk_favour* and *Falk_future* positively associated with helping. However, *Falk_giving* and *Falk_intentions* have negative coefficients, suggesting more complex interactions between individual preferences and the cost structure.

Overall, these results highlight the importance of both reputational considerations and individual social preferences in shaping helping behaviour within the GS mechanism. Participants are responsive to the status of others and to opportunities to improve their own status, and individual differences further modulate these decisions.

Image Scoring (IS)

Table 2.5 presents the probit regression results for the IS treatment under both homogeneous and heterogeneous cost conditions. The dependent variable remains the binary choice to help. We have a relatively smaller number of observations compared to the GS because in some sessions some participants took more time to have a full score (i.e., composed of 5 previous active histories).

In the homogeneous cost condition (columns 1–3), the coefficients on *other_image_X* are generally negative and significant, indicating that active players are less likely to help non-active players with lower image scores compared to those with the highest score (image score 5). For example, the coefficient on *other_image_1* is -0.816^{***}, suggesting a substantial reduction in the likelihood of helping non-active players with an image score of 1.

The variables *lagY_player_helped* capture the influence of an active player's own past helping behaviour on their current decision to help. The positive and significant coefficients on these variables indicate that participants who have helped in previous rounds are more likely to help again. This persistence in helping behaviour suggests the presence of individual dispositions towards cooperation.

In model (3), we include the *Falk_* variables to account for individual preferences. The coefficient on *Falk_favour* is positive and significant, indicating that participants who value favouring others are more inclined to help. The coefficients on *Falk_revenge* and *Falk_risk* are negative and significant at the 10% level, suggesting that individuals with higher tendencies towards revenge or risk aversion are less likely to help.

Table 2.5: Probit models for the help choice of active players in IS

Individual fixed effects are used in all regressions. Clustered standard errors take the dependence of the data within sessions, cohorts, and rounds into account. Only data from round 11 to round 34 are included.

	Homogeneous Cost:			*Heterogeneous Cost:*		
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	-0.534^{***}	-0.430^{**}	-0.666^{*}	-0.705^{***}	-0.386^{**}	-0.105
	(0.134)	(0.182)	(0.365)	(0.114)	(0.176)	(0.411)
other_image_0	-0.363^{*}	-0.342	-0.354^{*}	-0.965^{***}	-0.922^{***}	-0.989^{***}
	(0.212)	(0.212)	(0.211)	(0.152)	(0.152)	(0.156)
other_image_1	-0.816^{***}	-0.808^{***}	-0.824^{***}	-0.753^{***}	-0.806^{***}	-0.844^{***}
	(0.188)	(0.188)	(0.196)	(0.248)	(0.255)	(0.263)
other_image_2	-0.425^{***}	-0.419^{***}	-0.441^{***}	-0.424^{***}	-0.372^{**}	-0.385^{**}
	(0.138)	(0.137)	(0.141)	(0.150)	(0.153)	(0.156)
other_image_3	-0.350^{***}	-0.355^{***}	-0.350^{***}	-0.375^{***}	-0.355^{***}	-0.374^{***}
	(0.115)	(0.115)	(0.119)	(0.120)	(0.120)	(0.123)
other_image_4	-0.036	-0.035	-0.035	-0.128	-0.124	-0.138
	(0.117)	(0.118)	(0.122)	(0.118)	(0.120)	(0.123)
lag5_player_helped	0.581^{***}	0.583^{***}	0.540^{***}	0.670^{***}	0.675^{***}	0.650^{***}
	(0.089)	(0.089)	(0.091)	(0.092)	(0.093)	(0.094)
lag4_player_helped	0.173^{*}	0.171^{*}	0.123	0.277^{***}	0.275^{***}	0.252^{***}
	(0.091)	(0.091)	(0.093)	(0.096)	(0.096)	(0.097)
lag3_player_helped	0.337^{***}	0.333^{***}	0.290^{***}	0.275^{***}	0.255^{***}	0.229^{**}
	(0.090)	(0.090)	(0.091)	(0.099)	(0.099)	(0.098)
lag2_player_helped	0.427^{***}	0.421^{***}	0.374^{***}	0.390^{***}	0.362^{***}	0.339^{***}
	(0.089)	(0.090)	(0.092)	(0.092)	(0.092)	(0.093)
lag1_player_helped	0.271^{***}	0.263^{***}	0.218^{**}	0.487^{***}	0.448^{***}	0.422^{***}
	(0.091)	(0.091)	(0.093)	(0.093)	(0.093)	(0.093)
round		-0.004	-0.005		-0.012^{**}	-0.012^{**}
		(0.004)	(0.004)		(0.005)	(0.005)
male		0.011	0.040		0.051	0.100
		(0.083)	(0.094)		(0.086)	(0.095)
Falk_favour			0.102^{***}			0.048
			(0.032)			(0.032)
Falk_future			-0.004			0.024
			(0.027)			(0.026)
Falk_giving			0.007			-0.016
			(0.021)			(0.023)
Falk_intentions			0.001			0.008
			(0.018)			(0.019)
Falk_revenge			-0.037^{*}			-0.050^{**}
			(0.020)			(0.022)
Falk_risk			-0.049^{*}			-0.080^{**}
			(0.029)			(0.033)
Observations	812	812	812	820	820	820
Log Likelihood	-410.656	-404.457	-370.632	-442.432	-442.065	-405.343
Akaike Inf. Crit.	827.312	818.913	763.263	890.865	894.130	832.686

Note: *p<0.1; **p<0.05; ***p<0.01

In the heterogeneous cost condition (columns 4–6), the patterns are broadly similar. The negative coefficients on *other_image_X* remain significant, with even larger magnitudes for *other_image_0*, indicating a stronger reluctance to help those with the lowest image scores under heterogeneous costs.

The lagged helping variables continue to be positive and significant, reinforcing the idea that past helping behaviour influences current decisions. The variable *round* is negative and significant in models (5) and (6), suggesting a slight decline in helping behaviour over time in the heterogeneous cost condition.

The *Falk_* variables in model (6) show that *Falk_revenge* and *Falk_risk* have negative and significant coefficients, indicating that individuals with higher tendencies towards revenge or risk aversion are less likely to help. The coefficient on *Falk_favour* is positive but not significant in this condition.

Overall, these results indicate that in the IS treatment, helping behaviour is influenced by the image score of the non-active player and by the active player's own history of helping. Participants are less inclined to help those with lower image scores and more likely to help if they have previously helped others.

Discussion

The regression analyses reveal distinct patterns in the determinants of helping behaviour under the GS and IS mechanisms.

In the GS treatment, the decision to help is strongly influenced by the non-active player's status and by opportunities for the active player to improve their own status. The large and significant coefficients on *other_status* and *help_change_status* highlight the central role of reputational incentives within the GS mechanism. Participants are motivated to help those in good standing and to take actions that enhance their own reputation.

In contrast, in the IS treatment, while the non-active player's image score influences helping decisions, the effect is less pronounced than in the GS treatment. The negative coefficients on *other_image_X* indicate a reluctance to help those with lower image scores, but the magnitude of these effects is smaller than the corresponding effects of *other_status* in the GS treatment.

Additionally, the significant influence of lagged helping behaviour in the IS treatment suggests that individual tendencies towards helping are persistent over time. This may reflect a personal norm or habit of cooperation that is less directly tied to the reputational incentives provided by the mechanism.

The varying effects of the *Falk_* variables across treatments and cost conditions further suggest that individual social preferences interact differently with the GS and IS mechanisms. In the GS treatment, preferences related to favouring others and considering future consequences are more influential, while in the IS treatment, tendencies towards revenge and risk aversion play a more significant role.

These findings support our earlier conjectures regarding the effectiveness of the GS mechanism in fostering reciprocal cooperation. The stronger influence of the recipient's status and the active player's ability to improve their own status in the GS treatment indicates that this mechanism provides clearer and more direct reputational incentives for helping. In the IS treatment, helping behaviour appears to be influenced by a combination of reputational concerns and individual dispositions towards cooperation.

By examining these differences, we gain a deeper understanding of how reputational mechanisms can be designed to promote cooperative behaviour. The GS mechanism, by incorporating a norm of justified defection and providing more nuanced reputational information, seems better equipped to sustain reciprocal helping in environments where indirect reciprocity is essential.

2.6.2.3 Cost-benefit analysis of helping

We now present a series of tables that summarise the helping behaviour observed in each treatment and cost condition. These tables provide a detailed analysis of the choices made by participants and allow us to examine what would constitute the rational choice based on an expected cost–benefit analysis. In the homogeneous cost condition, all players face the same cost when deciding whether to help. In the heterogeneous cost condition, players are divided into two groups: those facing a low cost and those facing a high cost when deciding whether to help. We keep the analysis of these groups separate.

Homogeneous Cost in GS Treatment

Table 2.6 displays the data for the GS treatment under the homogeneous cost conditions. The scores for both active and non-active players are binary, taking the value of 0 or 1.

Table 2.6: Helping percentages in GS treatment in homogeneous cost games

		Non-active player		Proportion of active players
	scores	0	1	
Active player	0	10.88%	20.53%	39.01%
	1	11.90%	82.33%	61.00%
Proportion of non-active players		41.32%	58.68%	
Actual helped observed		11.48%	59.17%	
Expected experience		11.50%	58.23%	

The entries in table 2.6 below the scores represent the percentage of times an active player provided help, given their own score and the score of the non-active player. For example, an active player with a score of 1 helped a non-active player with a score of 1 in 82.33% of cases. The column labelled "Proportion of active players" indicates the distribution of scores among active players; 61.00% of active players had a score of 1.

The row "Proportion of non-active players" shows the distribution of scores among non-active players. The "Actual helped observed" row provides the overall percentage of times help was provided to non-active players with each score, regardless of the active player's score. For instance, non-active players with a score of 1 were helped in 59.17% of cases.

The "Expected experience" is a weighted average, calculated by considering the frequency of help given by active players with each score and the proportion of active players holding that score. Specifically, it is computed by summing, for each active player score, the product of the percentage of help given and the proportion of active players with that score (last column of the table). This metric reflects the expected probability of receiving help for a non-active player with a given score, considering the distribution of active players' scores.

Table 2.7 presents the cost–benefit analysis for the active players in the homogeneous cost GS treatment. Active players incur a cost of £4 when helping to have a score of 1, and a cost of £0 when no helping. The expected benefit is calculated by multiplying the "Expected experience" from Table 2.6 by £10, which is the benefit received when helped. The net gain is the expected benefit minus the cost.

We observe that active players with a score of 1 achieve a higher net gain (£1.82)

Table 2.7: Cost–benefit analysis in homogeneous cost games for GS treatment

Active Player's Score	Cost	Expected Benefit	Net Gain
0	£0	£1.15	£1.15
1	£4	£5.82	£1.82

compared to those with a score of 0 (£1.15). This suggests that, from a self-interested perspective, choosing to have or keep a score of 1 and incurring, when needed, the cost of helping is the rational choice, as it leads to greater expected returns.

Heterogeneous Cost in GS Treatment - Low cost players

Table 2.8 shows the helping behaviour for active players with a low cost in the heterogeneous cost games in the GS treatment.

Table 2.8: Helping percentages in GS treatment in heterogeneous cost games for low cost players

		Non-active player		Proportion of active players
	scores	0	1	
Active player	0	9.68%	29.09%	20.28%
	1	17.56%	91.30%	79.72%
Proportion of non-active players		38.21%	61.79%	
Actual helped observed		16.05%	78.24%	
Experienced experience		15.96%	78.68%	

Active players with a score of 1 helped non-active players with a score of 1 in 91.30% of cases, a high rate reflecting the lower cost of £2 for helping.
The expected benefit and net gain for these players are shown in Table 2.9.

Active Player's Score	Cost	Expected Benefit	Net Gain
0	£0	£1.60	£1.60
1	£2	£7.87	£5.87

Table 2.9: Cost–benefit analysis in heterogeneous cost games for GS treatment, low cost players

The net gain for active players with a score of 1 (£5.87) is substantially higher than for those with a score of 0 (£1.60), reinforcing the incentive to help when costs are low.

Heterogeneous Cost in GS Treatment - High cost players

Table 2.10 presents the helping behaviour for active players with a high cost in the GS treatment.

Table 2.10: Helping percentages in GS treatment in heterogeneous cost games for high cost players

		Non-active player		Proportion of active players
	scores	0	1	
Active player	0	9.68%	19.01%	46.36%
	1	4.76%	53.64%	54.13%
Proportion of non-active players		28.41%	71.60%	
Actual helped observed		7.20%	52.70%	
Experienced experience		7.04%	51.91%	

Here, the cost of helping is £6 for active players. The rate at which these players help non-active players with a score of 1 drops to 53.64%, significantly lower than in the low-cost condition.

The cost–benefit analysis in Table 2.11 reveals that the net gain for active players with a score of 1 is negative (£–0.81), making helping a less attractive option.

Table 2.11: Cost–benefit analysis in heterogeneous cost games for GS treatment, high cost players

Active Player's Score	Cost	Expected Benefit	Net Gain
0	£0	£0.70	£0.70
1	£6	£5.19	-£0.81

Comparing the three cost conditions in the GS treatment, we observe a clear pattern: helping behaviour is sensitive to the cost of helping. When the cost is low, active players with a score of 1 are highly likely to help, resulting in higher net gains. As the cost increases, the propensity to help decreases, and the net gains for helping diminish or even become negative.

This pattern suggests that participants are responding to the incentives embedded in the cost structure, adjusting their behaviour in a manner consistent with rational self-interest. The higher net gains associated with helping when costs are low encourage reciprocal behaviour, while high costs deter helping.

Homogeneous Cost in IS Treatment

We now turn to the IS treatment, where scores range from 0 to 5. Table 2.12 shows the helping percentages under homogeneous cost conditions.

Table 2.12: Helping percentages in IS treatment with homogeneous costs

Active player score	Non-active player score						Proportion of active players
	0	1	2	3	4	5	
0	0%	0%	0%	10.00%	9.09%	16.67%	4.98%
1	66.67%	NA	50.00%	50.00%	50.00%	28.57%	3.24%
2	55.56%	25.00%	61.54%	39.47%	75.00%	55.56%	12.27%
3	50.00%	69.23%	72.41%	71.67%	73.33%	57.14%	23.38%
4	83.33%	22.22%	62.50%	67.19%	77.05%	82.61%	27.89%
5	60.00%	71.43%	76.00%	82.93%	85.51%	95.65%	28.24%
Proportion of non-active players	4.75%	4.28%	13.43%	25.35%	26.74%	25.46%	
Actual helped observed	58.54%	45.95%	61.21%	63.47%	74.46%	79.09%	
Experienced experience	60.85%	45.62%	65.00%	65.87%	74.59%	71.98%	

Notes: NA: situation not happened.

In this table, each cell below the scores represents the percentage of times an active player with a given score helped a non-active player with a particular score. For example, an active player with a score of 5 helped a non-active player with a score of 5 in 95.65% of cases. The “Proportion of active players” column shows the distribution of scores among active players, whereas the “proportion of non-active players” raw the distribution of scores among non-active players.

The “Actual helped observed” row provides the overall percentage of times non-active players with each score were helped. The “Expected experience” row calculates the weighted average of helping probabilities, considering the distribution of active players’ scores.

Table 2.13 shows the cost–benefit analysis for active players in the homogeneous cost IS treatment. In this case, the cost of helping is calculated as cost of having one score. The latter varies as a function of how many times you need to help in order to get that score. For example, an active player needs to help 5 times in a row to have a score of 5, therefore paying the full cost (i.e., for the homogeneous cost games, £4). In case they want to have a score of 4, they have to help 4/5 times, therefore the cost that they’ve to pay £3.20. Formally, the cost for an active player to have a score S is calculated as $cost = (\frac{S}{5}) \times £4$.

Interestingly, the highest net gain is achieved with a score of 0 (£6.09). However, in Section 2.6.2.1, we observed that participants often maintained higher scores and helped more at higher rates. This suggests that motivations beyond immediate

Table 2.13: Cost–benefit analysis in homogeneous cost games for IS treatment

Active Player's Score	Cost	Expected Benefit	Net Gain
0	£0	£6.09	£6.09
1	£0.80	£4.56	£3.76
2	£1.60	£6.50	£4.90
3	£2.40	£6.59	£4.19
4	£3.20	£7.46	£4.26
5	£4.00	£7.20	£3.20

monetary returns, such as reputational concerns or cooperative norms, might play a role in this case.

Similar to the GS treatment, we examine the helping behaviour under heterogeneous cost conditions for the IS treatment both for high and low costs.

Heterogeneous Cost in IS Treatment - Low cost players

Table 2.14 presents the helping percentages for low-cost active players in the IS treatment.

Table 2.14: Helping percentages in IS treatment in heterogeneous cost games for low cost players

Active player	Non-active player						Proportion of active players
	0	1	2	3	4	5	
0	0%	100%	NA	0%	0%	NA	3.33%
1	50.00%	100%	0%	100%	0%	66.67%	2.62%
2	66.67%	100%	0%	62.50%	83.33%	NA	4.76%
3	61.54%	87.50%	71.43%	81.82%	94.44%	94.74%	22.38%
4	58.33%	11.11%	57.89%	81.08%	89.47%	90.24%	37.14%
5	76.92%	75.00%	66.67%	82.61%	92.68%	92.11%	29.76%
Proportion of non-active players	10.71%	5.71%	10.24%	23.57%	25.71%	24.05%	
Actual helped observed	62.22%	58.33%	58.14%	73.74%	87.04%	91.09%	
Experienced experience	62.81%	56.66%	57.33%	67.67%	85.92%	81.20%	

Active players with higher scores tend to help frequently, and the net gains, as shown in Table 2.15, remain positive across all scores. For low-cost players, the total cost per period is lower (e.g., £2). The cost calculations are adjusted accordingly.

In this case, the highest net gain is achieved with a score of 4 (£6.99). This is in line with the score that players keep as outlined previously. For low cost players it is a self-interest rational choice keeping a score of 4 and helping most of the time.

Heterogeneous Cost in IS Treatment - High cost players

Table 2.16 shows the helping behaviour for high-cost active players in the IS treatment.

Table 2.15: Cost–benefit analysis in heterogeneous cost games for IS treatment, low cost players

Active Player's Score	Cost	Expected Benefit	Net Gain
0	£0	£6.28	£6.28
1	£0.40	£5.67	£5.27
2	£0.80	£5.73	£4.93
3	£1.20	£6.68	£5.48
4	£1.60	£8.59	£6.99
5	£2.00	£8.12	£6.12

Table 2.16: Helping percentages in IS treatment in heterogeneous cost games for high cost players

Active player	Non-active player						Proportion of active players
	0	1	2	3	4	5	
0	0%	0%	0%	5.00%	0%	0%	15.76%
1	0%	NA	NA	42.86%	42.86%	25.00%	6.08%
2	15.38%	NA	100%	76.92%	75.00%	13.29%	12.30%
3	11.11%	100%	45.45%	46.67%	73.33%	63.64%	20.05%
4	0%	50.00%	50.00%	66.67%	62.50%	85.71%	21.40%
5	NA	66.67%	100%	82.35%	86.49%	92.86%	23.42%
Proportion of non-active players	9.01%	3.60%	6.98%	18.24%	31.53%	30.63%	
Actual helped observed	7.50%	31.25%	48.39%	50.62%	67.14%	68.38%	
Experienced experience	4.12%	46.36%	55.53%	55.77%	60.16%	56.00%	

The net gains for these players, presented in Table 2.17, are reduced and become negative for the highest score of 5, reflecting the higher costs.

Table 2.17: Cost–benefit analysis in heterogeneous cost games for IS treatment, high cost players

Active Player's Score	Cost	Expected Benefit	Net Gain
0	£0	£0.41	£0.41
1	£1.20	£4.64	£3.44
2	£2.40	£5.55	£3.15
3	£3.60	£5.58	£1.98
4	£4.80	£6.02	£1.22
5	£6.00	£5.60	-£0.40

In this case, the highest net gain is achieved with a score of 1 (£3.44). For high cost players it is a self-interest rational choice keeping a low score and helping rarely. This is not in line with that we observed in previous sections.

In general, the IS treatment shows a pattern of helping behaviour that is sensitive to costs, much like the GS treatment. However, helping remains more prevalent than one would predict from purely cost–benefit calculations: even when the net gains from maintaining a higher score are small or negative, participants continue to help at relatively high rates. This suggests that participants are not guided solely by short-term monetary returns. Indeed, as discussed in Section 2.6.2.1, participants in the IS treatment maintain higher scores by helping even under high-cost conditions, indicating that reputational concerns or social preferences may play a key role.

The observed differences between GS and IS are significant for understanding reciprocity and cooperation. In the GS treatment, helping behaviour closely aligns with the mechanism's incentives: participants adjust their actions in ways consistent with rational self-interest, striving to maximise net gains. The strong correlation between standing and being helped reflects how effectively the GS rule fosters reciprocity rooted in direct, tit-for-tat exchanges.

By contrast, the IS treatment seems to induce a form of reciprocity that goes beyond strict self-interest. Participants frequently help in situations where cost–benefit analyses alone would predict lower cooperation, and they even help individuals who themselves have not been helping. This pattern points to a sense of generalised reciprocity or cooperative norms that encourage helping regardless of immediate personal gains. Consequently, the higher levels of helping in the IS treatment underscore the potential influence of social preferences—such as fairness or altruism—in economic decision-making.

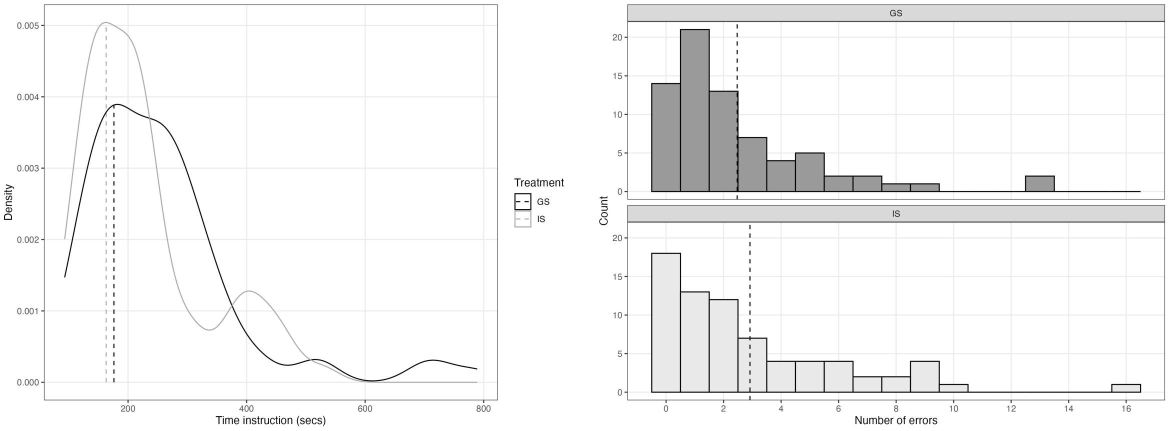
Overall, these results highlight the distinctive ways in which GS and IS mechanisms shape reciprocity and cooperation. While GS consistently incentivises reciprocal cooperation in line with individual cost–benefit reasoning, IS fosters cooperation that exceeds what self-interest would suggest, hinting at a broader range of social motives. Our comparative analysis thus shows that GS elicits higher levels of reciprocal cooperation (as indicated by stronger correlations between standing and helping), whereas IS promotes cooperation beyond immediate self-interest, likely driven by social norms and other-regarding preferences.

2.6.2.4 Analysis of control questions data

Finally, we examine the data pertaining to participants' comprehension of the game and the two distinct treatment modalities. To assess players' cognitive grasp, an analysis of response times and errors in the control questions for each treatment was conducted. The Image Scoring treatment comprised 13 questions, while the Good

Standing treatment consisted of 12 questions, with 8 questions common to both treatments, 5 specifically testing Image Scoring, and 4 specifically probing Good Standing.

Figure 2.13 try to provide a comparison of participants' comprehension between the two treatments, as reflected through response times and the number of errors made during control questions. The left panel 2.13a displays density plots representing the distribution of time taken to complete the instructions for each treatment. The Image Scoring (IS) treatment is shown in grey, while the Good Standing (GS) treatment is depicted in black. Notably, both density curves exhibit a peak around the 200-second mark, indicating that participants in both treatments generally required a similar amount of time to complete the instructions. However, the IS treatment's curve has a sharper peak and a more abrupt decline post-peak, suggesting that participants' engagement with the instructions might have been more uniform and concise compared to the GS treatment.



(a) Time in answering control questions by treatment

(b) Number of wrong answer to control questions by treatment

Figure 2.13: Analysis of control questions

The right panel 2.13b consists of histograms showing the distribution of errors made by participants in answering control questions across the two treatments. The top histogram represents the GS treatment, where the majority of participants made between 0 and 3 errors, with a noticeable concentration peaking around 2 errors. Conversely, the bottom histogram illustrates the distribution for the IS treatment, where a broader spread is observed. While many participants also made between 0 and 3 errors, the IS treatment features outliers with significantly higher numbers of errors, indicating a more varied level of comprehension.

These plots collectively suggest that although the time taken to process the instructions was similar between the two treatments, it would be useful to look at the time spent to answer the same questions between the two groups and the time spent to answer to the different questions. This would help to provide a better comparison and understanding of people's comprehension of the instructions.

2.7 Conclusion

In this chapter, we experimentally investigate the interplay between reputation mechanisms and helping behaviour, offering insights with potential applications to diverse economic challenges such as crowdsourcing systems, e-commerce, and knowledge-sharing networks. Our helping game establishes a novel experimental platform where helping incurs costs for the active player but gives benefit to the non-active player, aligning with scenarios involving non-monetary costs and benefits, as exemplified by forums where answering questions entails a temporal cost and an intrinsic pleasure from helping others.

Employing a between-subject design, we manipulate reputational mechanisms to create exogenous variation and assess their causal impact on helping behaviour. Our laboratory experiment shows that cooperation increases substantially under the Image Scoring (IS) mechanism, which proves empirically effective in sustaining higher levels of cooperation. Conversely, the Good Standing (GS) mechanism is effective in fostering reciprocal behaviour in this setting.

Our findings contribute to the literature on indirect reciprocity and reputation systems by providing empirical evidence on how different reputational mechanisms influence cooperative behaviour. Specifically, we replicate and extend the results of Engelmann and Fischbacher (2009), demonstrating that the average helping rates in our IS treatment with homogeneous costs closely align with their findings in treatments where both participants maintain public scores. This replication strengthens the external validity of our experimental design and supports the robustness of the IS mechanism in promoting cooperation.

Contrary to our initial belief related with Conjecture 1, which posited that the GS mechanism would induce higher levels of cooperation, our results indicate that the IS mechanism consistently generates higher average helping rates across both homogeneous and heterogeneous cost conditions. Statistical analysis at the cohort level reveals that the mean helping rates are significantly higher under IS than GS, with the difference being more pronounced in the homogeneous cost condition. These findings suggest that the simplicity and transparency of the IS mechanism make it more effective in sustaining cooperative behaviour among participants.

However, when examining reciprocity patterns, we find support for our anticipation regarding Conjecture 2. The GS mechanism induces stronger reciprocity in helping behaviour compared to the IS mechanism. Regression analyses show that in the GS treatment, the likelihood of an active player helping is strongly influenced by the non-active player's standing and the potential to improve one's own standing. The large and significant coefficients on these variables highlight the central role of reputational incentives within the GS mechanism. In contrast, while the IS mechanism also shows that the non-active player's image score influences helping decisions, the effect is less pronounced. Participants in the IS treatment exhibit helping behaviour that is influenced by a combination of reputational concerns and individual dispositions towards cooperation, as indicated by the significant impact of lagged helping behaviour.

Our cost–benefit analysis further reveals interesting contrasts between the two mechanisms. In the GS treatment, participants' helping behaviour aligns closely with the incentives provided by the mechanism itself. Active players adjust their actions in a manner consistent with rational self-interest, favouring choices that maximise their net gains. The strong correlation between standing and the frequency of being helped reinforces the effectiveness of the GS mechanism in promoting reciprocity based on direct exchanges.

In the IS treatment, however, we observe that participants help more than what would be predicted by immediate monetary returns. Even when the net gains from maintaining higher image scores are modest or negative, participants continue to help at relatively high rates. This suggests that factors beyond immediate self-interest, such as social preferences, altruism, or a desire to maintain a cooperative environment, influence decision-making in the IS treatment. The IS mechanism appears to foster a sense of generalised reciprocity or cooperative norms that encourage helping behaviour regardless of immediate personal gain.

Despite the insights gained, our study also raises intriguing questions that warrant further investigation. One such question is why the IS mechanism induces a high degree of cooperation or non-reciprocal help within our experimental helping game. Understanding the underlying motives driving this behaviour could provide valuable insights into how reputation systems influence social dynamics.

To unravel this phenomenon, we propose four potential directions for future research: (i) Providing players with more information: furnishing players with additional information about the helping rates associated with each image score could help them make more informed decisions. This might involve displaying statistical summaries or trends that highlight how helping behaviour correlates with different scores. (ii) Eliciting players' beliefs: gathering data on players' beliefs about the behaviour of others could shed light on how expectations influence cooperation. By understanding how players perceive the likelihood of being helped based on their own and others' scores, we can better comprehend the decision-making processes underpinning their actions. (iii) Conducting repeated dictator games: implementing repeated dictator games with fixed player roles could help isolate the effects of reputation mechanisms when the possibility of receiving help is not directly impacted by one's own score. This approach would allow us to assess how scoring dynamics evolve when the incentive structure is altered, providing insights into the role of self-interest and non-self-interest motives in cooperative settings. Finally, (iv) employing multiple sequences (supergames) may help us to observe behaviour across diverse repeated interactions and capture how learning unfolds over time.

These avenues promise to clarify motives behind image scoring, revealing dynamics of self-interest and non-self-interest in cooperative environments. Exploring variations in helping games, such as altering cost-benefit ratios or introducing uncertainty, could further illuminate how reputation mechanisms operate across diverse conditions.

In conclusion, our study demonstrates strengths and limitations of different reputation mechanisms and how they significantly influences cooperative and reciprocal behaviour. The GS mechanism promotes strong reciprocity by aligning reputational incentives with justified defection, while the IS mechanism achieves higher overall cooperation levels, potentially due to the influence of social norms and other-regarding preferences.

References

Alexander, R. D.: 1987, *The biology of moral systems*, New York: Aldine de Gruyter.

Andreoni, J. and Miller, J. H.: 1993, Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence, *The Economic Journal* **103**(418), 570–585.

Aoyagi, M., Bhaskar, V. and Fréchette, G. R.: 2019, The impact of monitoring in infinitely repeated games: Perfect, public, and private, *American Economic Journal: Microeconomics* **11**(1), 1–43.
URL: *https://www.aeaweb.org/articles?id=10.1257/mic.20160304*

Aoyagi, M. and Fréchette, G.: 2009, Collusion as public monitoring becomes noisy: Experimental evidence, *Journal of Economic Theory* **144**(3), 1135–1165.
URL: *https://www.sciencedirect.com/science/article/pii/S0022053108001555*

Arechar, A. A., Dreber, A., Fudenberg, D. and Rand, D. G.: 2017, "i'm just a soul whose intentions are good": The role of communication in noisy repeated games, *Games and Economic Behavior* **104**, 726–743.
URL: *https://www.sciencedirect.com/science/article/pii/S0899825617301112*

Axelrod, R.: 1984, *The Evolution of Cooperation*, Basic books, Basic Books.

Berg, J. E., Dickhaut, J. and McCabe, K.: 1995, Trust, reciprocity, and social history, *Games and Economic Behavior* **10**, 122–142.

Bigoni, M., Camera, G. and Casari, M.: 2020, Money is more than memory, *Journal of Monetary Economics* **110**, 99–115.

Bolton, G. E., Katok, E. and Ockenfels, A.: 2005, Cooperation among strangers with limited information about reputation, *Journal of Public Economics* **89**(8), 1457–1468.

Bolton, G. E. and Ockenfels, A.: 2000, Erc: A theory of equity, reciprocity, and competition, *The American Economic Review* **90**, 166–193.

Brandts, J., Saijo, T. and Schram, A.: 2000, How universal is behavior? a four country comparison of spite, cooperation and errors in voluntary contribution mechanisms, *Experimental & Empirical Studies eJournal* .

Camera, G. and Casari, M.: 2009, Cooperation among strangers under the shadow of the future, *American Economic Review* **99**(3), 979–1005.
URL: *https://www.aeaweb.org/articles?id=10.1257/aer.99.3.979*

Camera, G. and Casari, M.: 2018, Monitoring institutions in indefinitely repeated games., *Experimental Economics* **21**, 673–691.

Charness, G., Du, N. and Yang, C.-L.: 2011, Trust and trustworthiness reputations in an investment game, *Games and Economic Behavior* **72**(2), 361–375.

Chen, D. L., Schonger, M. and Wickens, C.: 2016, otree—an open-source platform for laboratory, online, and field experiments, *Journal of Behavioral and Experimental Finance* **9**, 88–97.

Cooper, R., Dejong, D., Forsythe, R. and Ross, T.: 1996, Cooperation without reputation: Experimental evidence from prisoner's dilemma games, *Games and Economic Behavior* **12**(2), 187–218.

Dal Bó, P.: 2005, Cooperation under the shadow of the future: Experimental evidence from infinitely repeated games, *American Economic Review* **95**(5), 1591–1604.
URL: *https://www.aeaweb.org/articles?id=10.1257/000282805775014434*

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

Engelmann, D. and Fischbacher, U.: 2008, Indirect reciprocity and strategic reputation building in an experimental helping game, *Working Paper 34*, Thurgauer Wirtschaftsinstitut.

Engelmann, D. and Fischbacher, U.: 2009, Indirect reciprocity and strategic reputation building in an experimental helping game, *Games and Economic Behavior* **67**(2), 399–407.

Engle-Warnick, J. and Slonim, R. L.: 2006, Learning to trust in indefinitely repeated games, *Games and Economic Behavior* **54**(1), 95–114.
URL: *https://www.sciencedirect.com/science/article/pii/S0899825604001678*

Erkut, H. and Reuben, E.: 2023, Social networks and organizational helping behavior: Experimental evidence from the helping game, *WZB Discussion Paper SP II 2023–203*, Wissenschaftszentrum Berlin für Sozialforschung (WZB).

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D. and Sunde, U.: 2018, Global evidence on economic preferences*, *The Quarterly Journal of Economics* **133**(4), 1645–1692.
URL: *https://doi.org/10.1093/qje/qjy013*

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

Fehr, E., Gachter, S. and Kirchsteiger, G.: 1997, Reciprocity as a contract enforcement device: Experimental evidence, *Econometrica* **65**(4), 833–860.

Fehr, E. and Schmidt, K. M.: 1999, A Theory of Fairness, Competition, and Cooperation*, *The Quarterly Journal of Economics* **114**(3), 817–868.

Friedman, J.: 1971, A non-cooperative equilibrium for supergames, *The Review of Economic Studies* **38**(1), 1–12.
URL: *https://EconPapers.repec.org/RePEc:oup:restud:v:38:y:1971:i:1:p:1-12.*

Fréchette, G. R. and Yuksel, S.: 2017, Infinitely repeated games in the laboratory: four perspectives on discounting and random termination, *Experimental Economics* **20**(2), 279–308.
URL: *https://EconPapers.repec.org/RePEc:kap:expeco:v:20:y:2017:i:2:d:10.1007_s10683-016-9494-z*

Fudenberg, D. and Maskin, E.: 1986, The folk theorem in repeated games with discounting or with incomplete information, *Econometrica* **54**(3), 533–554.
URL: *http://www.jstor.org/stable/1911307*

Fudenberg, D., Rand, D. G. and Dreber, A.: 2012, Slow to anger and fast to forgive: Cooperation in an uncertain world, *American Economic Review* **102**(2), 720–49.
URL: *https://www.aeaweb.org/articles?id=10.1257/aer.102.2.720*

Ghidoni, R. and Suetens, S.: 2022, The effect of sequentiality on cooperation in repeated games, *American Economic Journal: Microeconomics* **14**(4), 58–77.
URL: *https://www.aeaweb.org/articles?id=10.1257/mic.20200268*

Gong, B. and Yang, C.-L.: 2019, Cooperation through indirect reciprocity: The impact of higher-order history, *Games and Economic Behavior* **118**, 316–341.

Greiff, M. and Paetzel, F.: 2020, Information about average evaluations spurs cooperation: An experiment on noisy reputation systems, *Journal of Economic Behavior & Organization* **180**, 334–356.

Hopp, D. and Süß, K.: 2024, How altruistic is indirect reciprocity? — evidence from gift-exchange games in the lab, *Journal of Behavioral and Experimental Economics* **108**, 102127.

Irlenbusch, B. and Sliwka, D.: 2005, Incentives, decision frames, and motivation crowding out – an experimental investigation, *IZA Discussion Papers 1758*, Institute of Labor Economics (IZA).

Keser, C.: 2003, Experimental games for the design of reputation management systems, *IBM Systems Journal* **42**(3), 498–506.

Leimar, O. and Hammerstein, P.: 2001, Evolution of cooperation through indirect reciprocity, *Proc. R. Soc. Lond. B.* **268**, 745–753.

Levine, D.: 1998, Modeling altruism and spitefulness in experiment, *Review of Economic Dynamics* **1**(3), 593–622.

Lugovskyy, V., Puzzello, D., Sorensen, A., Walker, J. and Williams, A.: 2017, An experimental study of finitely and infinitely repeated linear public goods games, *Games and Economic Behavior* **102**, 286–302.
URL: *https://www.sciencedirect.com/science/article/pii/S0899825617300052*

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

Milinski, M., Semmann, D., Bakker, T. C. and Krambeck, H. J.: 2001, Cooperation through indirect reciprocity: image scoring or standing strategy?, *Proceedings of the Royal Society of London. Series B: Biological Sciences* **268**, 2495 – 2501.

Nowak, M. and Sigmund, K.: 1998a, The dynamics of indirect reciprocity, *Journal of Theoretical Biology* **194**(4), 561–574.

Nowak, M. and Sigmund, K.: 1998b, Evolution of indirect reciprocity by image scoring, *Nature* **393**, 573–577.

Palfrey, T. and Rosenthal, H.: 1994, Repeated play, cooperation and coordination: An experimental study, *The Review of Economic Studies* **61**(3), 545–565.
URL: *https://EconPapers.repec.org/RePEc:oup:restud:v:61:y:1994:i:3:p:545-565.*

Rabin, M.: 1993, Incorporating fairness into game theory and economics, *American Economic Review* **83**, 1281–1302.

Rand, D. G. and Nowak, M. A.: 2013, Human cooperation, *Trends in Cognitive Sciences* **17**(8), 413–425.

Roth, A. E. and Murnighan, J. K.: 1978, Equilibrium behavior and repeated play of the prisoner's dilemma, *Journal of Mathematical Psychology* **17**(2), 189–198.

Seinen, I. and Schram, A.: 2006, Social status and group norms: Indirect reciprocity in a repeated helping experiment, *European Economic Review* **50**(3), 581–602.

Sugden, R.: 1986, *The economics of rights, co-operation and welfare*, Oxford, UK: Basil Blackwell.

Swakman, V., Molleman, L., Ule, A. and Egas, M.: 2016, Reputation-based cooperation: empirical evidence for behavioral strategies, *Evolution and Human Behavior* **37**(3), 230–235.

Trivers, R.: 1971, The evolution of reciprocal altruism, *The Quarterly review of biology* **46**, 35–57.

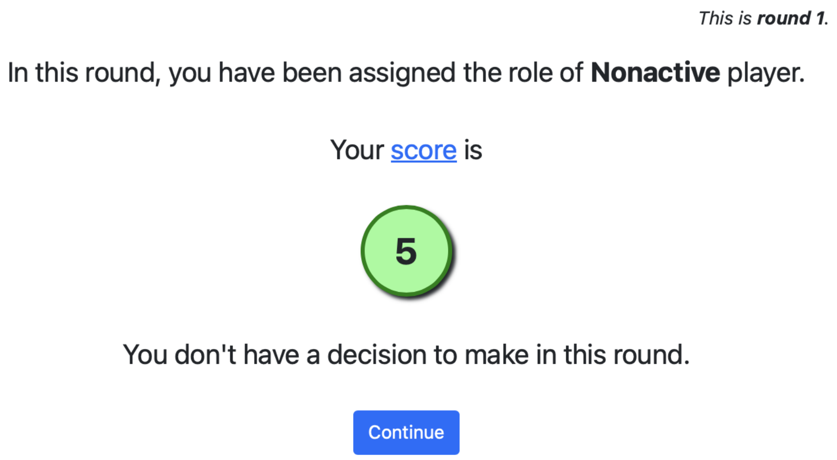
Ule, A., Schram, A., Riedl, A. and Cason, T. N.: 2009, Indirect punishment and generosity toward strangers, *Science* **326**(5960), 1701–1704.

Wedekind, C. and Milinski, M.: 2000, Cooperation through image scoring in humans, *Science* **288**(5467), 850–852.

Wibral, M.: 2015, Identity changes and the efficiency of reputation systems, *Exp Econ* **18**(5960), 408–431.

2.A Appendix

2.A.1 Additional Analysis and Figures



*This is **round 1**.*

In this round, you have been assigned the role of **Nonactive** player.

Your score is

5

You don't have a decision to make in this round.

Figure 2.14: Non-active player's screen in IS treatment

Table 2.18: Summary Statistics

	%
Gender:	
Female	50.69
Male	47.22
Other	1.40
Prefer not to say	0.69
Age:	
$\geq$27	14.08
26	2.11
25	7.04
24	3.52
23	14.08
22	13.38
21	12.68
20	15.49
19	14.79
18	2.83
Degree:	
Bachelor	59.03
Master	31.94
PhD	2.78
Other degree course or affiliation	2.78
Prefer not to say	2.09
INTO	0.69
Staff	0.69
Year(s) at UEA:	
1^{st} year	34.72
2^{nd} year	29.17
3^{rd} year	17.36
4^{th} year	9.03
More than 4 years	7.64
Prefer not to say	2.08
Faculty:	
Social Sciences	63.89
Sciences	22.22
Humanities	10.42
Others	3.47

Notes: The percentages are based on the total number of responses for each category. The category "Other" under Gender includes non-binary and other gender identities not specified. The age categories are divided to highlight different age groups. As per faculty, *Social Sciences:* NBS, ECO, DEV, LAW, EDU, PSY; *Sciences:* CMP, BIO, ENV, PHA, MED, HSC, MTH; *Humanities:* PPL, HIS, LDC, AMA, HUM.

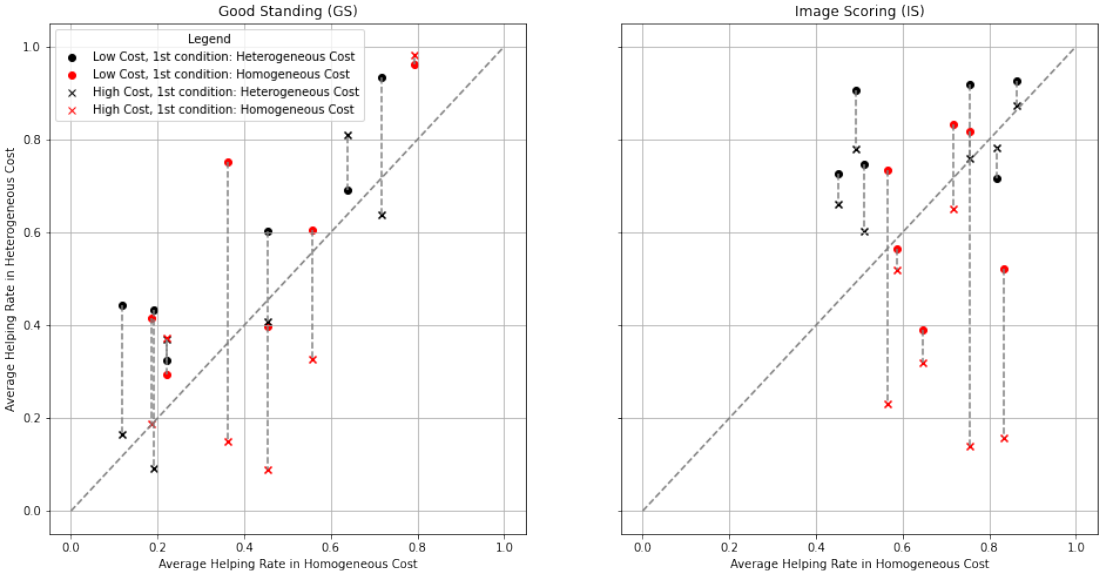


Figure 2.15: Correlation average helping rates GS vs IS

2.A.2 Experimental Instructions

Welcome

Welcome to today's experiment and thanks for coming. This is an experiment in the economics of decision-making. If you follow the instructions, complete the experiment, and make appropriate choices, you can earn an appreciable amount of money. This will be paid to you in private, in cash, at the end of the session, before you leave the laboratory.

It is important that you remain silent and do not look at other people's work. If you have any questions, or need assistance of any kind, please raise your hand and an experimenter will come to you. If you talk, laugh, exclaim out loud, etc., you will be asked to leave and you will not be paid. We expect and appreciate your cooperation.

All choices in today's experiment and any information you choose to give are recorded anonymously and will only be used in the analysis of the data from this experiment.

We will now describe the nature of the experiment in more detail.

Introduction

In this experiment, you will be assigned to a **group** of **six** people. The other people in your group will be five other people in this room. You will never find out which other people are the ones who are in your group.

The experiment has **two parts**. We will now describe Part 1. We will describe Part 2 after everyone has finished Part 1.

Part 1

The interaction

There will be a series of at least 40 **rounds**. At the beginning of each round, the computer will match you at random with another person in your group. You will **interact** with that person in that round. Because the computer will create a new matching at the beginning of each round, in different rounds you may interact with different members of your group. You will not find out which other person in your group you are matched with, nor whether or when you have been matched with them previously.

In each match, one of you will be randomly assigned the role of the **active player**. The other one will be assigned the role of the **non-active player**.

If you are the active player, you have an opportunity to help the non-active player. **Helping** is an action that has a monetary **cost of helping** for the active player and gives a monetary **benefit** to the non-active player.

You will find out your cost of helping for Part 1 before Round 1 begins. Your cost of helping will stay the same throughout Part 1. The cost of helping may be different for different members of your group. However, the cost of helping is no more than £6.00 for anyone.

In every round, you have an **account**. This account is separate for each round. At the start of each round, your account for that round will have an **endowment** of £7.00. The final value of the accounts of the active player and the non-active player depends on the choice the active player makes:

- If the active player **chooses to help**, the active player's cost of helping will be deducted from their account, and a benefit of £10.00 will be added to the non-active player's account.
- If the active player **chooses not to help**, no deduction will be made from the active player's account and no addition made to the non-active player's account.

After the active player makes their choice, any deductions or additions are made to the player's accounts for that round. The round then ends. When the next round begins, all players have a new account, with an endowment of £7.00.

Your earnings from Part 1

Only one of the rounds will be **for real**. The computer will randomly select which round this is, but you will not know which round was selected until the end of the experiment. Which round this is will be the same for all members of the group. For everyone in the group, their earnings from Part 1 will be equal to the final value of their account for the selected round.

Good Standing[17]

Scores

Through this part of the experiment, you will have a **score**. Your score summarises whether, in previous rounds, you chose to help or not to help. Your score can be 0 or 1.
In each round, you will see your own current score. In each round in which you are the non-active player, the active player will also see your current score. In each round in which you are the active player, you will see the current score of the non-active player.
At the start of Round 1, everyone has a score of 1.

The computer keeps a **record** of everyone's current score. At the end of each round in which you are the active player, your score is updated based on your choice and the score of the non-active player.

- If you choose to help, then your score at the start of the next round will be 1.
- If you choose not to help, and the non-active player's score was 1, then your score at the start of the next round will be 0.
- If you choose not to help, and the non-active player's score was 0, then your score at the start of the next round will be the same as your current score.

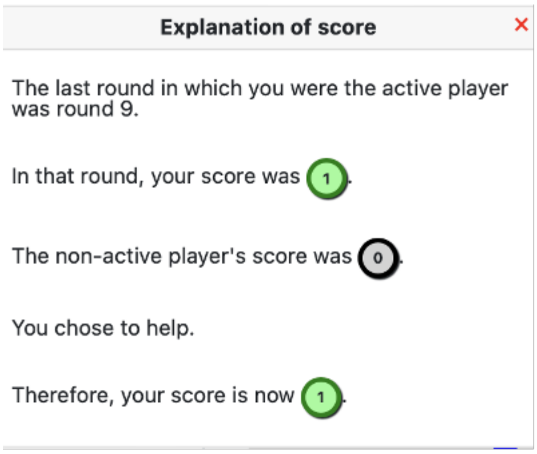
However, if this is your first round as the active player, your record will be filled in as if you had chosen to help in the previous round. So, your score in the first round of this part will be 1.

At the end of each round in which you are the active player, your score will be updated based on the choice you make in that round. In rounds in which you are

[17] This was not shown to the participants and it is displayed for differentiate the two versions of the instructions.

the non-active player, your score does not change.

For example, suppose in Round 11 you are the active player. The explanation of your score at the start of the round might be



Explanation of score

The last round in which you were the active player was round 9.

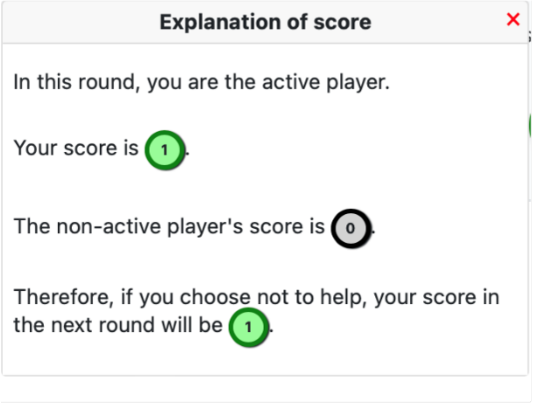
In that round, your score was 1.

The non-active player's score was 0.

You chose to help.

Therefore, your score is now 1.

This record shows that in the last round in which you were the active player, your score was 1 and the non-active player's score was 0. You chose to help, and therefore your score at the start of Round 11 is 1.
Suppose you are the active player in Round 11, and you are considering choosing not to help. The explanation of how this would affect your score would be



Explanation of score

In this round, you are the active player.

Your score is 1.

The non-active player's score is 0.

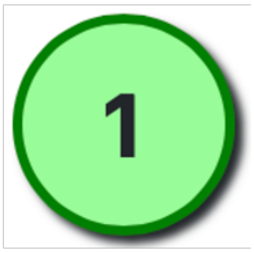
Therefore, if you choose not to help, your score in the next round will be 1.

Because the non-active player's score is 0, your score does not change. Therefore, your score at the start of Round 12 would be 1.

What you will see on the screen

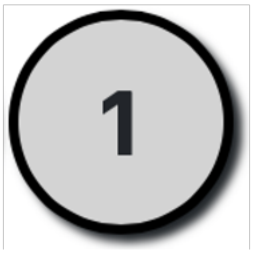
We will now show you what you will see on your screen during a round.

Scores will be displayed on the screen using **badges**, which are coloured circles showing the score. Your score will always be shown using a **green badge** like this



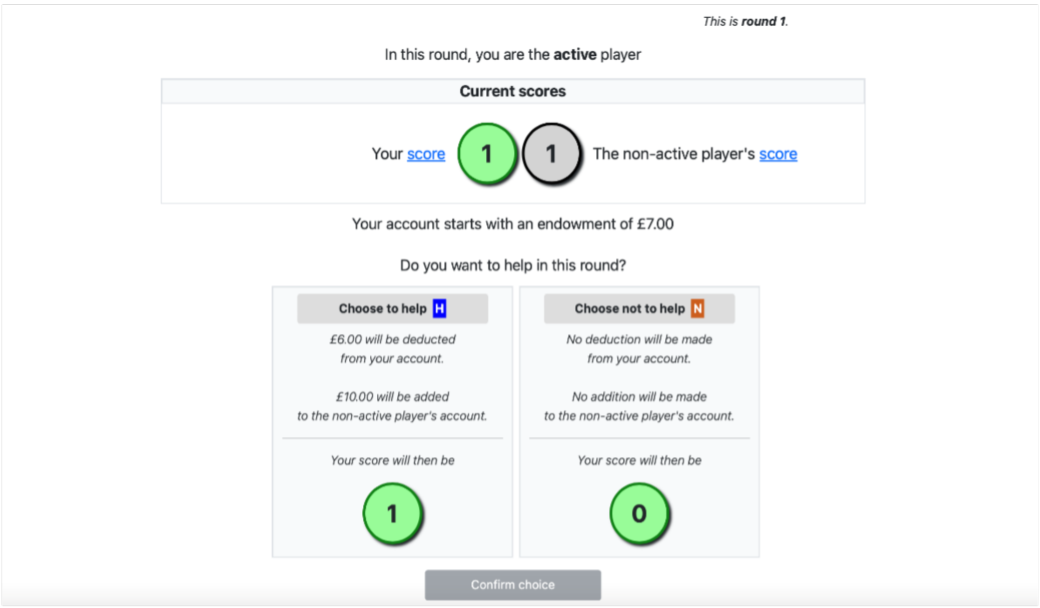
Any time you see a green badge representing your score, you can click on it to pop up a display which explains how your record was used to determine your score.

When you are the active player, you will see the score of the non-active player. The non-active player's score will be shown using a **grey badge** like this



You will only be able to see the current score of the non-active player, and not their record; therefore, clicking on a grey badge has no effect.

When you are the active player, you will see a screen like the one below



*This is **round 1**.*

In this round, you are the **active** player

Current scores

Your score

The non-active player's score

Your account starts with an endowment of £7.00

Do you want to help in this round?

£6.00 will be deducted from your account.

£10.00 will be added to the non-active player's account.

Your score will then be

No deduction will be made from your account.

No addition will be made to the non-active player's account.

Your score will then be

The screen shows the round number, your score, the non-active player's score, and your endowment in this round.

The two boxes represent the two choices you can make: to help or not to help. Each box summarises the consequences for you and for the non-active player of the corresponding choice.

To choose to help, click on the button labelled **Choose to help**. To choose not to help, click on the button labelled **Choose not to help**. When you click on one of those buttons, the background will change from grey to green. If you change your mind, you can change your choice simply by clicking on the button corresponding to the other choice. When you are satisfied with your choice, click the button labelled **Confirm choice**.

How Part 1 ends

Exactly how many rounds will be played is not fixed in advance. There will be at least 40 rounds. At the end of Round 40, the computer will simulate the roll of a six-sided die. If the roll is 1 or 2, then Part 1 will end; otherwise, Part 1 will continue to Round 41. This process will be repeated after each subsequent round, until the simulated die roll results in a 1 or 2. In other words, in each round starting from Round 40, there is a 1 in 3 chance that Part 1 will end after that round.

Image Scoring

Scores

Through this part of the experiment, you will have a **score**. Your score summarises whether, in previous rounds, you chose to help or not to help. Your score can be 0, 1, 2, 3, 4 or 5.

In each round, you will see your own current score. In each round in which you are the non-active player, the active player will also see your current score. In each round in which you are the active player, you will see the current score of the non-active player.

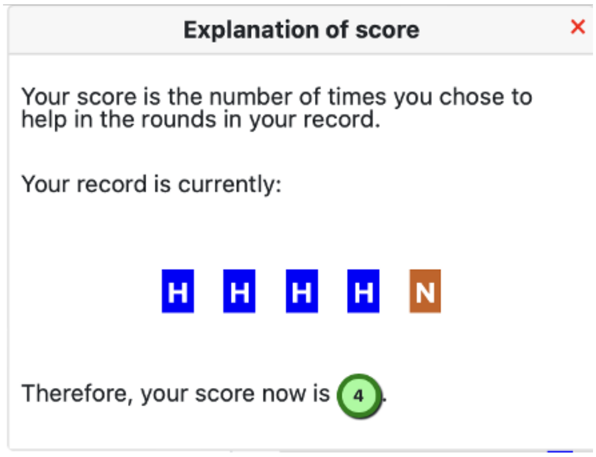
At the start of Round 1, everyone has a score of 5.

The computer keeps a **record** of your choices for the most recent five rounds in which you were the active player. Your score is the number of rounds out of those five rounds in which you had chosen to help.

However, if there are fewer than five previous rounds in which you were an active player, your record will be filled in as if you chose to help in the missing rounds. So, your score in the first round of this part will be 5.

At the end of each round in which you are the active player, your score is updated based on your choice and the score of the non-active player. In rounds in which you are the non-active player, your score does not change.

For example, suppose in Round 11 you are the active player. The explanation of your score at the start of the round might be



Explanation of score

Your score is the number of times you chose to help in the rounds in your record.

Your record is currently:

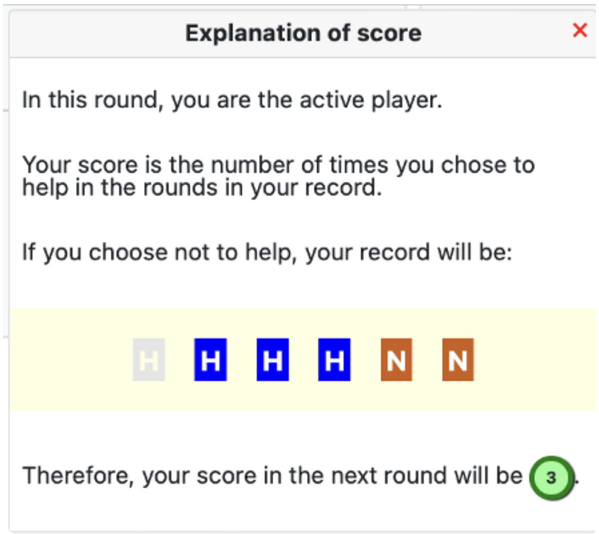
H H H H N

Therefore, your score now is 4.

This record shows that you chose not to help (N) in the most recent round in which you were the active player. You chose to help (H) in each of the four rounds previous to that. So, your score at the start of the round is 4.

When you make your choice for Round 11, your choice in the earliest round in your record will be removed, and your choice in Round 11 will be added.

For example, suppose you are the active player in Round 11, and you are considering choosing not to help. The explanation of how this would affect your score would be



Explanation of score ×

In this round, you are the active player.

Your score is the number of times you chose to help in the rounds in your record.

If you choose not to help, your record will be:

H H H H N N

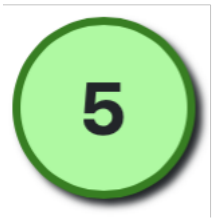
Therefore, your score in the next round will be 3.

There would now be only three rounds in your record in which you chose to help. So, your score at the start of Round 12 would be 3.

What you will see on the screen

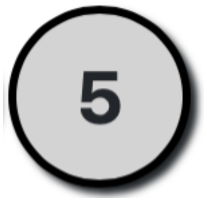
We will now show you what you will see on your screen during a round. Scores will be displayed on the screen using **badges**, which are coloured circles showing the score.

Your score will always be shown using a **green badge** like this



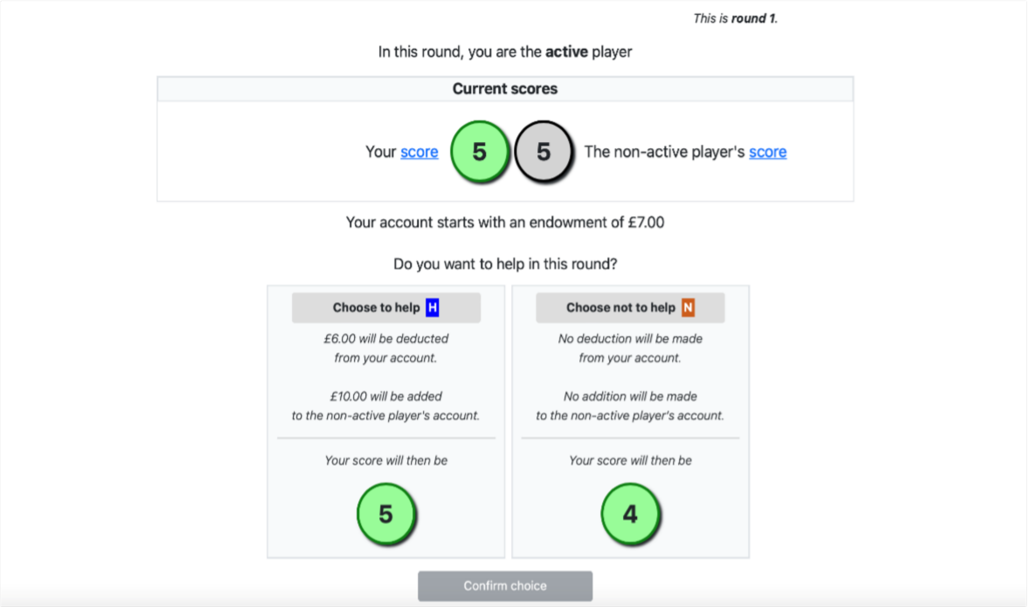
Any time you see a green badge representing your score, you can click on it to pop up a display which explains how your record was used to determine your score.

When you are the active player, you will see the score of the non-active player. The non-active player's score will be shown using a **grey badge** like this



You will only be able to see the current score of the non-active player, and not their record; therefore, clicking on a grey badge has no effect.

When you are the active player, you will see a screen like the one below



This is ***round 1.***

In this round, you are the **active** player

Current scores

Your score

The non-active player's score

Your account starts with an endowment of £7.00

Do you want to help in this round?

£6.00 will be deducted from your account.

£10.00 will be added to the non-active player's account.

Your score will then be

No deduction will be made from your account.

No addition will be made to the non-active player's account.

Your score will then be

The screen shows the round number, your score, the non-active player's score, and your endowment in this round.

The two boxes represent the two choices you can make: to help or not to help. Each box summarises the consequences for you and for the non-active player of the corresponding choice.

To choose to help, click on the button labelled **Choose to Help**. To choose not to help, click on the button labelled **Choose not to help**. When you click on one of those buttons, the background will change from grey to green. If you change your mind, you can change your choice simply by clicking on the button corresponding to the other choice. When you are satisfied with your choice, click the button labelled **Confirm choice**.

How Part 1 ends

Exactly how many rounds will be played is not fixed in advance. There will be at least 40 rounds. At the end of Round 40, the computer will simulate the roll of a six-sided die. If the roll is 1 or 2, then Part 1 will end; otherwise, Part 1 will continue to Round 41. This process will be repeated after each subsequent round, until the simulated die roll results in a 1 or 2. In other words, in each round starting from Round 40, there is a 1 in 3 chance that Part 1 will end after that round.

Part 2

In Part 2, you will be in the **same group** of **six** participants as in Part 1.

Part 2 has the same structure as Part 1. The only difference between Part 1 and Part 2 is that your cost of helping may not be the same as in Part 1. Likewise, for each of the other people in your group, the cost of helping may not be the same as in Part 1. However, as in Part 1, the cost of helping will be no more than £6.00 for anyone. You will be told your cost of helping for Part 2 before the first round begins. You will not find out the cost of helping for anyone else in your group.

As in Part 1, there will be a series of at least 40 **rounds**. At the beginning of each round, you will be matched at random with another person in your group. One of you will be randomly assigned the role of the **active player**. The other person will be the **non-active player**. The accounts of both players initially have an endowment of £7.00. The active player will have the opportunity to help the non-active player. If the active player chooses to help, the active player's cost of helping will be deducted from their account, and £10.00 will be added to the account of the non-active player. If the active player chooses not to help, no deduction or addition will be made to either player's account.

One of the rounds will be **for real**. The computer will select at random which round is for real, but you will not know which round that is until the end of the experiment. This round will be the same for all members of the group. For everyone in the group, their earnings from Part 2 will be equal to the final value of their account for the selected round. These will be added to your earnings from Part 1 to determine your earnings for the experiment as a whole.

Good Standing

As in Part 1, at the start of every round, you will have a **score** of 0 or 1 which summarises whether, in previous rounds, you chose to help or not to help. Scores will be computed and updated exactly as they were in Part 1.

Image Scoring

As in Part 1, at the start of every round, you will have a **score** of 0, 1, 2, 3, 4 or 5 which summarises whether, in previous rounds, you chose to help or not to help. Scores will be computed and updated exactly as they were in Part 1.

Exactly how many rounds will be played is not fixed in advance. There will be at least 40 rounds. At the end of Round 40, the computer will simulate the roll of a six-sided die. If the roll is 1 or 2, then Part 2 will end; otherwise, Part 2 will continue to Round 41. This process will be repeated after each subsequent round, until the simulated die roll results in a 1 or 2. In other words, in each round starting from Round 40, there is a 1 in 3 chance that Part 2 will end after that round. The number of rounds in Part 2 may not be the same as in Part 1 depending on the outcomes of the simulated die rolls.

Experimenter

[Read Instruction Part 1]
[After reading Instruction till the end of Part 1]

Please raise your hand if you have any questions.

Before starting to make choices, we ask you to answer some questions in the next several screens. The purpose of these questions is to check whether you have understood these instructions. Any mistake you may make will not affect your final monetary earnings.

Click on "Advance slowest players" button on monitor

[After everyone completed control questions]

Everyone has now completed the questions. We will now start Part 1.

On your screen you will see your cost for Part 1. Your cost will remain the same throughout this part.

Wait approximately 10 seconds, and then click "Advance slowest players" button on monitor.

[At the end of Part 1]

Part 1 has now ended. The instructions for Part 2 are now being circulated.

[Wait for all participants to have copy of Part 2 instructions]

[Read Part 2 instruction]

[After reading Part 2 instruction]

Please raise your hand if you have any questions.

We will now start Part 2.

On your screen you will see your cost for Part 2. Your cost will remain the same throughout this part.

Wait approximately 10 seconds, and then click "Advance slowest players" button on monitor.

[At the end of Part 2; wait for all participants to be on "SurveyWelcome" screen.]

Part 2 has now ended.

We now have a short survey which we invite you to complete.

Your answers to this survey are recorded anonymously, and will be used only in the analysis of the data from this experiment.

Click "advance slowest players" on monitor.

[Wait for all participants to reach Earnings page in final app]

(Read standard outro text)

This bring today's experiment to an end. All that remains is for us to take care of paying you your earnings from the session.

We will soon call you one by one to the payment station to receive your earnings from today's session. The payment station is at the back of the laboratory.

When you receive your payment, we have a receipt for you yo sign. On the receipt you should please fill in your name, your UEA IT login (three letters-two digits-two letters), and sign where indicated.

Recall that we promised that all payments will be done privately. You can help us to ensure everyone's privacy by waiting at your desk until you are called to go to the payment station. Please listen for your desk number to be called.

When you leave your desk, please make sure to take any personal belongings you have with you. You can leave any experiment materials at the desk; we will tidy these up.

Thank you again for participating in today's experiment, and we hope to see you again at a future session here at LEDR!

Comprehension Quiz[18]

1. In each round, there is an active and a non-active player.
   Select which of the following statements is true:
   The active player has a decision to make.
   The non-active player has a decision to make.
   Both players have a decision to make.

2. In each round, there is an active and a non-active player.
   Select which of the following statements is true:
   Whether you play as the active or non-active player will be determined randomly.
   You will definitely play as the active player.
   You will definitely play as the non-active player.

3. In each round, the active player can incur a cost to help the non-active player.
   Select which of the following statements is true:
   The non-active player′s benefit is always greater than the active player′s cost.
   The active player's cost is always greater than the non-active player's benefit.
   Sometimes the non-active player's benefit is greater than the active player's cost, other times the active player's cost is greater than the non-active player's benefit.

4. Recall that the endowment for all players is £7. Assume that the active player has a cost of £5, and the non-active player has a benefit of £10. If the active

[18] To ask before Part 1 starts. The numerated ones are for all. IS only for "Image Scoring" treatment. GS only for "Good Standing" treatment. The underlined answer is the correct one.

player **helped** in a particular round, what are the payoffs for this round?
Select which of the following statements is true:
The active player′s payoff is £2, and the non-active player′s payoff is £17.
The active player's payoff is £17, and the non-active player's payoff is £2.
Both players' payoff are £7.

5. Recall that the endowment for all players is £7. Assume that the active player has a cost of £5, and the non-active player has a benefit of £10. If the active player **did not help** in a particular round, what are the payoffs for this round?
   Select which of the following statements is true:
   Both players′ payoff are £7.
   The active player's payoff is £2, and the non-active player's payoff is £17.
   The active player's payoff is £17, and the non-active player's payoff is £2.

6. Your earnings from Part 1 depend on your payoffs for the rounds.
   Select which of the following statements is true:
   Your earnings from Part 1 are equal to your payoff from one randomly selected round.
   Your earnings from Part 1 are equal to the sum of your payoffs from all the rounds.
   Your earnings from Part 1 are equal to the sum of your payoffs from 5 randomly selected rounds.

7. Suppose that in round 16 you were the active player and chose to help, and that round 16 is the one selected to be for real for you.
   Select which of the following statements is true:
   Round 16 is definitely the round selected to be real for the non-active player you helped.
   Round 16 is definitely not the round selected to be real for the non-active player you helped.
   Round 16 may or may not be the round selected to be real for the non-active player you helped.

8. Recall that in each round, every player has a score which summarises his/her previous helping behaviour, and scores are updated at the end of the round.
   Select which of the following statements is true:
   The active player′s score may change, but the non-active player′s score will stay the same.
   The non-active player's score may change, but the active player's score will stay the same.

Both players' scores may change.
Neither player's score may change.

IS Recall that a player's score summarises his/her previous helping behaviour and that it takes values from 0 to 5. Suppose that the record of a player's decisions in the last 5 rounds in which he/she was the active player is: *Don't Help, Help, Don't Help, Help, Help.*
Select which of the following statements is true:
This player′s score is 3.
This player's score is 1.
This player's score is 4.
This player's score does not depend on his/her record.

IS Recall that a player's score summarises his/her previous helping behaviour and that it takes values from 0 to 5. Suppose that the active player has a record of *Help, Help, Don't Help, Help, Help*, and therefore has a score of **4**. If the active player **does not help** in this round, what will this player's score be at the end of the round?
Select which of the following statements is true:
The active player′s score will be 3.
The active player's score will be 4.
The active player's score will be 2.
The active player's score will depend on the score of the non-active player.

IS Recall that a player's score summarises his/her previous helping behaviour and that it takes values from 0 to 5. Consider a round in which both players have scores of **5**. If the active player **helps** in this round, what will this player's scores be at the end of the round?
Select which of the following statements is true:
The active player′s score will be 5.
The active player's score will be 6.
The active player's score will be 4.

IS Recall that a player's score summarises his/her previous helping behaviour and that it takes values from 0 to 5. Suppose that the active player has a score of **3**. Suppose that the record of the active player's decisions in the last 5 rounds in which where they were an active player is: *Don't Help, Help, Don't Help, Help, Help.* If the active player **helps** in this round, what will this player's score be at the end of the round?
Select which of the following statements is true:

The active player′s score will be 4.

The active player's score will be 3.

The active player's score will be 2.

GS Recall that a player's score summarises his/her previous helping behaviour and that it takes values 0 or 1. Suppose that the active player has a score of **1**. If the active player **helps** in this round, what will the active player' score be at the end of the round?

Select which of the following statements is true:

The active player's score will definitely be 1.

The active player's score will be 1 if the non-active player's score was 0, and 0 if the non-active player's score was 1.

The active player's score will definitely be 0.

The active player's score will be 0 if the non-active player's score was 0, and 1 if the non-active player's score was 1.

GS Recall that a player's score summarises his/her previous helping behaviour and that it takes values 0 or 1. Suppose that the active player has a score of **1**. If the active player **does not help** in this round, what will the active player' score be at the end of the round?

Select which of the following statements is true:

The active player′s score will be 1 if the non-active player′s score was 0, and 0 if the non-active player′s score was 1.

The active player's score will definitely be 0.

The active player's score will definitely be 1.

The active player's score will be 0 if the non-active player's score was 0, and 1 if the non-active player's score was 1.

GS Recall that a player's score summarises his/her previous helping behaviour and that it takes values 0 or 1. Suppose that the active player has a score of **0**. If the active player **helps** in this round, what will the active player' score be at the end of the round?

Select which of the following statements is true:

The active player′s score will definitely be 1.

The active player's score will be 1 if the non-active player's score was 0, and 0 if the non-active player's score was 1.

The active player's score will definitely be 0.

The active player's score will be 0 if the non-active player's score was 0, and 1 if the non-active player's score was 1.

Final Questionnaire[19]

Please, answer the following questions.
When you have finished, please remain seated and wait patiently until you are paid.

Gender → "What gender do you identify as?"
Age → "How old are you?"
Ethnicity → "Which country (or countries) were you a citizen of when you were born?"
Residence → "In which country is your current permanent residence? (If you are in the UK on a student visa, this is the country where "home' is.)"
School → "In which School are you currently enrolled/affiliated? (For example, BIO, ECO, EDU...)"
Degree → "What type of degree course are you currently enrolled on?"
Years → "How long have you been at UEA?"

From "**Global Evidence on Economic Preferences**", Falk et al. (2018):
Answer using the Likert Scale from 0 to 10, where 0 means you are "completely unwilling to do so" and a 10 means you are "very willing to do so".

1. *Patience* ⟶ How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future?
2. *Risk taking* ⟶ Please tell me, in general, how willing or unwilling are you to take risks?
3. *Altruism* ⟶ How willing are you to give to good causes without expecting anything in return?

Answer using the Likert Scale from 0 to 10, where 0 means "Does not describe me at all" and a 10 means "Describe me perfectly".

4. *Positive reciprocity* ⟶ When someone does me a favour, I am willing to return it.
5. *Negative reciprocity* ⟶ If I am treated very unjustly, I will take revenge at the first occasion, even if there is a cost to do so.
6. *Trust* ⟶ I assume that people have only the best intentions. How well do the following statement describe you as a person?

[19] To add at the end of Part 2.

Chapter 3

Binary Mechanisms in the Helping Game

This paper investigates the efficacy of different reputational mechanisms in promoting cooperative behaviour in helping games. Building on theoretical foundations, we focus on binary rating rules that encapsulate helpfulness, particularly Sugden's Good Standing (SGS) and Binary Image Scoring (BIS), alongside a baseline condition without reputational incentives. Through controlled laboratory experiments involving 216 participants under both homogeneous and heterogeneous cost conditions, we examine how these mechanisms influence individuals' propensity to help others. Our findings reveal that the SGS mechanism significantly outperforms BIS and the baseline in fostering cooperation. Participants under SGS consistently exhibit higher helping rates, with cooperation being more stable and persistent over time. The effectiveness of SGS is attributed to its structure, which considers both the helper's action and the recipient's standing, thereby encouraging cooperation within a network of trustworthy individuals.

3.1 Introduction

Cooperation is a foundational element of human society, underpinning interactions that range from everyday social exchanges to complex economic collaborations. Examples abound: neighbours sharing resources, colleagues assisting one another, and nations engaging in trade agreements that foster global prosperity. Despite its centrality, cooperation does not always emerge spontaneously, especially in contexts where individual incentives conflict with collective welfare. Economists are thus keen to understand the mechanisms that promote cooperative behaviour, particularly in settings where direct reciprocity is insufficient due to infrequent repeated interactions among the same individuals.

Indirect reciprocity has emerged as a key concept in explaining how cooperation can be sustained in large populations where direct reciprocity is limited. Reputation-based mechanisms play a crucial role in facilitating indirect reciprocity by allowing individuals to condition their cooperative behaviour on the observed actions of others. Two prominent reputation systems are Image Scoring (IS) (Nowak and Sigmund, 1998b) and Good Standing (GS) (Sugden, 1986; Leimar and Hammerstein, 2001). These mechanisms rely on publicly available information about past behaviours to inform future interactions, thereby fostering cooperation even among strangers.

In our previous work, we experimentally investigated the efficacy of the Image Scoring (IS) mechanism and Leimar and Hammerstein's (L&H) version of the Good Standing (GS) mechanism within a helping game — a setting where individuals decide whether to incur a cost to benefit another. We can divide all observations into two groups based on the median score: individuals with scores of 0, 1, 2, or 3 are classified as the lower-score group, and those with scores of 4 or 5 as the higher-score group. A two sample t-test between these two groups reveals a significant difference in the propensity to help between these two groups, favouring recipients with higher scores.

However, this difference is less pronounced than what we observe under the GS mechanism and less than what would be expect from theories of reciprocity. Under the IS mechanism, participants exhibits limited discrimination between recipients with different scores. While they are more likely to help recipients with a score of

4 compared to those with a score of 0, there are no significant difference in helping behaviour toward recipients with scores close to the median, such as 3 or 5. This suggests that participants reduce their help to those with very low scores but do not proportionally increase their cooperation toward those with slightly higher scores.

Consequently, the numerical scores in IS did not fully encourage participants to direct their cooperation according to the recipients' past cooperative behaviour. Both self-interested individuals and those with moral considerations — who tended to have disproportionately high scores — did not adjust their helping behaviour as strongly as predicted by reciprocity theories. Therefore, while IS facilitated some level of cooperation, it did not enhance reciprocity to the extent observed under the GS mechanism.

A potential explanation for this outcome lies in the asymmetry between the IS and GS scoring systems employed in our previous experiment. The IS mechanism utilised a six-level numerical score, whereas the GS mechanism was binary. This disparity in the granularity of reputational information may have influenced participants' behaviour, complicating direct comparisons between the two mechanisms. Furthermore, the six-level structure of Image Scoring, as implemented in Engelmann and Fischbacher (2008), introduces significant theoretical complexities. Their theoretical model necessitates specific assumptions regarding score updates, such as randomly selecting actions to discard rather than removing the oldest actions, to simplify the analysis. Additionally, the effectiveness of IS inherently relies on some degree of non-self-interest among participants. Such dependencies introduce additional layers of complexity, as individuals' willingness to cooperate may be influenced by intrinsic motivations beyond mere reputation management. Under IS, the broader range of scores might have prompted participants to aim for maintaining a particular score level, focusing more on their own reputation rather than reciprocating others' cooperation. In contrast, the binary nature of GS may have limited participants' ability to adjust their behaviour based on subtle differences in others' reputations.

To address this issue, the present study introduces a binary image scoring mechanism, thereby eliminating the asymmetry between the IS and GS mechanisms. By simplifying IS to a binary score — assigning a value of 1 if an individual helped in their last opportunity and 0 if they did not — we align the IS mechanism with the binary structure of Sugden GS. This adjustment allows for a clearer and more direct comparison of the two mechanisms' effectiveness in promoting cooperation and reciprocity. By examining both mechanisms under the same binary framework, we aim to determine whether the differences observed previously were attributable to the scoring asymmetry or to inherent differences in how the mechanisms influence cooperative behaviour.

Furthermore, we introduce Sugden version of the Good Standing (GS) mechanism, which aligns with L&H GS in the criteria for maintaining good standing when an individual currently has it but differs in the criteria for getting it. Specifically, under Sugden GS, an individual can get good standing only by helping someone who is already in good standing, whereas L&H GS allows individuals to get good standing by helping any other person, regardless of that person's standing. This distinction makes Sugden GS more stringent in how individuals are punished if they do not help cooperative individuals and how they can obtain good standing. Therefore, it heightens the incentive to help cooperative individuals — reinforcing reciprocal behaviour consistent with notions of fairness and deservingness — while simultaneously reducing the likelihood that individuals with a score of 0 will receive help. In contrast, L&H GS is more lenient, permitting individuals to gain good standing more easily — even after failing to help — by simply helping in any subsequent opportunity. This difference implies a trade-off between the two mechanisms: Sugden GS may foster stronger incentives for reciprocity and deter uncooperative behaviour more effectively, but it could also be less forgiving of occasional mistakes or unintended defections. By implementing Sugden GS, we aim to investigate whether this stricter mechanism can more effectively promote reciprocity and sustain cooperative behaviour compared to both the traditional IS and L&H GS.

To robustly evaluate the impact of these mechanisms, we also include a control condition without any reputational information. This allows us to establish baseline levels of cooperation and assess the extent to which reputation mechanisms influence behaviour. By comparing the reputational treatments with the control condition, we can isolate the effects of reputational cues on participants' willingness to help.

The inclusion of a control condition is particularly pertinent for examining the IS mechanism. In our previous study, we observed that participants often acted against their immediate self-interest by striving to maintain high scores, even when this resulted in monetary losses. This suggests that reputational concerns can motivate cooperative behaviour beyond straightforward payoff maximisation. By introducing a setting without reputational feedback, we aim to disentangle whether such non-self-interested behaviour is driven by the reputation mechanism itself or by intrinsic prosocial preferences. This approach also aligns with the methodology of Engelmann and Fischbacher (2009), who examined cooperation in the absence of reputational information. By incorporating a no-score condition, we can draw more precise comparisons and deepen our understanding of how reputational mechanisms influence cooperative dynamics.

An important aspect of our study is the elicitation of participants' beliefs regarding others' cooperative behaviour. In experimental economics, belief elicitation is a standard method for understanding how individuals form expectations about the actions of others, which in turn can significantly influences their own decision-making processes. By capturing these beliefs, researchers try to disentangle the extent to which observed behaviours are influenced by participants' preferences versus their expectations about others' behaviour. This is particularly pertinent in studies of cooperation and reciprocity, where an individual's willingness to help often depends not only on material payoffs but also on their perceptions of others' propensity to cooperate.

In our experiment, we implemented a new incentive-compatible belief elicitation procedure to measure participants' expectations about the frequency of helping among their peers. This design choice was motivated by the findings from our previous study, where we observed high overall levels of cooperation under the IS mechanism but limited reciprocal helping. While participants did show a significant difference in helping behaviour toward recipients with low scores compared to those with high scores, they did not proportionally adjust their cooperation across the full range of recipient scores. Participants tended to maintain high scores by helping others but did not significantly discriminate between recipients based on differences in scores. This behaviour could be interpreted in three ways.

First, it may reflect altruistic preferences. Participants might be helping others to maintain a high score, even at a personal cost, because they derive utility from being perceived as cooperative or because they value the act of helping itself. This aligns with models of prosocial behaviour such as that proposed by Levine (1998), where individuals are motivated to help those they consider altruistic. In this context, maintaining a high score becomes a means of signalling one's own altruism, and helping behaviour is less contingent on the recipient's past actions.

Second, the issue of false or incorrect beliefs warrants attention. It is essential to recognise that during the game, each player has access only to specific information: their own actions and scores when they are active players, and whether they were helped based on their score when they are non-active players. Therefore individuals might have incorrect beliefs regarding which is the rational score to maintain.

Third, the lack of reciprocity might result from misunderstandings or mistakes in interpreting the reputation mechanisms. Participants may not have fully grasped how their own scores or those of others were calculated and updated, leading them to make decisions that do not align with the intended strategic incentives of the mechanism. If participants misinterpret how their actions affect their scores, or how others' scores reflect past behaviour, they may fail to condition their helping on recipients' cooperativeness, thereby undermining reciprocity.

By eliciting participants' beliefs about others' cooperative behaviour, we aim to disentangle these possibilities. If participants accurately perceive high levels of cooperation among others and still choose to help indiscriminately, this would support the altruism interpretation. Conversely, if participants have incorrect beliefs about others' behaviour or misunderstand how the reputation mechanisms work, this could explain the lack of reciprocity observed due to mistakes or misperceptions.

Moreover, understanding participants' beliefs allows us to assess whether misunderstandings could be undermining the effectiveness of the reputation rules. If participants do not respond appropriately to reputational cues because of misperceptions, the mechanisms may fail to promote the desired cooperative and reciprocal behaviours. By rewarding accurate estimates in our belief elicitation task, we encouraged participants to provide thoughtful and sincere beliefs about the cooperative behaviour of others. This approach enables us to analyse the relationship between participants' beliefs and their own helping behaviour, providing insights into the cognitive processes underpinning cooperation.

For example, if participants who believe that others are generally cooperative are more likely to help themselves, this suggests that positive expectations foster prosocial behaviour. Conversely, if there is a disconnect between beliefs and actions, this may indicate that other factors, such as misunderstandings of the mechanisms or intrinsic preferences, are at play. By incorporating belief elicitation into our study, we aim to provide a more comprehensive understanding of the factors that promote or hinder cooperation in economic interactions.

This not only enriches the analysis of our experimental results but also has important implications for the design of institutions and mechanisms intended to foster cooperation in real-world settings. Understanding participants' beliefs helps ensure that reputation rules are not only theoretically sound but also practically effective, as they align with how individuals perceive and interpret social information. It also allows us to refine the mechanisms to address potential misunderstandings, thereby enhancing their capacity to sustain cooperative and reciprocal behaviours.

Finally, we contribute to the theoretical literature by developing an axiomatic framework that examines binary reputation mechanisms in helping games. This framework provides a formal foundation for identifying which mechanisms are both feasible and effective in promoting cooperation within the helping game. By introducing specific axioms that any viable binary mechanism should satisfy, we narrow down the set of plausible mechanisms to those most likely to sustain cooperative behaviour.

Our work builds upon the study of Camera and Gioffré (2022) who addressed the existence of cooperative equilibria supporting efficient allocations in indefinitely repeated helping games under private monitoring. While they provided general proofs and characterised efficient equilibria, there remained an open question regarding which specific reputation mechanisms could facilitate cooperation to emerge. By systematically analysing the properties of binary reputation mechanisms through our axiomatic approach, we reduce the number of candidate mechanisms to those that are theoretically sound and practically implementable. This reduction is important because it focuses both theoretical and experimental efforts on mechanisms with the greatest potential to promote cooperation. By eliminating mechanisms that do not satisfy essential criteria, we offer clearer guidance for designing reputation rules in economic environments where helping games are applicable.

Ohtsuki and Iwasa (2006) conducted a comprehensive analysis of reputation dynamics in the context of indirect reciprocity. Using simulations, they identify the “leading eight” social norms capable of maintaining high levels of cooperation. These norms share common characteristics: (i) cooperation is sustained within a population adhering to the norm; (ii) defectors are immediately recognised and labelled as “bad”; (iii) individuals with a “bad” reputation are denied cooperation, and those who refuse to help them can be regarded as “good”; and (iv) accidental defections due to errors can be rectified, allowing individuals to regain a “good” reputation through appropriate actions. Their evolutionary approach find these eight norms effectively promote cooperation, offering clear and intuitive reasons based on the dynamics of reputation.

Building on their evolutionary insights, our axiomatic approach formalises these common aspects into precise criteria for reputation mechanisms. While both Ohtsuki (2004) and Ohtsuki and Iwasa (2006) focused on evolutionary stability and reputation dynamics, we extend their findings by translating these principles into an axiomatic framework that identifies mechanisms which are both theoretically robust and practically implementable. Our axioms encapsulate the essential features of the leading eight norms, replicating their key elements within a mathematical structure. This allows us to systematically analyse and reduce the set of candidate reputation mechanisms to those that align with these foundational principles.

Furthermore, our proposed axioms are closely linked to the ongoing work by Fischbacher et al. (2024), who explore the social norms of peer punishment. They identify four simple rules organizing punishment behaviour in the lab: (i) do not punish cooperators; (ii) defectors should not punish; (iii) punish those who violate (i) or (ii); and (iv) punishing defectors is generally considered a right of cooperators, though some view it as a duty. These rules resonate with our axiomatic approach, as they outline fundamental norms that govern cooperation and punishment. By aligning our axioms with these experimentally observed social codes, we strengthen the relevance and applicability of our theoretical framework.

By integrating these evolutionary and experimental perspectives into our axiomatic approach, we not only refine the set of mechanisms to be considered but also enhance our understanding of how specific reputation rules can effectively foster cooperation. Our mathematical analysis, based on prior findings, contributes to a more comprehensive theoretical foundation, guiding both theoretical exploration and practical implementation of reputation mechanisms in economic environments.

This chapter aims to provide both theoretical and experimental answers on how binary reputation mechanisms shape cooperation and reciprocity.

From a theoretical standpoint, we seek to determine whether there exist binary rating rules beyond those currently established in the literature that effectively encapsulate helpfulness. Additionally, we identify the underlying principles that define such rating rules. This exploration attempts to create a theoretical framework surrounding binary reputation systems and understand the conditions under which such systems can sustain cooperative behaviour.

On the experimental front, our investigation is designed to address several questions. Primarily, within the scope of this chapter alone, we examine whether binary image scoring leads to higher levels of cooperation and reciprocity compared to a baseline condition without a reputation mechanism. Furthermore, we assess how Sugden Good Standing (SGS) mechanism compares with binary image scoring in promoting cooperation and sustaining reciprocal relationships. These questions are symmetric to the ones presented in Chapter 2, providing a comprehensive empirical evaluation of the tested mechanisms. Another critical aspect of our experiment involves understanding how participants' beliefs about reputation scores influence their cooperative behaviour, particularly in scenarios where misunderstandings of the scoring mechanisms may occur.

In addition to these experimental questions, this chapter also serves as a follow-up to the experiment analysed in Chapter 2 by enabling further comparative analyses. Specifically, we investigate how cooperation rates and reciprocity patterns differ between L&H GS and SGS mechanisms. Moreover, we evaluate the differences in cooperative behaviour between the control condition (without any reputation mechanism) and the reputational treatments, which include IS, binary image scoring, and GS mechanisms. Lastly, we compare binary image scoring with non-binary image scoring mechanism to discern their respective impacts on promoting cooperation and reciprocity. These additional comparisons aim to elucidate the effects of different reputation systems, thereby providing deeper insights into their efficacy in fostering and sustaining cooperative and reciprocal behaviour.

Our findings reveal that the SGS mechanism significantly outperforms BIS and the baseline in fostering cooperation. Participants under SGS consistently exhibit higher helping rates, with cooperation being more stable and persistent over time.

The effectiveness of SGS is attributed to its structure, which considers both the helper's action and both players' standing, thereby encouraging cooperation within a network of trustworthy individuals.

The results suggest that reputational systems incorporating social context and previous interactions, like SGS, are more successful in sustaining cooperative behaviour compared to mechanisms focusing solely on the most recent action. This has important implications for the design of institutions and platforms aiming to enhance cooperative interactions. By recognising the interplay between individual actions and the reputations of others, policymakers and designers can better foster environments where cooperation thrives.

The remainder of the chapter is organised as follows. Section 3.2 presents our axiomatic approach of binary reputation mechanisms in helping games. Section 3.3 outlines the experimental design and procedures, including the treatments and belief elicitation methods. Section 3.4 reports the experimental results, analysing cooperation rates, reciprocity patterns, and the impact of participants' beliefs. Section 3.5 concludes.

3.2 Theory

3.2.1 Definition

There is a finite set of players $N = \{1, ..., n\}$, where n is an even number. The game unfolds over a sequence of discrete time periods T.

There is a finite set of ratings R. At time $t \in T$, each player i has a rating $r_i^t \in R$. For the purpose of this analysis, we focus on binary ratings, so $R = \{0, 1\}$.

In every period t, players are randomly matched. Within each match, one player is randomly assigned the role of the active player, and the other the role of the non-active player. The assignment is independent and random for each match in every period. The matching is anonymous: the active player knows only their rating and the non-active player's rating. After seeing the ratings, the active player makes a decision: to help or not help the non-active player. Let $A = \{0, 1\}$ denote the set of such decisions, where 1 represents the action "help" and 0 represents "not help". Non-active players see only their own rating and do not make any decisions.

Consider a given match at period t, let i denote the active player, and $j \neq i$ denote the non-active player. Let $a_i^t \in A$ denote the action that active player makes in period t. The evolution of the active player's rating is governed by a *rating rule* $\gamma\colon R \times R \times A \to R$. This function is common knowledge among all players and determines an active player's rating in the next period according to

$$r_i^{t+1} = \gamma(r_i^t, r_j^t, a_i^t)$$

Table 3.1 presents a compact representation of the rating rule as a 2×2 matrix, where the rows correspond to the active player's current rating r_i^t, and the columns correspond to the non-active player's current rating r_j^t. Each cell contains a pair of values representing the active player's next-period rating when $a = 1$ (help), and when $a = 0$ (not help), respectively.

Table 3.1: Representation of the rating rule

		Non-active player rating	
		$r_j^t = 1$	$r_j^t = 0$
Active player rating	$r_i^t = 1$	$\gamma(1,1,1)$, $\gamma(1,1,0)$	$\gamma(1,0,1)$, $\gamma(1,0,0)$
	$r_i^t = 0$	$\gamma(0,1,1)$, $\gamma(0,1,0)$	$\gamma(0,0,1)$, $\gamma(0,0,0)$

There are $2 \times 2 \times 2 = 8$ possible combinations of (r_i^t, r_j^t, a). Each combination can be mapped to either 0 or 1 in R, leading to $2^8 = 256$ possible rating rules.

Let Γ denote the set of all these possible rating rules.

3.2.2 Examples

In situations where individuals repeatedly interact, the history of past actions can become remarkably complex. When faced with the decision of whether to help another, one must consider this potentially intricate history to inform one's choice. Rating rules serve as a means to distil the most essential information from past interactions, thereby simplifying the decision-making process.

In Figure 3.1 we present three such rating rules that have been studied in the literature.

(a) BIS

	1	0
1	$1, 0$	$1, 0$
0	$1, 0$	$1, 0$

(b) SGS

	1	0
1	$1, 0$	$1, 1$
0	$1, 0$	$0, 0$

(c) L&H GS

	1	0
1	$1, 0$	$1, 1$
0	$1, 0$	$1, 0$

Figure 3.1: Examples of rating rules

BIS: Binary Image Scoring; L&H GS: Leimar and Hammerstein Good Standing; SGS: Sugden Good Standing. Rows are active player's ratings, columns are non-active player's ratings. Each entry summarises active player's score in $t+1$ for both actions: if they help, and if they don't help at t.

One approach is to focus solely on the history of an individual's actions. This is exemplified in the concept of *Image Scoring* (IS) introduced by Nowak and Sigmund (1998a). We propose a simplified approach, *Binary Image Scoring* (BIS), where only the most recent decision matters, reducing the rating set to two levels: $R = \{0, 1\}$. Individuals receive a rating of 1 if they helped in their last interaction, and 0 otherwise (see Figure 3.1a). This method emphasises the immediacy of recent behaviour, capturing the extreme limit of recency by prioritising the latest action as the most informative indicator of future cooperation, disregarding other decisions.

While IS and BIS focus on the history of an individual's own actions, they may overlook significant aspects of the social context. In his seminal work, Sugden (1986) introduces the concept of *Good Standing* (GS), which adds an element of reputation to the mechanism. In Sugden's model, an individual's rating depends not only on their own actions but also on both their rating and the rating of the recipient. Specifically, an individual gains a good rating only if they help someone who has also a good rating (refer to Figure 3.1b). This introduces a layer of conditionality based on the recipient's reputation, thereby fostering cooperation within a circle of individuals who have themselves demonstrated cooperative behaviour. The mechanism is recurrent in nature; the long history of interactions is encapsulated within the ongoing reputational status of individuals.

Building upon this concept, Leimar and Hammerstein (2001) offer a modification to the conditions under which an individual's rating changes. In their version of GS, an individual can gain or maintain a good rating by helping, regardless of the recipient's rating (refer to Figure 3.1c). The key difference lies in the conditions for maintaining or improving one's own rating: their model is less stringent than

Sugden one. In fact, it does not require the recipient to have a particular rating for the helper to gain good standing. This adjustment broadens the scope of potential cooperation, as it lowers the barriers to improving one's reputation within the group. At the same time, it diminishes the strategic urgency to maintain a good rating, since players can restore their rating in the next period as active player and may still receive help even with a score of bad rating.

These varying approaches to rating rules highlight the balance between simplicity and the richness of information considered. While BIS offers a minimalist summary of an individual's recent behaviour, Sugden and Leimar and Hammerstein GS models incorporate additional layers of social information, potentially leading to different dynamics of cooperation within the population.

The examples in Figure 3.1 illustrate a subset of all the possible binary rating rules (Γ). Looking at these examples, we can see how the specific rating rules update the ratings differently. However, a common feature among these examples is the ordering of ratings, where 1 indicates "helpful" or "worthy" individuals, and 0 indicates the opposite. This ordering implies that a rating of 1 is considered "good" and a rating of 0 is considered "bad".

Examining these examples raises important questions: are there other binary rating rules that effectively encapsulate helpfulness? What principles define such rating rules? To address these questions, we introduce a set of axioms that encapsulates desirable properties for rating rules.

3.2.3 Axiomatic Approach

The axiomatic approach provides a foundation for designing rating rules that encourage cooperative behaviour in the helping game. Below, we outline four key axioms that any effective rating rule should satisfy, along with the intuition and principles motivating them.

Building upon the insights from evolutionary and experimental studies — such as those by Ohtsuki (2004) and Fischbacher et al. (2024) — we formalise the principles underlying effective reputation mechanisms through an axiomatic framework. This approach allows us to rigorously identify the essential properties that any desirable rating rule should possess in the context of the helping game.

The first axiom ensures that helping behaviour is not penalised. If a player's rating of 1 reflects their helpfulness, then helping others should not diminish it. This principle captures the idea that the rating system should align with their observable helpfulness of a player.

Axiom 3.1. *"Helping Monotonicity"*
A rating rule γ satisfies "Helping Monotonicity" if and only if

$$\forall (r_i, r_j) \in R \times R, \gamma(r_i, r_j, 0) = 1 \quad \Rightarrow \quad \gamma(r_i, r_j, 1) = 1 \tag{3.1}$$

The second axiom provides consistency in how the active player's decision not to help is evaluated, irrespective of the non-active player's rating.

Axiom 3.2. *"Worthiness"*
A rating rule γ satisfies "Worthiness" if and only if

$$\forall r_i \in R, \gamma(r_i, 1, 0) = 1 \quad \Rightarrow \quad \gamma(r_i, 0, 0) = 1 \tag{3.2}$$

For example, a social norms that permit not helping in some situations without unfairly penalising the active player is consistent with this axiom.
The third axiom ensures that a rating rule must meaningfully reflect player behaviour. This axiom rules out trivial rating mechanisms that fail to distinguish between helpful and unhelpful actions.

Axiom 3.3. *"Reputation Responsiveness"*
A rating rule γ satisfies "Reputation Responsiveness" if and only if

$$\forall r_i \in R, \exists r_j \in R, a_1, a_2 \in A, a_1 \neq a_2 : \gamma(r_i, r_j, a_1) = 0 \quad \wedge \quad \gamma(r_i, r_j, a_2) = 1. \tag{3.3}$$

The forth axiom provides consistency in how the active player's decision of helping is evaluated, irrespective of the non-active player's rating. This is the "mirror" version of "Worthiness".

Axiom 3.4. *"Worthiness 2"*

A rating rule γ satisfies "Worthiness 2" if and only if

$$\forall r_i \in R, \gamma(r_i, 0, 1) = 1 \quad \Rightarrow \quad \gamma(r_i, 1, 1) = 1 \tag{3.4}$$

Having established the axioms that desirable rating rules should satisfy, we now characterise the set of rating rules that meet these criteria. Our goal is to identify all rating rules $\gamma \in \Gamma$ that satisfy Axioms 3.1 (Helping Monotonicity), 3.2 (Worthiness), 3.3 (Reputation Responsiveness), and 3.4 (Worthiness 2).

Theorem 3.1. *Let $\Gamma^\star \subseteq \Gamma$ denote the set of rating rules that satisfy all four axioms. $\Gamma^\star$ contains exactly six rating rules represented in Figure 3.2.*

(a) Leimar and Hammerstein (L&H GS)

	1	0
1	$1, 0$	$1, 1$
0	$1, 0$	$1, 0$

(b) Sugden (SGS)

	1	0
1	$1, 0$	$1, 1$
0	$1, 0$	$0, 0$

(c) Binary Image Scoring (BIS)

	1	0
1	$1, 0$	$1, 0$
0	$1, 0$	$1, 0$

(d) "Modified" Binary Image Scoring

	1	0
1	$1, 0$	$1, 0$
0	$1, 0$	$0, 0$

(e) Group Lenient

	1	0
1	$1, 0$	$0, 0$
0	$1, 0$	$1, 0$

(f) Group Harsh

	1	0
1	$1, 0$	$0, 0$
0	$1, 0$	$0, 0$

Figure 3.2: $\Gamma^\star$: set of ratings that satisfy axioms A1, A2, A3, A4

To prove Theorem 3.1, we first establish two lemmas that will help in the characterisation of $\Gamma^\star$.

Lemma 3.1. *For each active player's current rating, there is at least one scenario where choosing not to help ensures the active player's rating is (or becomes) 0. Formally,*

$$\forall r_i \in R, \gamma(r_i, 1, 0) = 0 \quad \vee \quad \gamma(r_i, 0, 0) = 0 \tag{3.5}$$

Proof. This proof is by contradiction. Suppose that for some $r_i \in R$, both $\gamma(r_i, 1, 0) = 1$ and $\gamma(r_i, 0, 0) = 1$.

Step 1. Since $\gamma(r_i, 1, 0) = 1$, Axiom 2 implies $\gamma(r_i, 0, 0) = 1$, which is consistent with our assumption.

Step 2. For both $r_j = 0$ and $r_j = 1$, since $\gamma(r_i, r_j, 0) = 1$, Axiom 1 requires $\gamma(r_i, r_j, 1) = 1$. Therefore, for all $r_j \in R$: $\gamma(r_i, r_j, 0) = 1$ and $\gamma(r_i, r_j, 1) = 1$.

Step 3. Since $\gamma(r_i, r_j, 1) = 1$ for all $a \in 0, 1$ and $r_j \in R$, there is no such r_j and $a_1 \neq a_2$ where $\gamma(r_i, r_j, a_1) = 0$. This contradicts Axiom 3. □

Lemma 3.2. *For each active player's current rating, there is at least one scenario where choosing to help ensures the active player's rating is (or becomes) 1.*

$$\forall r_i \in R, \gamma(r_i, 1, 1) = 1 \quad \vee \quad \gamma(r_i, 0, 1) = 1. \tag{3.6}$$

Proof. This proof is by contradiction. Suppose that for some $r_i \in R$, both $\gamma(r_i, 1, 1) = 0$ and $\gamma(r_i, 0, 1) = 0$.

Step 1. If $\gamma(r_i, r_j, 1) = 0$, then there is no requirement on $\gamma(r_i, r_j, 0)$. However, since $\gamma(r_i, r_j, 1) = 0$ for both $r_j = 0$ and $r_j = 1$, the active player's rating remains 0 regardless of whom they help.

Step 2. Axiom 3 requires that there exists some $r_j \in R$ and actions $a_1 \neq a_2$ such that $\gamma(r_i, r_j, a_1) = 0$ and $\gamma(r_i, r_j, a_2) = 1$. Since $\gamma(r_i, r_j, 1) = 0$ for all r_j, the only way ro satisfy axiom 3 is if $\gamma(r_i, r_j, 0) = 1$ for some r_j.

Step 3. If $\gamma(r_i, r_j, 0) = 1$ and $\gamma(r_i, r_j, 1) = 0$, this directly violates axiom 1 because you cannot reward "not helping" with a 1 but then punish "helping" with a 0 in the same (r_i, r_j). □

Having established the two lemmas, we now prove Theorem 3.1.

Proof of Theorem 3.1. Necessary condition

$\Gamma^\star$ entails rating rules that that satisfy the following conditions:
For all $r_i \in R$:

1. $\gamma(r_i, 1, 1) = 1$ and $\gamma(r_i, 1, 0) = 0$.

2. If $\gamma(r_i, 0, 0) = 0$, then $\gamma(r_i, 0, a) = \{0, 1\}$, for all a.

3. If $\gamma(r_i, 0, 0) = 1$, then $\gamma(r_i, 0, 1) = 1$.

To characterize $\Gamma^\star$, we systematically eliminate rating rules that violate the axioms. The proof proceeds in five steps.

Step 1: Show that $\gamma(r_i, 1, 0) = 0$ for all r_i.

Suppose by contradiction that $\gamma(r_i, 1, 0) = 1$ for some r_i. Then, by Axiom 2, it must be that $\gamma(r_i, 0, 0) = 1$. By Axiom 1, since $\gamma(r_i, 1, 0) = 1$, it follows that $\gamma(r_i, 1, 1) = 1$. Similarly, since $\gamma(r_i, 0, 0) = 1$, Axiom 1 implies $\gamma(r_i, 0, 1) = 1$. Therefore, under this supposition, $\gamma(r_i, r_j, a) = 1$ for all $r_j \in R$ and $a \in A$. This would mean that the active player's rating is always 1, regardless of their action, violating Axiom 3, which requires the action to affect the rating in some scenario. Therefore, our supposition must be false, and thus:

$$\gamma(r_i, 1, 0) = 0 \text{ for all } r_i \in R \tag{3.7}$$

Note: Axiom 4 excludes as well the possibility $\gamma(1, 0, 1) = 1$.

Step 2: Determine possible values for $\gamma(r_i, 1, 1) = 1$ for all r_i.

Given that $\gamma(r_i, 1, 0) = 0$ from Step 1, Lemma 1 requires that either $\gamma(r_i, 0, 0) = 0$ or $\gamma(r_i, 0, 0) = 1$. If $\gamma(r_i, 0, 0) = 1$, then by Axiom 1, $\gamma(r_i, 0, 1) = 1$. However, this does not violate any axioms, so $\gamma(r_i, 0, 0)$ can be either 0 or 1, provided that if $\gamma(r_i, 0, 0) = 1$, then $\gamma(r_i, 0, 1) = 1$.

From Lemma 2, since $\gamma(r_i, 1, 1) = 1$ or $\gamma(r_i, 0, 1) = 1$ must hold, and given that $\gamma(r_i, 1, 1) = 1$ provides a clear compliance with Axiom 3, we have:

$$\gamma(r_i, 1, 1) = 1 \text{ for all } r_i \in R. \tag{3.8}$$

Note: Also Axiom 4 forbids the scenario where $\gamma(r_i, 0, 1) = 1$ but $\gamma(r_i, 1, 1) = 0$.

Step 3: Determine possible values for $\gamma(r_i, 0, a)$ for all $r_i \in R$ and $a \in A$

We analyse the pairs $(\gamma(r_i, 0, 1), \gamma(r_i, 0, 0))$ for all $r_i \in R$:

$$\big(\gamma(r_i, 0, 1),\, \gamma(r_i, 0, 0)\big) \quad \text{for } r_i \in \{0, 1\}.$$

In principle, there are four ways to assign the pair $\big[\gamma(r_i, 0, 1),\, \gamma(r_i, 0, 0)\big]$:

$$(1, 1), \quad (1, 0), \quad (0, 0), \quad (0, 1).$$

However, by "Helping Monotonicity" (Axiom 3.1), the combination $(0, 1)$ is impossible. Indeed, having $\gamma(r_i, 0, 1) = 0$ but $\gamma(r_i, 0, 0) = 1$ would reward not helping a "bad" opponent while punishing helping the same opponent. Hence the only possible pairs are:

- $\big(\gamma(r_i, 0, 1), \gamma(r_i, 0, 0)\big) = (1, 1)$.
- $\big(\gamma(r_i, 0, 1), \gamma(r_i, 0, 0)\big) = (1, 0)$.
- $\big(\gamma(r_i, 0, 1), \gamma(r_i, 0, 0)\big) = (0, 0)$.

Step 4: Eliminate impossible pairs due to Axiom 3.1.

From the discussion above, the only forbidden outcome pair is

$$\big(\gamma(r_i, r_j, 1),\ \gamma(r_i, r_j, 0)\big) = (0, 1), \quad \forall (r_i, r_j) \tag{3.9}$$

Indeed, "Helping Monotonicity" (Axiom 3.1) states that if $\gamma(r_i, r_j, 0) = 1$, then $\gamma(r_i, r_j, 1) = 1$. Thus $(0, 1)$ would violate Axiom 3.1. This rules out any rule that punishes helping while rewarding not helping in the exact same (r_i, r_j) circumstance.

Step 5: Enumerate possible rating rules consistent with the axioms

Using the constraints established in Steps 1–4, we enumerate all valid combinations:

- The left column (for $r_j = 1$) is fixed as $(1, 0)$ for both $r_i = 1$ and $r_i = 0$: this takes us down from 256 possibilities to 16.
- The right column (for $r_j = 0$) can take one of three pairs: $(1, 1)$, $(1, 0)$, or $(0, 0)$: this takes us down from 16 possibilities to 9.
- Eliminating combinations that fail Axiom 3.3, we are left with 6 valid rating rules, as shown in Figure 3.2.

Therefore, the necessary conditions are established.

Sufficient condition

The sufficiency of these conditions is straightforward by inspection. □

While all six rating rules satisfy Axioms 3.1—3.4, not all are equally effective in promoting cooperation. Specifically, the "Group Lenient" (3.2e) and "Group Harsh" (3.2f) rules may undermine the effectiveness of the reputation system by allowing situations where helping leads to a decrease in the helper's rating. This contradicts the principle that cooperative behaviour should not be penalised.

We think that an additional axiom would be helpful in strengthening the notion that the rating system should promote and reward cooperative behaviour.

Axiom 3.5. *"No penalty for helping"*
A rating rule γ satisfies "No penalty for helping" if and only if

$$\forall r_i \in R, \forall r_j \in R, \gamma(r_i, r_j, 1) \geq r_i \tag{3.10}$$

Axiom 3.5 can be interpreted as helping another player should not decrease the active player's rating.

Theorem 3.2. *Let $\Gamma^{\star\star} \subseteq \Gamma$ denote the set of rating rules that satisfy all five axioms. $\Gamma^{\star\star}$ contains exactly four rating rules. These are the top four (3.2a, 3.2b, 3.2c, 3.2d) represented in Figure 3.2.*

This refined set of rating rules $\Gamma^{\star\star}$ captures the essential features of the "leading eight" norms identified by Ohtsuki and Iwasa (2006), which emphasise that cooperation should be sustained within the population, defectors are recognised and penalised, and cooperators maintain favourable reputations. Our axiomatic approach formalises these principles, providing a rigorous mathematical foundation for designing and analysing reputation systems.

Moreover, our framework resonates with the findings of Fischbacher et al. (2024), who identify fundamental norms governing acceptable punishment in experimental settings. By aligning our axioms with these experimentally observed social codes, we enhance the relevance and applicability of our theoretical model.

In conclusion, our axiomatic approach has identified four binary rating rules that both theoretically and practically encapsulate helpfulness. Between Chapter 2 and 3 we test three of these rules in lab experiments thinking that they effectively capture the essence of helpfulness without introducing any undesirable properties.

3.2.4 Analysis of BIS Equilibrium

In the preceding analysis of this section, we developed a theoretical framework centred on reputation mechanisms and we introduced the Binary Image Scoring (BIS) mechanism, which simplifies reputations to a binary scale based solely on an individual's most recent action. While this represents a significant shift from the standing strategies examined in Chapter 1, we still have to analyse whether equilibria can emerge within this simplified framework and how these equilibria might compare to those previously identified.

To maintain consistency and facilitate comparison, we adopt the same analytical framework used in Chapter 1. We examine the incentives of self-interested players and derive equilibrium conditions within this modified game structure, focusing on whether helping behaviour can be sustained under BIS.

The helping game retains its structure as a repeated asymmetric interaction, where players alternate roles in random matches. In each period, one player (the active player) decides whether to incur a cost c to provide a benefit b to the other player (the non-active player), under the condition $b > c > 0$. Players are drawn from a finite set N of agents, and matches occur in discrete time $t = 1, 2, \dots, \infty$, continuing with probability δ in each period. Matches are random, and here we assume everyone is matched ($\omega = 1$). The random pairing ensures that every player has a probability of $\frac{1}{2}$ of being chosen as a helper in each round. Importantly, we assume perfect recall of previous interactions, allowing strategies to depend on the history of play.

In Chapter 1, we demonstrated that a good standing equilibrium can exist under certain conditions in the helping game. Players sustain helping behaviour by adhering to the Sugden good standing strategy, which rewards helpers with a good reputation and penalises those who fail to help. The conditions for equilibrium were derived following the following steps.

Consider the case in which all other players follow the good standing strategy. Let U_G denote the expected payoff from adhering to the good standing strategy indefinitely, given that one is the active player in the current period. While the game continues, the series of expected payoffs is:

$$-c, \quad \frac{b-c}{2}, \quad \frac{b-c}{2}, \quad \ldots$$

The probability that the game will end in any period reduces the expected value of future payoffs. Thus, the total expected payoff from the good standing strategy is:

$$U_G = -c + \frac{(b-c)}{2} \cdot \frac{\delta}{1-\delta}. \tag{3.11}$$

Now consider the alternative strategy of following myopic self-interest (the "bad standing" strategy), which we label U_B. Under this strategy, the player never incurs the cost of helping and therefore receives an expected payoff of $U_B = 0$.

The good standing strategy is sustainable as an equilibrium if $U_G > U_B$, which simplifies to:

$$-c + \frac{(b-c)}{2} \cdot \frac{\delta}{1-\delta} > 0. \tag{3.12}$$

Rearranging, we find:

$$\delta > \frac{2c}{b+c}. \tag{3.13}$$

Thus, there is a good standing equilibrium with 100% helping if and only if $\delta > \frac{2c}{b+c}$ (same as formula 1.10 in Chapter 1). It is worth noting that this is not the only equilibrium: there also exists an equilibrium in which no player ever helps.

This method of comparing "good standing forever" with "bad standing forever" is legitimate because the model assumes stationarity. If the good standing strategy is better than the bad standing strategy in the current period, it will remain the better choice whenever there is an opportunity to help. Consequently, players have no incentive to deviate from the equilibrium path.

Returning now to the Binary Image Scoring mechanism, we apply the same reasoning. Under BIS, the scoring rule is that if a player helped in the last round in which they were the active player, they have a score of 1; if they did not help, they have a score of 0. In any round t, what it is rational for a self-interested active player to do is independent of the non-active player's score – because the active player's score at the end of round t depends only on whether he helped or not (this is true of the Engelmann and Fischbacher (2008)'s IS game too[1]).

However, there may be an equilibrium in which each active player is indifferent between helping and not helping but is more likely to help a non-active player if their score is 1 rather than 0 (this approach is similar to the Engelmann and Fischbacher's game).

Consider the strategy: if the non-active player has a score of 0, help with probability p; if the non-active player has a score of 1, help with probability q, with $0 \leq p < q \leq 1$.

Suppose that your co-players always follow this strategy.

Let U_G be the expected payoff from following this strategy forever, given that you are the active player in the current period.

As long as the game continues, this will give you the series of expected payoffs: $-c, \frac{qb-c}{2}, \frac{qb-c}{2}, \ldots$. Because of the probability that the game will end,

$$U_G = -c + \left(\frac{qb-c}{2}\right) \cdot \frac{\delta}{1-\delta} \tag{3.14}$$

Let U_B be the expected payoff from following myopic self-interest (the B for bad standing strategy) forever, given that you are the active player in the current period.

As long as the game continues, this will give you the series of expected payoffs: $0, \frac{pb}{2}, \frac{pb}{2}, \ldots$. Because of the probability that the game will end,

$$U_B = \left(\frac{pb}{2}\right) \cdot \frac{\delta}{1-\delta} \tag{3.15}$$

So, $U_G > U_B$ if:

$$-c + \left(\frac{qb-c}{2}\right) \cdot \frac{\delta}{1-\delta} > \left(\frac{pb}{2}\right) \cdot \frac{\delta}{1-\delta} \tag{3.16}$$

$$\delta > \frac{2c}{c+b(q-p)} \tag{3.17}$$

What are the implications this inequality (3.17)? There can not be a self-interest equilibrium with $U_G > U_B$, in which helping is uniquely optimal, because the best reply to such a strategy is not to help at all (since one would receive help regardless

[1] See Appendix 1 of their Working Paper, page 32.

of one's own actions). Conversely, there can be a self-interest equilibrium with $U_G < U_B$, but that would be one in which no one helps.

Hence, the only type of self-interest equilibrium in which there is helping is one with $U_G = U_B$. Thus, for any given values of c, b and δ, $(q - p)$ must be such that:

$$\delta = \frac{2c}{c + b(q - p)} \tag{3.18}$$

$$q - p = \frac{c(2 - \delta)}{\delta b} \tag{3.19}$$

This result indicates that the probability difference $(q - p)$ is determined by the parameters of the game (c, b) and the continuation probability δ. Moreover, helping behaviour would sustained not because it is uniquely optimal but because players are indifferent between helping and not helping, with their choices influenced by the scores of their co-players.

Compared with results of Chapter 1, we observe that the standing strategy creates stronger incentives for cooperation through a more robust reputation system. The requirement for $\delta > \frac{2c}{b+c}$ under the standing strategy ensures that players value future interactions sufficiently to maintain helping at 100%. Under BIS, the equilibrium condition $\delta = \frac{2c}{c+b(q-p)}$ depends on the difference $q - p$, reflecting the degree to which players discriminate based on scores.

The binary IS game has similar properties to the Engelmann and Fischbacher's game, i.e., helping behaviour is induced by a combination of self-interest and, for tie-breaking, non-selfish preferences for reciprocity. Therefore, by using a binary scale rather than a six-point scale, we are not changing the fundamental structure of the game. But the binary game is much easier to analyse.

3.3 Experimental Design

Our experiment builds upon the framework of our previous study, introducing key modifications to examine the effects of two different reputational mechanisms and belief elicitation on cooperative behaviour. Specifically, we implement three treatments — Binary Image Scoring (BIS), Sugden Good Standing (SGS), and a control condition with no reputational information — to investigate their efficacy in fostering cooperation and reciprocity within a helping game context. Additionally, we incorporate a belief elicitation task at the end of the second sequence to assess participants' perceptions of others' cooperative behaviour.

3.3.1 The Game

In each session, participants are randomly assigned to one of the three treatments in a between-subjects design. They engage in two sequences of an indefinitely repeated helping game, encountering first a homogeneous cost condition and then a heterogeneous cost condition, or vice versa. The order of these sequences is counterbalanced across sessions to control for any potential order effects.

Participants are organised into cohorts of six players, interacting anonymously and being randomly matched into pairs in each round. Within each pair, one participant is randomly assigned the role of the active player, while the other assumes the role of the non-active player. Roles are reassigned randomly in each round, ensuring that all participants had equal opportunities to act as active or non-active players throughout the experiment.

Active and non-active players receive an account endowed with £7 at the start of each round. The active player then decides whether to help the non-active player at a personal cost, providing a benefit to the recipient. The non-active player does not make any decisions during the round. Only one round for each sequence is paid.

In the homogeneous cost condition, the cost of helping is the same for all active players: choosing to help incurred a cost of £4, deducted from the active player's account, while the non-active player receives a benefit of £10, added to their account. If the active player chooses not to help, both players retain their £7. In the heterogeneous cost condition, active players are randomly assigned either a high cost (£6) or a low cost (£2) at the beginning of the sequence, with this assignment remaining constant throughout that sequence. The benefit to the non-active player remains £10 in all cases.

Participants are informed of their own cost of helping and that the benefit of being helped is £10, consistent across all sessions. They are also told that others might have different costs, not exceeding £6, but are not informed of the specific costs of other players. This maintained some uncertainty about others' incentives, reflecting realistic social environments where individuals may not fully know others' circumstances.

We utilise the terms "active player" and "non-active player" to describe the roles to keep the same terminology of the previous experiment.

In the BIS treatment, participants' reputations are represented by binary scores: a score of 1 if they have helped in their last opportunity as an active player, and 0 if they have not. All participants begin with a score of 1. The active player can see their own score and that of the non-active player before making their decision. The non-active player see only their own score. This simplified reputation mechanism focused on the most recent behaviour, potentially making reputational cues more salient and easier to interpret.

In the SGS treatment, participants' reputations are also represented by binary scores, but the updating rule differed. Everyone starts with a score of 1. In every round, the score updates according to the following rule: if the active player helps, their score depends on the non-active player's score: it becomes 1 if the non-active player had a score of 1, or remains unchanged if the non-active player had a score of 0. If they do not help, their score again depends on the non-active player's score: it becomes 0 if the non-active player had s core of 1, or remains unchanged if the non-active player had a score of 0.

In the control treatment, participants receive no reputational information; the active player does not see any scores before making their decision. The screens are otherwise identical across treatments, ensuring that any differences in behaviour can be attributed to the presence or absence of reputational cues.

At the end of each round, participants receive a summary of their role, decisions, earnings, and, where applicable, any changes to their scores. This feedback aims to keep participants informed about their progress and the consequences of their actions.

Within each cohort, subjects were randomly re-matched into pairs at the start of each round, resulting in a $\frac{1}{5}$ probability of meeting the same participant in two consecutive rounds. Subjects did not know with whom they were paired, nor did they know who was in their matching cohort in any sequence. Each round, the computer randomly assigned one subject to the non-active role and the other to the active role, with equal probability. Hence, in every round, half the subjects were non-active players and half were active players.

A random continuation rule determined the duration of each sequence (Roth and Murnighan, 1978). Each sequence had 40 fixed rounds, after which the sequence continued with a probability of 0.67. This design ensures a finite but indeterminate duration of interaction; beginning with round 40, the sequence is expected to continue for three further rounds. In the experiment, a computer simulated the roll of a six-sided die. If the roll was 1 or 2, the sequence would end; otherwise, the sequence continued to round 41. At the end of each round, all subjects observed the number drawn, which informed them about the end or continuation of the sequence and also served as a public coordination device. Sequences terminated simultaneously for both cohorts in every session.

3.3.2 Belief Elicitation

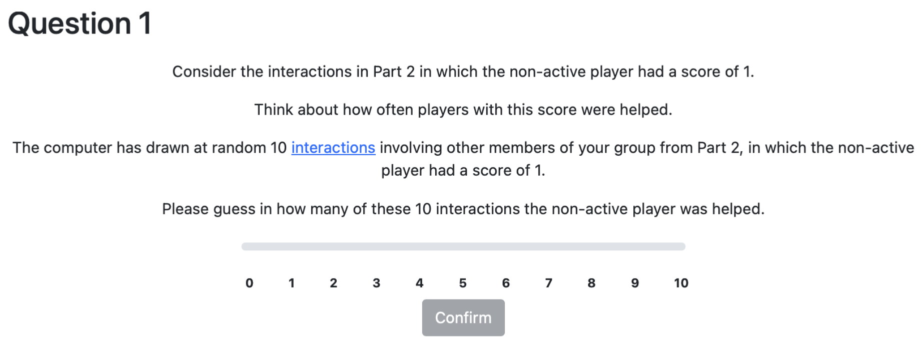
Following the second sequence of the helping game, we implemented a belief elicitation task to measure participants' expectations about helping behaviour among their peers. Specifically, for each participant, the computer randomly selected 10 games from the second sequence, excluding any interactions involving that participant (both as active and non-active player). In the treatments where scores were assigned (i.e., the IS and GS treatments), these selected games specifically included interactions where non-active players had scores of 0 or 1.

We asked each participant to guess in how many of those 10 games the non-active players with a particular score were helped by active players. This approach made the question and the incentive more concrete and easier for participants to understand. For example, in the IS and GS treatments, participants were presented with the prompt shown in Figure 3.3.

In the baseline treatment, participants were still asked to estimate the number of times non-active players were helped out of the 10 randomly drawn interactions as described above, without reference to scores[2].

To incentivise accurate reporting, we employed a novel incentive-compatible mechanism where participants received additional payments based on the accuracy of their estimates. Specifically, participants were paid according to the actual out-

[2] See Figure 3.18 in the Appendix 3.A



Question 1

Consider the interactions in Part 2 in which the non-active player had a score of 1.

Think about how often players with this score were helped.

The computer has drawn at random 10 interactions involving other members of your group from Part 2, in which the non-active player had a score of 1.

Please guess in how many of these 10 interactions the non-active player was helped.

Figure 3.3: Belief elicitation for non-active players with score of 1

comes in those 10 randomly selected games for their belief elicitation task, not according to behaviour in the experiment as a whole.

Participants reported an integer $k \in \{0, 1, \ldots, N\}$ representing their estimate, and their loss was given by $L(k, i) = |k - i|$, where i is the actual number of times non-active players were helped.

Under this loss function, a risk-neutral participant minimises their expected loss by reporting the median of their subjective belief distribution over possible outcomes. The intuition behind this result is that the expected loss $L(k)$ is minimised only if the cumulative probability of outcomes less than or equal to k is at least one-half. If a participant reports a value lower than the median, there is a greater probability that the actual outcome will be higher than their estimate, leading to an increased expected loss due to underestimation. Conversely, reporting a value higher than the median increases the expected loss due to overestimation, since there is a higher chance the actual outcome will be lower than their estimate. By reporting the median, participants balance these risks, minimising the expected absolute deviation from the actual outcome.

Our method differs from commonly used belief elicitation mechanisms, such as the quadratic scoring rule (Brier, 1950), which typically require participants to report probabilities for single events. The quadratic scoring rule is incentive-compatible and theoretically sound but can be cognitively demanding, as it asks participants to assign precise probabilities to uncertain events — a task that can be abstract and unintuitive. Psychological research suggests that individuals find it easier to reason about frequencies with a common base than about abstract probabilities (Gigerenzer and Hoffrage, 1995; Cosmides and Tooby, 1996). By asking participants to estimate frequencies rather than probabilities, we align the elicitation task more closely with natural cognitive processes, potentially leading to more accurate and reliable data.

While a single probability is a unique summary of a belief about a binary event, summarising a distribution over multiple outcomes with a single number is inherently less precise. However, our loss function leverages this by making the median the optimal report for a rational, risk-neutral participant. This provides participants with a clear and psychologically intuitive reporting strategy. Even if participants are not fully rational, the simplicity and naturalness of reporting a frequency make the task more accessible, reducing cognitive load and potential biases.

By structuring the incentives in this manner, we ensure that participants have a clear financial motivation to report their true (median) belief, thereby eliciting accurate representations of their expectations about others' cooperative behaviour. The simplicity of reporting a frequency, combined with the directness of the loss function, enhances the psychological plausibility of truthful reporting. The formal proof demonstrating that reporting the median minimises the expected loss is provided in the Appendix. This proof establishes the theoretical foundation for our belief elicitation method and justifies its application within our experimental design.

Furthermore, by comparing each participant's elicited belief to the session-wide median, we assess the accuracy of their perceptions. This comparison is in line with our proposed incentive mechanisms, which are designed to encourage truthful reporting of beliefs relative to the median.

Our approach thus offers a psychologically grounded and methodologically sound alternative to traditional probability-based elicitation methods. By focusing on frequencies and utilising a loss function that aligns with natural reasoning processes, we aim at obtaining more reliable data on participants' beliefs, which in turn allows for a deeper analysis of how these beliefs correlate with cooperative behaviour in strategic settings.

3.3.3 Implementation

Participants were recruited using SONA Systems from the student population at the University of East Anglia (UEA), consistent with the recruitment method employed in our previous experiment. To maintain comparability between studies, the only eligibility criterion was that participants had not taken part in the prior experiment, ensuring the underlying population remained the same.

Each session included twelve participants, divided into two cohorts of six. Participants remained in the same cohort throughout the session but were randomly rematched within their cohort in each round. We conducted a total of eighteen sessions, with an equal number assigned to each treatment: six sessions (twelve cohorts) each for the Binary Image Scoring (BIS), Sugden Good Standing (SGS), and control treatments. The order of the homogeneous and heterogeneous cost conditions was counterbalanced across sessions to mitigate potential order effects. This counterbalancing ensured that any systematic effects due to the order of cost conditions were evenly distributed across treatments, enhancing the internal validity of our findings.

To ensure anonymity and reduce potential biases, participants interacted via computer terminals, and communication was prohibited during the experiment. Instructions were provided in neutral language, informing participants about the structure of the game, their potential decisions, and how their earnings would be calculated. This standardised approach minimised the influence of extraneous variables and allowed us to attribute observed differences in behaviour to the experimental manipulations.

Prior to conducting the experiment, we preregistered our experimental design and analysis plan, which included a stopping rule for the control treatment based on statistical power considerations. We hypothesised that the presence of a reputation mechanism — either BIS or SGS — would increase helping behaviour compared to the baseline control condition without a reputation mechanism.

Initially, we planned to run three sessions per condition (control, BIS, SGS), resulting in a total of nine sessions and eighteen independent cohorts (since each session comprised two cohorts). After these initial sessions, we intended to assess whether the control condition exhibited significantly lower helping rates compared to the reputation mechanisms. Specifically, our stopping rule stipulated that if five or more of the bottom six cohorts (ranked by average helping rate) belonged to the control condition, we would discontinue the control treatment.

The intuition behind this stopping rule is rooted in statistical significance and power. By requiring that at least five of the bottom six cohorts are from the control group, we set a stringent criterion that would be unlikely to occur by chance under the null hypothesis of no difference between treatments. Under the null hypothesis, the probability of this event is approximately 8.3% (see Appendix 3.A.2 for detailed calculations and a comparison with a stricter hypergeometric calculation.). This threshold balances the risk of Type I and Type II errors, ensuring that we have sufficient power to detect a meaningful effect while avoiding unnecessary continuation of the control condition if it is indeed inferior.

If the stopping rule was met, we would proceed to conduct additional sessions only for the reputation mechanisms, aiming for a total of six sessions (twelve cohorts) per reputation mechanism and three sessions (six cohorts) for the control. This would result in 180 participants: 72 assigned to BIS, 72 to SGS, and 36 to the control group. Conversely, if the stopping criterion was not satisfied, we planned to conduct three additional sessions per condition, leading to a total of eighteen sessions (thirty-six cohorts), with six sessions (twelve cohorts) per condition and 216 participants overall.

In practice, the stopping criterion was not satisfied after the initial nine sessions, as the control cohorts did not consistently rank among the lowest in terms of helping rates. For instance, while some control cohorts exhibited lower helping rates, they did not constitute five of the bottom six cohorts. Consequently, we proceeded to conduct the full set of eighteen sessions, with six sessions (twelve cohorts) allocated to each condition: BIS, SGS, and the control.

This experimental design allows us to rigorously assess the impact of simplified reputation mechanisms on cooperation and reciprocity within helping games. By comparing behaviour across the BIS, SGS, and control treatments, and by analysing the relationship between participants' beliefs and their actions, we aim to provide insights into the effectiveness of these mechanisms in promoting prosocial behaviour.

3.4 Results

We begin our presentation of results by examining whether the preregistered stopping rule was met. The stopping rule stipulated that if five or more of the bottom six cohorts, ranked by average helping rates, were from the baseline treatment, we would discontinue this condition. Table 3.2 displays the ranked average helping rates for both homogeneous and heterogeneous cost conditions. The rankings indicate that this criterion was not satisfied.

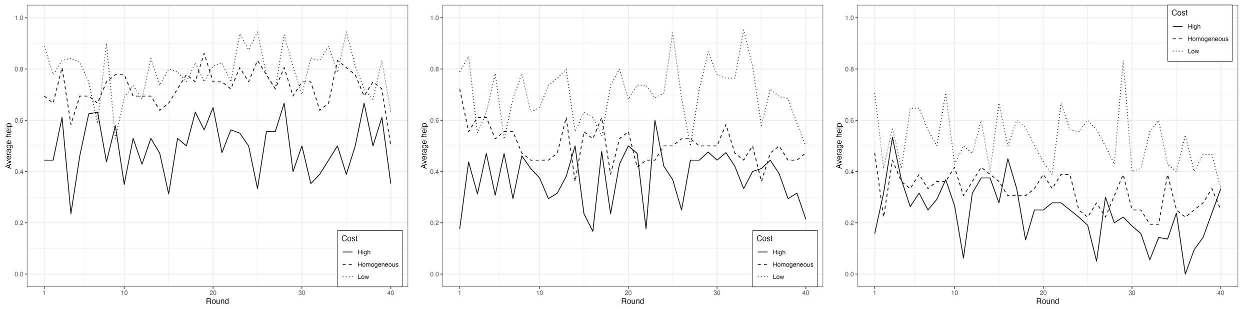
Table 3.2: Ranked average helping rates for homogeneous and heterogeneous cohorts
from lowest to highest

Rank	**Homogeneous Cost**			**Heterogeneous Cost**		
	Cohort	**Treatment**	**Avg. Help**	**Cohort**	**Treatment**	**Avg. Help**
1	1	B	0.0078	2	B	0.2417
2	2	B	0.0620	13	IS	0.2602
3	5	IS	0.1163	8	GS	0.3500
4	13	IS	0.2302	1	B	0.3667
5	15	B	0.2460	15	B	0.3739
6	10	B	0.2778	9	B	0.3833
7	9	B	0.3492	12	IS	0.3917
8	8	GS	0.3651	10	B	0.4083
9	12	IS	0.4286	7	GS	0.4500
10	6	IS	0.5349	6	IS	0.5000
11	7	GS	0.5556	5	IS	0.6250
12	4	GS	0.5891	3	GS	0.6333
13	16	B	0.6269	16	B	0.6422
14	11	IS	0.6269	4	GS	0.6750
15	18	B	0.6349	11	IS	0.7750
16	14	IS	0.7539	18	GS	0.8049
17	17	GS	0.9286	14	IS	0.8537
18	3	GS	1.0000	17	GS	0.9106

In the homogeneous cost condition, only four of the six lowest-ranked cohorts (Cohorts 1, 2, 15, and 10) belonged to the baseline treatment, with the remaining two cohorts from the IS and GS treatments. Similarly, in the heterogeneous cost condition, four of the bottom six cohorts (Cohorts 2, 1, 15, and 9) were associated with the baseline treatment. This confirms that the baseline treatment did not occupy the lowest 5 out of 6 ranks in helping behaviour. Consequently, we proceeded to conduct six sessions for each treatment condition, ensuring a balanced experimental design and sufficient data for robust comparisons.

Our study comprises a total of 216 participants[3], evenly allocated across three treatments: Sugden Good Standing (SGS), Binary Image Scoring (BIS), and a baseline condition without reputational mechanisms. Each treatment involved a consistent group of 36 active players in each round, ensuring comparability across conditions.

Figure 3.4 illustrates the average helping rates per round across the three treatments: SGS (3.4a), BIS (3.4b), and the baseline condition (3.4c). Different line styles distinguish the cost conditions.



(a) SGS treatment (b) BIS treatment (c) Baseline treatment

Figure 3.4: Trends average helping per round
Dashed lines display homogeneous cost. Thick and dotted lines display heterogeneous costs.

A salient feature across all treatments is the substantial variability in helping rates among high-cost individuals within the heterogeneous cost condition. These participants consistently exhibit lower helping rates compared to other cost conditions. In contrast, the homogeneous cost condition demonstrates greater stability, with helping rates levelling off after an initial decline—a pattern reminiscent of that observed in public goods games (e.g., Herrmann et al., 2008).

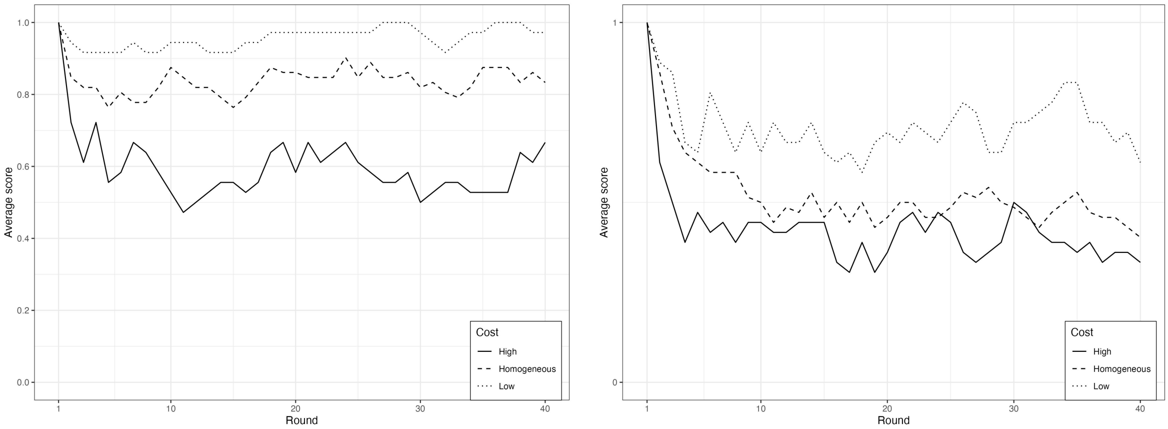
Notably, low-cost individuals in the heterogeneous cost condition achieve the highest average helping rates across all treatments. This divergence between high-cost and low-cost participants aligns with our theoretical expectation that the SGS mechanism would create a bifurcation in cooperative behaviour, widening the difference as the game progresses.

The SGS mechanism appears to foster higher average helping rates compared to

[3] See Table 3.3 in the *Appendix*.

the BIS and baseline treatments. This suggests that the SGS rule, by facilitating discrimination based on standing, allows individuals to form cooperative clusters or clubs — a phenomenon consistent with the formation of cooperative norms and selective reciprocity observed in previous studies (e.g., Fehr et al., 2002).

Figure 3.5 presents data on players' scores under the SGS and BIS treatments; scoring trends are not applicable for the baseline condition.



(a) SGS treatment

(b) BIS treatment

Figure 3.5: Trends average score per round
Dashed lines display homogeneous cost. Thick and dotted lines display heterogeneous costs.

In the SGS treatment (3.5a), both the homogeneous and heterogeneous cost conditions demonstrate a clear pattern of stabilisation after an initial decline in the first few rounds. This stabilisation suggests that players' standings settle as they adapt to the game's structure, reflecting consistent strategies. The mechanism itself, being recursive and updating automatically, may facilitate an adjustment process leading to a more stable cooperative environment.

In contrast, the BIS treatment (3.5b) exhibits stabilisation of scores slightly later, possibly reflecting differences in the learning dynamics or behavioural adjustments required under this system. The average scores under the SGS treatment are higher overall compared to the BIS treatment, particularly in the homogeneous cost condition. This difference implies that the SGS mechanism may better incentivise helping behaviour or more effectively encompass the information about who is forth of being helped.

Moreover, the contrast in behaviour between high-cost (full line) and low-cost players (dotted line) within the heterogeneous cost condition is striking. High-cost players tend to exhibit consistently lower average scores compared to their low-cost counterparts, highlighting the influence of cost structure on individual performance and cooperative dynamics across treatments.

The observed stabilisation of scores in both treatments provides a strong rationale for focusing our main analysis on rounds 5 to 35. By round 5, players have likely moved past the initial adjustment phase, and their scores reflect consistent and meaningful strategies. Analysing rounds beyond round 35 may introduce end-game biases, where participants alter their behaviour in anticipation of the game's conclusion (Andreoni, 1988). Focusing on rounds 5 to 35 allows us to examine participants' steady-state behaviour, minimising the confounding effects of early learning and late-game strategic shifts[4].

We proceed with a cohort-level analysis, as this method offers independent data points that enable a transparent and reliable examination of overall helping behaviour. We analyse the average helping rates across cohorts for each treatment in Section 3.4.1. Subsequently, we delve into detailed behavioural patterns using individual-level data in Section 3.4.2, allowing for a thorough examination of decision-making processes in the helping game. In Section 3.4.3, we analyse the beliefs data collected at the end of each session.

3.4.1 Cohort-level Analysis

In this section, we evaluate rounds 5 to 35 of each game, using the average choices of each cohort as the unit of observation. By excluding rounds characterised by initial learning and potential end-game effects, we obtain 36 independent data points corresponding to the 36 cohorts in our experiment[5].

3.4.1.1 Cooperation rates between SGS, BIS, and B in helping games

A preliminary examination of the data reveals substantial differences in helping rates across cohorts. Figure 3.5 presents box plots and dot plots of the average helping rates for each cohort in the homogeneous cost helping games. Each dot represents a cohort, and the box encompasses the interquartile range (IQR), representing the

[4] This analysis is similar if we follow the specification used in Chapter 2: rounds 10 to 35.

[5] Results are similar if we use the full dataset or the specification used in the previous chapter: rounds 10 to 35.

middle 50% of the data. Figure 3.6 provides the corresponding information for the heterogeneous cost helping games.

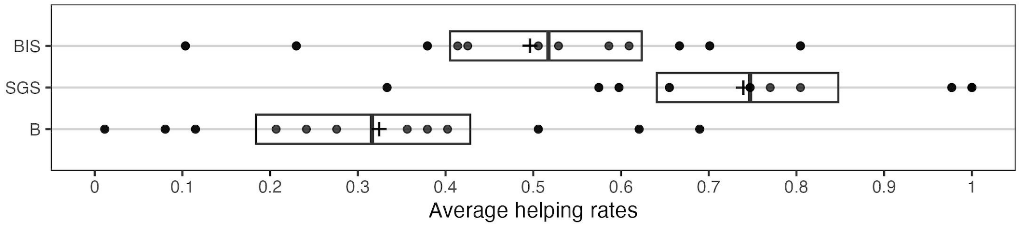


Figure 3.6: Average helping rate in homogeneous cost helping games
Each dot represents one cohort. Each cross represents the average for each treatment.

In the homogeneous cost condition (Figure 3.6), the SGS mechanism exhibits a median helping rate of 0.747, the BIS mechanism shows a lower median of 0.517, and the baseline condition displays an even lower median of 0.316. This ordering suggests that the SGS mechanism is most effective at sustaining cooperation, consistent with our initial expectations.

In the heterogeneous cost condition (Figure 3.7), the SGS mechanism has a median helping rate of 0.638, the BIS mechanism a median of 0.534, and the baseline treatment a median of 0.374. The reduction in median helping rates compared to the homogeneous condition reflects the impact of cost heterogeneity on cooperation.

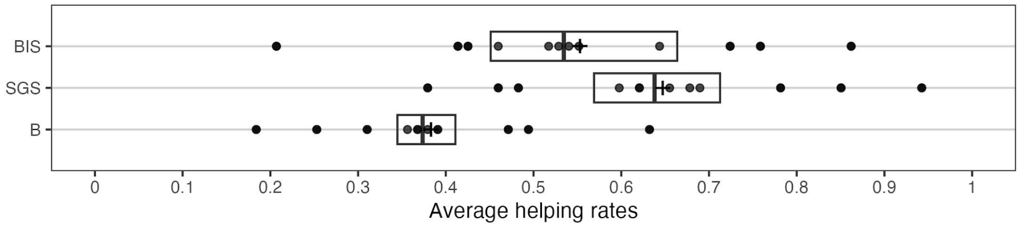


Figure 3.7: Average helping rate in heterogeneous cost helping games
Each dot represents one cohort. Each cross represents the average for each treatment.

Importantly, the variances in helping rates are similar across treatments in both conditions. Specifically, in the homogeneous cost condition, the variance for SGS (0.039) is similar to BIS (0.040) ($\tilde{\chi}^2 = 0.227$, $p = 0.634$), SGS is similar to the baseline (0.044) ($\tilde{\chi}^2 = 0.196$, $p = 0.658$), and BIS is similar to the baseline ($\tilde{\chi}^2 = 0.101$, $p = 0.750$).

In the heterogeneous cost condition, the variance for SGS (0.026) i similar to BIS (0.031)($\tilde{\chi}^2 = 0.025$, $p = 0.874$), SGS is similar to the baseline (0.013) ($\tilde{\chi}^2 = 1.999$,

$p = 0.157$), and BIS and the baseline also have comparable variances ($\tilde{\chi}^2 = 1.268$, $p = 0.260$).

This suggests that the SGS mechanism increases average cooperation without necessarily increasing variability in behaviour.

Result 1: *The Sugden Good Standing mechanism induces higher levels of cooperation compared to both Binary Image Scoring and a baseline treatment.*

The mean helping rates differ significantly between the treatments in the homogeneous cost condition. Specifically, the mean helping rate for SGS is 0.739, compared to 0.496 for BIS and 0.324 for the baseline. Pairwise comparisons reveal significant differences: SGS exhibits higher mean helping rates than BIS ($Z = 2.583$, $p = 0.007$) and baseline ($Z = -3.494$, $p < 0.001$), while BIS and baseline show marginally different mean helping rates ($Z = -1.926$, $p = 0.054$).

In the heterogeneous cost condition, the mean helping rates are 0.647 for SGS, 0.553 for BIS, and 0.383 for the baseline. SGS has significantly higher rates compared to the baseline. Pairwise comparisons indicate that SGS and BIS have similar helping rates ($Z = 1.336$, $p = 0.192$), while SGS has significantly higher rates compared to the baseline ($Z = -3.358$, $p < 0.001$). BIS also demonstrates higher mean helping rates than the baseline ($Z = -2.453$, $p = 0.011$).

These findings suggest overall that the SGS mechanism generally promotes greater cooperation than both the BIS and baseline treatments. The effectiveness of the SGS rule may stem from its capacity to facilitate selective helping based on standing, thereby reinforcing cooperative norms and mitigating free-riding—a mechanism absent in the baseline condition and less pronounced in the BIS treatment.

3.4.1.2 High cost and low cost players in heterogeneous game

We further analyse helping behaviour under heterogeneous cost conditions to explore how cost differences affect cooperation within each treatment.

For the low-cost players (Figure 3.8), the average helping rate is highest in the SGS treatment (0.786), followed by BIS (0.712) and the baseline (0.539). Pairwise comparisons show a little difference between SGS and BIS ($Z = 1.136$, $p = 0.262$), while SGS exhibited higher helping rates than the baseline ($Z = -2.699$, $p = 0.005$). BIS exhibits as well higher helping rates than the baseline ($Z = -1.876$, $p = 0.060$).

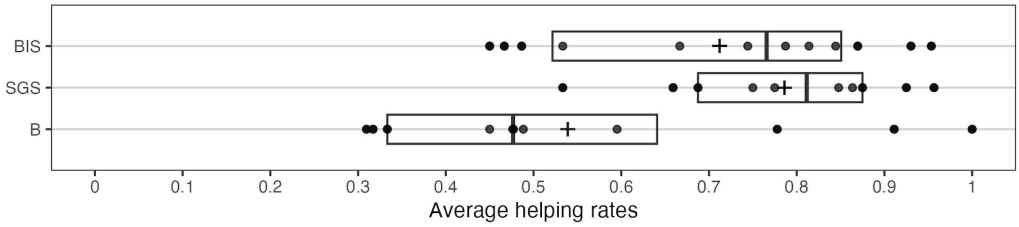


Figure 3.8: Average helping rate in heterogeneous cost helping games for player with a low cost

Each dot represents one cohort. Each cross represents the average for each treatment.

The variance in helping rates was lowest in SGS (0.016), followed by BIS (0.034) and the baseline (0.056). Pairwise comparisons of variances indicate no significant differences between SGS and BIS ($\tilde{\chi}^2 = 1.725$, $p = 0.189$), SGS and the baseline ($\tilde{\chi}^2 = 1.137$, $p = 0.286$), or BIS and the baseline ($\tilde{\chi}^2 = 0.012$, $p = 0.914$).

For the high-cost players (Figure 3.9), the average helping rate was highest in the SGS treatment (0.502), followed by BIS (0.397) and the baseline (0.243). Pairwise comparisons show that the mean helping rate for SGS is comparable to BIS ($Z = 1.052$, $p = 0.301$), while SGS has a higher mean helping rate compared to the baseline ($Z = -2.754$, $p = 0.003$). The difference in mean helping rates between BIS and the baseline is less pronounced ($Z = -1.687$, $p = 0.092$).

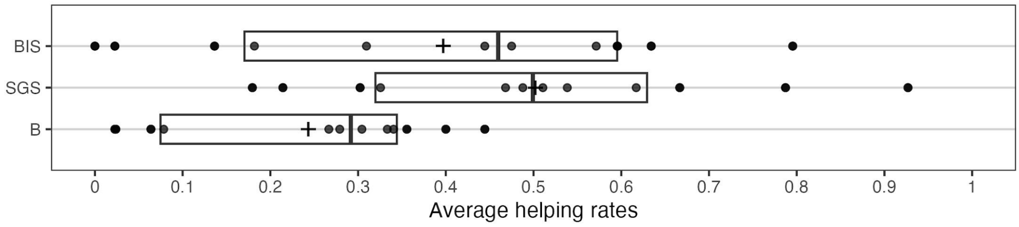


Figure 3.9: Average helping rate in heterogeneous cost helping games for player with a high cost

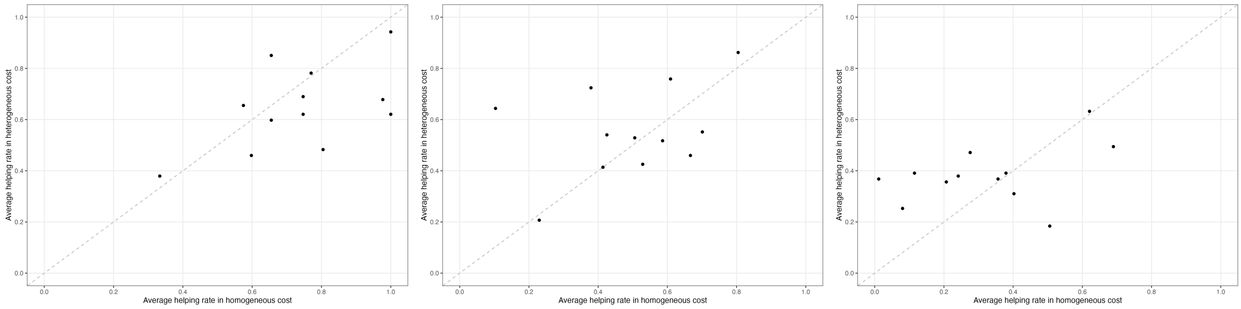
Each dot represents one cohort. Each cross represents the average for each treatment.

The variance in helping rates for high-cost players was 0.032 for SGS, 0.068 for BIS, and 0.023 for the baseline. Pairwise variance comparisons indicate no notable differences between SGS and BIS ($\tilde{\chi}^2 = 0.767$, $p = 0.381$), SGS and the baseline ($\tilde{\chi}^2 = 1.187$, $p = 0.276$), or BIS and the baseline ($\tilde{\chi}^2 = 3.223$, $p = 0.073$).

These findings support the overall trend reported in the homogenous cost condition, where the Sugden GS mechanism leads to higher helping rates than both the binary image scoring mechanism and the baseline treatment. This tendency remains across different cost conditions, particularly among low-cost participants, implying that SGS may be more successful at inducing helping.

3.4.1.3 Correlation between average helping rates within IS and GS

Figure 2.10 present three scatter plots comparing average helping rates under homogenous and heterogeneous cost situations for the SGS and BIS mechanisms, as well as the baseline treatment. We compute Pearson correlation coefficients for each treatment to assess the persistence of helping behaviour across cost structures.



(a) SGS treatment (b) BIS treatment (c) Baseline treatment

Figure 3.10: Correlation between average helping rates in homogeneous and heterogeneous cost condition
Each dot correspond to a cohort.

For the SGS mechanism, there is a moderately positive correlation between helping rates in the two cost conditions (Pearson correlation,$r(12) = 0.550$, $p = 0.064$) This suggests that cohorts with higher cooperation in one cost condition tend to maintain higher cooperation in the other, indicating a consistent pattern of cooperative behaviour facilitated by the SGS mechanism.

In contrast, the BIS mechanism (Pearson correlation, $r(12) = 0.371$, $p = 0.235$) and the baseline treatment (Pearson correlation, $r(12) = 0.400$, $p = 0.197$) show weaker and statistically insignificant correlations. This implies that helping behaviour in these treatments is more context-dependent and less persistent across different cost structures.

The positive correlation within SGS cohorts may reflect the influence of participants' prior experiences and expectations brought from outside the laboratory.

Individuals come with varying degrees of reciprocity and beliefs about others' behaviour, leading to different mixes of types within a cohort. The SGS mechanism, by providing a clear rule for good standing, may reinforce these tendencies and create a positive correlation between cohorts—a phenomenon observed in public goods experiments with conditional cooperators (Fischbacher et al. (2001)).

3.4.2 Individual-level Analysis

We now shift our focus to individual-level data to further explore the dynamics of helping behaviour. Our analysis includes 1,044 individual helping decisions for each treatment condition, corresponding to 36 active players over 29 rounds.

3.4.2.1 Analysis of frequency of being helped

Figures 3.11 and 3.12 analyse how a non-active player's standing (in the SGS mechanism) and image score (in the BIS mechanism) influence their likelihood of being helped under homogeneous and heterogeneous cost structures.

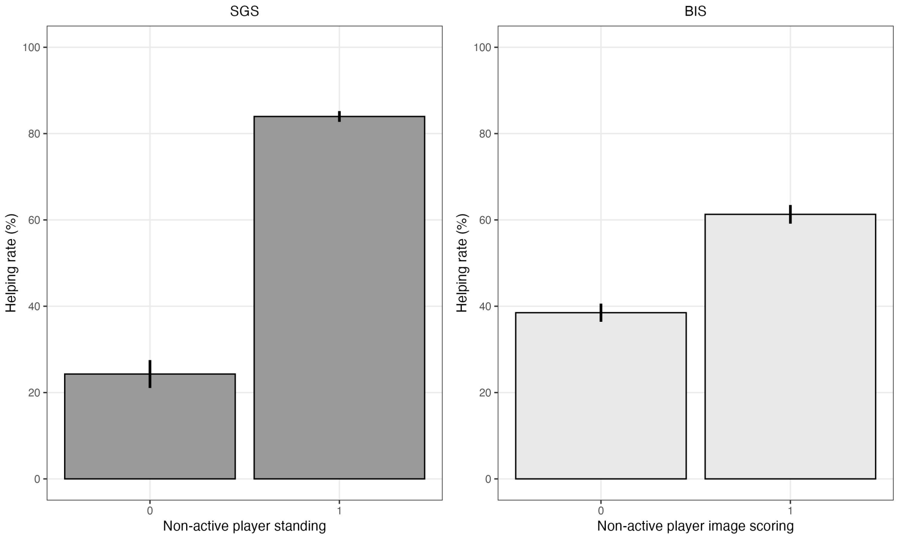


Figure 3.11: Homogeneous cost, frequency of being helped by score
Left panel: Sugden Good Standing. Right panel: Binary Image Scoring.

In the SGS mechanism (Figure 3.11, left panel), there is a strong positive relationship between a player's standing and the frequency of being helped. Non-active players in good standing are helped 83.9% of the time, compared to 24.3% for those in bad standing. The Pearson correlation between standing and being helped is high ($r(1044) = 0.51$, $p < 0.001$).

In contrast, the BIS mechanism (Figure 3.11, right panel) shows a weaker relationship. Players with a high image score are helped 61.3% of the time, compared to 38.5% for those with a low score. The correlation is moderate ($r(1044) = 0.228$, $p < 0.001$).

Under heterogeneous costs, the pattern is similar. In the SGS mechanism (Figure 3.12, left panel), players in good standing are helped 77.7% of the time, compared to 24.9% for those in bad standing ($r(1044) = 0.478$, $p < 0.001$). The BIS mechanism (Figure 3.12, right panel) again shows a weaker relationship, with high-score players helped 61.7% of the time versus 47.3% for low-score players ($r(1044) = 0.144$, $p < 0.001$).

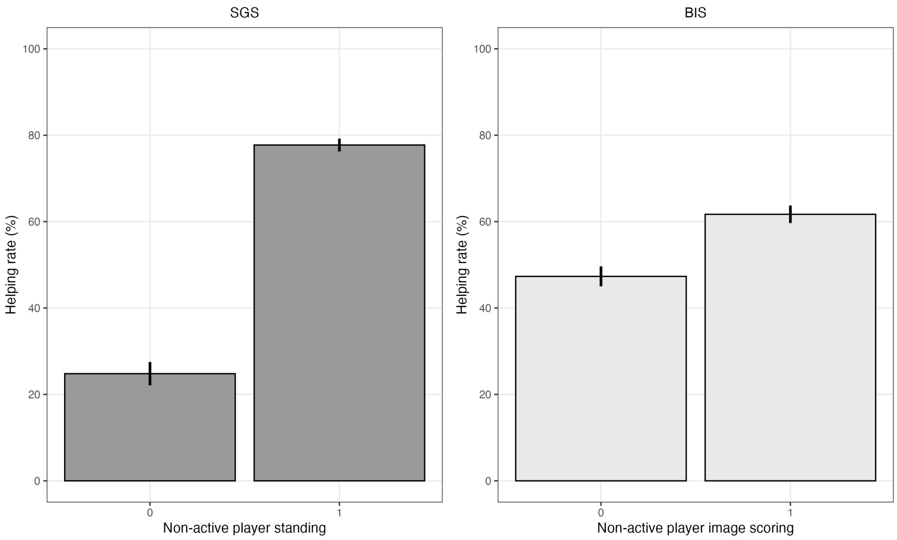


Figure 3.12: Heterogeneous cost, frequency of being helped by score
Left panel: Sugden Good Standing. Right panel: Binary Image Scoring.

These results highlight the effectiveness of the SGS mechanism in fostering cooperative behaviour through reputational incentives. The strong correlation between standing and being helped suggests that participants are discriminating in their helping decisions based on the clear rule provided by the SGS system. In contrast, the BIS mechanism, providing the information of the last action, may lead to less pronounced discrimination and, consequently, lower overall cooperation.

In the baseline condition (Figure 3.13), the average helping rates are similar across cost conditions, with a slightly higher rate under the heterogeneous cost condition (38.3%) compared to the homogeneous cost condition (32.4%). This indicates that, in the absence of reputational mechanisms, cost conditions alone do not lead to substantial differences in helping rates. The presence of a reputational system, particularly one with clear behavioural prescriptions and an equilibrium with 100% helping like SGS, appears crucial in enhancing cooperation.

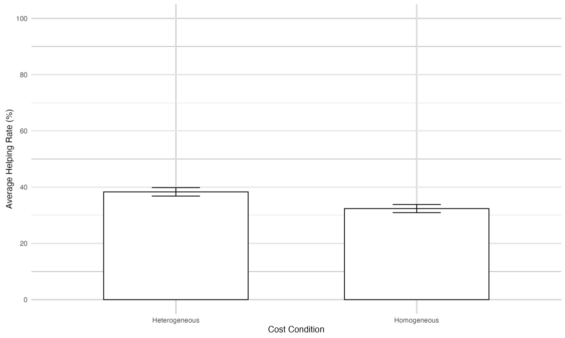


Figure 3.13: Frequency of being helped by cost condition in baseline treatment

3.4.3 Beliefs

In this subsection, we analyse the beliefs data collected from participants across different treatments: Baseline, Binary Image Scoring (BIS), and (SGS). Our primary aim is to assess whether our belief elicitation method effectively captures participants' perceptions of others' helping behaviour and how these perceptions differ under various scoring mechanisms.

We begin with the Baseline treatment, where participants were asked to predict the overall helping rate without any score-based conditioning. Figure 3.14 presents a scatter plot where each dot represents one of the 72 participants in this treatment. The horizontal axis denotes the actual average helping rate in the participant's cohort, while the vertical axis indicates the participant's prediction (belief) of this rate.

The strong positive correlation observed in Figure 3.14 suggests that participants have a reasonably accurate sense of the overall helping behaviour in their cohort. This alignment between predictions and actual helping rates indicates that our belief elicitation method is effective in capturing participants' perceptions in the absence of scoring mechanism.

We next examine the BIS and SGS treatments, where participants' beliefs were elicited conditional on the non-active player's score—either 0 or 1. Figures 3.15 and 3.16 illustrate the scatter plots for these treatments, with separate panels for score

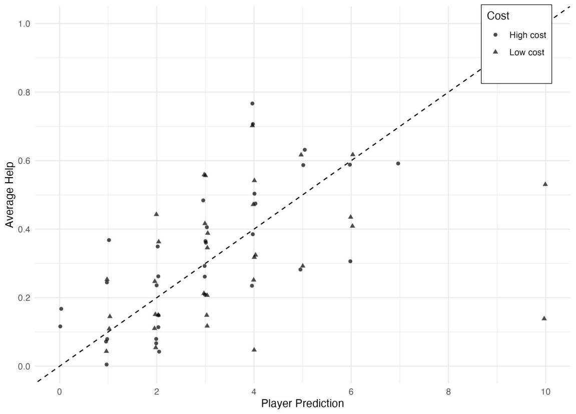
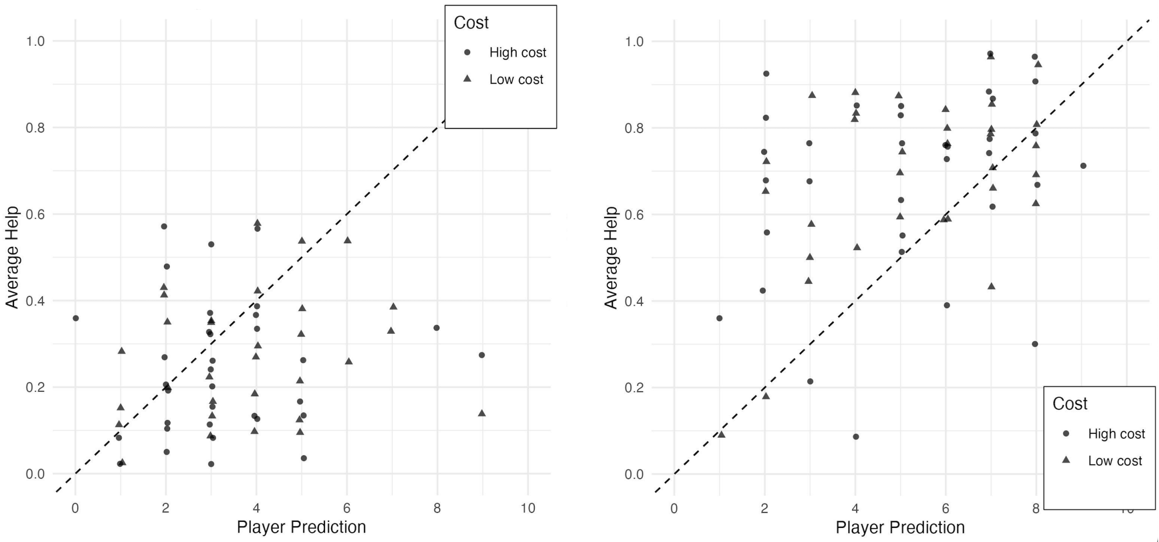


Figure 3.14: Correlation between player's prediction (belief) and average help in Baseline treatment

0 (left) and score 1 (right).

The scatter plots illustrate the relationship between participants' predictions of helping behaviour and the actual average helping behaviour observed within their cohorts. Each point in the plots corresponds to a participant, with the horizontal axis representing the predicted helping rate and the vertical axis representing the average observed helping rate. The dashed 45-degree line indicates perfect accuracy, where predictions align exactly with observed outcomes.

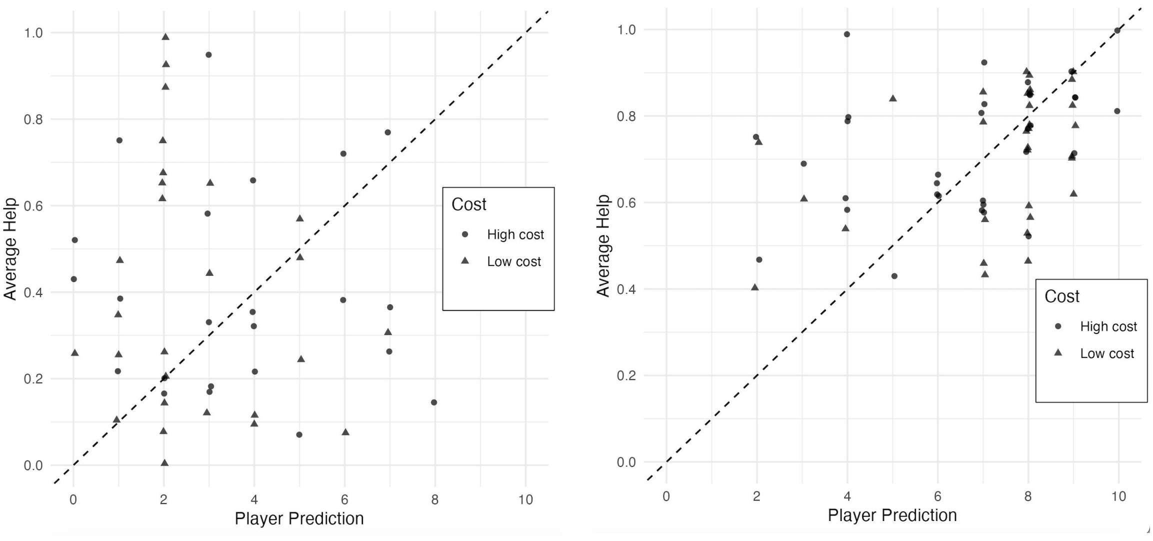


(a) BIS, score 0

(b) BIS, score 1

Figure 3.15: Correlation between player's prediction (belief) and average help in IS treatment

In the SGS treatment, this positive correlation persists but varies across score



(a) SGS, score 0

(b) SGS, score 1

Figure 3.16: Correlation between player's prediction (belief) and average help in GS treatment

conditions. When the non-active player has a score of 1, participants' predictions are notably accurate, clustering tightly around the line of perfect accuracy. This suggests that participants recognise the influence of a higher score on encouraging helping behaviour. However, in the score 0 condition, predictions are less aligned with actual helping rates, indicating some uncertainty or variability in anticipating others' actions when the score is low.

The BIS treatment displays a weaker correlation between predictions and helping rates. The scatter plots reveal significant dispersion around the 45-degree line, especially in the score 0 condition. This suggests that participants' predictions in the BIS treatment are less sensitive to the actual helping behaviour observed, potentially reflecting confusion or misperceptions about how scores affect others' decisions in this context.

Assessing the responsiveness of participants' predictions involves we examining the slope of the regression line between predictions and helping rates. Baseline and SGS treatments, particularly in the SGS score 1 condition, the slope approaches 1. This indicates that participants adjust their predictions proportionally with changes in observed helping behaviour.

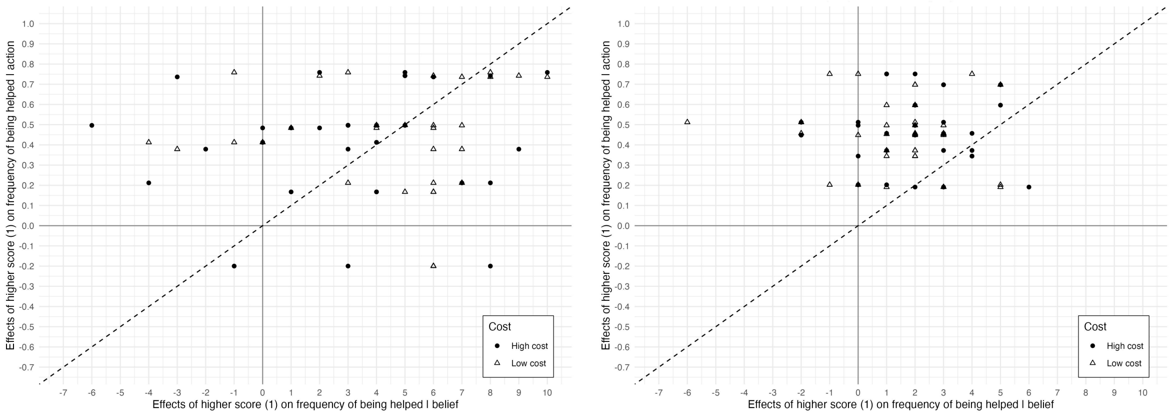
In contrast, the BIS treatment has a flatter regression slope, with a value less than 1. This flatter slope reflects a lower sensitivity of participants' predictions to variations in actual helping rates, suggesting that participants in the BIS condition may not fully appreciate the impact of score differences on helping behaviour.

Moreover, in the BIS treatment, there is a tendency to over-predict helping when the score is 0 and under-predict when the score is 1. This is evidenced by predictions lying predominantly above the 45-degree line in the score 0 condition and below it in the score 1 condition. Such patterns indicate that participants underestimate the benefit of having a higher score in the BIS setting.

The SGS treatment exhibits a more balanced pattern. While there is slight over-prediction in the score 0 condition, it is less pronounced than in BIS. In the score 1 condition, participants' predictions align closely with actual helping rates, with no systematic bias towards over- or under-prediction. This suggests that participants in the SGS treatment have a more accurate understanding of how scores influence helping behaviour.

Finally, Figure 3.17 reveals some noteworthy patterns. Both panels depict the relationship between the predicted and actual effects of holding a higher score on the likelihood of receiving help, as observed under the Sugden good standing (3.17a) and the binary image scoring (3.17b). The x-axis represents the predicted difference in the belief of being helped between individuals with a score of 1 compared to those with a score of 0, referred to as the effect of having a higher score on belief. In contrast, the y-axis represents the actual observed difference in the frequency of help received between individuals with a score of 1 versus those with a score of 0, capturing the effect of a higher score on action. The data points are categorised into two conditions: filled circles indicate the high cost condition, while open triangles represent the low cost condition. A dashed diagonal line represents a perfect 1:1 correlation, where predicted and actual values would be in complete agreement.

Under the binary image scoring model (Figure 3.17b), all data points are positioned above the horizontal line at zero. This indicates that every participant received more help when they had a score of 1 compared to when they had a score of 0. In contrast, this is not universally true under the Sugden good standing model (Figure 3.17a), although only a few participants experienced the opposite. When examining the data to the left of the vertical line at zero, these points correspond to participants who predicted they would receive less help with a score of 1 than with a score of 0. This aligns with previous findings in the literature, which suggest that



(a) Sugden Good Standing

(b) Binary Image Scoring

Figure 3.17: Correlation between difference in predictions (1 - 0) and difference in actions (1 - 0)

self-reported beliefs can seldom differ to a small extent from actual experiences.

Figure 3.17 ultimately suggests that participants under the binary image scoring model tend to underestimate the true frequency of help they receive. Conversely, this tendency appears to be less pronounced under the Sugden good standing model, where the data points are more dispersed.

3.5 Conclusion

This study set out to examine the efficacy of binary reputational mechanisms in fostering cooperative behaviour within helping games. Our theoretical framework identified four binary rating rules that satisfy key axioms intended to encapsulate helpfulness — Leimar and Hammerstein Good Standing (L&H GS), Sugden Good Standing (SGS), Binary Image Scoring (BIS), and Modified Binary Image Scoring. While these mechanisms align theoretically with principles of indirect reciprocity, we experimentally test the most interesting three and our experimental findings reveal significant differences in their practical effectiveness.

The SGS mechanism emerged as particularly effective in promoting cooperation. Across both homogeneous and heterogeneous cost conditions, SGS consistently elicited higher levels of helping behaviour compared to BIS and the baseline condition devoid of reputational incentives. This superiority was evident not only in higher average helping rates but also in the stability and persistence of cooperative behaviour over time.

The superior performance of SGS can be attributed to its structural properties. Unlike BIS, which updates an individual's reputation based solely on their most recent action, SGS incorporates both the action and the standing of the players. An individual gains good standing only by helping someone who is also in good standing. This mechanism fosters a club-like structure where cooperation is selectively directed towards those who have themselves demonstrated cooperative behaviour.

Philosophically, this distinction underscores important considerations about how reputations are formed and maintained. Scoring mechanisms that focus exclusively on recent actions, such as BIS, effectively erase the influence of past behaviours. While this simplicity may reduce cognitive demands, it disregards the accumulation of consistent cooperative or uncooperative actions over time. Such an approach can lead to reputations that are fragile and easily swayed by singular events, undermining the stability of cooperative relationships. In contrast, SGS acknowledges that reputations are built through a history of interactions, promoting a more robust and enduring form of cooperation.

Our findings are further reinforced when considering our previous experiment conducted with the same underlying population, as detailed in Chapter 2. In these studies, each subject encountered only one mechanism. Aggregating the data from these experiments, SGS still outperforms all other mechanisms tested. This consistency suggests that SGS not only excels in isolated scenarios but also maintains its effectiveness across different experimental conditions.

One key difference between SGS and Leimar and Hammerstein's Good Standing (L&H GS) lies in the potential for cooperative recovery. In SGS mechanism, there exists an absorbing state where, if both players are in bad standing, they cannot escape this condition—cooperation becomes unsustainable . L&H GS avoids this pitfall by allowing individuals to regain good standing by helping. They argue this feature should prevent the permanent breakdown of cooperation and sustains the viability of the cooperative club.

However, mechanisms like L&H GS or even BIS permit individuals to postpone cooperative behaviour, knowing they can "rehabilitate" their reputation with a single future act. This flexibility can undermine immediate cooperation, as individuals may choose to defect now and repair their reputation later. SGS eliminates this loophole, thereby reinforcing immediate cooperative incentives.

From a methodological perspective, our beliefs' elicitation method proved effective in capturing people understanding of the surrounding cooperation in their cohort.

To draw more substantial conclusions from our research, it is imperative to conduct further statistical analyses that encompass all the data presented in both Chapter 2 and Chapter 3. This should involve a detailed examination of the datasets specific to Chapter 3, as well as an integrated analysis of the combined data from both chapters. By thoroughly analysing these datasets individually and collectively, we can identify overarching patterns that may not be apparent at the current stage of this chapter or looking at each chapter in isolation. Such comprehensive analysis will enhance the validity of our findings and contribute to a deeper understanding of the subject matter.

Overall, our study highlights the critical role of reputational mechanisms in sustaining cooperative behaviour. Sugden Good Standing mechanism, by integrating both an individual's actions and the reputations of others, effectively fosters stable and enduring cooperation. These insights have significant implications for understanding indirect reciprocity and designing systems—whether in economic markets, organisational structures, or social networks—that aim to promote cooperative interactions. By recognising the importance of both individual behaviour and social context in shaping reputations, we can better cultivate environments where cooperation thrives.

References

Andreoni, J.: 1988, Why free ride?: Strategies and learning in public goods experiments, *Journal of Public Economics* **37**(3), 291–304.
URL: *https://www.sciencedirect.com/science/article/pii/0047272788900436*

Brier, G. W.: 1950, Verification of forecasts expressed in terms of probability, *Monthly weather review* **78**(1), 1–3.

Camera, G. and Gioffré, A.: 2022, Cooperation in indefinitely repeated helping games: Existence and characterization, *Journal of Economic Behavior 'I&' Organization* **200**, 1344–1356.
URL: *https://www.sciencedirect.com/science/article/pii/S0167268119303579*

Cosmides, L. and Tooby, J.: 1996, Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty, *Cognition* **58**(1), 1–73.
URL: *https://www.sciencedirect.com/science/article/pii/0010027795006648*

Engelmann, D. and Fischbacher, U.: 2008, Indirect reciprocity and strategic reputation building in an experimental helping game, *Working Paper 34*, Thurgauer Wirtschaftsinstitut.

Engelmann, D. and Fischbacher, U.: 2009, Indirect reciprocity and strategic reputation building in an experimental helping game, *Games and Economic Behavior* **67**(2), 399–407.

Fehr, E., Fischbacher, U. and Gächter, S.: 2002, Strong reciprocity, human cooperation, and the enforcement of social norms, *Human Nature* **13**(1), 1–25.

Fischbacher, U., Gächter, S. and Fehr, E.: 2001, Are people conditionally cooperative? evidence from a public goods experiment, *Economics Letters* **71**(3), 397–404.
URL: *https://www.sciencedirect.com/science/article/pii/S0165176501003949*

Fischbacher, U., Wolff, I. and Hussien, M.: 2024, Rights, duties, and taboos: The social codex of peer punishment. Manuscript in preparation.

Gigerenzer, G. and Hoffrage, U.: 1995, How to improve bayesian reasoning without instruction: Frequency formats, *Psychological Review* **102**(4), 684–704.

Herrmann, B., Thöni, C. and Gächter, S.: 2008, Antisocial punishment across societies, *Science* **319**(5868), 1362–1367.
URL: *https://www.science.org/doi/abs/10.1126/science.1153808*

Leimar, O. and Hammerstein, P.: 2001, Evolution of cooperation through indirect reciprocity, *Proc. R. Soc. Lond. B.* **268**, 745–753.

Levine, D.: 1998, Modeling altruism and spitefulness in experiment, *Review of Economic Dynamics* **1**(3), 593–622.

Nowak, M. and Sigmund, K.: 1998a, The dynamics of indirect reciprocity, *Journal of Theoretical Biology* **194**(4), 561–574.

Nowak, M. and Sigmund, K.: 1998b, Evolution of indirect reciprocity by image scoring, *Nature* **393**, 573–577.

Ohtsuki, H.: 2004, Reactive strategies in indirect reciprocity, *Journal of Theoretical Biology* **227**(3), 299–314.
URL: *https://www.sciencedirect.com/science/article/pii/S0022519303004260*

Ohtsuki, H. and Iwasa, Y.: 2006, The leading eight: social norms that can maintain cooperation by indirect reciprocity, *Journal of theoretical biology* **239**(4), 435–444.

Roth, A. E. and Murnighan, J. K.: 1978, Equilibrium behavior and repeated play of the prisoner's dilemma, *Journal of Mathematical Psychology* **17**(2), 189–198.

Sugden, R.: 1986, *The economics of rights, co-operation and welfare*, Oxford, UK: Basil Blackwell.

3.A Appendix

3.A.1 Incentive Compatibility of Belief Elicitation[6]

We elicit beliefs about an outcome which is an integer $0 \leq i \leq N$. Suppose the Decision Maker (DM) has beliefs π_i that state i will occur. We ask the DM to report a single value of k. If the DM reports k and outcome i actually happens, their loss function is $|k - i|$.

Given their beliefs (and assuming risk-neutrality) the expected loss function is

$$L(k) = \sum_{i=0}^{k-1} (k - i)\pi_i + \sum_{i=k+1}^{N} (i - k)\pi_i, \tag{3.20}$$

where we interpret sums with ending index less than starting index as being equal to zero. The DM wants to minimise (3.20); we aim to characterise $K^\star$, the set of reports which minimise the loss.

Proposition 3.1. *For all* $0 < k < N$*,*

$$\Delta(k) \equiv L(k+1) - L(k) = \sum_{i=0}^{k} \pi_i - \sum_{i=k+1}^{N} \pi_i. \tag{3.21}$$

Proof. By straightforward algebra we have

$$\begin{aligned}
L(k+1) &= \sum_{i=0}^{k} (k + 1 - i)\pi_i + \sum_{i=k+2}^{N} (i - k - 1)\pi_i, && (3.22)\\
&= \sum_{i=0}^{k-1} (k - i)\pi_i + (k - k)\pi_k + \sum_{i=0}^{k} \pi_i \\
&\quad + \sum_{i=k+1}^{N} (i - k)\pi_i - \pi_{k+1} - \sum_{i=k+2}^{N} \pi_i, && (3.23)\\
&= \sum_{i=0}^{k-1} (k - i)\pi_i + \sum_{i=k+1}^{N} (i - k)\pi_i + \sum_{i=0}^{k} \pi_i - \sum_{i=k+1}^{N} \pi_i, && (3.24)\\
&= L(k) + \sum_{i=0}^{k} \pi_i - \sum_{i=k+1}^{N} \pi_i. && (3.25)
\end{aligned}$$

□

[6] This proof has been done by Theodore T. Turocy.

Observe that $\Delta(k)$ is non-decreasing in k. This implies the following:

Proposition 3.2. *If for some k, $\sum_{i=0}^{k} \pi_i > \frac{1}{2}$, then $L(k') > L(k)$ for all $k' > k$.*

Proof. We write

$$L(k') - L(k) = \sum_{j=k}^{k'-1} L(j+1) - L(j) = \sum_{j=k}^{k'-1} \Delta(j). \tag{3.26}$$

Because $\sum_{i=0}^{k} \pi_i > \frac{1}{2}$, the first term in the sum is positive. Because $\Delta(j)$ is non-decreasing in j, all further terms in the sum are likewise positive. Therefore the sum is positive, and $L(k') > L(k)$ as claimed. □

Proposition 3.3. *If for some k, $\sum_{i=k}^{N} \pi_i > \frac{1}{2}$, then $L(k') > L(k)$ for all $k' < k$.*

Proof. Analogous to previous Proposition. □

These results motivate the definition of two quantities,

$$\overline{k} \equiv \arg\min \left\{ k : \sum_{i=0}^{k} \pi_i > \frac{1}{2} \right\} \tag{3.27}$$

$$\underline{k} \equiv \arg\max \left\{ k : \sum_{i=k}^{N} \pi_i > \frac{1}{2} \right\} \tag{3.28}$$

It is immediate that $\underline{k} \leq \overline{k}$, and it must be the case that for any $k^\star \in K^\star$, $\underline{k} \leq k^\star \leq \overline{k}$. There are a few possibilities:

If there exists some i such that $\pi_i > \frac{1}{2}$, then $\overline{k} = \underline{k} \equiv \hat{k}$, and therefore $K^\star = \{\hat{k}\}$. This implies that the median of the distribution occurs at this outcome.

If $\overline{k} = \underline{k} + 1$, then $\overline{k} \in K^\star$ if and only if $\sum_{i=\overline{k}}^{N} \pi_i \geq \sum_{i=0}^{\underline{k}} \pi_i$, and $\underline{k} \in K^\star$ if and only if $\sum_{i=\overline{k}}^{N} \pi_i \leq \sum_{i=0}^{\underline{k}} \pi_i$. That is, which is the optimal response can be determined by comparing the "captive" tail probabilities. The element(s) of $K^\star$ are at an end of the interval which brackets the median.

If $\overline{k} - \underline{k} > 1$, then it must be the case that $\sum_{i=\overline{k}}^{N} \pi_i = \sum_{i=0}^{\underline{k}} \pi_i = \frac{1}{2}$. In this case, again the median must be in the interval, but the best response is not unique: $K^\star = \{\underline{k}, ..., \overline{k}\}$.

Observe that we can rule out the last (somewhat pathological) case if the distribution is unimodal.

In terms of the proposed experiment, if the value i is realised by drawing several observations from a population, it is (probably) reasonable to rule out this pathological case. Because it is a sampling process, the only way to get a bimodal distribution would be if the participant held beliefs that there are two states of the world. In one state, non-active players are helped fewer than $N/2$ times; in another state, non-active players are not helped fewer than $N/2$ times. Further, both of these states are exactly equally likely. Then, you can create a situation where there is a zero-probability region at the median. Any relaxation of those beliefs, combined with the belief that sampling is done as we state it is, would lead to there being probability mass around the median, pinning down the report (to within ± 1).

3.A.2 Statistical Rationale and Calculations for the Stopping Rule

We provide the formal proof and detailed statistical calculations underlying the stopping rule employed in our experimental design.

Our stopping rule was designed to ensure that we would have sufficient statistical power to detect a significant difference between the control and reputation treatments. The criterion of having five or more control cohorts among the bottom six cohorts (ranked by average helping rate) was chosen based on the probabilities associated with such an outcome under the null hypothesis of no treatment effect.

Under the null hypothesis, the ranks of the cohorts are randomly assigned with respect to the treatment conditions. We calculated the probability of observing five or more control cohorts in the bottom six purely by chance, which would be unlikely if there were truly no difference between the treatments.

3.A.2.1 Calculation of the Probability for the Stopping Rule

This appendix shows how we arrive at an approximate probability of **8.3%** for observing at least five control cohorts among the bottom six, under the null hypothesis of "no difference" between treatments. We also contrast it with a simpler *hypergeometric* approach that yields $\approx 0.39\%$ for a *strict* bottom-six event. The discrepancy arises from different ways of counting borderline/tied ranks and defining "worst-case" placements.

Hypergeometric Calculation (Strict Cutoff)

We consider a total of $N = 18$ cohorts: $n_C = 6$ control cohorts (denoted by C) and $n_S = 12$ cohorts with reputation mechanisms (denoted by S). The cohorts are ranked from highest to lowest based on their average helping rates.

Our stopping rule specifies that if $k \geq 5$ of the bottom $b = 6$ cohorts are control cohorts, we will discontinue the control treatment. We aim to calculate the probability P of this event occurring under the null hypothesis H_0 that there is no difference in helping rates between the control and reputation treatments.

Under H_0, the assignment of cohorts to ranks is random with respect to the treatment conditions. The number of control cohorts X in the bottom b cohorts follows a hypergeometric distribution with parameters:

- Population size: $N = 18$
- Number of control cohorts in the population: $n_C = 6$
- Sample size (number of bottom cohorts): $b = 6$
- Number of control cohorts in the sample: X

The probability mass function (pmf) of the hypergeometric distribution is:

$$P(X = k) = \frac{\binom{n_C}{k}\binom{n_S}{b-k}}{\binom{N}{b}},$$

where $\binom{n}{k}$ is the binomial coefficient "n choose k."

We calculate the probability of observing $k = 5$ or $k = 6$ control cohorts in the bottom $b = 6$ cohorts:

$$P(X \geq 5) = P(X = 5) + P(X = 6).$$

Calculating $P(X = 5)$:

$$P(X = 5) = \frac{\binom{6}{5}\binom{12}{1}}{\binom{18}{6}} = \frac{6 \times 12}{18564} = \frac{72}{18564} \approx 0.003879.$$

Calculating $P(X = 6)$:

$$P(X = 6) = \frac{\binom{6}{6}\binom{12}{0}}{\binom{18}{6}} = \frac{1 \times 1}{18564} = \frac{1}{18564} \approx 0.000054.$$

Total Probability:

$$P(X \geq 5) = P(X = 5) + P(X = 6) = \frac{72 + 1}{18564} = \frac{73}{18564} \approx 0.003933.$$

Thus, the probability of observing five or more control cohorts in the bottom six purely by chance is approximately 0.3933%.

Interpretation. This $\approx 0.39\%$ figure is for a *strict* cutoff in which exactly ranks $1, 2, 3, 4, 5, 6$ are counted as the "bottom 6," with no allowance for ties or borderline cases. In many practical designs, however, rank ties or "near ties" might group multiple cohorts into or out of those bottom positions, effectively increasing the chance of seeing "at least 5 Control cohorts near the bottom."

Rank-Sum Perspective and 8.3% Probability

In our actual stopping rule, we use a version of the Mann–Whitney approach to account for situations where the 6th Control cohort might be only marginally above the rank-6 threshold, or where some cohorts tie. We treat any arrangement that places ≥ 5 Control cohorts at or near the bottom as a "worst-case" scenario for continuing the Control treatment.

The U statistic for the control group is calculated as:

$$U_C = R_C - \frac{n_C(n_C + 1)}{2}, \tag{3.29}$$

where R_C is the sum of the ranks for the control cohorts, and $n_C = 6$ is the number of control cohorts.

Under H_0, the expected value of U_C is:

$$E(U_C) = \frac{n_C n_S}{2} = \frac{6 \times 12}{2} = 36. \tag{3.30}$$

The standard deviation of U_C is:

$$\sigma_{U_C} = \sqrt{\frac{n_C n_S(n_C + n_S + 1)}{12}} = \sqrt{\frac{6 \times 12 \times 19}{12}} \approx 8.485. \tag{3.31}$$

Suppose in the worst-case scenario for our stopping rule, the control cohorts occupy the ranks such that five control cohorts are in the bottom six ranks, and one control cohort is in a higher rank. The sum of the ranks for the control cohorts, R_C, can be calculated accordingly.

Assuming the five control cohorts are ranked $r = 1, 2, 3, 4, 5$ (with rank 1 being the lowest helping rate), and the sixth control cohort is ranked somewhere among the top ranks, the sum R_C would be at least:

$$R_C = 1 + 2 + 3 + 4 + 5 + r_{\text{top}}, \tag{3.32}$$

where r_{top} is the rank of the sixth control cohort.

The minimum possible sum R_C in this scenario would be:

$$R_C = 1 + 2 + 3 + 4 + 5 + 6 = 21, \tag{3.33}$$

if the sixth control cohort is ranked 6th.

Using

$$U_C = R_C - \frac{n_C(n_C + 1)}{2} = 21 - \frac{6 \times 7}{2} = 21 - 21 = 0. \tag{3.34}$$

This is the minimum possible U statistic for the control group.

We can calculate the z-score:

$$z = \frac{U_C - E(U_C)}{\sigma_{U_C}} = \frac{0 - 36}{8.485} \approx -4.243. \tag{3.35}$$

The corresponding p-value (one-tailed test) is extremely small.

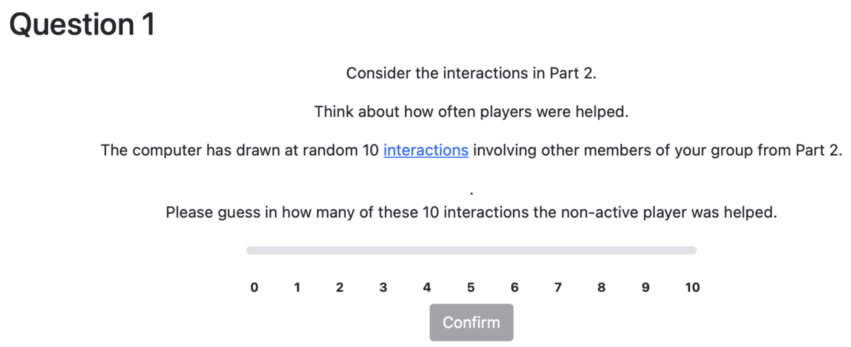
However, in our stopping rule, we consider the worst-case p-value when five of the bottom six cohorts are control cohorts. The actual p-value in this scenario, considering possible rank arrangements and using the Mann–Whitney U test, is approximately 8.3%.

This estimation accounts for the fact that the sixth control cohort may not necessarily be at the very top rank, and there may be ties or variations in ranks.

Thus the final figure **8.3%** matches our stated stopping condition more precisely than the hypergeometric $\approx 0.39\%$ does. In the main text, we therefore report the $\approx 8.3\%$ probability as the more relevant measure for our design.

3.A.3 Additional Tables and Figures

3.A.3.1 Baseline Belief Elicitation



Question 1

Consider the interactions in Part 2.

Think about how often players were helped.

The computer has drawn at random 10 interactions involving other members of your group from Part 2.

.

Please guess in how many of these 10 interactions the non-active player was helped.

Figure 3.18: Baseline beliefs' elicitation

3.A.3.2 Summary Statistics

Table 3.3: Summary statistics

	%
Gender:	
Female	49.91
Male	47.50
Other	1.72
Prefer not to say	0.87
Age:	
≥27	14.23
26	2.01
25	7.17
24	3.64
23	14.06
22	13.54
21	12.50
20	15.33
19	14.70
18	2.92
Degree:	
Bachelor	59.59
Master	31.53
PhD	2.92
Other degree course or affiliation	2.63
Prefer not to say	2.30
INTO	0.61
Staff	0.72
Year(s) at UEA:	
1^{st} year	34.93
2^{nd} year	29.05
3^{rd} year	17.29
4^{th} year	8.91
More than 4 years	7.51
Prefer not to say	2.41
Faculty:	
Social Sciences	65.81
Sciences	19.33
Humanities	11.06
Others	3.80

Notes: The percentages are based on the total number of responses for each category. The category "Other" under Gender includes non-binary and other gender identities not specified. The age categories are divided to highlight different age groups. As per faculty, *Social Sciences:* NBS, ECO, DEV, LAW, EDU, PSY; *Sciences:* CMP, BIO, ENV, PHA, MED, HSC, MTH; *Humanities:* PPL, HIS, LDC, AMA, HUM.

3.A.4 Experimental Instructions

Welcome

Welcome to today's experiment and thanks for coming. This is an experiment in the economics of decision-making. If you follow the instructions, complete the experiment, and make appropriate choices, you can earn an appreciable amount of money. This will be paid to you in private, in cash, at the end of the session, before you leave the laboratory.

It is important that you remain silent and do not look at other people's work. If you have any questions, or need assistance of any kind, please raise your hand and an experimenter will come to you. If you talk, laugh, exclaim out loud, etc., you will be asked to leave and you will not be paid. We expect and appreciate your cooperation.

All choices in today's experiment and any information you choose to give are recorded anonymously and will only be used in the analysis of the data from this experiment.
We will now describe the nature of the experiment in more detail.

Introduction

In this experiment, you will be assigned to a **group** of **six** people. The other people in your group will be five other people in this room. You will never find out which other people are the ones who are in your group.

The experiment has **two parts**. We will now describe Part 1. We will describe Part 2 after everyone has finished Part 1.

Part 1

The interaction

There will be a series of at least 40 **rounds**. At the beginning of each round, the computer will match you at random with another person in your group. You will **interact** with that person in that round. Because the computer will create a new matching at the beginning of each round, in different rounds you may interact with different members of your group. You will not find out which other person in your group you are matched with, nor whether or when you have been matched with them previously.

In each match, one of you will be randomly assigned the role of the **active player**. The other one will be assigned the role of the **non-active player**.

If you are the active player, you have an opportunity to help the non-active player. **Helping** is an action that has a monetary **cost of helping** for the active player and gives a monetary **benefit** to the non-active player.

You will find out your cost of helping for Part 1 before Round 1 begins. Your cost of helping will stay the same throughout Part 1. The cost of helping may be different for different members of your group. However, the cost of helping is no more than £6.00 for anyone.

In every round, you have an **account**. This account is separate for each round. At the start of each round, your account for that round will have an **endowment** of £7.00. The final value of the accounts of the active player and the non-active player depends on the choice the active player makes:

- If the active player **chooses to help**, the active player's cost of helping will be deducted from their account, and a benefit of £10.00 will be added to the non-active player's account.
- If the active player **chooses not to help**, no deduction will be made from the active player's account and no addition made to the non-active player's

account.

After the active player makes their choice, any deductions or additions are made to the player's accounts for that round. The round then ends. When the next round begins, all players have a new account, with an endowment of £7.00.

Your earnings from Part 1

Only one of the rounds will be **for real**. The computer will randomly select which round this is, but you will not know which round was selected until the end of the experiment. Which round this is will be the same for all members of the group. For everyone in the group, their earnings from Part 1 will be equal to the final value of their account for the selected round.

Baseline[7]

No additional text for this part.

Sugden Good Standing

Scores

Through this part of the experiment, you will have a **score**. Your score summarises whether, in previous rounds, you chose to help or not to help. Your score can be 0 or 1.

In each round, you will see your own current score. In each round in which you are the non-active player, the active player will also see your current score. In each round in which you are the active player, you will see the current score of the non-active player.

At the start of Round 1, everyone has a score of 1.

The computer keeps a **record** of everyone's current score. At the end of each round in which you are the active player, your score is updated based on your choice and the score of the non-active player.

- If you choose to help:

[7] This was not shown to the participants and it is displayed for differentiate the three versions of the instructions.

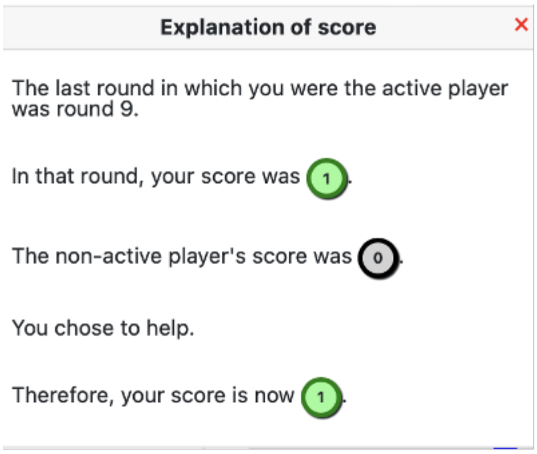
  - If the non-active player's score was 1, then your score at the start of the next round will be 1.
  - If the non-active player's score was 0, then your score at the start of the next round will be the same as your current score.

- If you choose not to help:
  - If the non-active player's score was 1, then your score at the start of the next round will be 0.
  - If the non-active player's score was 0, then your score at the start of the next round will be the same as your current score.

However, if this is your first round as the active player, your record will be filled in as if you had chosen to help in the previous round. So, your score in the first round of this part will be 1.

At the end of each round in which you are the active player, your score will be updated based on the choice you make in that round. In rounds in which you are the non-active player, your score does not change.

For example, suppose in Round 11 you are the active player. The explanation of your score at the start of the round might be



Explanation of score

The last round in which you were the active player was round 9.

In that round, your score was 1.

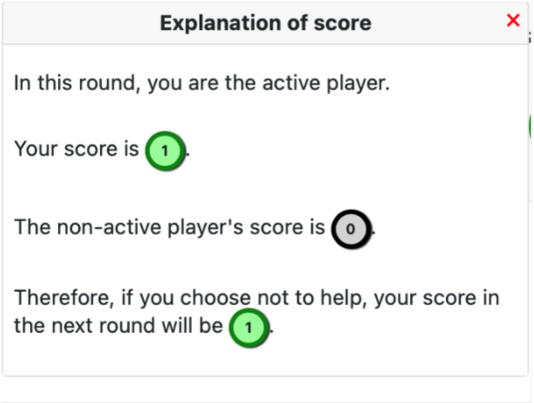
The non-active player's score was 0.

You chose to help.

Therefore, your score is now 1.

This record shows that in the last round in which you were the active player, your score was 1 and the non-active player's score was 0. You chose to help, and therefore your score at the start of Round 11 is 1.

Suppose you are the active player in Round 11, and you are considering choosing not to help. The explanation of how this would affect your score would be



Explanation of score

In this round, you are the active player.

Your score is 1.

The non-active player's score is 0.

Therefore, if you choose not to help, your score in the next round will be 1.

Because the non-active player's score is 0, your score does not change. Therefore, your score at the start of Round 12 would be 1.

What you will see on the screen

We will now show you what you will see on your screen during a round.

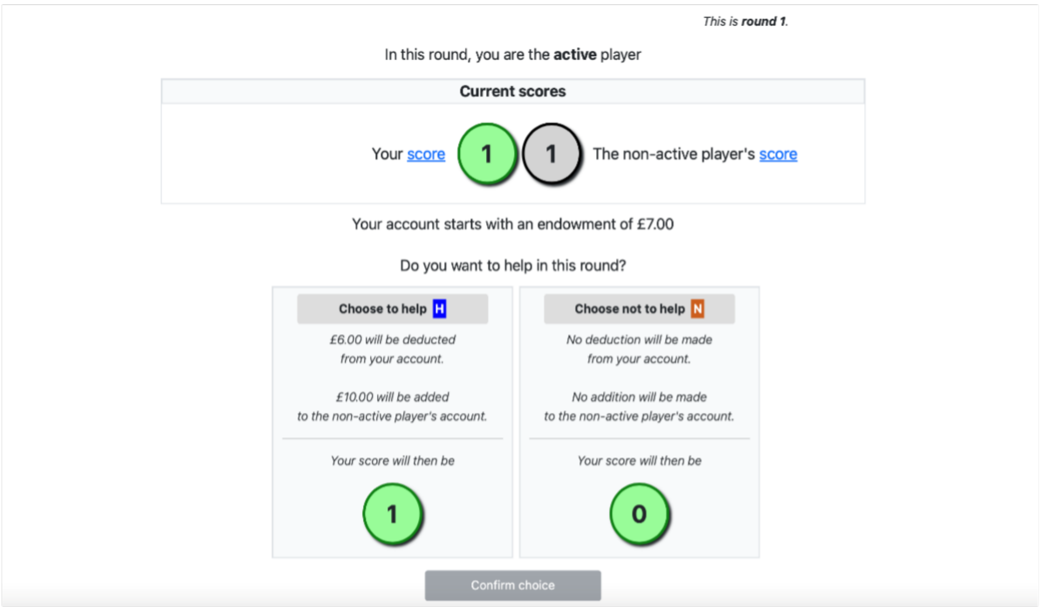
Scores will be displayed on the screen using **badges**, which are coloured circles showing the score. Your score will always be shown using a **green badge** like this

Any time you see a green badge representing your score, you can click on it to pop up a display which explains how your record was used to determine your score.

When you are the active player, you will see the score of the non-active player. The non-active player's score will be shown using a **grey badge** like this

You will only be able to see the current score of the non-active player, and not their record; therefore, clicking on a grey badge has no effect.

When you are the active player, you will see a screen like the one below



*This is **round 1**.*

In this round, you are the **active** player

Current scores

Your score The non-active player's score

Your account starts with an endowment of £7.00

Do you want to help in this round?

Choose to help H	Choose not to help N
£6.00 will be deducted from your account. *£10.00 will be added to the non-active player's account.*	*No deduction will be made from your account.* *No addition will be made to the non-active player's account.*
Your score will then be 1	*Your score will then be* 0

Confirm choice

The screen shows the round number, your score, the non-active player's score, and your endowment in this round.

The two boxes represent the two choices you can make: to help or not to help. Each box summarises the consequences for you and for the non-active player of the corresponding choice.

To choose to help, click on the button labelled **Choose to help**. To choose not to help, click on the button labelled **Choose not to help**. When you click on one of those buttons, the background will change from grey to green. If you change your mind, you can change your choice simply by clicking on the button corresponding to the other choice. When you are satisfied with your choice, click the button labelled **Confirm choice**.

How Part 1 ends

Exactly how many rounds will be played is not fixed in advance. There will be at least 40 rounds. At the end of Round 40, the computer will simulate the roll of a six-sided die. If the roll is 1 or 2, then Part 1 will end; otherwise, Part 1 will continue to Round 41. This process will be repeated after each subsequent round, until the simulated die roll results in a 1 or 2. In other words, in each round starting from Round 40, there is a 1 in 3 chance that Part 1 will end after that round.

Binary Image Scoring

Scores

Through this part of the experiment, you will have a **score**. Your score summarises whether, in previous rounds, you chose to help or not to help. Your score can be 0 or 1.

In each round, you will see your own current score. In each round in which you are the non-active player, the active player will also see your current score. In each round in which you are the active player, you will see the current score of the non-active player.

At the start of Round 1, everyone has a score of 1.

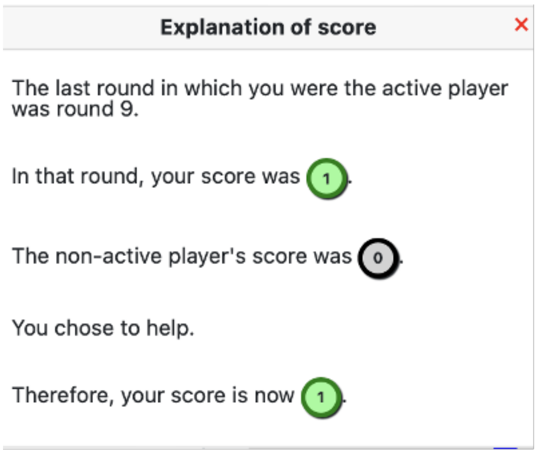
The computer keeps a **record** of everyone's current score. At the end of each round in which you are the active player, your score is updated based on your choice.

- If you choose to help, your score in the next round will be 1.
- If you choose not to help, your score in the next round will be 0.

However, if there are fewer than five previous rounds in which you were an active player, your record will be filled in as if you chose to help in the missing rounds. So, your score in the first round of this part will be 1.

At the end of each round in which you are the active player, your score is updated based on the choice you make in that round. In rounds in which you are the non-active player, your score does not change.

For example, suppose in Round 11 you are the active player. The explanation of your score at the start of the round might be



Explanation of score

The last round in which you were the active player was round 9.

In that round, your score was 1.

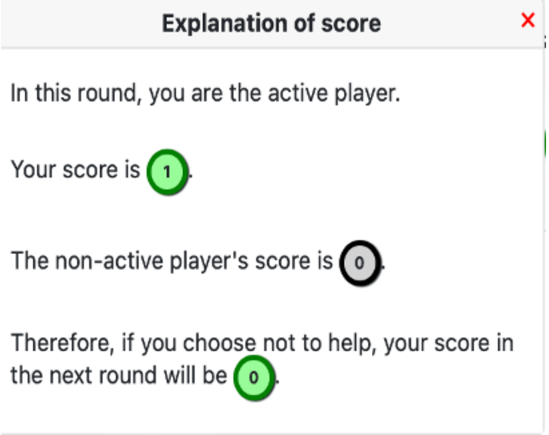
The non-active player's score was 0.

You chose to help.

Therefore, your score is now 1.

This record shows that in the last round in which you were the active player, your score was 1 and the non-active player's score was 0. You chose to help, and therefore your score at the start of Round 11 is 1.

Suppose you are the active player in Round 11, and you are considering choosing not to help. The explanation of how this would affect your score would be



Explanation of score

In this round, you are the active player.

Your score is 1.

The non-active player's score is 0.

Therefore, if you choose not to help, your score in the next round will be 0.

If you choose not to help in this round, your score in the next round would be 0.

What you will see on the screen

We will now show you what you will see on your screen during a round.

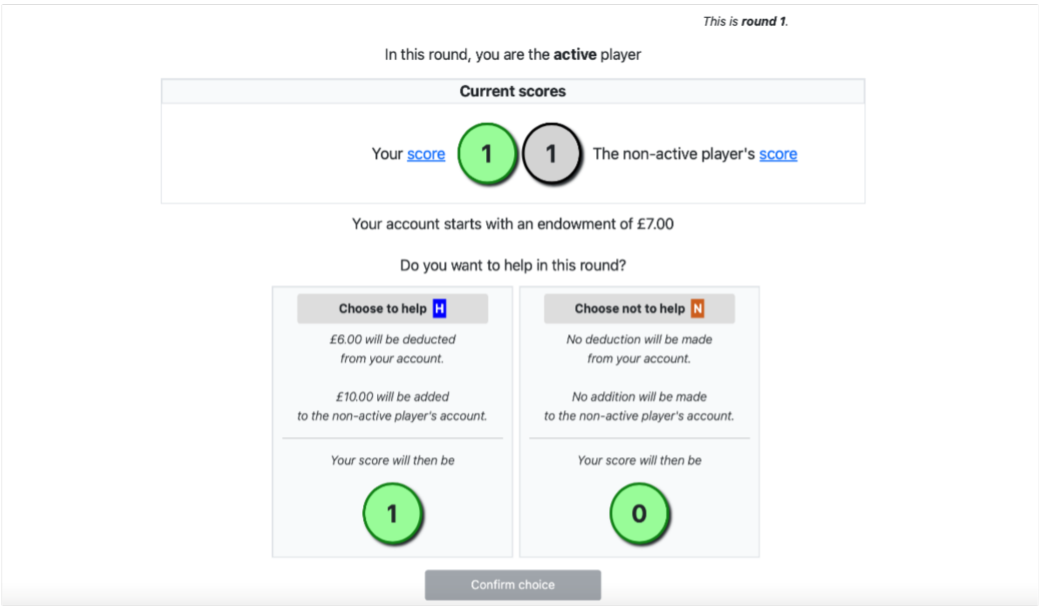
Scores will be displayed on the screen using **badges**, which are coloured circles showing the score. Your score will always be shown using a **green badge** like this

Any time you see a green badge representing your score, you can click on it to pop up a display which explains how your record was used to determine your score.

When you are the active player, you will see the score of the non-active player. The non-active player's score will be shown using a **grey badge** like this

You will only be able to see the current score of the non-active player, and not their record; therefore, clicking on a grey badge has no effect.

When you are the active player, you will see a screen like the one below



*This is **round 1**.*

In this round, you are the **active** player

Current scores

Your score 1 1 The non-active player's score

Your account starts with an endowment of £7.00

Do you want to help in this round?

£6.00 will be deducted from your account.

£10.00 will be added to the non-active player's account.

Your score will then be

No deduction will be made from your account.

No addition will be made to the non-active player's account.

Your score will then be

The screen shows the round number, your score, the non-active player's score, and your endowment in this round.

The two boxes represent the two choices you can make: to help or not to help. Each box summarises the consequences for you and for the non-active player of the corresponding choice.

To choose to help, click on the button labelled **Choose to help**. To choose not to help, click on the button labelled **Choose not to help**. When you click on one of those buttons, the background will change from grey to green. If you change your mind, you can change your choice simply by clicking on the button corresponding to the other choice. When you are satisfied with your choice, click the button labelled **Confirm choice**.

How Part 1 ends

Exactly how many rounds will be played is not fixed in advance. There will be at least 40 rounds. At the end of Round 40, the computer will simulate the roll of a six-sided die. If the roll is 1 or 2, then Part 1 will end; otherwise, Part 1 will continue to Round 41. This process will be repeated after each subsequent round, until the simulated die roll results in a 1 or 2. In other words, in each round starting from Round 40, there is a 1 in 3 chance that Part 1 will end after that round.

Part 2

In Part 2, you will be in the **same group** of **six** participants as in Part 1.

Part 2 has the same structure as Part 1. The only difference between Part 1 and Part 2 is that your cost of helping may not be the same as in Part 1. Likewise, for each of the other people in your group, the cost of helping may not be the same as in Part 1. However, as in Part 1, the cost of helping will be no more than £6.00 for anyone. You will be told your cost of helping for Part 2 before the first round begins. You will not find out the cost of helping for anyone else in your group.

As in Part 1, there will be a series of at least 40 **rounds**. At the beginning of each round, you will be matched at random with another person in your group. One of you will be randomly assigned the role of the **active player**. The other person will be the **non-active player**. The accounts of both players initially have an endowment of £7.00. The active player will have the opportunity to help the non-active player. If the active player chooses to help, the active player's cost of helping will be deducted from their account, and £10.00 will be added to the account of the non-active player. If the active player chooses not to help, no deduction or addition will be made to either player's account.

One of the rounds will be **for real**. The computer will select at random which round is for real, but you will not know which round that is until the end of the experiment. This round will be the same for all members of the group. For everyone in the group, their earnings from Part 2 will be equal to the final value of their account for the selected round. These will be added to your earnings from Part 1 to determine your earnings for the experiment as a whole.

Sugden Good Standing and Binary Image Scoring

As in Part 1, at the start of every round, you will have a **score** of 0 or 1 which summarises whether, in previous rounds, you chose to help or not to help. Scores will be computed and updated exactly as they were in Part 1.

Exactly how many rounds will be played is not fixed in advance. There will be at least 40 rounds. At the end of Round 40, the computer will simulate the roll of a six-sided die. If the roll is 1 or 2, then Part 2 will end; otherwise, Part 2 will continue to Round 41. This process will be repeated after each subsequent round, until the simulated die roll results in a 1 or 2. In other words, in each round starting from Round 40, there is a 1 in 3 chance that Part 2 will end after that round. The number of rounds in Part 2 may not be the same as in Part 1 depending on the outcomes of the simulated die rolls.

Part 3

[In parenthesis Good Standing and Binary Image Scoring]
In this part, we have one [two] question[s] for you to answer. You have the opportunity to earn an additional amount from your answers to this [these] question[s], over and above the amounts you have earned from Part 1 and Part 2.

Question [1]

Consider the interactions in Part 2 [in which the non-active player had a score of 1]. Think about how often players with this score were helped.

The computer has drawn at random 10 interactions involving other members of your group from Part 2[, in which the non-active player had a score of 1]. That is, these are interactions in which you were not involved either as the active or the non-active player.

For this Question [1], we would like you to guess in how many of these 10 interactions the non-active player was helped.

In this Question [1], if you guess exactly the number of times the non-active player was helped in these 10 interactions, then you will earn an additional payment of

£5.00. If your guess is off by one – that is, the actual number is one more or one fewer than your guess – you will earn £4.50. If your guess is off by two, you will earn £4.00, and so on.

Question 2

[This was present only in BIS and SGS.]
Now, still considering Part 2, consider the interactions in which the non-active player had a score of 0. Think about how often players with this score were helped.

The computer has drawn at random 10 interactions involving other members of your group from Part 2, in which the non-active player had a score of 0. That is, these are interactions in which you were not involved either as the active or the non-active player.

For Question 2, we would like you to guess in how many of these 10 interactions the non-active player was helped.

In Question 2, If you guess exactly the number of times the non-active player was helped in these 10 interactions, then you will earn an additional payment of £5.00. If your guess is off by one – that is, the actual number is one more or one fewer than your guess – you will earn £4.50. If your guess is off by two, you will earn £4.00, and so on.

ASPREDICTED

Wharton
CREDIBILITY LAB
UNIVERSITY of PENNSYLVANIA

Reputations in Helping Games, May 2024 (#174378)

Author(s)
Andrea Marietta Leina (University of East Anglia) - andrea.mariettaleina@univr.it
Amrish Patel (University of East Anglia) - amrish.patel@uea.ac.uk
Robert Sugden (University of East Anglia) - r.sugden@uea.ac.uk
Theodore Turocy (University of East Anglia) - t.turocy@uea.ac.uk

Pre-registered on: 05/09/2024 08:51 AM (PT)

1) Have any data been collected for this study already?
It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.

2) What's the main question being asked or hypothesis being tested in this study?
Reputation mechanisms have been proposed as effective tools to sustain indirect reciprocity in repeated games. These mechanisms are information structures that encode and publicly display an individual's past behaviour. The general inquiry of this study is to understand which reputation mechanism, Binary Image Scoring (BIS) (Nowak and Sigmund, 1998) or Sugden's concept of Good Standing (SGS) (Sugden, 1986), serves as a better explanatory framework for indirect reciprocity within repeated helping games.

Specifically, we aim to answer three questions:
1)�Which mechanism, BIS or SGS, leads to higher helping rates?
2)�Does the efficacy of these mechanisms vary according to whether cost of helping is homogeneous or heterogeneous?
3)�In each mechanism, to what extent are subjects' beliefs regarding the chance of being helped correct?

While we hypothesise that reputational mechanisms outperform scenarios without any reputation (the baseline), we do not formulate explicit hypotheses regarding the other mechanisms.

3) Describe the key dependent variable(s) specifying how they will be measured.
The experiment is divided in 3 parts.
Parts 1 and 2 are the repeated helping games. In these parts, cohorts of six subjects are matched at random in couples every round (40 finite rounds + 2/3 chance a new round will occur after each subsequent round). In every couple, one subject is the non-active player and has no choices to make. The other is the active player and has the choice whether to help the non-active player, paying a cost c that provides a benefit b to the non-active player (b>c>0), or not to help. In one of the two parts, everyone will have the same cost. In the other part, a fixed half of the subjects will have a higher cost (ch) and the other half a lower cost (cl), with b>ch >cl >0). All subjects in BIS and SGS are endowed with a binary score after each round, the reputation mechanism, that summarises their previous helping behaviour.
In part 3, subjects have to guess in how many interactions out of 10 randomly selected ones from their cohort in Part 2, the non-active player was helped, conditional on their score (in the baseline, these beliefs are elicited just as chances of being helped).
The dependent variables will be operationalized as follows:
Average helping rates: For each subject, this rate is calculated as the proportion of rounds in which they chose to help their partners out of the total number of rounds played as active players. We will also condition these variables on players' scores.
Beliefs about likelihood of being helped: For each subject, this likelihood is their reported guess of the proportion of sampled interactions in which the non-active player was helped.
Proportion of correct beliefs: For each subject, this proportion is the ratio between their guess and the median rate of helping in their cohort. This would tell us how far their beliefs are in line with the true value. We will compare aggregate proportions at the condition level.

4) How many and which conditions will participants be assigned to?
Each reputational mechanism is presented to subjects as a binary score (1 or 0) that summarises previous helping behaviour. Participants will be assigned to three different conditions in a between-subjects design:
1. BIS Condition: The reputational mechanism is based on the subject's action in the most recent round in which they were an active player. The subject has a score of 1 if they helped in that round and 0 if they did not. All subjects start with a score of 1.
2. SGS Condition: The reputational mechanism is based on the subject's action in the most recent round in which they were an active player and in which the non-active player had a score of 1. The subject has a score of 1 if they helped in that round and 0 if they did not. All subjects start with a score of 1.
3. Control Condition: Subjects will not be endowed with any reputational mechanisms. This condition serves as a baseline to observe how individuals behave without any reputation in the games.
In each condition, we will conduct two scenarios in a within-subject design:
1. Homogeneous Cost Scenario: All subjects will have the same cost for helping the non-active player.
2. Heterogeneous Cost Scenario: Subjects will have two different costs for helping non-active players. Specifically, a fixed half of the population will have a high cost (ch), and the other fixed half will have a low cost (cl).

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

To answer questions 1 and 2, we will use as dependent variable the average helping rates.

We will aggregate the data for all rounds, and, following Engelman & Fischbacher (2009), we will mainly use in our analysis data from rounds 10 to 35.

1. Comparison of Cooperation Levels: conduct a series of Fisher-Pitman test to compare the average levels of helping between the three conditions, separately for both the homogeneous and heterogeneous costs scenarios.
2. Interaction Analysis: perform a 2x2 factorial ANOVA (or its non-parametric equivalent) with factors being the reputation mechanism (BIS vs. SGS) and the cost scenario (homogeneous vs. heterogeneous).
3. Comparison with Control Condition: conduct independent samples Fisher-Pitman tests to compare the cooperation levels between each reputation mechanism condition (BIS and SGS) and the control condition, separately for both the homogeneous and heterogeneous costs scenarios.

For part 3, we will use beliefs about likelihood of being helped and proportion of correct beliefs:

First, we will explore beliefs about the mechanisms or participant characteristics, using OLS regressions to understand the underlying mechanisms driving cooperation levels.

Second, we will analyse more carefully beliefs conditional on different scores. We will do that using descriptive statistics and Fisher-Pitman to determine if there are significant differences in beliefs conditional on scores between the BIS and SGS conditions.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.

All subjects will be required to complete all the experiment. We will include observations from all subjects who complete all the parts of the experiment.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.

Subjects will be organised into distinct cohorts, each comprising six subjects. Within each session, 12 subjects will be assigned, forming two such cohorts. Six sessions will be conducted for each reputation mechanism, totalling 12 independent cohorts, with 6 dedicated to each mechanism. Additional to these, we will have some baseline sessions. We hypothesise that the presence of a reputation mechanism will increase helping behaviour. Consequently, we will start running 3 sessions per condition at random (control, BIS, SGS), thereby ensuring 18 independent cohorts, evenly distributed across the conditions. If, after these sessions, the bottom 6 cohorts (judged by average helping rate) encompass 5 or more cohorts from the control group, the control condition will be discontinued. In such an event, only the remaining 6 sessions (3 per reputation mechanism) will proceed. Should the stopping rule not be met, an additional 3 sessions per condition will be administered. Thus, in adherence to the stopping rule, 180 subjects will be recruited, with 72 assigned to Binary Image Scoring (BIS), 72 to Sugden's Good Standing (SGS), and 36 to the control group. This entails 6 sessions allocated to each reputational mechanism and 3 to the control condition. Conversely, if the stopping criterion is not satisfied, recruitment will extend to 216 subjects, with 72 assigned to each condition, amounting to 6 sessions per condition.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)

We have already collected data from six sessions. I, Andrea Marietta Leina, forgot to press send on this preregistration, but I have not looked at the data, and I have had this preregistration drafted on aspredicted.org since April 15, 2024.

In general, we aim to replicate and compare our analysis with two studies in the literature: Bolton & al. (2006) and Engelman & Fischbacher (2009). Moreover, we will also compare the results with a previous experiment we have already conducted. The latter differs from this study on the reputation mechanisms tested. We tested an "Image scoring" machanism, similar to Engelman & Fischbacher (2009) based on the definition of image scoring by Nowak and Sigmund (1998), i.e. a score based on 5 previous histories, and Hammerstein's concept of Good Standing (Leimar & Hammerstein, 2001). The only difference from SGS is that an active player with a score of 0 can gain a score of 1 by helping a non-active player with a score of 0.